Universiteit Leiden

Master Computer Science

Multi-aspect monitoring of
machine learning models in operation

Name:	Jan Schillemans
Student ID:	2382903
Date:	27/01/2021
Specialisation:	Business Studies
1st supervisor:	Prof.dr.ir. Joost Visser
2nd supervisor:	Dr. Andreas Zolnowski

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Background As machine learning is increasingly used in many economically critical areas in companies, monitoring of those machine learning applications becomes much more important. Most companies using machine learning in operation have little experience and lack of guidance in monitoring machine learning applications.

Aim In this thesis project, this problem will be addressed and worked on, so that in the end a sophisticated concept will be developed of how to ensure the quality of a machine learning application in operation.

Method We did literature research, conducted surveys, talked to professionals and performed case studies on a given use case to propose methods which are feasible and useful to obtain increased control over business applications of machine learning. With the help of a use case for a German insurance company, which tries to anticipate and prevent contract terminations by using machine learning, we validated our findings.

Results To satisfy the needs of a wide range of stakeholders, we created metrics for machine learning models in the areas of 1. general metrics, 2. machine learning specific metrics, 3. static machine learning metrics, 4. business metrics and 5. regulatory metrics in this thesis. However, in this thesis project, we focused on items 2, 4 and 5, whereas the rest was only developed and added to the appendix.

For the three items on which we focused, we created a total of 44 metrics. 25 machine learning specific metrics, 13 business metrics and 6 regulatory metrics. Of the 44 metrics we did not implement 6, of which 3 did not provide sufficient benefit and 3 were too complex for the time-frame of the thesis. Later on, many of these metrics turned out to be extremely valuable, from which we finally picked out 6 Key-Metrics, that we found to be essential.

Conclusion In the end, we have shown that our proposed monitoring approach allows simple and effective monitoring of machine learning applications in operation. We also believe that our concept is transferable to further machine learning projects.

Contents

Chapter 1

Introduction

In our first chapter we introduce the topic of multi-aspect monitoring of machine learning models in operation by starting with explaining our motivation and why we see an opportunity to improve. Afterwards, we present the objectives we want to achieve during this thesis project. Next, we present four research questions which we will focus on in this thesis and introduce the approach we will use to answer these questions. Finally, we present an overview over the different thesis chapters to guide the reader through the thesis paper.

1.1 Motivation and problem statement

Machine learning applications are powerful tools that companies and researchers alike often use to solve complex problems. Especially in companies the ambition and the usage of machine learning applications in critical business areas is rising every year.

This development can be seen by comparing the yearly studies of the IDG Research Services concerning the current usage of machine learning and deep learning in German companies [25] [26] [27]. It can be observed that from year to year there is a constant increase in the number of companies using machine learning. Moreover, there is also a constant increase in the interest of companies to engage in machine learning. The multitude of topics that these companies work on or would like to work on using machine learning are topics such as improving internal processes, improving customer relations, optimizing manufacturing processes and many others. A lot of these topics are crucial for the business.

In addition to such market analyses, this tendency is also supported by other sources. In 2016, for example, a study shown in Figure 1.1 was conducted in cooperation between Statista and OMDIA | TRACTICA, which shows estimated revenues from artificial intelligence for enterprise applications market worldwide

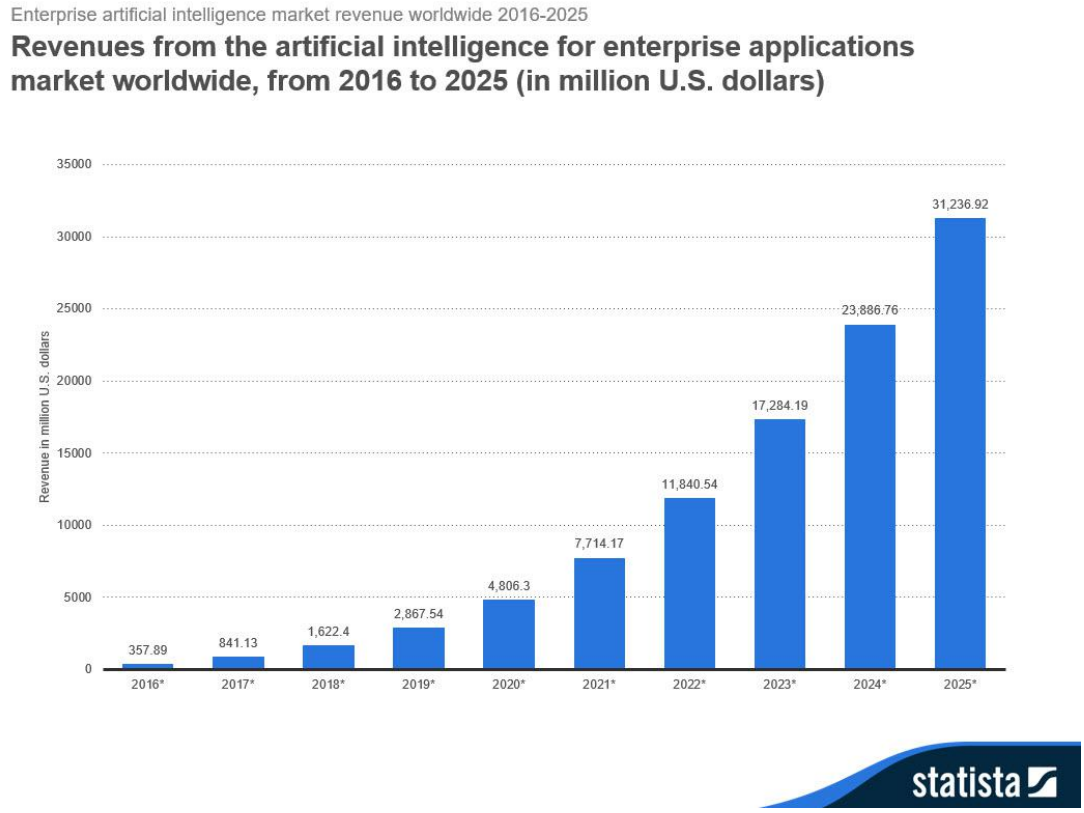


Figure 1.1: Revenues from the artificial intelligence for enterprise applications market worldwide, from 2016 to 2025 [30]

from 2016 to 2025 on basis of the then current worth of the market in 2016. An exponential increase is clearly visible here.

The situation is similar for the study presented in Figure 1.2, which is about global funding for AI startups from 2015 to 2019. Again, there is a clear increase, although not as strong as in the previous statistic. However, it proves that this topic is gaining in importance every year.

This increase in the importance of the area around AI and the topics involved generates a need for controllability. Especially business-critical areas or generally areas where a lot of money is invested, mistakes often have expensive consequences. Therefore, it must be ensured that the machine learning model used always meets the highest quality standards.

This assurance begins with the conception of the model and must be carried out right through to operation. However, monitoring does not stop here. Whereas traditional software delivers a static product that still delivers the same results years later, machine learning models react to changes in the environment. This

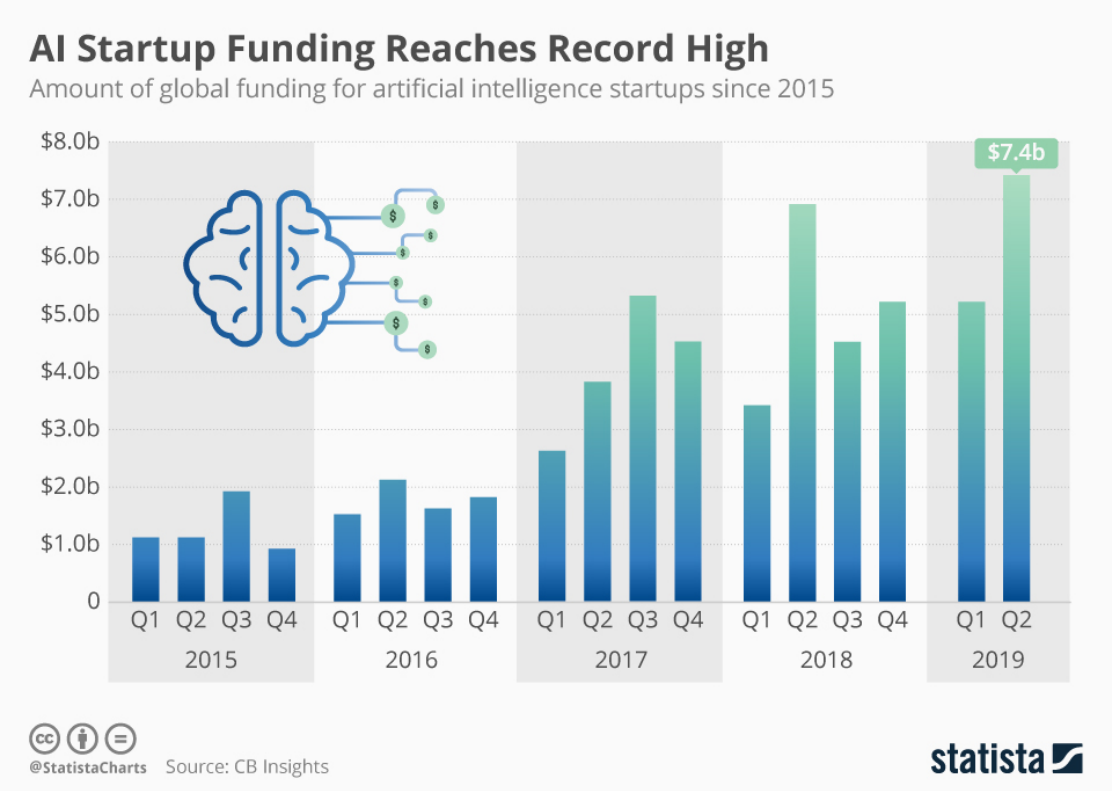


Figure 1.2: AI Startup Funding Reaches Record High [29]

means that active monitoring must be carried out in order to be able to react quickly to quality-reducing events.

The SIGNAL IDUNA Group, with which we are working together on this thesis project, also uses machine learning and must ensure that their models and applications, whether internal or external, always meet the highest quality standards. The integration of monitoring machine learning models in operation, however, at the SIGNAL IDUNA Group is in its initial phase and not yet widely available.

The SIGNAL IDUNA Group is a German insurance company with headquarters in Dortmund and Hamburg. The context of this thesis is placed, due to the cooperation with SIGNAL IDUNA, in the German insurance sector and is therefore particularly interesting for regulatory requirements by laws of the EU and the Bundesrepublik Deutschland. Together, we work on the thesis topic by using the software “Stornoprophylaxe”, which tries to predict cancellations of risk accident insurances. Particularly difficult is that we have a large time offset of 3 months between prediction and prediction result. Here we will put a special focus to be able to assess at the present time how well the machine learning model is performing without having the results or getting them in near time. Three months is a

long period.

Next to that, monitoring of machine learning models in operation appears to be not broadly used in the industry. This can be deduced from our experiences and discussions within and outside the company as well as from the scientific literature, which does not provide a large database in this area. Especially research about machine learning pipelines often stops after the artifact is in operation. Only blog entries, which are considered as gray literature, deal increasingly with this topic. However, there is less reporting on sophisticated concepts, but more shared experiences. Accordingly, the theory of monitoring is missing here.

As already mentioned, the topic of monitoring machine learning models in operation is not much discussed in the scientific literature at this time. Currently, such papers seem to focus on the development of more effective and efficient algorithms and on the maintenance and monitoring of machine learning pipelines. This may be due to the fact that some application scenarios require a frequent retraining of the model like the machine learning model clusters of Uber Eats [19]. Therefore, monitoring of proper predictions during operation may not be essential, since the probability of the model deteriorating over these short periods of time is very low. However, this only applies to some scenarios, whereas other scenarios do not need retraining until the model quality decreases. Here monitoring becomes essential again. Another possibility could be, that this area is considered to be of secondary importance after the machine learning model pipelines and that the focus here could emerge in the coming years.

Nevertheless, we consider monitoring to be highly important also due to the debates around responsible AI. For instance, the high-level expert group on AI (AIHLEG) released a report through the European Commission about ethics guidelines for trustworthy AI where they describe how an AI software should behave throughout its whole lifecycle to be considered as trustworthy [16]. Although Commission staff facilitated the preparation of this document, the views expressed herein reflect the opinion of the AIHLEG and is not reflecting an official position of the European Commission. Nevertheless, trustworthy AI has three components defined by the AIHLEG:

1. It should be lawful, complying with all applicable laws and regulations.
2. It should be ethical, ensuring adherence to ethical principles and values.
3. It should be robust, both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm.

For those presented reasons, we believe that there is almost always a necessity for monitoring machine learning models especially after deployment. With this thesis we would like to present a concept that is generally transferable and can be

used by other parties. Nevertheless, we also want to point out that this topic is huge and probably cannot be covered in this thesis project only.

As mentioned earlier, the context of this thesis project will be provided by the German insurance sector. Here, especially many regulations apply. Present are a lot of regulations regarding data protection, data security and other topics, which often does not allow those companies in that sector to achieve a high degree of agility. Alongside this, they must ensure that all regulations are followed at all time, which declines the room for agility again. This includes, for example, the handling of personal data, discrimination, respect for privacy or transparency in decision-making. This makes monitoring the quality and the behavior of our published products even more important.

Next to this, at SIGNAL IDUNA many different groups of people are involved in machine learning projects. Each of these stakeholder groups have their own requirements concerning monitoring machine learning models. Naturally, we will try to cover all specifics of all groups while developing metrics for monitoring machine learning models. A more detailed explanation of which groups of people those are, is covered in Chapter 3 Design Objectives.

In the end, Ori Cohen summarizes it well in a blog entry on Towards Data Science, in which he describes that we want a system that knows how to monitor and generates alerts when it detects failures or predict future failures ahead of time [7].

1.2 Research objectives

During the work on the thesis project, we focused on various goals. On the one hand the design objectives, covered in Chapter 3, but also the research objectives. Our research objectives are entirely about the adaptation and development of metrics for machine learning models that are in operation. We distinguish between classical metrics, which are already known from training models, and extended metrics, which are only applicable in operation. The objective is to identify known metrics that are a good source of information for our use case, but also to develop complementary metrics that can provide further insights into the quality level of the model while operating in a live environment.

To identify and develop metrics we use the Goal Question Metrics approach [2], which is presented and used in Chapter 4. In parallel, we also assign appropriate thresholds to these metrics to define when metrics imply a needed intervention. Afterwards we are expecting to have identified suitable metrics and to have assigned reasonable thresholds to these metrics, which will send a notification if they are breached.

Afterwards we evaluate the performance of our metrics based on different sce-

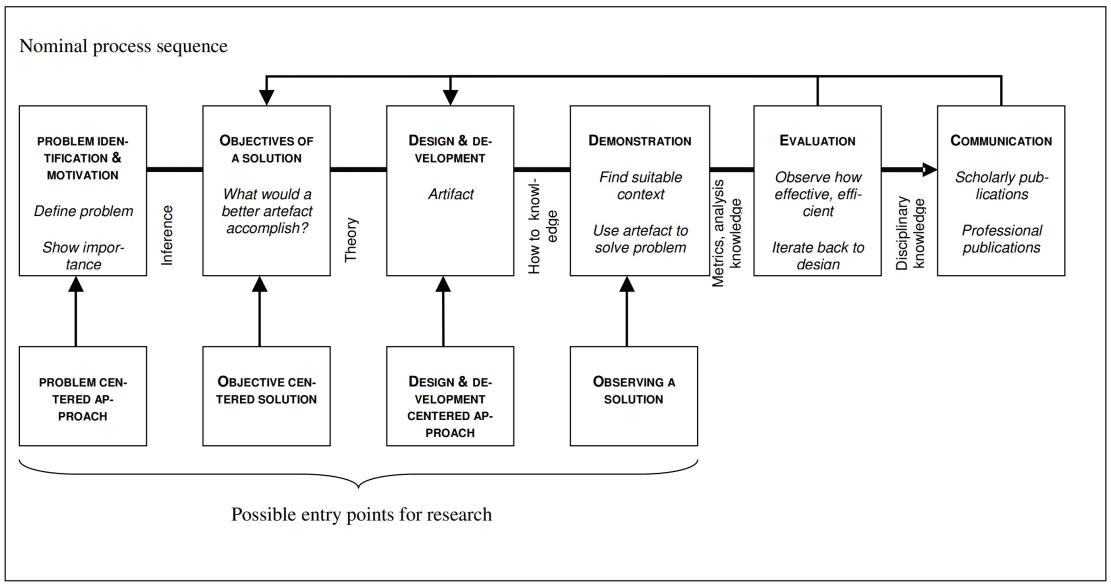


Figure 1.3: Design science research process (DSRP) model [21]

narios, which are performed in Chapter 5. We hope to gain insights which metrics generate the most value and can most accurately determine whether the model deteriorates or will deteriorate. In addition to confirming our selection of metrics and thresholds, we hope to gain insights into the practicability of our monitoring concepts. After all, a complicated, complex and correspondingly impracticable concept can be as good as it may be, if it is not used or cannot be used due to its impracticability. The general applicability of our developed concepts is tried to be assured.

As can be seen from the description of our research objectives, we have chosen an applied research approach to work on this thesis project. This approach is adapted from the design science research process (DSRP) model from the research paper written by Peffers et al. [21]. In this model, a theoretical hypothesis is replaced by an objective. The model consists of six activities which can be seen in Figure 1.3.

In the first activity, the problem is identified and the motivation for the solution is presented. Afterwards, it is described why a new solution would be better than the existing ones. Subsequently, an artifact is created for the new solution in the "Design & Development" activity. This artifact is then used to demonstrate how it operates and solves the reported problem. The fifth phase evaluates how effective and efficient the developed artifact is. Finally, a discussion is held in the form of a conclusion.

We have almost completely adopted these six activities, which is also reflected

in the structure of our thesis. However, the activities have been adapted to our use case.

1.3 Research questions

As described in the previous section we hope to gain several insights about how to monitor machine learning models in operation best in combination with our applied research process model. With that in mind we came up with four key research questions, which we will answer individually in Chapter 7.1.

RQ.1 What features are necessary for a monitoring tool for machine learning models in operation?

RQ.2 What defines a good metric for assessing the quality of machine learning models in operation?

RQ.3 Which of those good metrics are best suited to evaluate a machine learning model in operation?

RQ.4 What are possible thresholds where metrics indicate a necessary change of a machine learning model?

With these four research questions we try to answer fundamental questions about monitoring machine learning models in operation. We start with the monitoring tool, where we try to identify what features are needed to serve the largest possible user group, since a tool will only be used if it is well structured and does not overwhelm the user. After that we want to define appropriate metrics for this tool to monitor machine learning models in operation. To do this, however, we first need to define what a good metric is and what it is characterized by. Then we want to create and evaluate the metrics based on the found requirements. After developing these metrics, we only have to find appropriate thresholds where these metrics indicate a necessity for maintenance.

1.4 Thesis overview

After this introductory chapter, the thesis is divided into six further chapters. In Chapter 2, we will discuss related work and explain the topic of monitoring around machine learning in operation and machine learning pipelines in detail. In Chapter 3 we describe the objectives we have determined for the final result and which components should be included. We also explain why these components are important. Subsequently, in Chapter 4, a sophisticated monitoring concept

is designed and developed, which is then demonstrated in Chapter 5 and reflects the results of our work. After the demonstration we evaluate our implemented concept in Chapter 6 and show how effective and efficient our concept performs, which metrics turned out to be most valuable and how the final product relates to the design objectives. Finally, in Chapter 7 we summarize our work and findings, answer the research questions, point out the contributions and discuss possible further work.

Chapter 2

Related Work

In this section we will present some of previous and similar work in the field of monitoring machine learning models in operation and monitoring machine learning pipelines and how they relate to our research. Doing so, we will focus on the questions "What did others do to solve our problem?" and "Why do we need to improve and change them?".

2.1 Monitoring of ML-Models in operation

Monitoring machine learning models in operation is a major challenge. In contrast to the training of a model, where we can directly compare the predictions with the outcomes, we always have a time lag with operational models, depending on the use case of the model. This can range from a few minutes to days or months. In our case it is even three months, as we will describe later. In addition, the external environment of the model is in constant flux. This is also different in training. All this makes it particularly difficult to make an accurate assessment of the current model quality.

Therefore, we have tried to find sources that deal exactly with this topic. However, monitoring machine learning models in operation is not mentioned a lot in research papers as main topic. It seems that research focuses more on how to build a machine learning application with pipeline support and why it is necessary to monitor operational applications, instead of how to monitor it.

An example is the research paper "Hidden Technical Debt in Machine Learning Systems" [24]. This research paper deals with technical debt while developing ML-Applications. Technical debt is a metaphor introduced by Ward Cunningham in 1992 to help reason about the long term costs incurred by moving quickly in software engineering, whereas here the researchers apply it onto machine learning

model engineering. Doing so, they state that both, known and unknown (hidden) debt need to be serviced. Within the paper, it is extensively described which technical debt could occur while training and how to mitigate it. That includes issues like “feedback loops”, “glue code”, “dead experimental codepaths” and similar. Next to issues that usually only occur in developing and training the model, the paper also states that comprehensive live monitoring of system behavior in real time combined with an automated response is critical for long-term system reliability. Besides, they also mention, that external changes occur in real time and thus the response to those changes has to be in real time as well. Relying on human intervention can be brittle for time sensitive issues. However, while mentioning how important and brittle monitoring operational machine learning models can be, they only propose three metrics that should be checked:

1. prediction bias
2. action limits
3. up-stream producers

This is a start, but is not comprehensive enough to monitor a whole machine learning application, its surroundings, and its specifics. There is a lack of measurements that informs us if the model still works as intended and how well it performs. Moreover, the proposed monitoring metrics still needs humans in the loop. For example, the prediction bias might indicate that there is something wrong with the model, but to confirm that, one still needs to check that. It could also be an expected shift. This interaction can be further improved so that human investigations are reduced as much as possible.

Another paper from Google [5] summarizes that testing and monitoring are important strategies for improving reliability, reducing technical debt, and lowering long-term maintenance cost. As consequence, it is crucial to know that your machine learning system continues to work correctly over time. Doing so, they propose multiple monitoring tests for machine learning applications like detect data invariants in training and serving inputs or the age of the deployed model. In total they present seven different monitoring tests:

1. Dependency changes result in notification
2. Data invariants hold in training and serving inputs
3. Training and serving features compute the same values
4. Models are not too stale

5. The model is numerically stable
6. The model has not experienced a dramatic or slow-leak regressions in training speed, serving latency, throughput, or RAM usage
7. The model has not experienced a regression in prediction quality on served data

One might think now that everything is covered because of the large number of diversified monitoring approaches. In fact, they covered the model itself, its environment, and its dependencies. Moreover, they also addressed the issue of getting the result labels of the predictions not shortly after serving them. However, if we take a more detailed look at monitor test number seven it can be recognized, that this metric only indicates that the model quality is decreasing and that there might be a problem. Therefore, it is way more concrete than the previously presented paper, but still not sophisticated enough to be sure that there is a problem, which needs to be handled. This is one one the points we try to address. We will try to give notifications not only that something changed, but also that an issue occurred.

Besides the few interesting research papers we found, we also would like to present how developers without a scientific background are introduced to monitoring machine learning applications in operation. For this we would like to point out that the following presented documentation is gray literature and no scientific claim was pursued. Nevertheless, it is worth a look at the developer page of Google, where (future) developers are advised into this specific topic [12]. It becomes clear relatively quickly, as this is an introduction guide, that this topic has been presented in a significantly reduced form and is limited to the most important metrics determined by Google. However, it is still possible to get an impression of what important metrics could be. They propose to track the training-serving skew, meaning the difference for the input data between training and serving, dividing between schema and feature skew. Next to this, they propose to monitor the age of the model, since the serving data evolves with time but if your model is not retrained regularly, then it is most likely that the model quality is declining. Besides machine learning specific monitoring, they also recommend to monitor the general performance of the model by versions of code, model and data, API response times, number of queries answered per second and similar. It can be stated that the metrics presented can be found in the previously presented papers as well. Furthermore, it is clear that this limited collection of metrics is not sufficient to monitor a complex machine learning application. Nevertheless, we believe that the presented metrics are a start for this thesis project.

We now summed up some literature which deals with the problem we are addressing in this paper. Nonetheless, the proposed methods within those papers leave room for improvement. In particular, none of the papers describes accurate enough how to deal with models where the outcome of the predictions becomes known much later and thus the evaluability of the standard metrics loses significance, which is the case for our test model "Stornoprohylaxe". As already mentioned in the introduction we want a system that knows how to monitor and generates alerts when it detects failures or predict future failures ahead of time [7]. This is one of our objectives, which is given a high priority in our project.

2.2 Monitoring of ML-Pipelines

Of course, it is not only important to monitor and understand what happens in operation. It is equally important to be beware of what happens within the process of developing and training a model. Accordingly, it is equally important to monitor and understand what happens from data collection to deployment.

The first step to successfully run a machine learning application in operation is to develop it. For this purpose, a machine learning pipeline is usually created, which, in a nutshell, receives the test data, builds the model and finally deploys it into operation. This model has to fulfill predefined quality standards, which vary from model to model. Furthermore, it must be ensured that the machine learning pipeline operates reliably and faultlessly at all times. All this must or should be monitored according to the principle identifying and resolving errors in early project phases saves time and money. This is called error cost escalation through the project life cycle.

Typical errors that can occur in the development phases and lead to a reduction in model quality have already been discussed in detail in the Google research paper "Hidden Technical Debt in Machine Learning Systems" [24]. Among others this includes unstable data dependencies, hidden feedback loops, glue code, pipeline jungles and more. But most important is that not all debt is bad, but all debt needs to be serviced. Especially hidden debt is dangerous because it compounds silently.

Another paper of Google is dealing with this kind of issue [3]. They are attempting to develop a machine learning platform that simplifies the development and construction of machine learning applications and increases reliability and maintainability. In doing so, they deal with the following phases, which are additionally covered by tests to ensure that the desired result is achieved:

1. Data analysis, transformation and validation

2. Model training

3. Model evaluation and validation

4. Model serving

They especially focus on the monitoring of the data. This is the most detailed part and the most important part in model development as machine learning models are only as good as their training data.

A similar approach has been followed by IBM [17]. They present a cloud-based framework and platform for end-to-end development and lifecycle management of artificial intelligence (AI) applications which also includes machine learning applications. Doing so they use the principles of software lifecycle management to enable automation, trust, reliability, traceability, quality control, and reproducibility of AI / ML pipelines.

First, they present multiple use cases, which they use to validate their work, and challenges, which must be serviced while developing within an AI pipeline. Here they also state the necessity to monitor specific behaviors to ensure the quality of the resulting model. Moreover, they must be easy to plug in or plug out, depending on the use cases.

Afterwards, they discuss the system design, which however does not focus on the specification of the monitoring. Only a brief description is given of why it is necessary and how it could be implemented. Accordingly abstract are the metrics, which are for example “drop in performance” or “a bias detection algorithms detected a bias”. Likewise, it is only mentioned that event triggers are useful, but not how they should be formulated.

At last the abstract implementation is described.

Chapter 3

Design Objectives

In this chapter we describe the objectives we want to achieve by creating our artifact, which are metrics for monitoring that are embedded in a monitoring tool we create. We distinguish between general objectives and SIGNAL IDUNA specific objectives. The latter focuses on objectives that are only applicable to SIGNAL IDUNA rather then the insurance industry or to the development of artificial intelligence in general. The general objectives, on the other hand, are intended to be transferable to companies in the insurance business and possibly also partially to the general development of artificial intelligence. We classified the general objectives into business, data protection and technical objectives.

3.1 Preparations and Stakeholder Analysis

First, we conducted an extensive literature research, presented in Chapter 2, to discover what approaches for machine learning monitoring in operation have already been developed and how they have been implemented. In addition, we reviewed literature that deals with similar aspects, such as monitoring of machine learning pipelines, in order to use or adapt techniques for the monitoring of machine learning models in operation.

After we had gathered initial insights, we wanted to focus more on target groups in order to verify whether our insights also reflect the general usefulness for different target groups.

For this purpose, we held four interviews with employees of the cloud and machine learning department at SIGNAL IDUNA, which each took around 30 minutes. No protocols were included in the appendix due to confidential and corporate information the interviewed gave. Since we conducted individual interviews, there was no fixed structure. Nevertheless, we followed the following procedure. This procedure we prepared ourselves before the interviews.

First, we briefly introduce what the topic and goal of the thesis project is and with what help we will evaluate our outcomes. This includes, among other things, the usage of the use case "Stornoprophylaxe". After that, we briefly address the scope of the project, where it will be conducted and what are limitations. Next, we asked general questions how the people interviewed could imagine a monitoring tool for "Stornoprophylaxe" and which metrics and functions should be included. Finally, we gave the opportunity to ask questions that we answered the rest of the time.

In addition, we conducted a survey directed at SIGNAL IDUNA employees and at students to verify our initial insights with further input. For this survey we used the survey tool Qualtrix Core XM [22], which is provided to students in cooperation with Leiden University. The survey was available for filling in between 16$^{\text{th}}$ March 2020 and 10$^{\text{th}}$ April 2020 before it was closed and analyzed by us. It was distributed by us through direct messages on social media and by mail.

The initial part of the survey dealt with general questions relating to metric creation, but also to the development of a monitoring tool. Accordingly, this survey aimed to reveal the conditions under which people with expertise in the field of machine learning would use such a monitoring tool. It was therefore an exploratory survey. In total we obtained 26 usable responses, of which 16 responses came from our employees at SIGNAL IDUNA and 10 from students at Leiden University. As unusable we classified responses, which were only 50% or less completed or obviously could not be considered as a thematically suitable answer. In the following subsections, the results are presented and a full report of the survey can be found in the appendix.

During the analysis of the survey, we first grouped the responses of the respondents accordingly to their job group and analyzed them separately. Subsequently, we considered all responses without grouping and finally merged them with several overlapping responses into a table which was inserted as a figure (Figure 3.1).

First of all, it is important to mention that most of the participants in the survey answered all multiple choice questions, but for the concrete free text questions on necessary metrics and on expectations of a monitoring tool, about 60% of the text fields were left empty or filled with standard metrics already known from the training of machine learning models.

This led to the assumption that all participants of this survey consider monitoring of machine learning models in operation to be important, but are not yet able to determine how it could be implemented in practice. In addition to this important insight, we were also able to derive various target groups, which in some cases can only be found in the insurance sector.

Again, a distinction must be made between target groups of SIGNAL IDUNA, which will also be present in other insurance companies, and overall target groups

in the area of machine learning. During the implementation of projects with the aim of an effective and efficient monitoring of machine learning model in operation, a wide range of people in insurance companies deal with the target product.

1. Managers
2. Actuaries
3. Compliance officers
4. Data scientists & Data analysts
5. Software engineers & Machine learning engineers

On the one hand there is the manager perspective. These are usually project managers who have a more general view of the projects and are less interested in the technical details. For such persons it is more important that the product works without problems and fulfills its desired benefits. These would be metrics like prediction accuracy or the number of contracts saved.

Next to the managers perspective, there are also the actuaries. As scientifically educated experts in the insurance industry, they deal with the composition of insurance contracts based on mathematical and statistical methods. As a result, they also deal with the area of machine learning at SIGNAL IDUNA and are involved in technical projects. Due to their mathematical background, they are therefore more interested in mathematical metrics that describe the current technical quality of the model.

A further target group is formed by the compliance officers, who, as the name suggests, are responsible for compliance. Accordingly, regulatory requirements and the adherence to them by the model are most relevant for monitoring. Sometimes the regulations are very complex and difficult to evaluate. At this stage no specific measurements surfaced and we will look deeper into this in Section 4.2.2.

Subsequently, the target group of data scientists and data analysts was identified. We defined these groups of people as one target group because they have partially overlapping activities and their focus is mainly on data. Therefore, their focus in a monitoring tool for machine learning models in operation is on the data basis and how good the data quality still is or how the data influences the model.

As the last target group we grouped the software engineers and the machine learning engineers. Our experience shows that for both job-profiles activities overlap strongly as well and it is consequently a valid assumption that both groups of people want metrics about the prediction quality of the model. This is also reflected in our survey.

Mapping between target groups and design objectives		Managers	Actuaries	Compliance officers	Data scientists & Data analysts	Software engineers & Machine learning engineers
Business objectives	BO-1	X				
	BO-2	X				
	BO-3	X				
Data protection objectives	DPO-1			X	X	X
	DPO-2			X		
	DPO-3	X		X		
Technical objectives	TO-1	X	X	X	X	X
	TO-2	X	X		X	X
	TO-3		X		X	
	TO-4	X		X		X
	TO-5	X	X	X	X	X
	TO-6	X		X	X	
	TO-7	X	X	X	X	X
	TO-8			X	X	X
	TO-9		X		X	X
	TO-10	X	X	X	X	X
SIGNAL IDUNA specific objectives	SISO-1	X				
	SISO-2				X	X
	SISO-3		X		X	X
	SISO-4		X		X	

Figure 3.1: Mapping between target groups and design objectives

3.2 General objectives

In the following three subsections we describe different goals that can be assigned to different target groups. On the one hand, there are the business objectives, which are the most relevant for the management perspective. Secondly, there are the data protection objectives, which are important for compliance managers. Finally the technical objectives for the remaining target groups, under which we have sorted data aspects as well as machine learning aspects and surrounding areas.

In addition to specific objectives, however, there are also general objectives that are desired by all groups equally. Still, we divided the objectives into three different groups, but this does not mean that the different target groups can be mapped directly onto them. For example, project managers have, in addition to business objectives, also technical objectives that might be relevant for them.

3.2.1 Business objectives

With the business objectives we primarily include the objectives of the project managers. The reason for this is that it is important, especially for this group, how well the model fulfills the expected benefits. The business objectives can be seen in Table 3.1.

Since our application has the purpose of predicting whether customers would cancel their accident insurance contract within the next three months, the objectives will be based on this use case, but are nevertheless transferable to other applications.

Both, the survey and the interviews, gave the impression that the most relevant business objectives are how well the model performs and thereby, as a result, how much monetary value has been saved. This is applicable to additional insurance premiums for SIGNAL IDUNA and hence also monthly stock commissions for the field service partners (ADPs) by the contracts whose cancellation was prevented, which were predicted to cancel by the application. Besides the monetary benefits the model could bring, another objective emerged.

As well as preventing cancellations, the insurance stock also plays an important role. Therefore, it is also a very important objective to have the largest possible insurance stock in order to have the largest percentage of the market compared to the competitors. It is not always crucial that you earn the most money, as market dominance is at stake.

There are also other business objectives, but these are not specifically applied to our model. For example, if a customer cannot be prevented from canceling his contract, another insurance may become relevant. An example would be the retirement age. The probability that a person will need an accident insurance at the age of retirement is strongly reduced, since accident insurance is mainly focused on self-employed persons or freelancers who no longer perform their previous activities when they retire. In this case, offering a possible alternative insurance in order to avoid losing the customer would also be a possible option, but this is not intended for this use case since we are focusing on saving the contract and not replacing it. That might be a future project to enhance the accident churn prediction model.

3.2.2 Data protection objectives

Another important objective of this thesis project is the monitoring of model activities within the limits of laws and regulations. This is ensured in insurance companies and other companies equally by the compliance department or the compliance manager. The list of data protection objectives can be found in Table 3.2.

Especially in Europe, the GDPR (General Data Protection Regulation) re-

Business Objectives	
BO-1: Maximize insurance premiums	The objective is to maximize the overall contract premiums by preventing contract cancellations. This is done by intervening when our model predicts positively, trying to proactively convince the customer not to cancel the contract.
BO-2: Maximize stock commission	This objective is similar to the upper one, but here the stock commission of the service field partners is maximized. This is also achieved by preventing cancellations.
BO-3: Maximize contract stock size	Here the objective aims to maximize the size of the contract stock, whereas the market share is maximized as well, which sometimes may be more important then earning the most.

Table 3.1: Identified business objectives for the monitoring artifact.

stricts the use of personal data, which is necessary for machine learning models in the insurance industry almost everywhere. However, there are also other regulations that deal, for example, with discrimination and transparency of decision-making processes. In addition, the GDV (Gesamtverband der Deutschen Versicherungswirtschaft, which translates to German Insurance Association) stated in its presentation "Künstliche Intelligenz in der Versicherungswirtschaft" (Artificial Intelligence in the Insurance Industry) [10] from 31-10-2019, that due to legal regulations the processing of raw data is halved again. Moreover, they have found a total of twelve different regulatory documents that limit the use of AI in the insurance industry [10]. Even if these twelve documents are initially limited to the German market, the GDPR, for example, is already valid throughout Europe.

Due to this large number of limitations, when using artificial intelligence and in our case machine learning, it is necessary to take a very careful check whether the machine learning application adheres to all regulations. In addition to the enormous reputational damage, a high fine would also be the consequence of violating the regulations, regardless of whether it was done intentionally or accidentally. For this reason, compliance managers at SIGNAL IDUNA and, in high likelihood, at many other companies are very interested in having a monitoring tool to support them in their work. Consequently, the main objective here will be to make as much information as possible available to compliance managers in the form of metrics, be it to make the model's decision-making process more transparent or to check the decision for discriminatory behavior.

Data Protection Objectives	
DPO-1: Revealing of information	In this objective we try to reveal as much information as possible about how the model operates. This includes workflows as well as prediction decisions.
DPO-2: Support of the compliance work	Here, an attempt is made to support the compliance managers in their work on the basis of the additional information revealed previously. For this purpose, the revealed information must be presented accordingly.
DPO-3: Adherence to all regulations	All regulations should be monitored, partly automated partly semi-automated, by our monitoring artifact.

Table 3.2: Identified data protection objectives for the monitoring artifact.

3.2.3 Technical objectives

Since this thesis project is supervised by the IT department of SIGNAL IDUNA, the technical objectives have a majority part in the developments, whereas Table 3.3 contains a listing of all technical objectives.

As it can be derived from the survey most people want to have monitoring of machine learning models in operation to ensure the quality of the model and, next to this, see the current status of the machine learning model. Furthermore, 15 respondents would like to have detailed information about any metrics of the machine learning model. What that means in detail, with respect to the different target groups, stays open for now. Another objective would be that the monitoring tool should provide a clear and accurate summary of the current status. To combine this with a detailed view of all metrics possible, will be a challenge.

In addition, the software should be easy to use. This means that, in parallel to a simple and straightforward usability, it should also be simple to install, preferably according to an "install & go" principle. Subsequently, standard views must be created that meet the expectations of all target groups. Furthermore, there should be help menus in the form of "?" buttons that explain the metrics displayed. This should make it easier to get started with the monitoring tool and also provide information to groups of people, who are not familiar with machine learning. However, what was not often mentioned in the survey is the time offset when displaying the metrics. Only for two respondents a larger time offset was not acceptable. The updates for the metric values should be presented in short term, but it remains unclear what short term defines. However, it became quite clear that when thresholds are exceeded, a notification should only be sent to the

relevant target group for the respective metrics and not to everyone. Especially compliance managers are usually less interested in the performance of the machine learning model. Moreover, anomalies in the model or the data used should be detected and possible improvements should be suggested.

Most responses were relatively widely distributed concerning how metrics should be presented in the tool. Suggested were a traffic light system to show the status as fast as possible, a scale system to show in as much detail as possible how well the current metric is assessed and a combination of both. Because of the even distribution, we plan to implement a combination if this seems reasonable.

Which metrics will eventually be used by us and which ones we still have to define ourselves will finally be determined under section Conception. However, metrics that are mostly known from the training phase of machine learning models were mentioned a lot. These include accuracy, specificity, root mean square error, model certainty, data-drifts, precision and recall, area under curve and more. In our use case, where there is a time lag of three months between the prediction of the model and the evaluation of the prediction, we have to consider how we still create or use accurate metrics to describe the current state without prediction results.

In summary, we would like to develop a monitoring tool that is easy to use, immediately operational, meets the needs of its users, provides sufficiently accurate and detailed metrics, sends notifications to responsible target groups and reflects both the current status and possible future developments.

3.3 SIGNAL IDUNA specific objectives

In the last section of this chapter we present additional objectives, which were designed especially for SIGNAL IDUNA and are not directly transferable to other companies. These are objectives that are directly related to their software for the prevention of cancellations of accident insurance contracts and therefore only applicable for SIGNAL IDUNA. They are summarized in Table 3.4. Sometimes there may be objectives that are transferable to other use cases. However, the objectives which are defined here do not claim to be transferable.

In the specification phase of the “Stornoprophylaxe” (SP) (software for predicting possible insurance cancellations) the developing team at SIGNAL IDUNA chose to predict 10% of all terminated insurance contracts. In total, they have an average probability of 5% that a contract will be canceled in the case of accident insurances. As a result, we get a percentage of "saved" contracts that should be reached annually. These measured values should also be monitored. Another goal was to have as many true positive predictions as possible and at the same time as few false negative predictions as possible and therefore maximize the recall.

Technical Objectives	
TO-1: Ensuring the model quality	In any case, the reduction of the model quality should be prevented and, if possible, an increase in quality should be achieved.
TO-2: Current status	An overall status of the model should be available.
TO-3: Detailed information	It should be possible to obtain detailed information about the model. This information may vary for the respective target group.
TO-4: Ease of use	A simple and straightforward usability should be present, as well as a simple installation according to an "install & go" principle.
TO-5: Standard views	Standard views should be provided that meet the expectations of all target groups or at least is approximating to these.
TO-6: Metric help	Help menus in the form of "?" buttons that explain the metrics displayed should be present.
TO-7: Notifications	A notification should only be sent to the relevant target group for the respective metrics and not to everyone when a threshold is exceeded.
TO-8: Anomaly detection	A detection of anomalies in the model or in the (input) data should be present.
TO-9: Metric presentation	Metrics should be presented in a appropriate scale with additional traffic light if it is adding additional benefits for a faster evaluation.
TO-10: Time offset for metrics	Due to our time lag of three months between the prediction of the model and the evaluation of the prediction, we need to create metrics which describe the current state of the model without the prediction results.

Table 3.3: Identified technical objectives for the monitoring artifact.

SIGNAL IDUNA Specific Objectives	
SISO-1: Percentage of saved contracts	As specified, 10% of the annually cancellations should be determined and prevented. The annual cancellation amount is on average around 5% of the stock size.
SISO-2: Maximize the recall	The team aimed to get as many true positive predictions as possible and at the same time as few false negative predictions as possible. –> Recall = TP / (TP + FN)
SISO-3: Adjustments causing data drifts	Detect possible data drifts which are caused by new contract designs or adjustments to insurance conditions.
SISO-4: Demographic changes	It is desirable to detect demographic changes before they affect the model forecasts.

Table 3.4: Identified SIGNAL IDUNA specific objectives for the monitoring artifact.

In addition to key indicators that need to be monitored, it is also necessary to identify data drifts due to new contract designs or adjustments to insurance conditions. Furthermore, it would also be desirable to see demographic changes in the data before they affect the model or the forecasts. These would include, for example, increased retirement ages or developments in insured sectors that would make working conditions much more secure.

Chapter 4

Design and development

In this chapter we describe the entire development process of our metrics and how they would be integrated in the system as monitoring tool. Starting with a section about preconditions, which have to be clarified before developing and creating the metrics. Afterwards, we present the preparation we performed before developing in Section 4.2. Then we describe in Section 4.3 the conceptual design and the concrete integration in the given application. This includes concept decisions, technical designs, metric designs and finally the review of the design.

As our main focus is on the creation of suitable metrics, this is the main part of this thesis. In addition to the metrics, a monitoring tool was developed as a prototype with which we validate the metrics. We limit the description of the prototype and do not discuss it extensively, as we mainly use it to verifying our metrics.

4.1 Preconditions

Before we could start preparing the development, we defined four general preconditions that must be followed during implementation or must be present before implementation. On the one hand, this is intended to ensure a good quality standard and, on the other, to maximize user acceptance.

1. Data pipeline must be present
2. Machine learning pipeline must be present
3. Direct intervenability of the model
4. Employee availability to respond to notifications

Data pipeline As we develop a monitoring tool that sends notifications as soon as defined thresholds of the model metrics are passed, we have to make sure that one can react appropriately based on the notifications. This also includes retraining the model. Accordingly, it must be ensured that we have suitable data for the training at all times or can obtain it quickly and at short notice. For this we need a data pipeline. That must continuously receive data from production and process it. For example, personal data must be deleted / anonymized / pseudonymized or corrupt data must be fixed / removed. If special data is required, for example a fast data drift occurred, the pipeline must be able to provide customized data based on parameters.

Machine learning pipeline In addition, we also need a machine learning pipeline so that we can train, test and deploy the model on demand. The pipeline must as well be monitored for quality purposes. Whether automated or with a human in the loop must be decided by the respective development team. However, it must be ensured that no worse model is being deployed. In addition, it must be well structured and efficiently designed in order to be usable for everyone and to ensure that the DevOps process from data collecting to operation does not take too long.

Direct intervenability A further precondition is the ability to intervene in the model application. This is necessary because as soon as anomalies or rapid quality losses are detected, the application must be stoppable in order to avoid generating too many false predictions and to act accordingly. Besides the possibility of a stop, an additional roll-back is also conceivable. This depends on the functioning of the machine learning application. In our case of "Stornoprophylaxe" a stop of the application would be sufficient, as those false warnings would be sent to the field service partners after all predictions. Therefore a roll-back is not needed.

Employee availability During business hours, certain persons must be available all the time in order to be able to process alerts immediately by the monitoring tool. Since this alert already causes a reactive rather than a proactive activity, it must be processed quickly. It is crucial that the quality of the model is monitored and can be reacted to quickly if necessary. However, this also requires the creation of capacities in the workforce.

4.2 Preparations

After clarifying the requirements we started with the preparation for the tool implementation. In the process, we obtained authorizations for systems in use, set up our development environment, requested test data, dealt with regulatory issues

(see Section 4.2.2), received information about the development of the model and established a group of experts for various topics, which we could ask for advise and clarifications.

Permission preparations:

- Obtain access to the model's git repository.
- Obtain access to the Jenkins build server for the ML-model.
- Obtain access to the OpenShift environment were the model application is deployed.
- Obtain domain access rights to the data warehouse to be able to perform CRUD operations in our domain.

Development environment:

- Install and configure IntelliJ IDEA as IDE for the model application.
- Install Insomnia Core (REST Client) to test the model.
- DBMS QE2 to access database data outside of the model application.
- Install Anaconda incl. Jupyter Notebook as DS platform.
- Install PyCharm as IDE for developing our monitoring tool.

Other preparations:

- Obtain test data access.
- Established a group of experts as contact persons for various topics.

First we had to identify which other systems we require and where we need to obtain authorizations. The most central part of this was the Git code repository, where the model and its components are versioned and stored. We also needed access to the build server to test our developments, which we uploaded to the Git repository. In our case this is a Jenkins build server. The Jenkins server builds the projects and deploys the artifacts via docker to an OpenShift environment. In order to access this environment, we have requested the permissions for this as well. Furthermore we still have to save our monitoring data. For this we have requested access rights to SIGNAL IDUNA's data warehouse to be able to create,

edit and delete tables in our own domain. Other related and required systems are not known at this point in time and are therefore not requested.

Next, we needed to set up our development environment. For this we have decided to use IntelliJ IDEA for the further development of the machine learning model, because our model is implemented in Java and the preprocessor in R. Both languages are supported by IntelliJ Idea. In addition, a Git client is integrated so that we can combine all aspects of model development in one software with this IDE. To use or test the model we need a REST client, which in our case is Insomnia Core. In the later development, however, we use automated Python scripts to demonstrate our solutions in chapter 5. We also need a database management system to manage the monitoring tables in the data warehouse. For this we use the internal software QE2, which is well suited for working with DB2 databases which were designed by IBM. Additionally we decided to install Anaconda including the Jupyter Notebook as data science platform. We hope to have an efficient and effective platform for first transformations, tests and experiments. However, the Jupyter notebook is not sufficient as the final Python development environment to us. For this reason we decided to use the IDE PyCharm, which also integrates Jupyter Notebook and has an excellent integration of Python. The main development will take place there.

On top of that we had to obtain test data. This data should be as realistic as possible, so that we can carry out our development as close as possible to a real scenario. For this we had the advantage that we already had test data available from the development of the model, which was enriched with some more recent data. Nevertheless, we had problems to get access to this data, because it also contains sensitive data sets. In addition, it had to be clarified to what extent we were allowed to use this kind of data in productive operations. More on this in section 4.2.2. The test data is data over a period of 5 years, where we tried to find different patterns, such as data drifts or anomalies, to test our developments. For test scenarios not covered by the test data, we synthesized test data by adjusting the real data for testing purposes. In detail, this means that we analyze how a specific real world scenario would manifest itself and what effects it will have on the test data used and the model. Based on this information, we then consider how the test data can be adjusted to reflect the behavior of the scenario. This is a viable practice, as it is important to us how well our metrics work and how accurately our thresholds are chosen. Therefore, the focus is not on the data representing the real world, but rather on the data representing possible events which could occur in the real world and showing that our monitoring responds appropriately. How we adjusted the test data for our two scenarios are described at length in Section 5.2 and Section 5.3.

Finally, we informed ourselves about the model itself in the company. While

our objective is to ensure that the metrics are generally applicable to any machine learning model, there will also be metrics that are application specific. Further, there may be created metrics that are generally transferable, but not so well suited or informative regarding our model. Regardless of the development situation, a scientific project requires the greatest possible knowledge about ones work. This includes knowing what we are working on and of course understanding the software or model we want to monitor in this case.

4.2.1 Stornoprophylaxe

The software that we will monitor later on is called "Stornoprophylaxe" and is a prediction model that tries to find accident insurance contracts that are going to be canceled by customers within the next three months and then informs the field service partner so that he can take preventive measures for customer preservation. The workflow of the model can be seen in figure 4.1. It begins with the prediction of one or more contracts at the end of each month. Each contract is evaluated separately and a result is predicted. If a prediction is positive, meaning that the customer is highly likely to cancel, the field service partner is informed so that he can intervene preventively. He will then try to bind the customer again for a longer period of time, for example with special offers or recommendations. If he does not succeed, the customer may cancel the contract somewhere between "Make prediction" and "Last moment at which the result is established". The contract then would continue for another three months until its cancellation is permanent.

In this figure, it can be seen particularly well that we have a large three-month time lag between the prediction and the prediction result. As previously stated, a large part of the metrics development focus is on this issue.

The objective of the model is to predict at least 10% of all actual cancellations. It is not relevant whether the contract has been saved or not. With a customer fluctuation of 5% per year, this corresponds to a percentage of 0.5% of the total contract stock. This value is so low because there are many different reasons for canceling an accident insurance contract and not every reason weighs the same for each customer. Accordingly, it is practically impossible to detect 100% of all customers who are willing to cancel using the machine learning model without drastically increasing the false positive values. Besides the prediction rate, another objective of the model, based on the prediction rate, is to predict as many true positive values as possible while minimizing the number of false negatives. In principle, both objectives describe the metric Recall, but the first objective sets a minimum limit for the value Recall which needs to be achieved and the second objective sets an optimal target without a direct score. Therefore, accuracy is not the most important goal, as it is often the case, but recall.

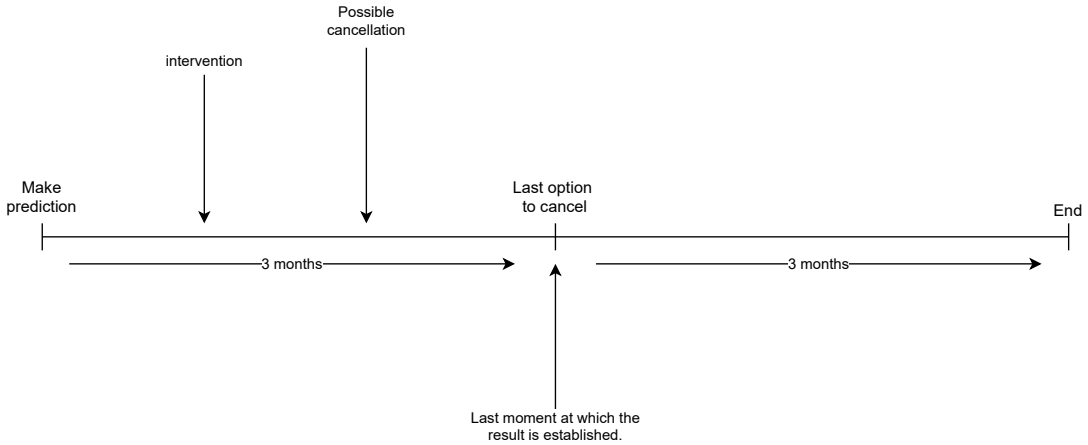


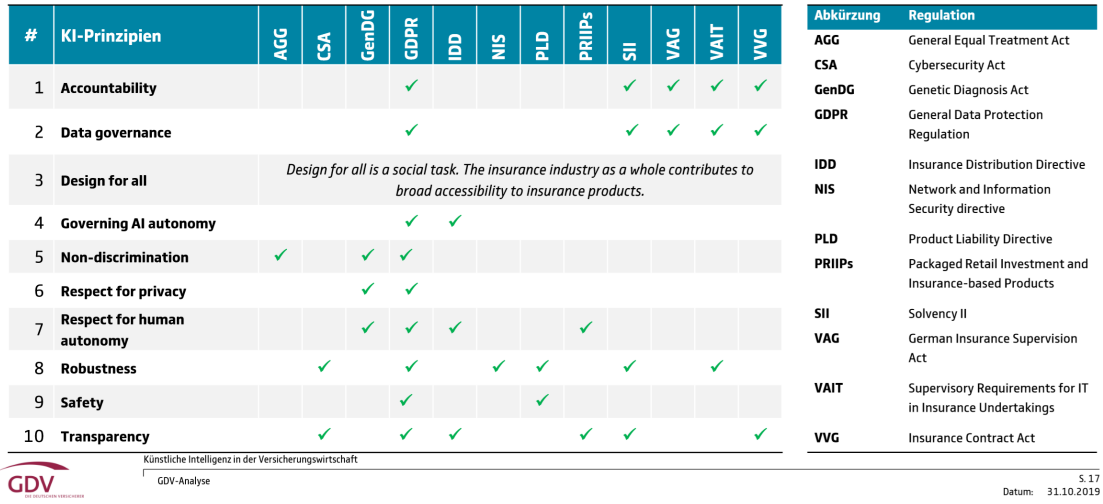
Figure 4.1: Model workflow

Model objectives:

- Predict 10% of all actual cancellations.
- Maximize the number of true positive predictions and at the same time minimize the number of false negative predictions.

4.2.2 Regulations in general and for insurance companies for AI / ML

In Germany, as in Europe, there are many regulations on the usage of Artificial Intelligence (AI) and thus also on machine learning. They are primarily intended to protect the users or those affected by such type of applications. In addition to regulations that apply to all companies, the regulations in the insurance industry are usually even stricter, since a lot of personal information is often used and processed here. It turned out that the use of AI and in our case, machine learning in insurance companies is so heavily restricted and limited that most applications have to work with trade-offs, if they are allowed at all. We came to this conclusion after several interviews with multiple compliance managers and the GDV (Gesamtverband der Deutschen Versicherungswirtschaft which translates to German Insurance Association), next to reading provided papers by the GDV concerning AI regulations. A lot of adjustment to the test data used and extra allowances for production data are required to be able to use machine learning applications. This statement is similar to what the GDV stated in its presentation “Künstliche Intelligenz in der Versicherungswirtschaft” (Artificial Intelligence in the Insurance Industry) from 31-10-2019 [10]. They found a total of twelve different regulatory documents that



#	KI-Prinzipien	AGG	CSA	GenDG	GDPR	IDD	NIS	PLD	PRIIPs	SII	VAG	VAIT	VVG
1	Accountability				✓					✓	✓	✓	✓
2	Data governance				✓					✓	✓	✓	✓
3	Design for all	*Design for all is a social task. The insurance industry as a whole contributes to broad accessibility to insurance products.*											
4	Governing AI autonomy				✓	✓							
5	Non-discrimination	✓		✓	✓								
6	Respect for privacy			✓	✓								
7	Respect for human autonomy			✓	✓	✓			✓				
8	Robustness		✓		✓		✓	✓		✓		✓	
9	Safety				✓			✓					
10	Transparency		✓		✓	✓			✓	✓			✓

Abkürzung	Regulation
AGG	General Equal Treatment Act
CSA	Cybersecurity Act
GenDG	Genetic Diagnosis Act
GDPR	General Data Protection Regulation
IDD	Insurance Distribution Directive
NIS	Network and Information Security directive
PLD	Product Liability Directive
PRIIPs	Packaged Retail Investment and Insurance-based Products
SII	Solvency II
VAG	German Insurance Supervision Act
VAIT	Supervisory Requirements for IT in Insurance Undertakings
VVG	Insurance Contract Act

Figure 4.2: Comprehensive regulatory overview for the insurance industry

limit the use of AI in the insurance industry. All those twelve regulations can be seen in figure 4.2. According to the GDV, the main issue in applying the regulations to machine learning is that a large number of regulations already exist in the area of insurance and that these are mostly formulated in a technology-neutral way. Consequently, they are also applicable to machine learning applications. This restricts the use of such components tremendously. In a nutshell, there are a lot of regulations that have to be respected and since this would go beyond the scope of the thesis, we decided to include the monitoring part of the regulations, but to limit it in complexity and amount of regulations, so that we are not only focusing on this topic and an automated or semi-automated evaluation is still possible. Nevertheless, the following regulations can be summarized:

General formulated regulations:

1. End-products must be continuously examined for possible risks of discrimination [15, 14, 11].

2. All additional processing of personal data is prohibited. It is collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes [15, Art. 5].

    - Accordingly, personalized data can only be used if it is absolutely necessary to fulfill the contract details. Consequently, the ML model must also be necessary to fulfill the contract if the customer has not agreed to further processing of his data.

3. Data should only be used adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed [15, Art. 5 § 1c].

4. End-products must be intervenable to prevent potential risks to the rights and freedoms of natural persons [9].

Data protection requirements for artificial intelligence summarized by independent federal and state data protection supervisory authorities: [8]

1. AI must not make people an object.
   (a) Fully automated decisions or profiling by AI systems are only permitted to a limited extent.
   (b) People need to intervene if needed.
2. AI may only be used for constitutionally legitimate purposes and may not be used for any other purpose than that for which it was intended.
   (a) In the case of AI systems, extended processing purposes must also be compatible with the original collection purpose. This also applies to the use of personal data for training purposes of AI systems.
3. AI must be transparent, comprehensible and explainable.
   (a) Transparency requirements must be met on an ongoing basis when AI systems are used to process personal data. The accountability of the person responsible applies (Art. 5 para. 2 GDPR).
4. AI must avoid discrimination.
   (a) Covert discrimination must also be avoided.
5. The principle of data minimization applies to AI.
   (a) The processing of personal data must therefore always be limited to what is necessary. The necessity test may show that the processing of completely anonymous data is sufficient to achieve the legitimate purpose.
6. AI needs accountability.
   (a) Those involved in the deployment of an AI system must identify and clearly communicate responsibilities.
7. AI needs technical and organizational standards.

(a) For example pseudonymization.

It can be seen that even an extract of the applicable regulations leaves almost no room for working with machine learning.

Nevertheless, according to our management consultant for data analytics governance for banks and insurance companies, the regulations can be divided into three important categories. This management consultant was recruited by SIGNAL IDUNA as an external consultant to advise on how to work with AI and regulations.

1. Personal Data
2. Discrimination
3. Transparency

First the personal data. These are the property of the individual and are never transferable in any way. Only parts can be made available for use, but only the minimum personal data, that is actually required to fulfill the agreed contractual service, can be requested. This is then called "contractual use of data". If data is no longer required to fulfill the contractual service, it must be deleted immediately and entirely. The same applies to all personal data in the event of cancellation of a contract. This is described in Article 5 in the GDPR [15]. Excluded are personal data, which are subject to a legal retention period. These are described in various law books, including the BDSG ("Bundesdatenschutzgesetz") or the HGB ("Handelsgesetzbuch"). Such periods are not uniform and may vary according to the law. In addition to the contractual use of personal data, it is still possible to request additional data from the user / customer. In case of approval these are usable, but on the other hand this approval can be revoked at any time. This would mean that this data would have to be deleted immediately. Therefore the request for personal data for AI systems on a voluntary basis makes little sense due to the immediate revocability of the data.

Moreover, discrimination of any kind must be avoided. It is irrelevant whether these are hidden or visible, intended or unintended. What is considered discrimination and when different treatment of individuals is permitted is described among other documents in the General Equal Treatment Act ("Allgemeines Gleichbehandlungsgesetz") [6]. For details about the types of discrimination, we would like to refer to the legal text and do not go into further detail. However, it is clear that in order to detect discrimination, the machine learning algorithm must be able to be analyzed transparently. To what extent this supports the recognition of hidden discrimination remains unclear. For this purpose, a clearing office is usually necessary, which looks at the decisions of the algorithm and assesses whether

discrimination has taken place or not. Such an office is often known as the "Data Analytics Governance Board".

As previously noted, the last category is transparency. This must be maintained at all time, which is very complicated for machine learning models which are considered to be black boxes. Nonetheless, many regulations, including Article 12 of the GDPR, require transparent data processing, information presentation and decision-making. For this reason there is a lot of research done on "explainable ML". Among other things also the Federal Ministry of Education and Research ("Bundesministerium für Bildung und Forschung") describes in an announcement why the explainability and transparency of machine learning and artificial intelligence is needed and wants to stimulate this [4]. Cynthia Rudin pursues another approach in her paper "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". She criticizes the benefits of black box machine learning models and introduces an alternative to the forced explainability of such models. Her way forward is to design models that are inherently interpretable and thus provide their own explanations instead of explaining the black box model afterwards, whereas often those explanations are not reliable [23].

As last note, we would like to point out that one of the currently most famous approaches for using personal data or data worthy of protection is pseudonymization through hash functions. However, the validity of this approach has recently been questioned by a collaboration between the Spanish Data Protection Agency (Agencia Española de Protección de Datos - AEPD) and the European Data Protection Supervisor (EDPS). However, under certain conditions it is still possible to use hash techniques to pseudonomise data. Requirements for this are for example to justify the used hash technique by a re-identification risk analysis, whereby this analysis must result in an objective evaluation [28].

4.3 Conception

In this section we present our process for creating our metrics, including the integration in a monitoring tool.

First of all, we discuss concept decisions, where we define decisions that are indispensable for us, which partly come from us, but also from the interviews, the literature research and the survey. These decisions are more general and relate to fundamental decisions, such as functionalities.

Afterwards, we present the technical design of our application within the scope of the model. We present the components of our machine learning services as well as the surrounding infrastructure for monitoring.

4.3.1 Concept decisions

As with many other decisions we have made, we have based our decisions on our own opinion, but also on the results of our research. We took especially strongly into consideration our survey, which gave us initial input by the future stakeholders. Here we intentionally asked general questions in order to get a feeling for what needs to be developed to achieve the highest level of user acceptance.

Especially the desire for a simple install and go solution was a major concern in the survey, which became especially relevant when selecting the monitoring components in Section 4.3.2.

In addition, a good overview must be ensured, in which the most relevant metrics are displayed. For this reason we decided to create an overall status, which summarizes the model. On closer analysis, it is then possible to see in the respective metrics why the model was no longer rated as flawless. Besides that, there will be a predefined overview with the most important metrics which we consider to be most important. Nevertheless, it is possible to customize the overview to suit the individual taste.

Since the survey results revealed a tendency for the metrics to be displayed in a scale representation, although a large number of respondents preferred a combination of a scale and traffic light representation, we decided to use scale values and added a traffic light representation where it seemed to be useful. The goal is to scale the scales to values between 0 and 10, whereas 10 is the highest value.

In addition, we categorize the metrics into three categories. If the model or the metrics are of the original or good quality, they are in the category “normal” . In the case of quality losses that do not yet require re-training, but where attention should still be paid, the model or the metrics are categorized in “attention required” . If however the quality decrease is noticeable, the model or the metrics are categorized in the category “severe impairment” . It is important to note that there may be metrics that provide relevant information, but are not influential in causing impairments to the model. Here the maximum category is “attention required” .

Alongside the presentation of the metrics and the model itself, there will be notifications of quality losses. It is particularly important that, depending on the type of metric (model metric, business metric or regulatory metric), the correct group of people is notified. In this way, it should be prevented that groups of people receive notifications which are not relevant for them.

Finally, there will be static metrics that are not assigned to any category. These metrics are neutral and intended to convey static information. This includes, for instance, the threshold, version numbers, model size, deployment date and similar information that could be relevant for error analysis / troubleshooting. These are not intended to measure the current model quality.

Summarized this results in the following list:

1. Provide a simple install and go solution. → immediately operational
2. Ensure a straightforward usability.
3. An overall status to summarize the models quality.
4. Display metrics as a scale representation and add a traffic light representation where useful.
5. Button or link as interface for explaining the displayed metric.
6. Three categories for metric values:
   (a) Normal
   (b) Attention required
   (c) Severe Impairment
7. Notification to the correct group of people based on the type of metric.
8. Static metrics to convey static information.
9. Default views provided as standard.

4.3.2 Technical design

The complete technical design is shown in Figure 4.3. Here the entire application architecture is shown, although our monitoring application is not explicitly identified. Fundamentally, the structure may be divided into four areas:

1. External Service Calls
2. Accident Churn Prediction Service
3. Monitoring Scope with Grafana and Prometheus
4. Data-Warehouse

External Service Call

The first area is represented as an abstract external service or abstract external application. This represents any kind of software that uses the machine learning model service. It can be a web application, a forwarding REST-Service or similar. The only requirement is that a valid JSON object is sent to the accident churn prediction service via the REST interface, so that the input is readable and not improperly formatted.

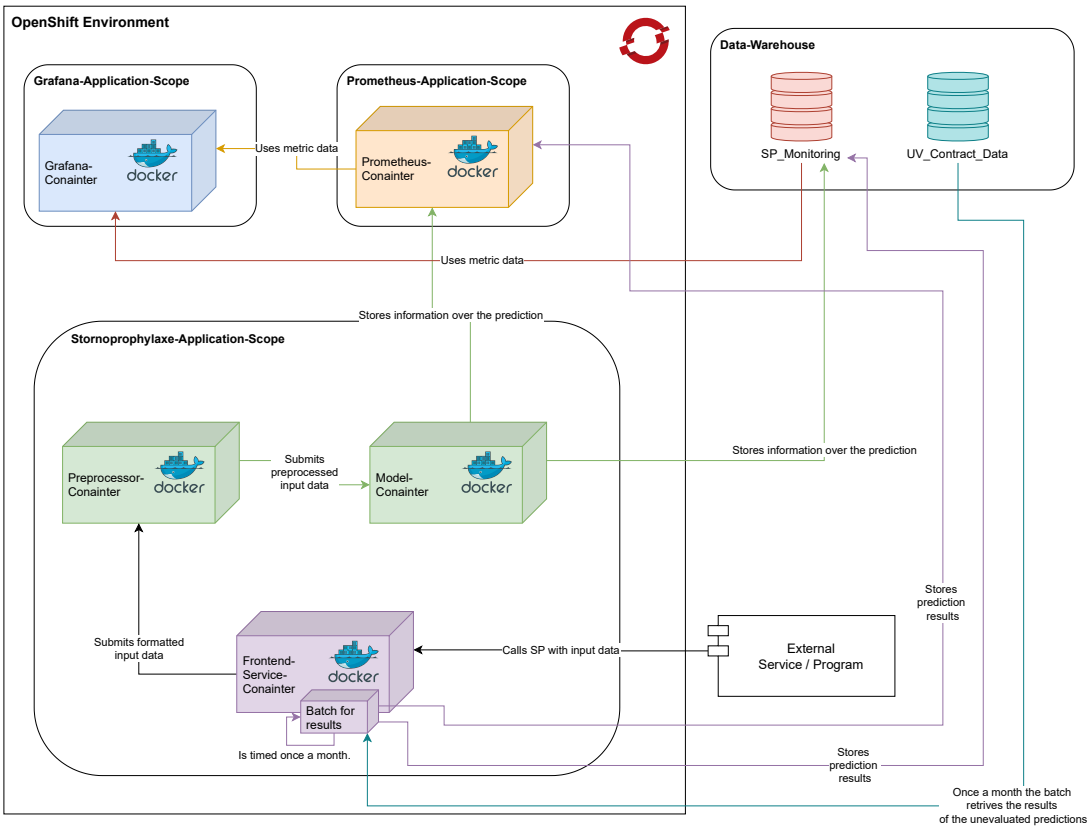


Figure 4.3: Architecture at SI

Accident Churn Prediction Service

This service consists of three separate containers. The service is therefore divided into

1. a frontend,
2. a preprocessor and
3. the actual model.

Before this thesis project, all three areas were already available, but were significantly expanded by us. The implementation of the batch for the result matching, for instance, was completely developed by us. In addition, the data receiving process was significantly expanded and the logging of all information regarding metrics was completely implemented by us.

The frontend provides a REST api, which can be used by any external service. This service receives the JSON sent using a POST call and attempts to extract the information transferred. It is important that a valid JSON is sent, otherwise the process would be terminated. The frontend service then sends the extracted information to the preprocessor.

This is our second container in the accident churn prediction service construct and it prepares the data before the prediction is calculated. Additional fields are calculated, fields such as NAN or None are cleaned up and further preparations are made. The preprocessor service then forwards the data to the model, which then performs a calculation for the prediction.

During this calculation, the model then also stores a large range of information about the prediction. A more detailed description of the stored information is given in the presentation of the "SP_Monitoring" database (4.4). In this phase, information is stored equally in Prometheus and in the "SP_Monitoring" database. In addition, the prediction including the additional information is sent back to the calling service via a JSON object. The reason for this is that the developers of the external service may wish to perform their own evaluations, away from our developed metrics, which we integrate into our monitoring.

Finally, the accident churn prediction service includes a batch for evaluating the predictions. This is part of the frontend service and is executed once a month. The reason for this is the previously described time offset between prediction and evaluation of the prediction. Therefore, this batch calculates the ID hash that is stored in the monitoring database for each contract in the accident insurance stock data, tries to locate it in the monitoring database, evaluates whether the prediction date implies an evaluation of the result and performs an evaluation if needed. Then, the result is stored again in Prometheus and in the monitoring database. Only after this it is possible to calculate metrics such as accuracy and precision and similar. This is also one of our biggest challenges in this project, as we have to distance ourselves from these metrics and still be able to evaluate the quality of the model predictions.

Monitoring Scope with Grafana and Prometheus

Most monitoring takes place in the third area. This area consists of two docker containers, a Grafana container and a Prometheus container. These two containers were also created by us. But as both of these software products can be downloaded directly as a ready-to-use container, it was more a configuration task than an implementation.

Grafana on the one hand is an open source platform to visualize metrics and alert by defined thresholds and events. It has a multiplicity of visualization options to help understand produced and stored data, which in our case are information of

our predictions. On their homepage Grafana advertise their open source product as the analytics platform for all your metrics [18].

Prometheus on the other hand is an open-source systems monitoring and alerting toolkit originally built at SoundCloud. Prometheus scrapes metrics from instrumented jobs, either directly or via an intermediary push gateway for short-lived jobs. It stores all scraped samples locally and runs rules over this data to either aggregate and record new time series from existing data or generate alerts. Moreover, it has a native support for Grafana, which makes it the perfect partner for visualizing data out of Prometheus [1]. Together they make an excellent team for monitoring software. The flexible data collection by Prometheus and the simple calculation and visualization of metrics in Grafana is a good combination for us to monitor our machine learning service in operation.

Data-Warehouse

At last we have a data storage, which is located in our data warehouse. On the one hand, this contains the stock data of the accident insurances, which are needed to match the prediction and the actual result, and on the other hand a database for the relational storing of all prediction information, which are necessary for the metric calculation. The first one was already preexisting, whereas the latter was created by us. The reason for the extra creation of this second database is that we considered it useful to use a relational database for information storage in addition to the logging tool Prometheus, in order to be able to perform high-performance queries and analyses. Furthermore, a relational database offers a good structuring of data and also makes it easier to analyze it. Both databases are IBM DB2 databases, which is the database system SIGNAL IDUNA uses for relational databases at their company and thus we had no alternative option available. Nevertheless, IBM DB2 databases are known for their fast data processing, but due to their low popularity they are not considered as data sink / data source by every external tool. As an example, even Prometheus has to use a third party plugin to use DB2 tables as data source, whereas MSSQL, MySql or PostgreSQL are supported by default.

Monitoring Database

The structure of the database tables used to store the prediction information can be viewed in Figure 4.4. The information is stored in four database tables.

First there is the PREDICTION_CORE_METRICS table. This table contains basic data for the respective predictions. It contains the prediction itself, which in our model is always a floating point number between 0 and 1.

This also applies to the threshold, which also lies between 0 and 1 and is also a floating point number. The reason for saving the threshold, which should be

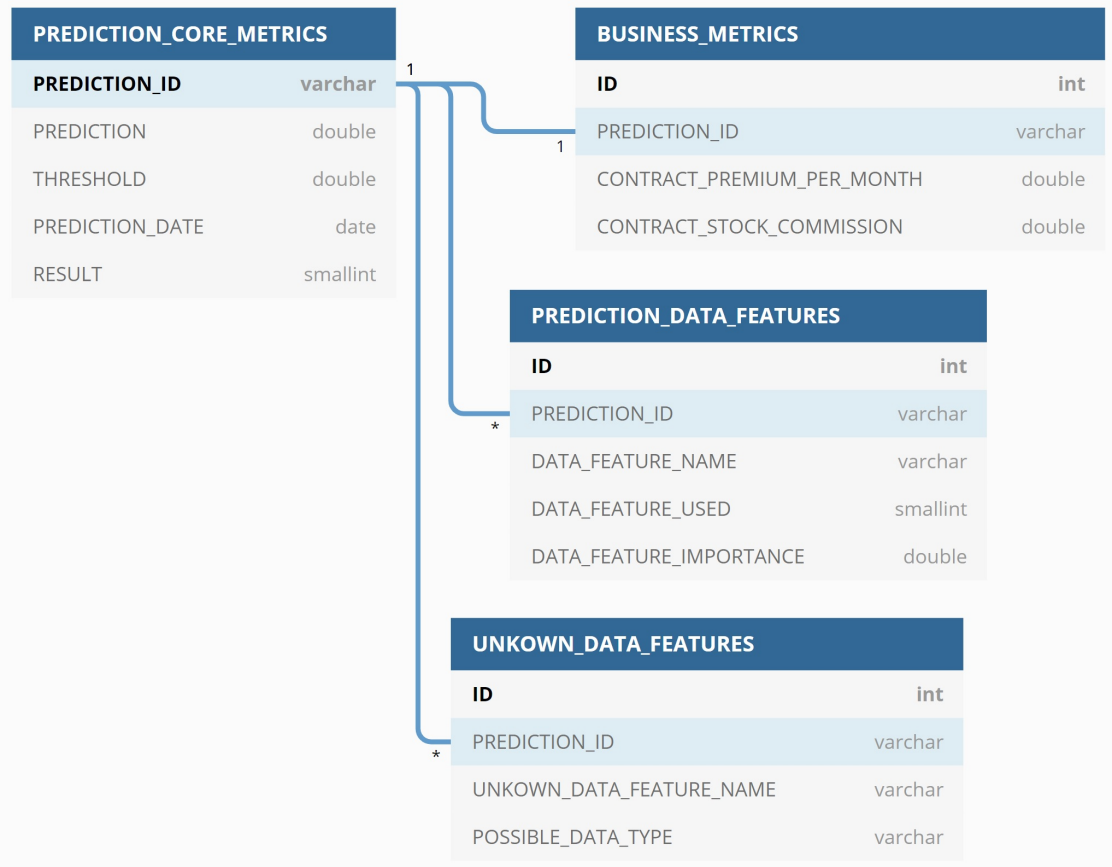


Figure 4.4: Architecture of the SP_Monitoring database

application wide known, is, that the optimal threshold may have changed during retraining of the model. If the new threshold would be applied to old prediction data it would distort the data information. Accordingly, we decided to store the respective threshold in addition to the prediction data.

In order to be able to evaluate afterwards how accurate the prediction was, we also added the field RESULT, which is filled with 0 or 1. Here 0 stands for "contract was not canceled" and 1 for "contract was canceled". Since there is a maximum time offset of 3 months between the prediction and the result, this field is initially empty for more recent predictions. In order to be able to evaluate the prediction at all, we need an ID with which we can identify the contract for the prediction and load the status from the accident insurance stock database.

We also need the ID to link the primary and foreign key of the four data tables. Due to significant data protection regulations, the PREDICTION_ID is a hash variable that can be computed by a batch in one-way as described above. This ensures that no information can be deduced from the monitoring data to the

contract.

Finally, we save the prediction date as well in order to be able to assign the prediction to a specific time and to know when we can trigger an evaluation of the prediction.

The second database table with the name BUSINESS_METRICS contains information about the respective contracts. It contains the monthly premiums for the contract and the stock commission, which is paid monthly to the ADP. This information is used to develop business metrics which can be presented in monetary fashion and can be understood without a great degree of technical comprehension. Since there is only one data record in this table for each contract, a 1:1 relationship is present here.

Following, we created a table that stores the importance of each data feature during the prediction. Since the presence of data features can change over time, some can be added, some can be removed, an 1:n relationship was implemented here. This means that for each prediction there are X data features with a data feature name, whether the data feature was used and a floating point number representing the importance. To calculate the importance of the data feature, we use a unified framework for interpreting predictions, SHAP (SHapley Additive exPlanations) [20]. Due to the large number of data features used, this is by far our largest table.

Our last database table contains all unknown data features. All incoming data features that are not considered expected are stored here. This means that unknown, new or simply undefined data features, as well as incorrectly labeled data features are stored here. Here we hope to be able to detect data anomalies and irregularities in the input data.

4.4 Metric design

In the now following section, we present the metric design. Here we have developed several metrics with the Goal Question Metric Approach [2], which are integrated in our monitoring. These will also be the basis for the demonstration in Chapter 5 and the evaluation in Chapter 6. The metrics were developed on the basis of prior knowledge of machine learning and software engineering, literature research, the survey and interviews with experts from the company.

4.4.1 Metric categories

The metric design process is the most important part of this thesis. It lays the foundation for a successful demonstration and evaluation. Accordingly, we carefully selected specific metrics. We used four different sources for the selection:

1. Metrics already known by us.
2. Metrics found and collected out of literature.
3. Metrics gathered from our initial survey.
4. Metrics developed in the course of this research.

For the presentation of the metrics, we first present the method we used to identify or develop the metrics we needed. Then we present the metrics themselves in detail.

The metrics we identified or have developed can basically be divided into five categories:

1. ML-specific metrics
2. Static (ML-)metrics
3. General metrics
4. Business metrics
5. Regulatory metrics

ML-specific metrics are the main part of the study and describe metrics with which the machine learning model or machine learning application can be directly evaluated. This also includes metrics that can be used to interpret the model and thus can also be metrics for governmental regulations concerning explainability.

Static (ML-)metrics, on the other hand, are of static nature. This means they try to find differences between static structures, which usually only change with new releases, like for instance model sizes. This can affect the model, the application, but also the environment, such as the infrastructure. Therefore releases of components that are not directly related to the machine learning project are also relevant.

The following category contains general metrics. These are metrics that are used to monitor applications in general. This includes for example CPU usage, action limits and similar. Of course these are not directly related to machine learning, but they are still important measures for stability, performance and reliability and should not be neglected.

Next to that, we defined business metrics. Such metrics are mainly used to assess the business benefit and are therefore less relevant for technicians, but all the more for managers, executives and people who are not familiar with the technical aspects of machine learning. Those metrics do not have specific boundaries. They could be ML-specific, but also of general or monetary nature.

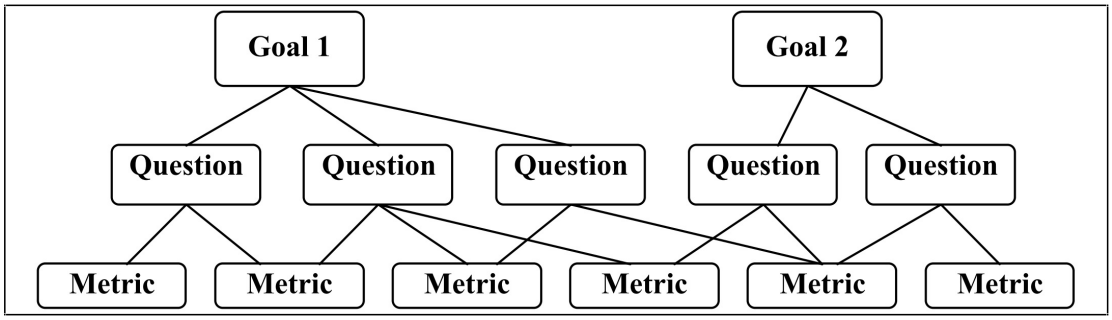


Figure 4.5: The Goal Question Metric Approach [2]

Finally, we have defined the category of regulatory metrics. These metrics attempt to assess the model in the context of regulations, namely whether it operates in compliance with legal regulatory obligations.

4.4.2 The Goal Question Metric Approach

As described above, we have used the "The Goal Question Metric Approach" from Basili, Caldiera, and Rombach to sharpen our metrics and match them to our goals [2]. We have chosen this framework because it follows a goal-driven approach. This means that the goals are placed in the focus and metrics are defined or selected based on these goals. The figure 4.5 presents the approach of the framework and how it is designed.

First of all, one defines goals that one wants to achieve. This is called conceptual level. A goal is defined for an object relative to a particular environment. Objects of measurement are a product, a process or a resource. The definition of a goal is usually composed of four elements. (1) A purpose, (2) an issue, (3) an object (process) and (4) a viewpoint.

Afterwards one defines at least one, but often several questions in order to characterize the goal with regard to certain quality criteria. This is the second level in the model, the operational level.

For these questions, metrics are then developed that can answer the questions about the quality of the goals. A metric does not have to be assigned to one question, rather it can be used to answer several different questions. The metrics should answer the questions in a quantitative way. Therefore the third level is called quantitative level. These metrics can be designed both objectively and subjectively. Subjective metrics depend on both, the object that is being measured and the viewpoint from which it is measured and thus an involvement of human judgment. Those metrics could be, for instance, a subjective view on an appropriate response time for problems (depending on the severity of the problem,

the response should be faster) or just the readability of a text. Whereas objective metrics are measurable within defined scales like the amount of true positive predictions.

The GQM approach can be well represented in the form of a table, as we have done in our metric design. (See Table 4.1, 4.2, 4.3 and 4.4)

In total we focused on three different goals, representing our design objectives, and built questions and metrics for them.

- If the cell background is light gray, the metric are not integratable into a monitoring tool, because they need a subjective assessment by a human.
- If the cell background is light yellow, the metrics are not implementable at the moment of the thesis project, for other reasons like too large complexity for this limited project.

4.4.3 Ensure the model accuracy

Our first goal is to "ensure the model accuracy of the model from the operators perspective". This goal reflects the model quality and is intended to ensure that our applications make the best possible predictions. A major issue remains that it takes about three months before we can evaluate whether the predicted result has been achieved or not.

Question 1: Accordingly, standard metrics, such as precision, recall, accuracy, specificity, F1-score and False Positive Rate (FPR), should be considered with caution for model evaluation. Nevertheless, we think it makes sense to integrate them. On the one hand, it can be determined at last after three months whether the model has lost quality, provided that the other metrics do not fail and do not report anything. This would at least provide a fail-safe in case of incorrect or poor complementary metrics, which of course should not happen nor be the case. On the other hand, they are still very informative, even if only limited. They are considered to be the most important standard metrics in machine learning model training and should not be ignored in any case, even if they are of limited significance due to the time lag in our use case. The extent to which certain standard metrics are more important than others strongly depends on the use case. For instance, Recall and FPR are more relevant to us than the other metrics, which is related to the objectives in Section 4.2.1.

The representation of the metrics should be made in a chart where the x-axis is a time scale and the y-axis is the score. As described in the question this chart should contain x evaluable prediction chunks. The size of the chunk is dependent of the use case. In our case one chunk is one month, since our application is making

Goal	Purpose Issue Object (process) Viewpoint	Ensure the model performance of the model from the operators perspective
Question 1	How accurate is the model over the last x evaluable prediction chunks? (X should be adjusted to a reasonable amount in contrast to the prediction amount in a reasonable time frame.)	
Metrics	$Precision = \frac{TruePositive}{TruePositive+FalsePositive} = \frac{TruePositive}{ActualResults}$	
	$TPR, Sensitivity, Recall = \frac{TruePositive}{TruePositive+FalseNegative}$	
	$Accuracy = \frac{TruePositive+TrueNegative}{Total}$	
	$Specificity = \frac{TrueNegative}{TrueNegative+FalsePositive}$	
	$F1 - Score = 2 * \frac{Precision*Recall}{Precision+Recall}$	
	$FPR(FalsePositiveRate) = \frac{FalsePositive}{TrueNegative+FalsePositive} = 1 - Specificity$	
Question 2	How accurate might the model be for the current predictions that cannot be evaluated yet?	
Metrics	Calculate a trend of general metrics such as accuracy, specificity and other over the last five time cycles to show a trend, as one time cycle is one month.	
	Use a distance measure for data sets and compare the input dataset of the current month with all past months, which are already evaluated by accuracy, specificity etc. The closest match could be a good indicator for the current values of such metrics. (This metric must be reset after retraining because the assumption would apply to two different models and may be wrong then)	
	Current quality probability: Compare the current input dataset to the last three former ones and compare characteristics. If characteristics changed a lot, the probability that the prediction quality differ to the last predictions is high.	

Table 4.1: Goal Question Metric - Model Accuracy (1)

predictions every month. The chunk size is therefore derived by the amount of prediction needed to be made in that specific month. As our data for the model is quite static and not changing a lot, the number of chunks should be set between one and two years, hence 12 to 24 months, to not lose the overview in the charts.

Question 2: An additional question, in our context, is therefore "How accurate might the model be for the current predictions that cannot be evaluated yet?". Thereby we pursued three approaches. First, we assumed that in most of the cases the environment, thus the input data, changes slowly and steadily, such as retirement age, demographics and the like. Correspondingly, we can assume that a data drift takes place slowly. This process is gradual and slow and can be illustrated by comparing the current results with a moving average of the standard metrics. A moving average cycle of five months is considered suitable by us for our use case. The observation by a moving average should prevent one-off quality anomalies, namely for example a month with a particularly large number of bad or good forecasts, from triggering an alarm immediately. Moreover, it shows a slow decrease of model quality more accurate. Since the model decrease in this case is considered slow and since we have a time offset of three month, we set the threshold 0.02 points away from the average result out of the test data set.

Furthermore, we assume that not only changes in the standard metrics are noticeable, but also in the data structure or the frequency of the occurring features. Thus, our approach would be to compare the input dataset from the current month with the datasets already evaluated. The data set that is most similar to the current one would then probably have a similar accuracy, etc. This is of course only an assumption, since the data characteristics may also occur with similar frequency, but in different combinations, which could lead to a completely different result. It remains to be seen to what extent this metric will prove valuable. Furthermore, those metrics can only be used unless the model is not retrained. After retraining the model it is not possible to compare the old input data with the new input data, due to possible changes in the structure, but also due to two different models. As the old data sets have been evaluated with the old model, the old evaluations can no longer be mapped to the new model and the new input data set. Consequently, the metric has to be reset and is then unusable for at least three months until an evaluated dataset with the new model is available again.

Our last approach follows a similar approach. Here a comparison between the last three input data sets and the current input data set will be made. The more drastically the current input data set differs from the previous ones, the greater the probability that the prediction quality will change. However, this metric has two major disadvantages. First, it does not say anything about whether the prediction quality deteriorates, but rather that it might change. Secondly, it can only be used

to react to drastic and rapid changes in the data sets. Gradual changes are not covered here.

Question 3: Now we have already reviewed the standard metrics and input data. Especially the standard metrics can give an insight into how good the model works, but not how certain it is. We assume that the certainty of the model is the preliminary stage of a quality development. This means that as soon as the model loses certainty, the probability of a correct prediction decreases. In fact, the model may even become a mere guesser. This is primarily applicable to models that make a prediction in a scale range, as is the case for our model with scales between 0 and 1. If the predictions move more towards the defined threshold for positive or negative predictions, the model becomes more uncertain. Therefore we have created five metrics.

First we would like to examine the distribution of the model predictions. In this case we are interested in the floating point value and not in the will cancel or will not cancel the contract result values. With the distribution of the respective prediction values we try to determine how certain the model is and if it loses certainty. Thereafter suitable thresholds should be defined. These thresholds are valid for both positive and negative predictions. For example, if the threshold is 0.5, below would be a negative and above would be a positive prediction, a threshold should be defined for positive and negative predictions. How to set these thresholds is dependent on the model and the use case. Let's assume 0.4 and 0.6. In addition, we need the maximum count of predictions which are allowed to be between 0.4 - 0.5 and 0.5 - 0.6. Again, this depends on the model and the use case. In the first instance we want to use this metric for information purposes and see how the distributions change. This is because we have both a non-static number of predictions, they vary per month, and the uneven distribution of predictions due to the static nature of the contracts (5% contract fluctuation per year). Accordingly, it is currently difficult for us to develop reasonable and plausible thresholds.

We then defined two more metrics. First the *average positive prediction over a time cycle* and the *average negative prediction over a time cycle*. These metrics are used to represent the average of the respective predictions for and against a cancellation of the monitored months. Afterwards, they are compared with the previous months. If the average prediction is then found to be moving towards the threshold, it can be seen that the model is becoming more uncertain in its predictions. At a certain level of uncertainty, a warning should be sent. This again depends on the use case and model. Nevertheless the warning should be taken seriously and not to be ignored. It does not yet have the status of an alert, because the model can still perform well even if it is uncertain, but it can be a preliminary stage before quality losses occur and therefore it makes sense to train

Goal	Purpose Issue Object (process) Viewpoint	Ensure the model performance of the model from the operators perspective
Question 3	How certain is the model while predicting results?	
Metrics	Distribution of the return values of the model predictions over a time cycle, whereas a time cycle is one month.	
	Average positive prediction over a time cycle = $\frac{\sum_{k=1}^{n} predpos_k}{n}$ whereas n = number of Predictions	
	Average negative prediction over a time cycle = $\frac{\sum_{k=1}^{n} predneg_k}{n}$ whereas n = number of Predictions	
	Standard deviation of positive predictions over a time cycle $\Rightarrow \delta = \sqrt{\frac{\sum_{k=1}^{n}(predpos_k - mean)^2}{n}}$	
	Standard deviation of negative predictions over a time cycle $\Rightarrow \delta = \sqrt{\frac{\sum_{k=1}^{n}(predneg_k - mean)^2}{n}}$	
Question 4	Is the bias of the predictions still in a plausible range?	
Metrics	Average of test predictions $\Rightarrow \frac{\sum_{k=1}^{n} pred_t}{n} \Rightarrow n$ = amount of predictions in test data	
	Average of predictions per month $\Rightarrow \frac{\sum_{k=1}^{n} pred}{n} \Rightarrow n$ = amount of predictions in month X	
	$Bias\,of\,model\,predictions_X = Avg\,test\,predictions - Avg\,predictions\,month_X$	
Question 5	How fast can the operators react to declining accuracy of the model and modify the model / restore the initial quality / performance?	
Metrics	Reaction time between a given alarm/notification and the operators acting.	
	Time used while finding & solving the performance issue.	
	Build-Pipeline-Speed	

Table 4.2: Goal Question Metric - Model Accuracy (2)

the model again in order to achieve a higher degree of prediction certainty. Again, we have chosen a threshold of 0.02 points difference to the average values of the test set. This seemed appropriate to us, since this threshold firstly produces a warning and no alarm, and secondly does not imply any immediate action. We can take our time to analyze why the model has become uncertain. However, this requires time, so we intentionally define the threshold at 0.02 points discrepancy to inform the monitors early.

In addition to the threshold which implies uncertainty, we add another threshold. This does not imply uncertainty, but it implies an anomaly or anomalies. This threshold is defined 0.02 points above or below the average value of the test data set, which theoretically would send a warning when the model becomes more certain. However, it is important to remember that you do not know the result of these predictions at this current moment and it is possible that the model will become more certain, but still deteriorate in quality because of an anomaly in the real world. Therefore, in addition to decreasing certainty, one should also monitor suddenly increasing ones, since they imply unforeseen events, which most likely also cause an unintended effect.

The last two metrics defined for answering question 3 are the standard deviations of the positive and negative predictions. By using these metrics, we try to discover whether the average predictions we have just defined are reliable. Since the standard deviation tells us what the spread of the values around their average is like, we can tell whether the predictions are close together and thus the model makes similar predictions, or whether it calculates very widely spread values and thus only the average looks good, but the actual values are far from the average. Here we try to avoid that unnatural behavior not recognized by the average calculation is caught. The following example explains what is meant:

Assuming that of the last 10 predictions five are 0.6 and five are 0.7, the average would be 0.65. The same result would be obtained if of the last 10 predictions five were 0.5 and five 0.8, but the standard deviation would detect the difference and still observe an unnatural behavior. This is because although the average of the predictions has remained the same, the predictions have drifted unusually far apart.

Question 4: Our fourth question focuses on the bias of the predictions and is adapted from the prediction bias. The prediction bias is a quantity that measures the distance of two averages. In machine learning terms that means:

$$prediction\,bias = average\,of\,predictions - average\,of\,labels\,in\,data\,set \quad (4.1)$$

However, instead of calculating the bias of a chunk or the entire prediction set, we want to compare the bias over the months, so we had to adjust the representation as presented on the Google Developer website [13]. Furthermore, we have

the previously described time offset between prediction and result. As the prediction bias thereby loses relevance or meaningfulness for predictions just made, we had to adapt our thinking. Therefore, we took the respective average of the positive or negative predictions, not the results, from the test set as the basis for *average of predictions* and subtracted the average positive or negative predictions from the current month.

The result cannot be named prediction bias directly, since no more predictions are compared to results, but since the calculation of this metric has the same basis and we only take other metrics, we decided to call it "bias of the model predictions".

We set the threshold to 0.02 and -0.02, because normally the bias should be very low and that seemed reasonable.

Question 5: Our last question "How fast can the operators react to declining accuracy of the model and modify the model / restore the initial quality / performance?" is intended to round off the first block of metrics and deals more with restoring a qualitatively high status and less with the current status. At first we want to measure the reaction time of the operators, which is needed until an alarm or a warning is processed. What a reasonable time is, is a very subjective decision and depends on the importance of the application for the business.

Subsequently we want to measure how much time is needed to find and solve the issue of worse performance. In doing so, we are orienting ourselves on the ticket system already used at SIGNAL IDUNA, where a reaction time of 12 hours, 24 hours or a working week is specified, depending on the severity of the ticket.

To conclude, we also think it makes sense to track the build pipeline speed for model creation and deployment. Care should be taken that the speed does not decrease too much, otherwise you would lose agility and flexibility.

Nevertheless, we do not intend to test these metrics in demonstration, firstly because two metrics involve human interaction each time, and secondly because they do not relate to the current state of the model, but to recovery. Nevertheless, we thought it would be useful to introduce them briefly.

4.4.4 Calculate the (monetary) value of the model

In addition to the quality of our machine learning model from the developer's or operator's perspective, we would also like to highlight other viewpoints. Another one would be the view of managers or product owners. These groups of people do not necessarily have knowledge of the machine learning model used in the application and are therefore unable to interpret technical metrics well or at all. For this purpose we have created the goal of monetary or business metrics. These are therefore also specifically focused on stock quantities and monetary targets.

However, before we go into the questions and answer them on the basis of metrics, we would like to briefly introduce the "ADPs". These are field service partners of SIGNAL IDUNA and are responsible for the acquisition and preservation of insurance contracts. They are comparable to sales managers, but even after the contract has been concluded, they continue to be the contact person for the insured person regarding other possible contracts and contract terms. However, insurance details are again the responsibility of SIGNAL IDUNA. Accordingly, when making a prediction regarding the cancellation of a contract, ADP must approach the policyholder, if permitted, and try to convince him or her to change his or her mind by means of offers or arguments.

Question 6: The first question in this block also refers directly to the ADP. Although the ADPs are not a direct stakeholder of the application, the amount of the saved stock commission for the ADPs is also interesting for the management. Accordingly, the metric for answering Question 6 calculates how much stock commission we were able to save by using our application for each month. We decided to show the monthly saved amount instead of the commulated amount, because we want to evaluate the monthly performance and not the performance over the whole lifetime of the application. Furthermore, we would also have to track whether the contract still exists in the following months and then adjust the saved amounts again. In this case, we believe that the effort required to calculate the amount does not reflect the value of the metric.

It is important to note that we need direct feedback from the respective ADPs to discover which contracts classified as subject to cancellation could be saved by an ADP. We need this in order to make a correct entry in the confusion matrix. This is because the rescued contracts would be entered under FP if we did not receive the information that they would have actually been canceled. Unfortunately, at the time of the project, we were not authorized to receive this information. Therefore, we can either only predict how many contracts have been rescued or this metric is not applicable. Nevertheless, we wanted to bring in the metric to give readers of this thesis a motivation to deal with such a situation and to think about how to avoid or deal with it.

Question 7: The logical next question is then how much premium income we have saved. However, this is not always parallel to the stock commission, since the stock commission is tied percentage-wise to the contract premium depending on the respective ADP. Accordingly, this metric is not redundant but reflects a further indicator for the monetary assessment of the product. In summary, the answer to this question is more important for SIGNAL IDUNA, as it directly reflects the financial status of the accident insurance contracts. Thereby the metric "premium

Goal	Purpose Issue Object (process) Viewpoint	Calculate the (monetary) value of the model for the business from the managers perspective
Question 6	How much stock commission have we saved for ADPs through SP?	
Metrics	$SavedCommission_i = \sum_{k=1}^{j} com_k$ - whereas j = contracts saved in month i and com_k = the stock commission of saved contract k	
Question 7	How much premium income have we saved through SP?	
Metrics	$SavedPremium_i = \sum_{k=1}^{j} prem_k$ - whereas j = contracts saved in month i and $prem_k$ = the premium payed for saved contract k	
	$SavedIncome_i = SavedPremium_i - SavedCommission_i$	
Question 8	How much stock commission has been lost? (This only includes all true positive results, where the ADP probably did not intervene.)	
Metrics	$LostCommission = \sum_{k=1}^{n} com_k$ - whereas n = contracts lost and com_k = stock commission of lost contract k	
	$LostPremiums = \sum_{k=1}^{n} com_k$ - whereas n = contracts lost and com_k = contract premium of lost contract k	
Question 9	How much larger is our stock through SP?	
Metrics	$RC = \{C\} \setminus \{SC\}$ – Relative complement, whereas SC = {Saved contracts} and C = {All contracts}	
	Increase of the stock = $\frac{\lvert C \rvert}{\lvert RC \rvert}$	
Question 10	What is the monthly rate of “saved” contracts?	
Metrics	$SavedContracts_i$ - Number of contracts saved in month i	
	$MonthlyRate = \frac{\sum_{i=1}^{n} SavedContracts_i}{n}$ - whereas n is the number of months where the model is in operation.	
Question 11	What is the ratio between correct predicted cancellations and all cancellations?	
Metrics	$TPC = \{TruePositiveContracts\}$	
	$FNC = \{FalseNegativeContracts\}$	
	$LC = \{LostContracts\} = TPC + FNC$	
	$PercentageRatio = \frac{\lvert\{TPC\}\rvert}{\lvert\{LC\}\rvert}$	

Table 4.3: Goal Question Metric - (Monetary) Value of the model

income" is the result of the calculation of $SavedIncome = SavedPremium - SavedCommission$.

Again, the restriction described in Question 6 applies here as well.

Question 8: The subsequent question follows a different approach. Instead of looking at how much stock commission we saved, which is difficult to evaluate, we look at how much stock commission we have lost. Here we should define an upper limit that should not be exceeded and the goal should be to keep the limit constant or even reduce it. The elegant thing about this metric is that it is evaluable in any case and cannot be manipulated negatively in our use case. All one has to do is count all TP and FN results and set them off against the corresponding stock commission of the respective contracts. Still it has a time offset of three month.

To complete that metric we also included the same metric for lost contract premiums and not only for lost stock commissions.

Question 9: Now we already presented metrics to show how to assess the monetary value or changes of the machine learning model. Nevertheless, we believe that not only income is relevant. For instance, in order to generate market dominance and a brand, it is often more important to hold as many contracts as possible instead of generating as much revenue as possible. Accordingly, our machine learning application directly contributes to increasing and maximizing the size of the stock. This leads to the question how much larger is our stock through SP. To answer this question we have created two metrics.

First, to build the first metric, the amount of contracts in the total stock and the amount of saved contracts are needed. With these two metrics the relative complement can be calculated. This results from $RC = \{C\} \setminus \{SC\}$. RC is then the set of all contracts that are in the total stock minus the saved contracts.

But the relative complement itself is not yet a very meaningful metric. We would like to know by what percentage the total portfolio has increased, assuming that the contracts that were saved had all been canceled. This means that the size of the total portfolio did not increase, but by comparing it with the relative complement, we can show an increase in size. This results in the last metric the increase of the stock.

Question 10: To answer question 10, we compare the saved contracts in the respective month with the rolling average, whereby the rolling average is always calculated from the beginning of the model deployment. The purpose is that we can see whether we can save a certain number of contracts on average over the entire model lifetime. Thus, for example, during the whole model lifetime, the rolling average should not fall below the value X, which would still have to be defined.

Furthermore, this allows us to identify slow trends that have occurred since the beginning of the model deployment, but we must be careful not to lose too much of the significance of this metric if too many values or months are included. Thus, this metric is more interesting for models that are re-trained at regular intervals. Otherwise, a rolling average with a fixed number of months is recommended, but the initial start of the model will be neglected at some point.

Question 11: In our last question about business metrics, we would like to answer the ratio of correctly predicted contracts to total lost contracts. The aim is to see if the main purpose of the application "Predict 10% of all actual canceled contracts" is achieved. Although this is a rather technical metric, it has been categorized into the business metrics, as the previously mentioned goal has been noted as extremely important in the business requirements. Nevertheless, it is actually just another notation of the metric "Recall". It is therefore debatable whether we should even write the metric down a second time, but since we each address a different stakeholder group, we did it anyway.

4.4.5 Fulfilling the governmental regulations

To describe how to monitor regulatory requirements, we have created three questions on the three most important topics "data usage", "discrimination" and "transparency".

Unfortunately, some of these regulations are so complex that it is not possible to automatically determine with metrics whether the model complies with all requirements or not. This refers, for instance, to the protection of personal data, which may only be used to the extent that it contributes to the fulfillment of the contract or the customer allows it. To check this on a continuous basis it requires a lot of effort and probably an extra program. Therefore, it does not seem to be possible to map it by one or more metrics.

In addition, it is also partly subjective how metrics are to be interpreted. So it is absolutely necessary that the processing of data in the machine learning model can be presented transparently. However, it is not defined when a sufficient degree of transparency is reached.

Question 12: Our first question deals with the topic of data usage. Since only data that is necessary for the direct fulfillment of the contract or additional granted data may be used, this must be ensured by controls. For this purpose we have created two metrics.

The first metric is the subjective assessment of the data owner(s). However, since this is a subjective assessment of a human component, it cannot be evaluated

Goal	Purpose Issue Object (process) Viewpoint	Fulfilling the governmental regulation for the use of AI in insurance companies in Germany from the company's perspective
Question 12	Does the model only use data for predictions that is necessary to fulfil the respective UV contract or that was approved by the customer and still is?	
Metrics	Subjective assessment by the data owner(s).	
	Compare the used data features for the predictions by the anonymous user id with the respective entry in the permissions database for data sensitivity.	
Question 13	Is the model free from negative discriminatory decisions?	
Metrics	Distribution of the importance of the used data features contributed to the model predictions over a time cycle, whereas a time cycle in this case is a month due to the frequency the model is predicting.	
	Subjective assessment by the case officer(s).	
Question 14	Are the model decisions traceable at all time? (transparency) (This is more logging, but important to fulfill this requirement. The idea is to give as much information as possible to the specific prediction to interpret why the model calculated the outcome)	
Metrics	Importance of used data features.	
	Prediction certainty	

Table 4.4: Goal Question Metric - Governmental regulations

automatically and integrated into the monitoring tool. Nevertheless, this is the first step to ensure correct data usage. One could now argue that a single person or a small group of people cannot check the data of thousands of contracts, but since insurance contracts are defined statically, the data owner or data owners do not have to check every contract, but only make sure that not more than the allowed features are used.

Our second metric takes a similar approach. It tries to implement the behavior of the data owner in a mechanical way. This metric returns whether not more data features are used than allowed, a true or false metric. In order to detect this, the data features used in the respective predictions are compared with the corresponding entries in a permissions database for data sensitivity. This database table must contain this information for each contract and must be maintained continuously. Furthermore, there must be a link between the permissions table and the predictions, but this is probably only possible with restrictions or not at all, because then the prediction could be used to draw direct conclusions about the contract. Furthermore, it is rather impractical to maintain such a database.

Question 13: The subsequent question is whether the model is free of negative discriminating decisions. This is a difficult question to answer, but an indicator of discrimination can be different characteristics of the data features. If there are features that are particularly noticeable that could be utilized for discrimination, this is an indicator that an investigation should be conducted. Unfortunately, it is not possible to find a clear answer to this question, because discrimination is not only objectively observable. Nevertheless, we have defined the distribution of the most important data features as a first metric to at least be able to recognize when anomalies indicating discrimination.

Independent of the previously defined metric, our second metric is the subjective assessment of a case officer. Nevertheless, the case officer can also use the previously presented metric as a help / advice. As mentioned before, a human interaction is required, which is why this metric cannot be implemented in the monitoring tool. We can only try to make it as easy as possible for these employees by providing helpful metrics like the distribution of data feature importances.

Question 14: Our last question is whether the model can be understood transparently at all times. Due to the black-box issue it is not possible to understand the complete decision process. For this reason we use the approach to log as much information as possible in metrics to make decisions more comprehensible or more understandable. This of course includes the distribution of data feature importances, but since we already noted this, we will not go into it again here. In addition, we use the prediction certainty again to see how certain the model

is. The more certain the model is, the easier it is to understand why the model has chosen this way, because the importance of the respective data features has to look accordingly.

Additional SIGNAL IDUNA specific metrics: In Chapter 3.3 we have presented SIGNAL IDUNA specific objectives, which should also be in a certain way recognizable in the metrics. However, we decided not to use metrics to cast these objectives.

This has the following reasons. First, we proposed to make a metric that shows whether we were able to save 10% of the 5% fluctuating contracts at the end of the year. However, in our opinion this metric does not add any value to the assessment of the current performance of the model. It is more of a KPI (Key Performance Indicator) that must be achieved and can also be calculated retroactively.

Further, we suggested to detect data drifts based on new accident insurance contracts in the system or adjustments of insurance conditions. However, we lacked the automatic input that would trigger such an analysis. Moreover, such a data drift would also be visible in a normal data drift analysis, since both analyses use the same data. Only the trigger to perform the analysis is different.

Finally, we wanted to analyze demographic changes before they would affect the model. In principle, we think this is an excellent approach that should be applied, but the effort to achieve this was ultimately too big to be implemented in this project. You first have to define which demographic attributes should be analyzed and then how to observe and automatically or semi-automatically analyze and assess them.

Nevertheless, we consider all three, but especially the last one, to be valid metrics and should be examined and implemented in future work.

Other metrics: Metrics that did not make it into our selection will be added as a table in the appendix. These are, for example, metrics that monitor container resources or metrics that seem to have less added value than those presented before.

4.5 Design review and evaluation

In the following section, we will discuss how the metrics created by us affect other people, namely, whether they would rate the respective metric as good or bad. In the process, we talked with colleagues about our metrics and created a second survey to obtain a broader pool of opinions.

We would leave out the assessments of our colleagues here, since they largely agreed with our findings and may also have had subjective influences by us.

For the survey, we took a selection of the metrics that we consider to be the most important and best conceivable, because otherwise we would have gotten a far too large survey, which would have gained much less acceptance.

The setting of the survey is comparable to that of the first survey. Again, we used the survey tool Qualtrix Core XM [22]. The availability of the survey was this time between the 2nd November 2020 and the 15th December 2020. It was again distributed by us through direct messages on social media and by mail.

First, we presented our use case in the survey and asked an open question as to which components of the model, technically and subject-related, should be monitored. By doing so, we wanted to make sure that we can determine whether the respondents have an interest in the survey and thus fill in sophisticated answers. We then had respondents sort the aspects we identified by importance to see what the most common order of importance is. We then asked the respondent to rate the metrics we selected for the survey on a scale of 0 to 10.

It turned out that this time we received fewer responses than in our first survey. We assume that due to the rather technical nature of the survey, many people felt that they could not contribute enough to it. This would mean that many people would consider the topic and the problem of monitoring machine learning models in operation as important and would work on it, but the general knowledge about the topic does not seem to be widespread enough yet.

4.5.1 Survey results:

In our second survey we received a total of 15 answers. The number of answers is not very large, but we think that there are still enough to be representative.

The age of the respondents is distributed between 18 and over 50 years, with 73% being between 18 and 30. The level of education is not as widely distributed. There, 14 out of 15 respondents have either a bachelor or master degree. Only one respondent has a doctorate. The job groups are broadly distributed, hence we expect a large variance in answers.

As described at the beginning, we started with an open question to check if the respondents have familiarized themselves with the described use case. A number of different answers were given, ranging from demographics and damage claim aspects to expected answers such as machine learning metrics like accuracy etc. Thus, most of the respondents seem to have at least partially dealt with the use case. However, some of them were only given short answers.

Afterwards we had the respondents sort the aspects we had identified. Although we received many different aspects in the previous question, data aspects and machine learning metrics were ranked highest in most cases. Especially the computational performance seem to be uninteresting. This may be regarded as given at any time and does not have to be the most important aspect in moni-

toring anymore. Regulatory and business aspects are quite well distributed, with regulatory aspects tending towards the lower ranks and business aspects towards the upper ranks.

After we got our general questions answered, we focused on our concrete metrics. The questions we had to answer were the quality of the metrics precision, recall, accuracy, specificity, the f1-score and the false positive rate with respect to our use case:

Precision The values here are between 7 and 10, so 13 respondents consider this metric to be quite good to very good.

Recall The metric recall looks similar, with 13 respondents between 6 and 10. Of these, 6 chose the value 10, thus very good, and three each rated this metric on the scale as 6 or 7.

Accuracy The following metric is very well distributed, but the tendency points more to an important metric with 7 respondents between 6 and 7, and another 5 between 9 and 10.

Specificity The specificity on the other hand is very central distributed with values between 3 and 8 with 13 respondents in total. Thus it can be said that opinions differ quite a lot here.

F1-Score The subsequent metric tends to have a higher ranking among the respondents. With 10 answers between 7 and 10 it seems to be considered very good.

FPR Remarkable is that the same distribution occurs for the FPR and therefore has the same significance as the F1 score.

In conclusion, Precision and Recall in particular are seen as very good metrics, F1 Score and FPR as quite good, and only Accuracy and Specificity are widely divided. Especially interesting was a comment on Accuracy, where one respondent stated: “You have to take into account the problem of unbalance. Accuracy is therefore rather problematic, whereas TPR is better.”

Unbalance refers to the unequal distribution between positive and negative predictions.

Rolling Averages In contrast to the individual evaluations of the standard training metrics, the use of this metrics in conjunction with a rolling average was found to be consistently good for our use case. No rating was below 5 and the largest proportion, 13 out of 15 responses, were 7 or higher.

We then addressed metrics that deal with the distribution of predictions in general and separately between positive and negative.

Comparison of the distribution of model predictions over the prediction months In the first metric, opinions seem to split into two different sides. Whereas 4 respondents think that this metric is very good. The remaining 11 respondents are in the midfield between 4 and 7.

Comparison of the average positive and negative predictions over the prediction months On the second metric, the trend is a little more positive. It remains with 4 respondents who also rate the second metric as very good, but the following 11 answers are this time between 4 and 8, with four of them choosing 8, meaning they chose it to be quite good.

All in all, it can be said that the latter metric is seen as better, since it shows a more positive tendency, but it also shows that there is disagreement about the quality of both metrics. However, neither of the two metrics is considered unsuitable, which is certainly a positive signal.

Prediction Bias The Prediction Bias, which has some similarities between the previous metrics, turns out to be an interesting case. In this Net Promoter Score assessment, we have one respondent who selected only a 3 and 14 respondents who chose relatively well distributed between 5 and 10. It is therefore not clear whether this metric can be seen as reasonably good or very good. The commentary on the metric makes it clear that not every respondent was aware of why the predictions were compared against the test predictions. This is specific to our use case and has to do with the fact that, as described several times before, we can only evaluate the predictions after 3 months and we consider this to be too late to assess the current status of the model. Possibly many answers were given around the center due to the lack of clarity.

Business Metrics We then asked about the business metrics of our application and have been able to draw a clear trend from the four questions. The favorite was the saved contracts per month, which received 12 answers between 8 and 10 and received great approval. For the following three questions we received the

same answer pattern every time. The opinions were very widely spread. We had at least one answer with a value of 0 and one with a value of 10. Basically they were well distributed over all possible answers. Sometimes a bit higher, sometimes a bit lower.

In summary, however, we can report that the saved contracts metric was the preferred one and the other three metrics received distributed approval. The experiments must therefore show which metrics are most suitable, since no conclusions can be drawn from the survey regarding the business metrics.

Regulatory Metrics For the assessment of the regulatory metrics, we gave the survey participants a selection of three metrics.

The metric that checks the data sensitivity against a permissions database received particularly high approval. With 14 out of 15 answers between 7 and 10, and 6 of these in position 10, the participants consider this metric to be the best. However, we also received a comment that is in line with our assessment of the feasibility of this metric. It was questioned to what extent the metric is feasible because it requires many external components and continuous maintenance.

After that, the subjective assessment of the data owner received the most acceptance, with a balanced distribution between 5 and 10. However, we also received a comment here as to how far this could represent a metric. In the case of the Goal-Question-Metric principle this is possible.

The lowest approval was given to the distribution of the data features for the predictions per month. Here, ratings from very good to completely unsuitable were given, of which 6 ratings are on 5 and 6. There seems to be more discrepancy about the usefulness and value of the metric. It is possible that the reference to the detection of discrimination has not become entirely clear.

Conclusion: In conclusion, it can be said that we feel partially confirmed in the selection and development of our metrics. There is agreement with our metrics in many parts, but also criticism and disagreement. In the experiments and the subsequent evaluation, it will then become visible whether or not the assessment of those surveyed is confirmed.

Chapter 5

Demonstration

In the chapter "Demonstration" we present how our chosen metrics perform in different scenarios, namely how well they can monitor machine learning applications in operation.

We begin by presenting how we prepared the experiments to show the effectiveness of the metrics. We also discuss limitations and tradeoffs that we had to make when preparing the experiments.

Afterwards we present two scenarios that reflect the behavior of our use case "Stornoprophylaxe" in operation. The first scenario "Pandemic" represents an abrupt and rapid change in the external world, in line with the current situation in 2020, while in contrast the second scenario "Slow changes" represents a slow change in the external world.

While presenting the two scenarios, we are also rating our used metrics in the context of our use case. However, Chapter 6 then summarizes all ratings and puts it in a structural order.

5.1 Experimental setup

For the experimental setup we first tried to use the real data we obtained for training and testing and tried to find certain behaviors in the data which represent a slow or fast transition of data aspects. Unfortunately, we have not been able to find both fast and slow changes in the data. On the one hand, this may be due to the fact that accident insurance contracts have a relatively static contract life, but it may also be due to the fact that we simply could not identify them.

Due to the lack of different data behaviors we had to use an alternative experimental setup. Here we followed the goal of this thesis, "multi-aspect monitoring of machine learning models in operation". Thus we want to show which metrics can effectively monitor a model in operation. However, this does not require an oper-

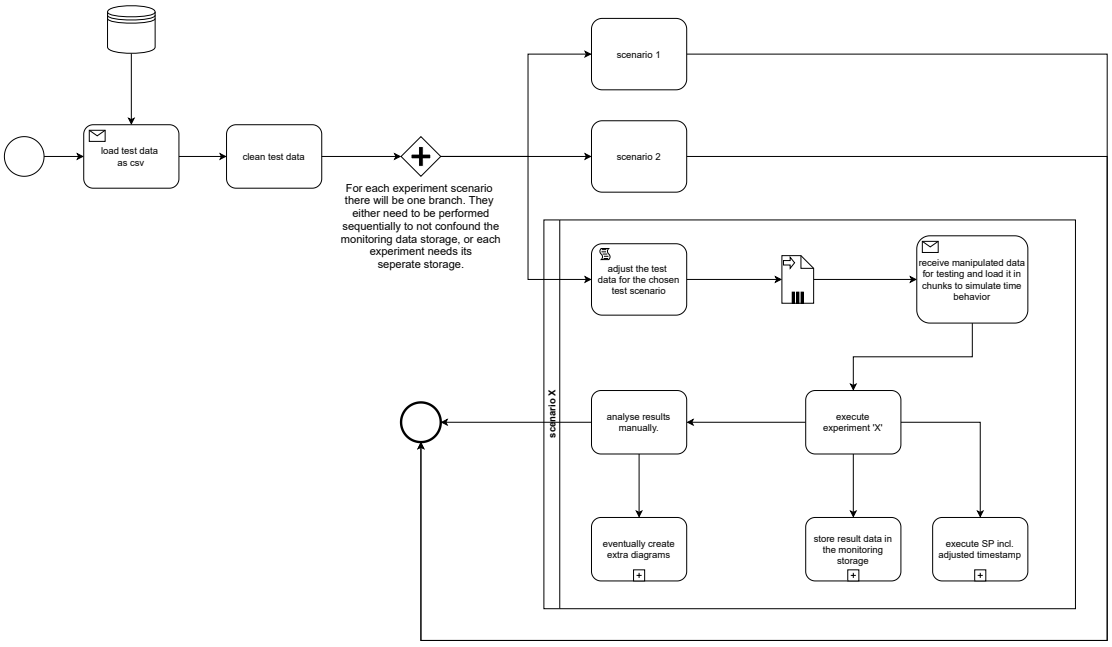


Figure 5.1: Architecture at SI

ational system, since the question is not whether monitoring in operation works, but how the metrics behave in certain scenarios and whether they reflect the correct information. For this reason, we decided to adjust the data we used during training and testing of the model to create two scenarios where the metrics have to show their effects. We then sent this data to our application with a manipulated timestamp. The reason for this is that we do not want to run each scenario for a year until we can evaluate it. The entries in the databases merged with the experiment data are then used to demonstrate the metrics.

For the entire experimental procedure we have written a modular Python program, which is shown in the flow diagram in Figure 5.1.

At the beginning of the process, the test data is loaded into the program from a database as a CSV file.

Afterwards we clean up the data. We assign default values to nonconforming fields where it makes sense, or delete the entire data entry which have nonrepairable fields. Since most of the data has no corrupt fields, the dataset is only marginally smaller afterwards.

The respective test scenarios are then executed. These are defined beforehand as separate and independent modules and contain seven steps.

First, the data is adjusted according to the scenario. Depending on the scenario, this is a different type of data adjustment. A detailed description of what was adjusted for which scenario and why is given in the following sections. After the

data has been adjusted, it is split into chunks to simulate a time behavior. The chunks are then loaded into a program module that performs the experiment. In doing so "Stornoprophylaxe" is called with a given timestamp and the respective prediction data including the result, which is already known from the test data, are stored in a database. Finally, after the experiment has been performed, the data is used to apply the metrics to the data and analyze it. If necessary, additional plots are generated.

5.2 Experiment 1 - Scenario Pandemic

Our first experiment is based on the current situation in 2020, whereas the virus COVID-19 suddenly appeared and influenced our all lives.

In the course of the first months of the pandemic, mainly March, April and May, the economy of small and medium-sized enterprises suffered a particularly severe collapse. This was due to reasons such as temporary forced closures of restaurants, hotels, retail and others, but also to a very poor level of orders in skilled crafts and trade businesses. As a large insurer of medium-sized businesses, the SIGNAL IDUNA Group also felt the effects of this.

For this reason, we made the assumption that due to the economic consequences, parts of the policyholders of the accident insurances will either not be able to pay their monthly premium or will pay it late. Payment reminders are then sent as a consequence. Such payment reminders are also a data feature in our machine learning model. As soon as this feature is marked as positive, meaning a payment reminder has been sent, the prediction breaks the threshold each time and a prediction of a possible cancellation is given. Just because some policyholders cannot pay the monthly insurance premium does not mean in consequence that they want to cancel the insurance. Furthermore, we assume that as soon as the unknown situation normalizes, the insurance business will also normalize as well.

For all these reasons, we have created the following test scenario based on our assumptions:

- As a trial period to test whether the predictions are relatively constant, we have chosen six months in which everything remains normal.
- In the following month, payment reminders suddenly rise sharply, but the actual cancellations remain normal. Here the metrics should already warn.
- These then remain high to very high until month 13, whereas the actual number of cancellations increases only slightly to medium. We described above for which reason we did that.

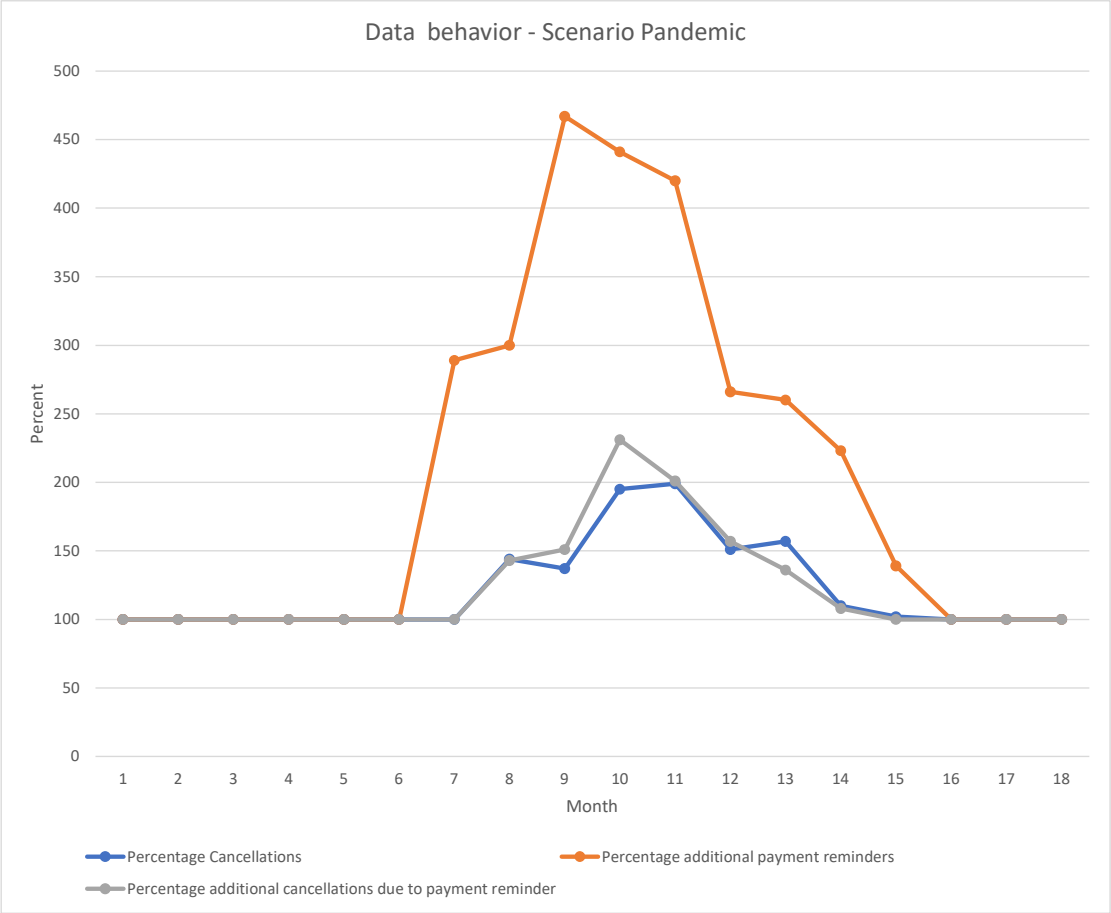


Figure 5.2: Scenario Pandemic - Data behavior

- In the last months, months 14 to 18, payment reminders are slowly returning to normal and contract cancellations are also on a normal level again.

The scenario can also be viewed in Figure 5.2.

5.2.1 Experiment 1 - Results

When conducting the first experiment we achieved the following results:

Question 1: Before we go into the results we want to remind the readers that the results of month 1 are known from month 4 on, because we have a time lag of 3 months due to our use case.

Nevertheless, we used the metrics precision, recall, accuracy, specificity, the f1-score and the false positive rate to evaluate the model over the last months.

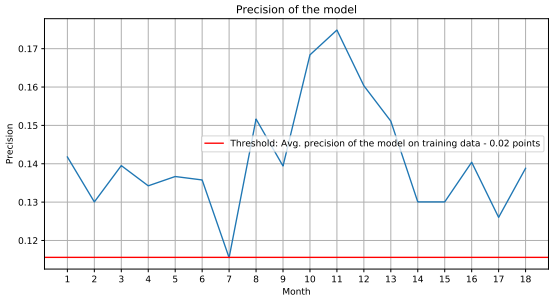


Figure 5.3: Scenario Pandemic: Precision

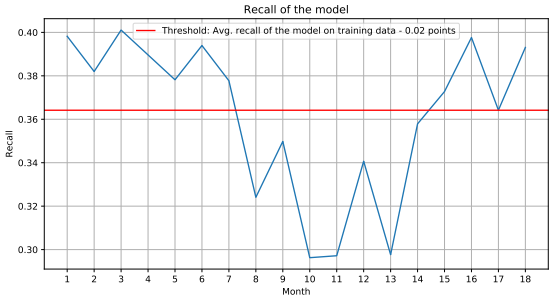


Figure 5.4: Scenario Pandemic: Recall

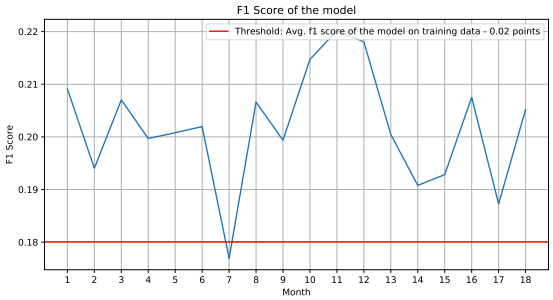


Figure 5.5: Scenario Pandemic: F1-Score

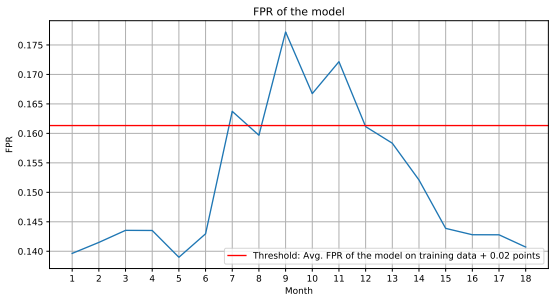


Figure 5.6: Scenario Pandemic: FPR

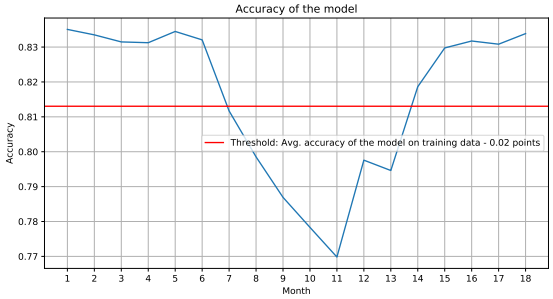


Figure 5.7: Scenario Pandemic: Accuracy

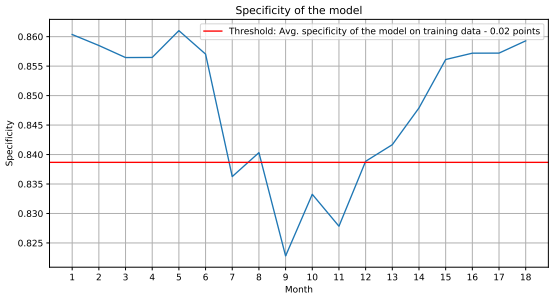


Figure 5.8: Scenario Pandemic: Specificity

As a threshold for decreasing quality, we took the average value of the respective metric from the training results and subtracted 0.02 points. The respective plots are shown in Figures 5.3, 5.4, 5.5, 5.6, 5.7 and 5.8.

It quickly becomes clear that due to the time lag, none of the metrics would hit fast enough in case of abrupt changes. Only after a delay of three months an alert would have been given. Accordingly, these metrics are not suitable to detect abrupt changes for our use case. Nevertheless, there are use cases where results

are available immediately or relatively quickly. Here we consider these metrics to be essential. Furthermore, we consider it useful to integrate the metrics in order to retrospectively see where the quality has deteriorated the most.

A retrospective view shows that accuracy, specificity, recall and the false positive rate achieve good results. In addition, the f1-score was also breached, which reflects the balance between precision and recall. Precision, on the other hand, did not just hit the threshold but did not breach it. The accuracy dropped particularly sharply, which is why the threshold was greatly overdrawn.

Question 2: We try to answer the second question by using the past results of the standard metrics and calculating a rolling average over five months. In doing so, we try to see trends and thus try to give an estimate of the current performance. The idea is that if the trend is negative for five months, it is quite likely to continue. The results can be seen in Figure 5.9.

It quickly becomes noticeable that these metrics do not add value in the event of abrupt changes, as they reflect the trend of recent months and therefore abrupt declines cannot be identified. It is interesting to see that they exceed the respective thresholds even later than if one simply uses for instance the metric fpr and waits three months until it is evaluable.

We also noticed that it might make sense to define the thresholds even more strictly, because changes in the rolling average are not noticeable that quickly. For example, if we would use -0.01 points instead of -0.02 points, we would recognize the abrupt change one or sometimes two months earlier than if we would wait for the evaluability of the respective metrics.

Alternatively, the rolling average could be limited to three rather than five months.

Nonetheless, we initially decided not to change either the thresholds nor the number of months in the rolling average because we felt that we would become vulnerable to external changes too quickly.

Question 3: With the metrics of Question 3 we want to see how certain the model is while prediction results, as we believe that the model quality and also the predictions are getting worse if the models prediction get uncertain. Doing so, we created five metrics, whereas the first metric is displayed in Figure 5.10, the second and third metric in Figure 5.11 and the last two in Figure 5.12

The course of the plot of the model certainty, if it runs optimally, should be such that most predictions are either at 0 or at 1. The least should be around the threshold. This results in a kind of $\cup$-shape. Then the model is very certain about its predictions.

In reality, it often looks different, as you can see in our plot. Our plot is more

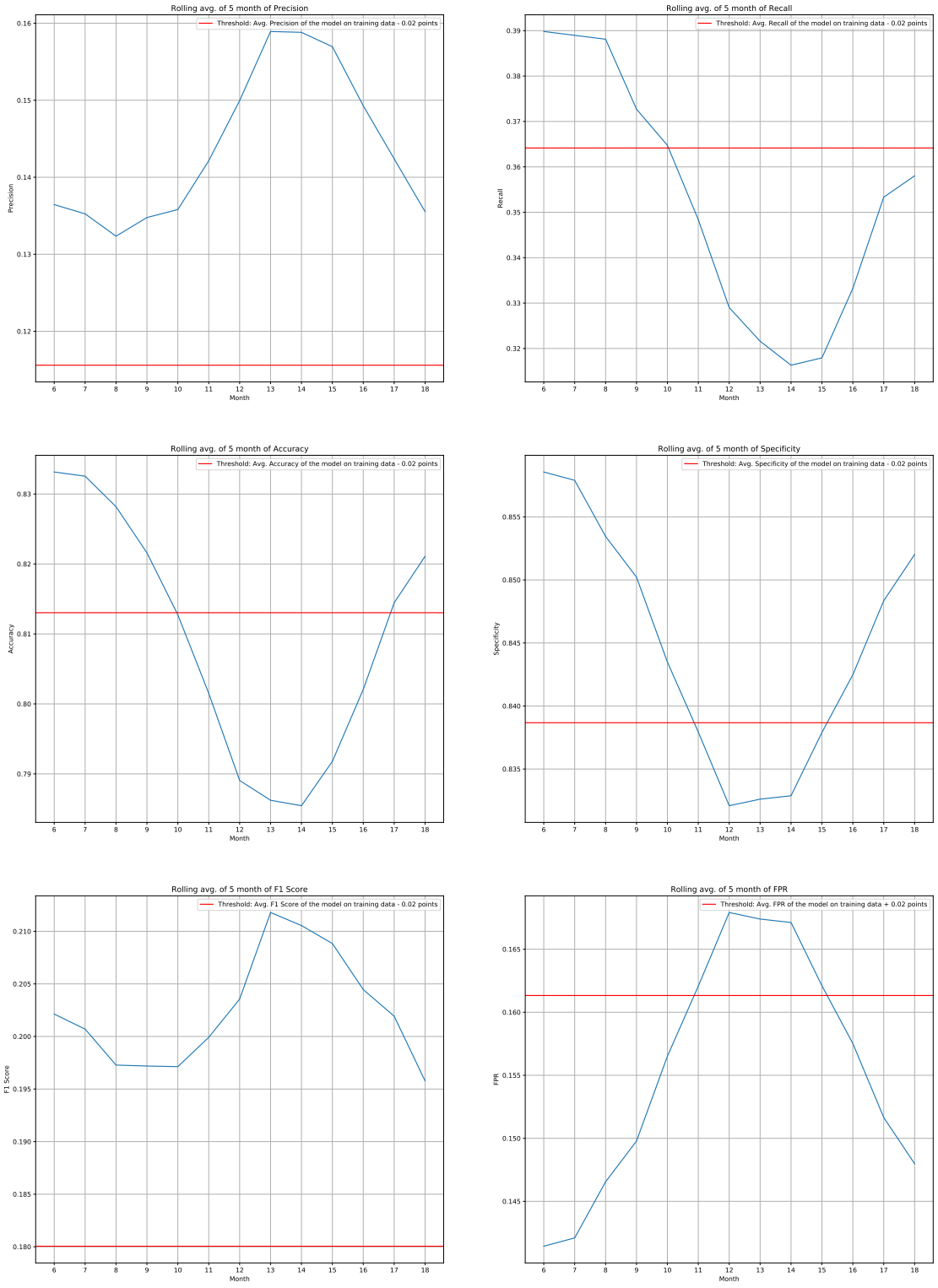


Figure 5.9: Scenario Pandemic: Standard Metrics Rolling Average

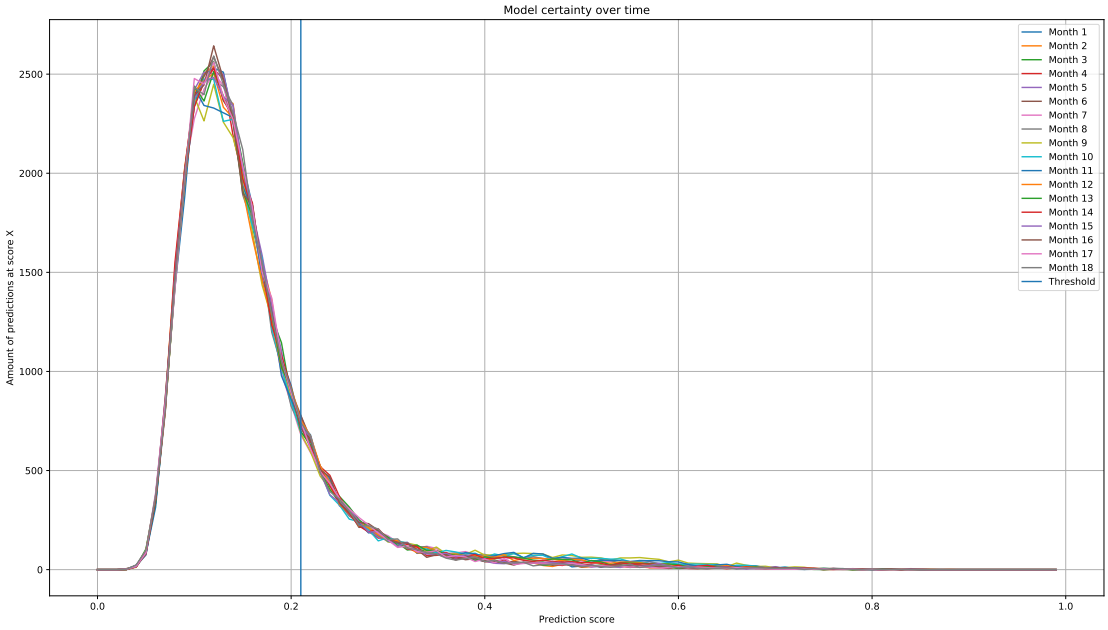


Figure 5.10: Scenario Pandemic: Model certainty

like a ∩-shape, which hints at an uncertain behavior. But in our case it is mainly because the majority of the predictions are always negative and the threshold is chosen quite low. Therefore, the positive predictions can be well distributed, although it would be better if they were all located at 1.

Nevertheless, only few negative predictions are located between 0 and 0.5.

The fact remains that over the 18 months, the distributions of the predictions mostly remain similarly well distributed and no month sticks out so strongly that an alarm could be raised. For this reason, we did not integrate a threshold at that time. It remains to be seen to what extent this metric offers added value, currently not that much.

As for our second and third metric, the plot looks much more meaningful. We again took a threshold of 0.02 points difference to the training value for positive and negative predictions and defined them as boundaries. The big advantage of this metrics are that no result is needed to evaluate the metrics, because they are based on the predictions but not the prediction results.

It is particularly nice to see that in month 7, when a particularly large number of payment reminders were sent, the threshold was directly broken upwards. So a warning was thrown about an abrupt change in the outside world.

The same is true for the last two metrics. Here the plots look similar to the averages of positive and negative predictions. The only difference is that we have set the threshold at 0.01 difference, because we thought that the standard deviation

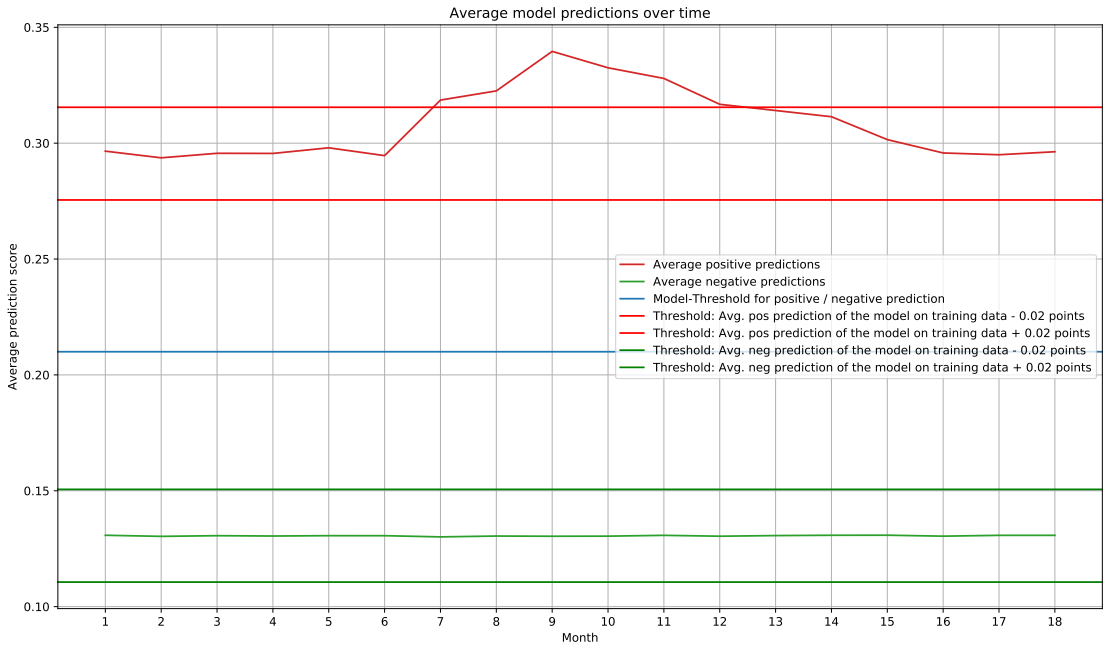


Figure 5.11: Scenario Pandemic: Averages of positive and negative predictions

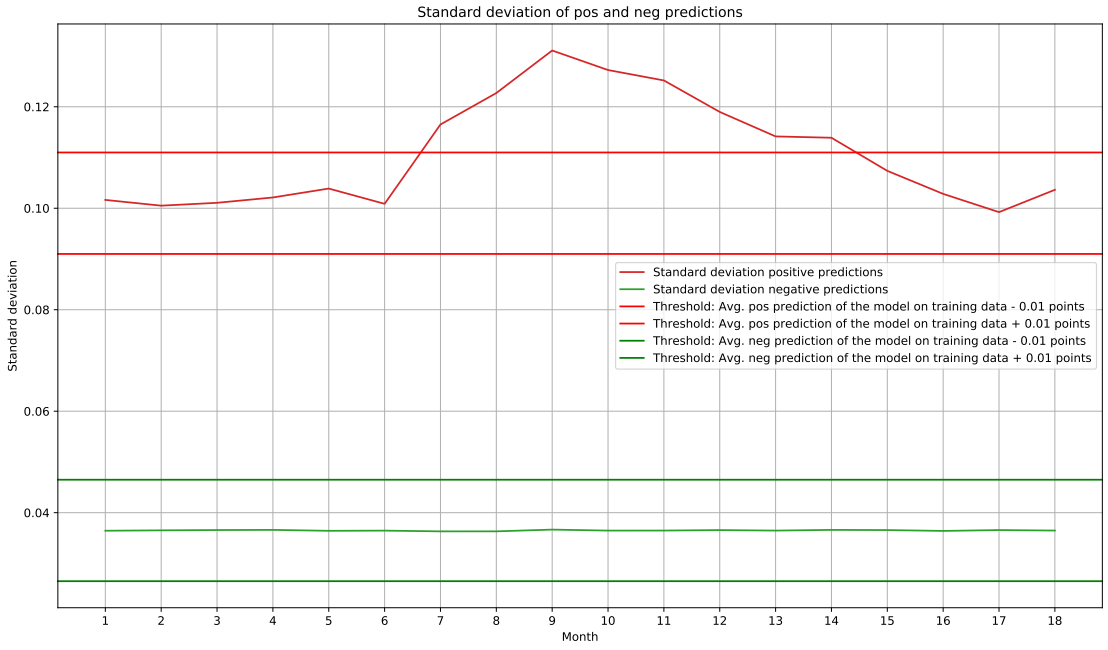


Figure 5.12: Scenario Pandemic: Standard Deviation of the positive and negative predictions

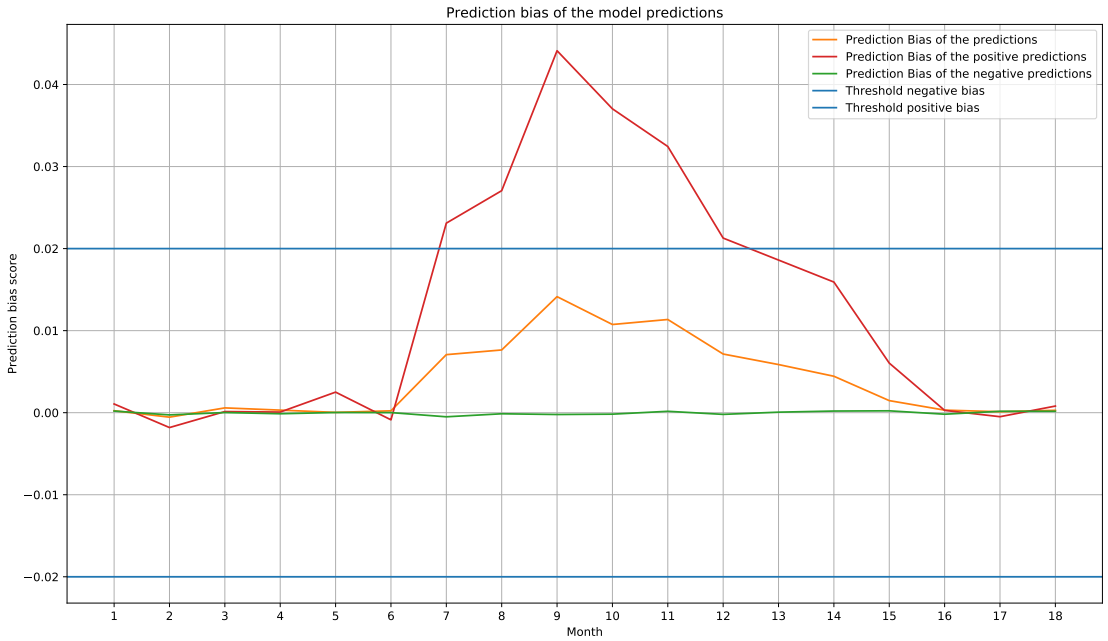


Figure 5.13: Scenario Pandemic: Prediction bias of the model predictions

of the scores should not be so different.

Both metrics for positive predictions threw a warning in the seventh month and recognized the abrupt change in the outside world.

Question 4: As our prediction bias, Figure 5.13, represents the difference between the complete average test predictions compared to the predictions of each month, it can be clearly seen, that the greatest difference is in month 9. Additionally, this metric also exceeded the threshold in month 7, as desired in the case of the abrupt changes caused by the pandemic.

Besides the positive and negative bias of the predictions, we also have a line for the cumulative predictions. We hoped to recognize whether both positive and negative predictions were shifting in the same direction or in the contrary. We also wanted to see how sharp these movements are if they move in opposite directions. After evaluating this plot, however, we concluded that this third line does not provide any added value. It can be calculated in the background at any time, if desired, but here it only shows the upward tendency weakened and if the lines would shift contrary, it only represents the balance between two movements and would never come close to a threshold.

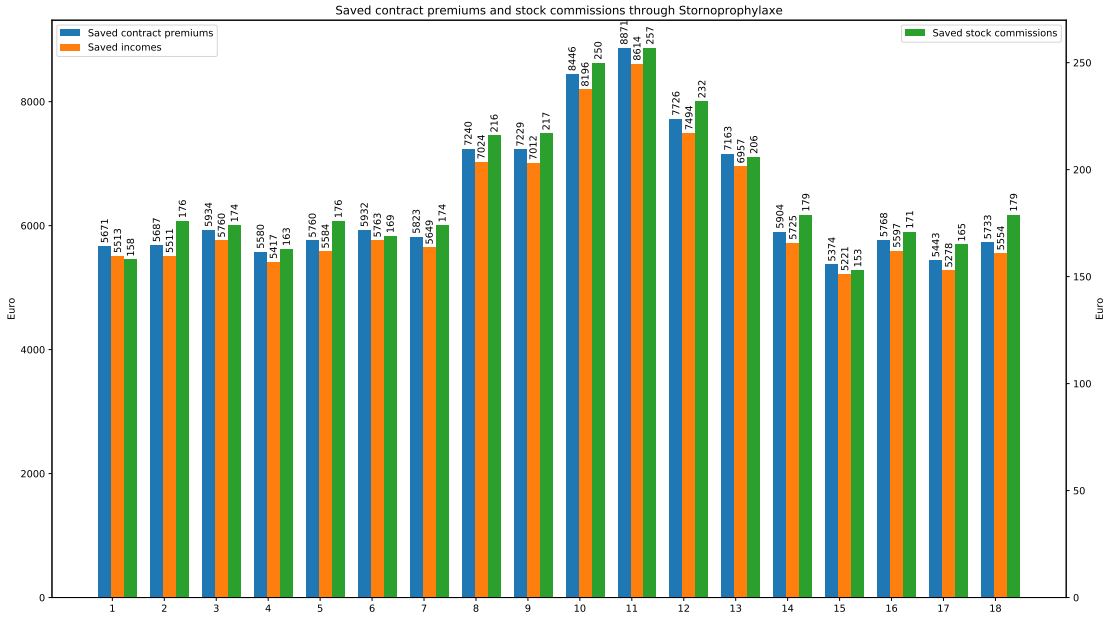


Figure 5.14: Scenario Pandemic: Saved contract premiums and stock commissions through Stornoprophylaxe

Question 5: Question 5 is not presented here as we explained beforehand in Section 4.4 Question 5.

Question 6 & 7: The following plot, Figure 5.14, represents the monthly contract premiums saved, the resulting stock commissions saved and the resulting revenue saved for the SIGNAL IDUNA Group.

The bar chart shows the respective values over the 18 months in Euro. It is noticeable that starting from month 8 more monetary worth could be saved, the highest ratio was reached in month 11 and starting from approximately month 14 or 15 again original values are measured. It is encouraging to see that this kind of development is also covered by the development of the first scenario and thus shows that our model works. We have more cancellations, but we also predict more of them. How good the ratio between saved and lost contracts is, however, cannot be seen here.

Nevertheless, this plot shows that the model is continuously working well and is improving temporarily, even if this is initiated by an increased number of cancellations due to the pandemic scenario.

Whether we should set thresholds remains open at this stage. For possible business SLAs (Service Level Agreements) in connection with respective KPIs (Key Performance Indicators), thresholds could be derived from them. If there is

no such thing, we consider this metric more informative than warning. The reason for this is that we need feedback from the ADPs for this metric to know whether a contract has been saved or not, but also because there is still the maximum time lag of 3 months until we know whether the contract has been lost or not. However, there it is not possible to quicken this information path.

Despite this, we consider this metric to be an enormously important criteria for assessing and representing the monetary success or failure of the model.

Question 8: In contrast to the previous question, where we need the feedback from ADP to answer, we can evaluate after 3 months whether the contract is still in stock or not. Accordingly, we think this metric is more interesting because of the decoupling from other parties.

Nevertheless, due to normal fluctuation in this segment we have consistently high values. In addition, we usually prefer to use metrics that show whether the model is successful and not where it can be improved.

We believe that this metric should be used secondary to the previously presented metrics, but should still be included, as it is a fall-back metric if information from the external party is not available.

Again we decided not to use a threshold for the same reasons as for question 6 & 7.

The course of the metrics from Figure 5.15 is similar to the course of the metrics

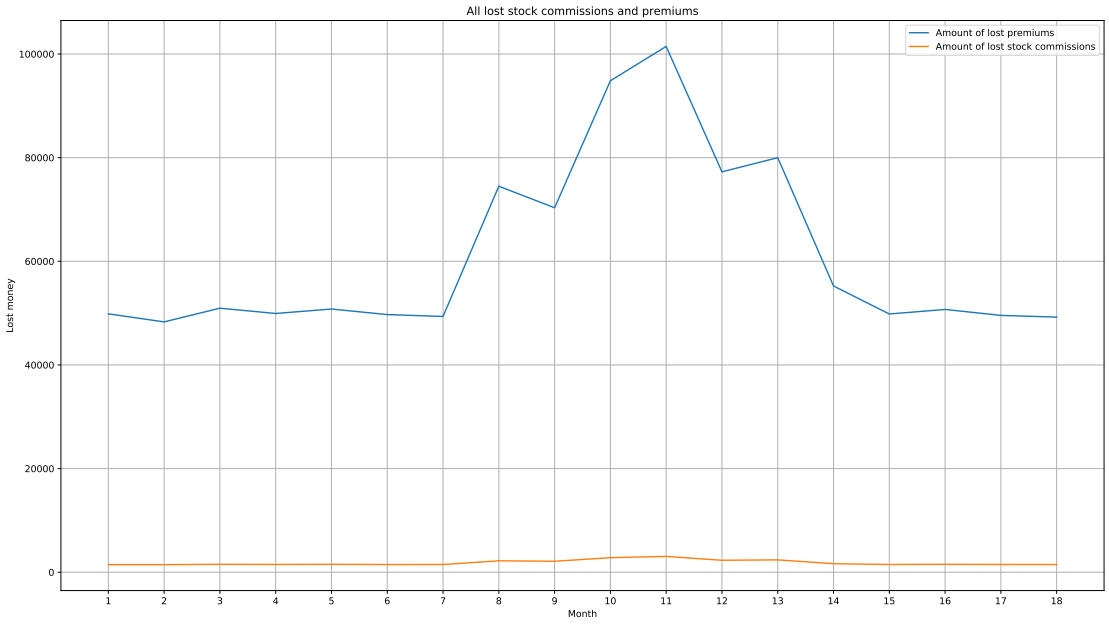


Figure 5.15: Scenario Pandemic: All lost stock commissions and premiums

from the previous question, Figure 5.14.

Question 9: The following plot, Figure 5.16, shows the theoretical increase of the insurance stock. It is noticeable that the percentage values are very small. This is due to the fact that we have a large stock and only a small portion of it cancels the contracts and only a part of those are predicted. Moreover, not all contracts are included for the prediction per month, but only selected contracts. This is why these values are so low, but normal.

The course is similar to the previous ones, which is also related to our scenario, where between month 7 and month 13 more contracts can be saved.

Question 10: The following metrics (Figure 5.17) show on the one hand the saved contracts again, but also the average of the saving rates over the current model lifetime. We intended to generalize individually occurring peaks by using the average. The motivation for this is that anomalies that only occur individually do not cause unnecessary warnings. Thus we assumed that as soon as an anomaly occurs and then the values return to normal afterwards, no re-training of the model is necessary.

However, it turned out that a moving average is more appropriate, as the metric becomes less significant after a longer run-time of the model.

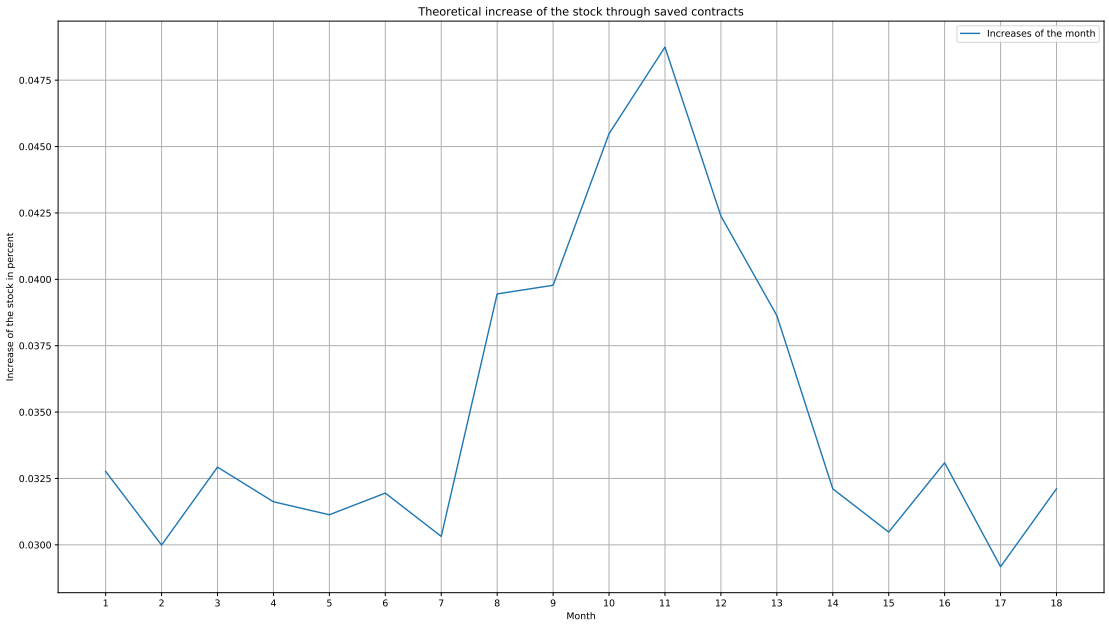


Figure 5.16: Scenario Pandemic: Theoretical increase of the stock through saved contracts

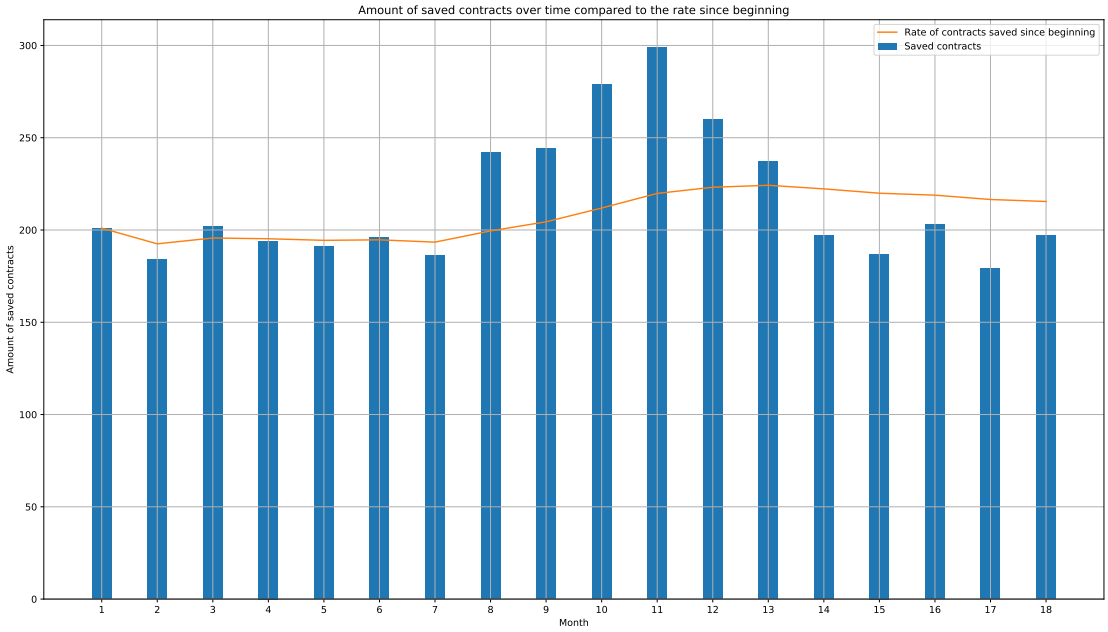


Figure 5.17: Scenario Pandemic: All cancellations (TP and FN) and therefore all lost contracts

Question 11: Since Question 11 only refers to the metric Recall, but with different terminology, we will not describe the plot or the benefit again here. It was a business goal to keep or reduce the value of Recall as low as possible. Accordingly, this would also be the goal here.

The metric is displayed in Figure 5.18 with the adjusted terminology.

Question 12: Question 12 is not presented here as we explained beforehand in Section 4.4 Question 12. A manual evaluation by a data owner and a creation and maintenance of the permission database is beyond the scope of this thesis.

Question 13: Just as in question 12, we have integrated a subjective assessment of an employee in question 13. This time it is a case officer. Accordingly, there is no plot for this metric.

The second metric tries to support the case officer in identifying discrimination by using the data feature importance. An automatic detection of discrimination is not possible, at least not at this stage. One reason for this is that it is not exactly defined for each case when discrimination starts and what is considered normal. Nevertheless, there are indicators that frequently point to discrimination. This metric aims to show these indicators and figures 5.19, 5.20, 5.21 and 5.22 try to visualize this as clearly as possible.

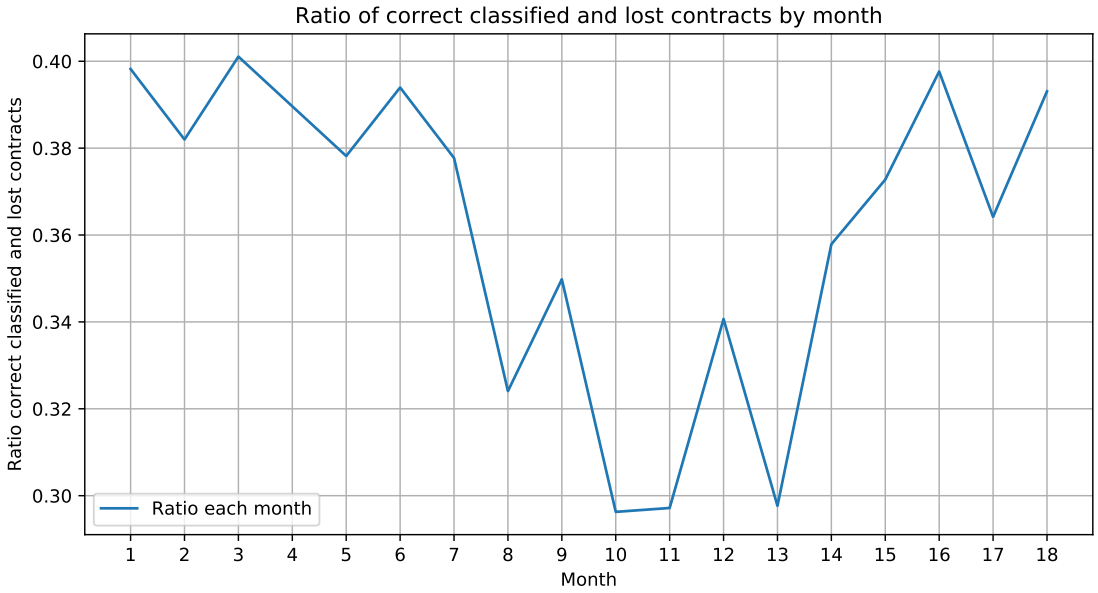


Figure 5.18: Scenario Pandemic: Ratio of correct classified and lost contracts by month

The first two plots (Figure 5.19 & 5.20) take the six data features that contribute most to a positive or negative prediction. A combination of the two plots only resulted in an unreadable compression. For this reason, they were split into two plots, one for positive contributions and one for negative contributions. However, there are no major fluctuations in the values, which can generally be seen as positive. It is important to note that the fluctuations of both plots are similarly strong, but are more noticeable in the negative contributing data features, because they are closer to each other and therefore the scale could be chosen smaller. Therefore the fluctuations look stronger than they actually are.

The last two plots (Figure 5.21 & 5.22) show the ten most contributing data features, but in a different way. They show how often each month the data feature has been scored a certain value. This allows the case officer to track changes in the plots, as well as to identify outliners and anomalies that would have been missed in the previous plots which used the average. Again, there is almost no change, which can be considered positive, as mentioned above. Nevertheless, it cannot be completely ruled out that there may be no visible or hidden discrimination in the predictions. For this purpose, we still need a person who is able to assess and evaluate this conclusively with knowledge of regulation and common sense.

Question 14: The last metrics try to answer the question whether a model is giving transparent predictions at all times. Since we did not develop a visualization

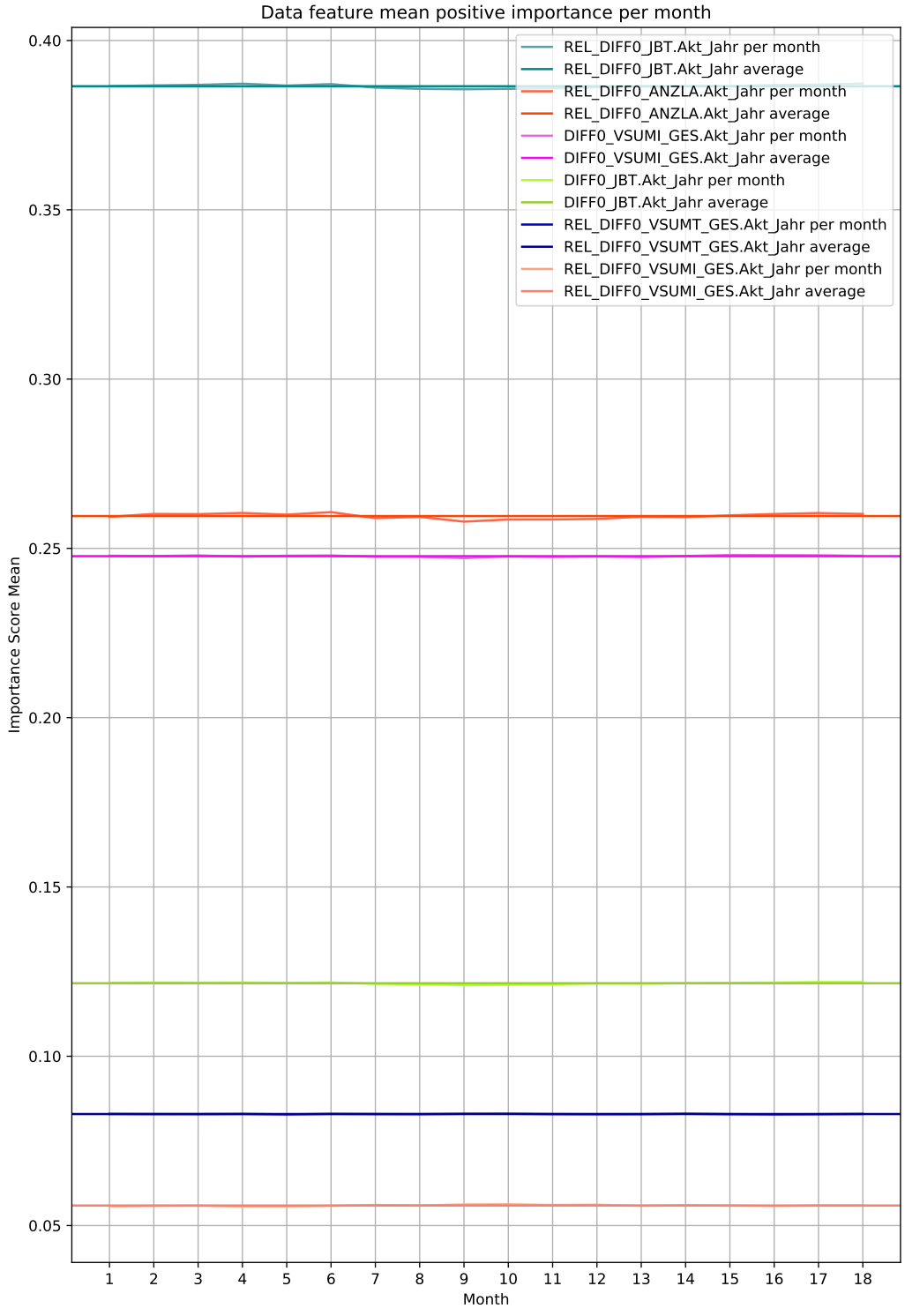


Figure 5.19: Scenario Pandemic: Data feature mean positive importance per month

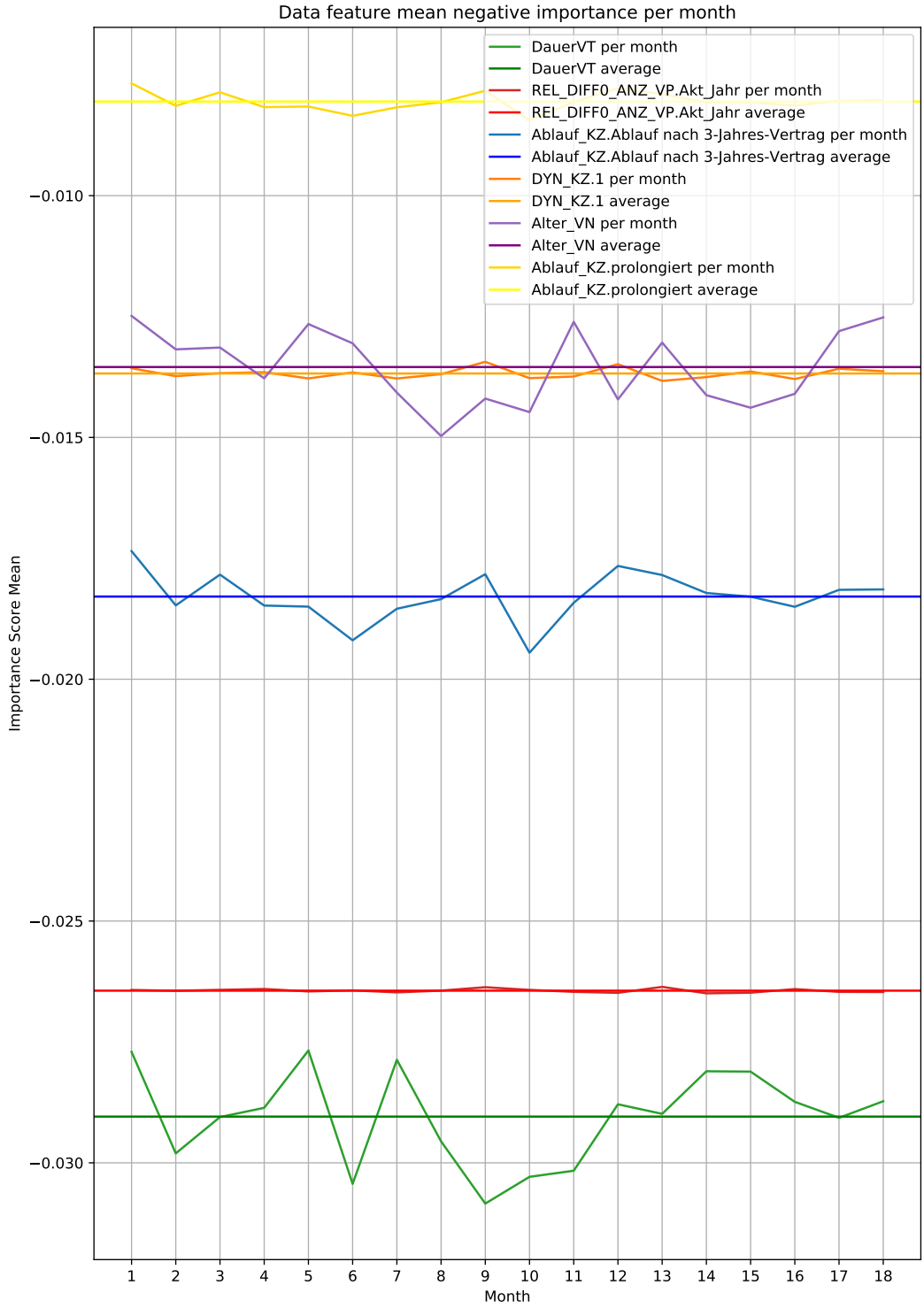


Figure 5.20: Scenario Pandemic: Data feature mean negative importance per month

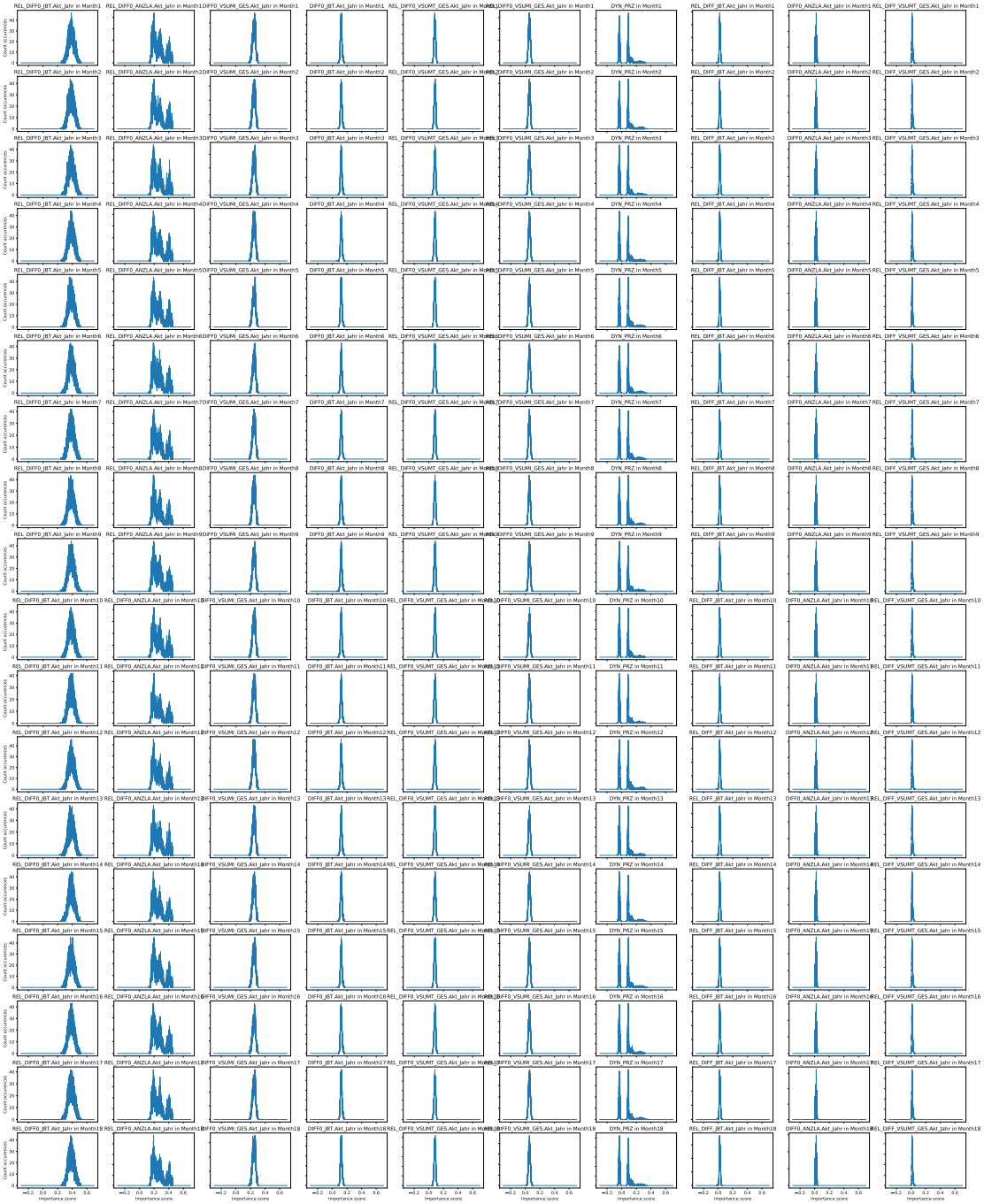


Figure 5.21: Scenario Pandemic: Data feature over time - positive

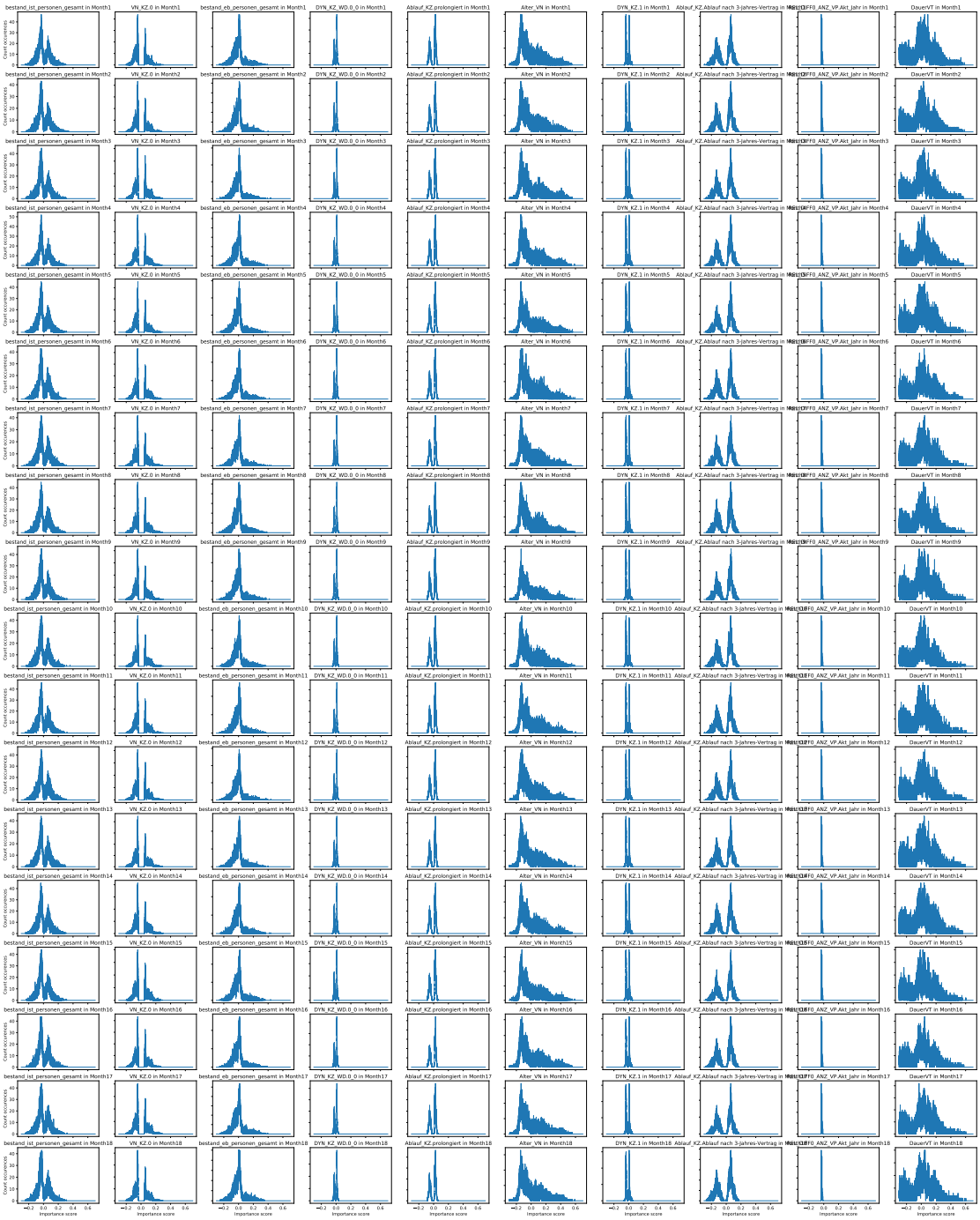


Figure 5.22: Scenario Pandemic: Data feature over time - negative

algorithm that shows how the model moves through the decision trees, in our case, to compute the prediction, we cannot answer this question with yes or no, nor with our metrics alone.

In the meantime, it has become apparent that the entire topic of regulatories generally has to cover very complex matters and therefore often requires a person with specialist and background knowledge to assess the given facts. This is again the case here. To support such persons we have created two metrics that provide valuable information to support transparency and make decisions made by the model explainable and therefore transparent.

We had already presented both metrics. The first is the model certainty (fig. 5.10) and the second is the data feature impontances (fig. 5.19, 5.20, 5.21 & 5.22). With model certainty, it should be made visible how the predictions behave in general, so that it can be assessed whether the examined prediction agrees with the majority of the predictions and seems to be rather okay or not. Similar is the idea for the data feature importance, but here it can also be seen, which data features of the prediction are how prominent and if this is normal.

5.3 Experiment 2 - Scenario Slow Changes

While the first experiment represents an abrupt and fast change of the external world, our second experiment represents a slow change of the surroundings. This can often be seen in categories such as demographics. The objective is to see if our metrics will alert us early enough to slow changes before there is a large loss of quality. An example for those slow changes would for instance be the annual raising pension age.

To perform the experiment we have created the scenario slow changes. In this scenario we almost linearly increase the cancellations by a certain percentage. For the increase of the cancellations we decided to achieve this by manipulating the values of one of two data features in the respective contracts. Those are the data features:

- bestand_ist_personen_uv → Development of the portfolio of insured persons in the accident insurance contract in the fiscal year.
- Inkasso_jn → Was a payment reminder or a dunning letter sent to the policyholder, yes or no?

We also let these increase linearly, but not as continuously. We hope to be able to identify trends in the metrics, whether the model changes or will change its behavior by that.

The process of the experiment is shown in Figure 5.23.

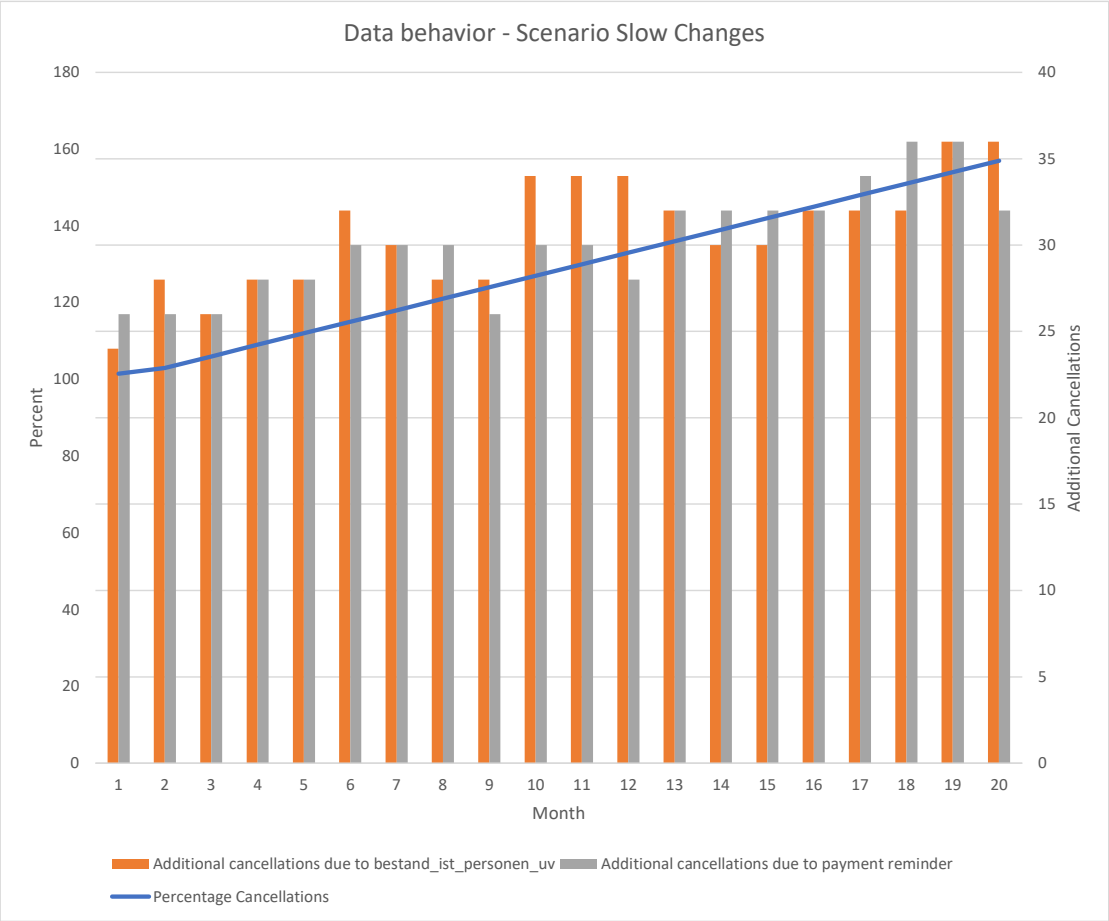


Figure 5.23: Scenario Slow Changes - Data behavior

5.3.1 Experiment 2 - First attempt

Before we describe the demonstration and our results in the second experiment, we would like to briefly mention the first failed attempt.

In the first attempt, we made the same assumption, but we had only set the data adjustments half as high. It turned out, however, that the adjustments were too small in relation to the data set size to allow the metrics to work. Thus, we could not distinguish which metrics would actually add value and which would be less suitable because most of them never exceeded the defined thresholds. So we doubled the values and got better results in the end.

5.3.2 Experiment 2 - Results

In the second trial of the second experiment we achieved the following results:

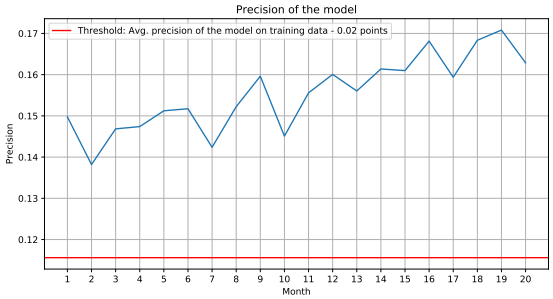


Figure 5.24: Scenario Slow Changes: Precision

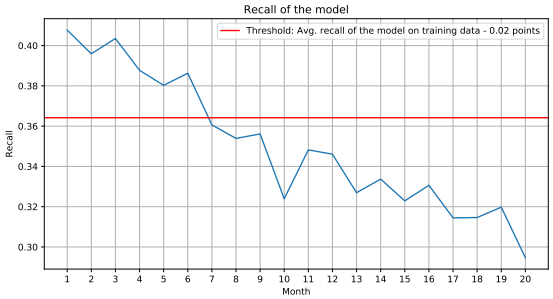


Figure 5.25: Scenario Slow Changes: Recall

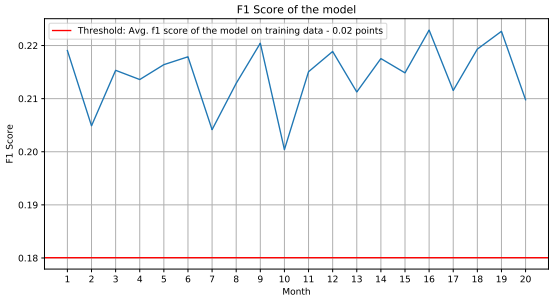


Figure 5.26: Scenario Slow Changes: F1-Score

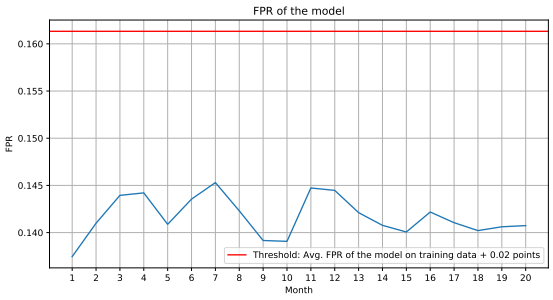


Figure 5.27: Scenario Slow Changes: FPR

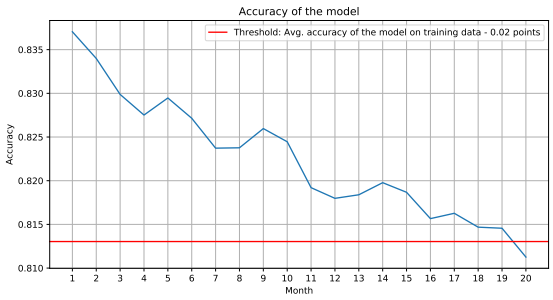


Figure 5.28: Scenario Slow Changes: Accuracy

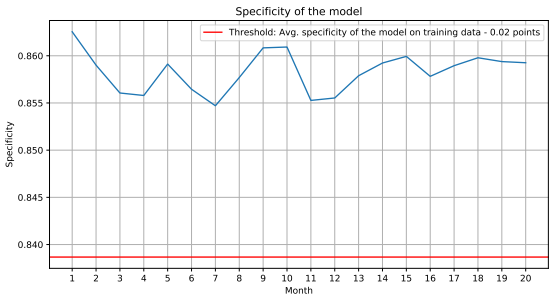


Figure 5.29: Scenario Slow Changes: Specificity

Question 1: In contrast to the first experiment, the results of the metrics precision, recall, f1-score, fpr, accuracy and specificity look very different. Only the metrics recall and accuracy passed the defined thresholds at some point. Recall in month 7 or more precisely with a delay of 3 months in month 10. Accuracy, on the other hand, in the last month, month 20, or including the delay in month 23. The remaining metrics remain above or below the thresholds.

A very interesting aspect is the behavior of the accuracy. While four out of six metrics do not show a trend, it is clear that accuracy is slowly but surely approaching the threshold and in the end finally did. Two assumptions can be made. Either the threshold is not chosen strictly enough so that the accuracy with a clearly recognizable trend is not passing it, or we still have not adjusted enough contracts for the scenario so that the simulation has changed too little to pass the thresholds. However, since we already doubled the adjustments and the recall metric is clearly passing the threshold quite early it might be the first assumption. This is supported by the fact that the accuracy is already deteriorating by 0.025 points which is on a scale from 0 to 1 not significant but still recognizable.

Another mentionable aspect is that the precision, unlike all other metrics, is improving instead of deteriorating.

In the end we would still have received an early enough alarm because one of the six metrics fell below the threshold in month 7.

Question 2: Similar to question 1, only the metric recall exceeded the threshold quite early. Not like in question 1 the accuracy is not passing the threshold, but got quite close to it. For the other four metrics it seems as if the threshold is very far away, but they are usually only about 0.02 to 0.04 points away. Therefore it only seems that way.

The bottom line is that the rolling average metrics look very similar to those without an rolling average, but with less noise.

As already mentioned in Question 1, it is debatable whether the threshold should be selected stricter, as the metric recall is passing the threshold only a month later then in question 1 (remember the 3 month delay). Therefore, we have the same situation as before.

More important would be to discuss if the rolling averages give a benefit compared to the normal metrics. A direct benefit of the rolling average metrics is unfortunately not recognizable immediately, as we already stated in Experiment 1. Thus, it might not be necessary to include them or it is necessary to adjust the rolling average to another time span like for instance three month.

Question 3: Just like in Experiment 1, no tendency can be seen in the plot shown in Figure 5.31. The predictions are similarly distributed over the months, so that no effects of the included slow changes are visible.

In contrast to Experiment 1, no big changes can be seen in Figure 5.32 and 5.33 this time. There is no trend in the metrics over all 20 months, thus the distribution of the average predictions is similar and does not indicate a need for intervention. Only the standard deviation fluctuates a little more, but is still marginal and therefore not large enough to exceed the thresholds.

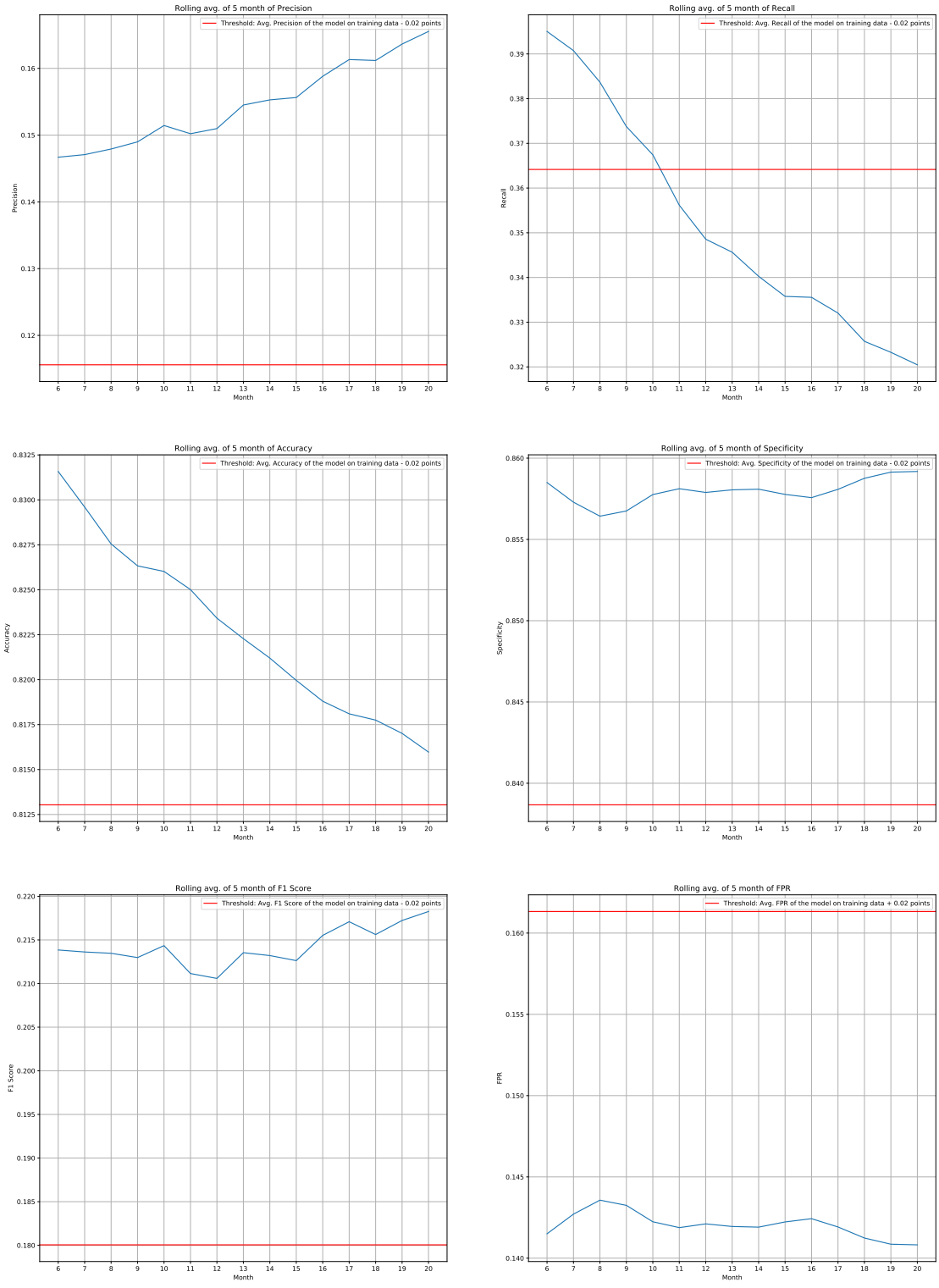


Figure 5.30: Scenario Slow Changes: Standard Metrics Rolling Average

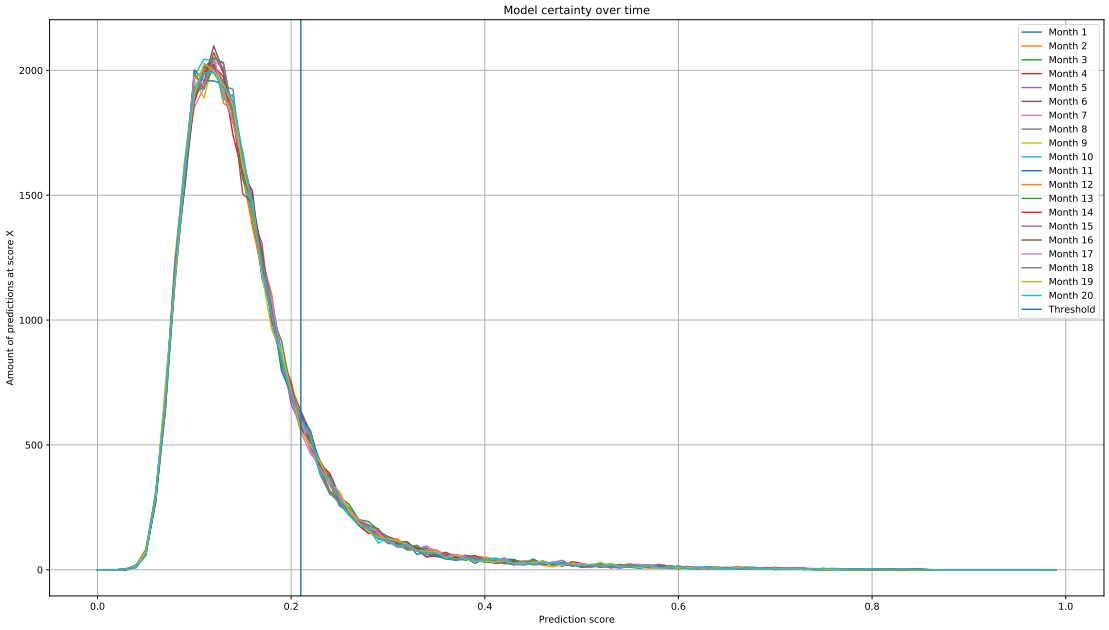


Figure 5.31: Scenario Slow Changes: Model certainty

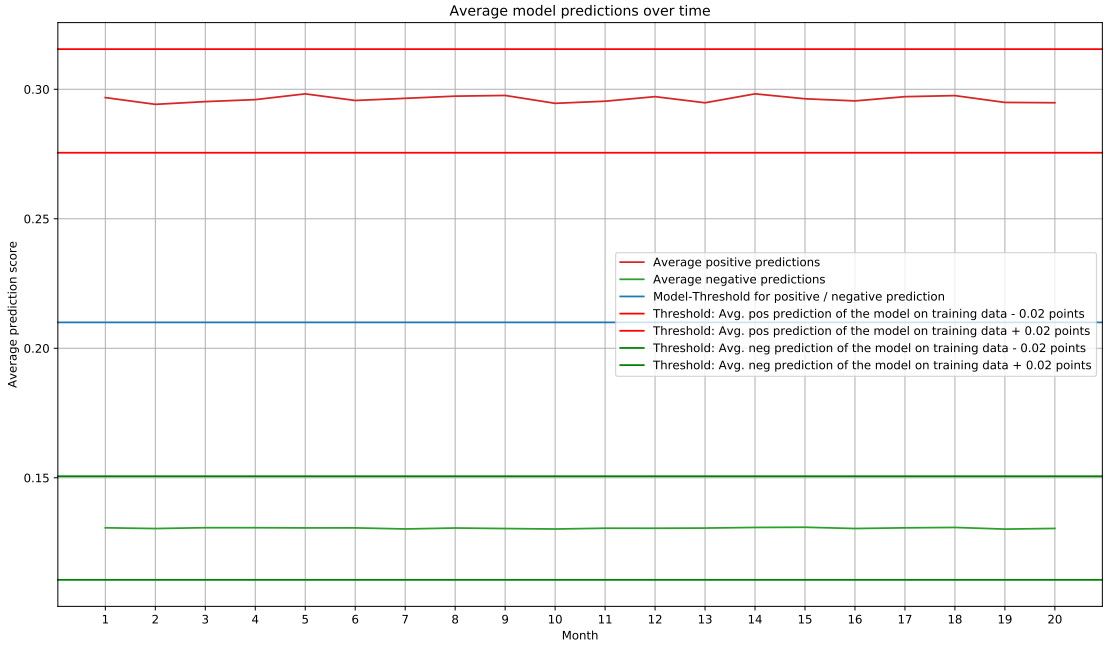


Figure 5.32: Scenario Slow Changes: Averages of positive and negative predictions

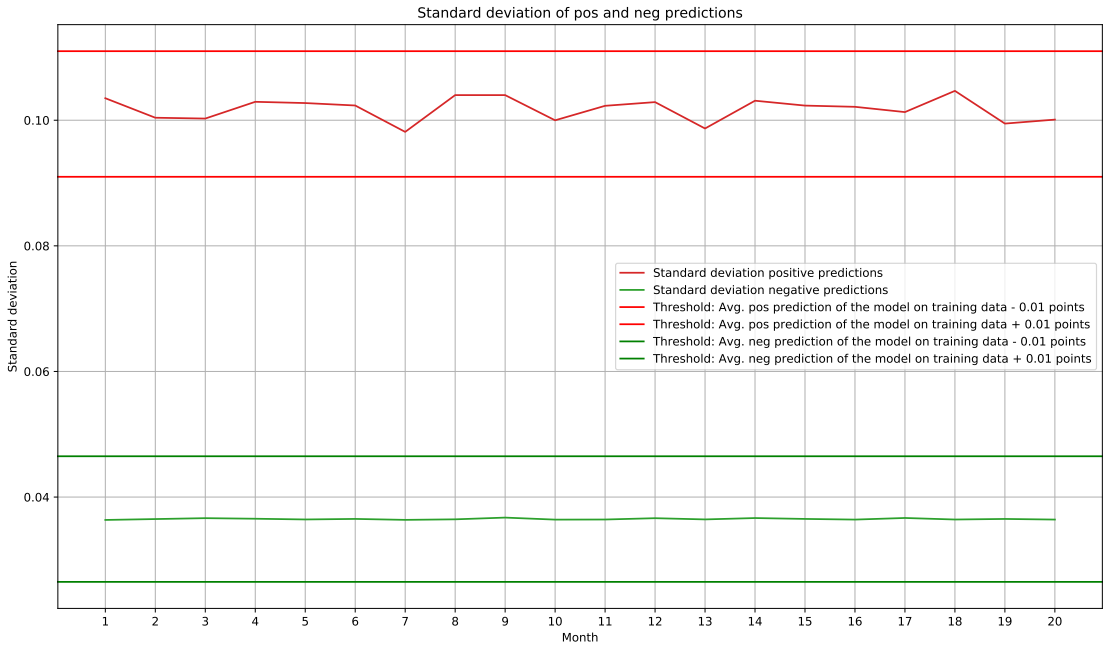


Figure 5.33: Scenario Slow Changes: Standard Deviation of the positive and negative predictions

Question 4: By answering the question, whether the bias of the predictions is still in a plausible range, no large fluctuation can be seen in the second experiment either. The score sometimes goes up a little bit, sometimes down a little bit, but remains around zero for the whole time. This is behaving completely contrary to the first scenario.

Accordingly, no slow transition of the predictions can be seen here as well. See Figure 5.34.

Question 5: Question 5 is not presented here as we explained beforehand in Section 4.4 Question 5.

Question 6 & 7: For the explanation of the plot to question 6 and 7, we would like to refer to 5.2 as we do not want to repeat ourselves all over again.

In contrast to the plot from the first experiment no significant change over time can be seen in this plot (Figure 5.35). If one would draw a line through the average values, one would see a small increase over the whole 20 months. However, this is so small that we would categorize it as too small in relation to our scenario. The values seem to be rather in an irregular fluctuation, since month 20 has similar values as month 1, but also since month 6 has similarly values as month 18.

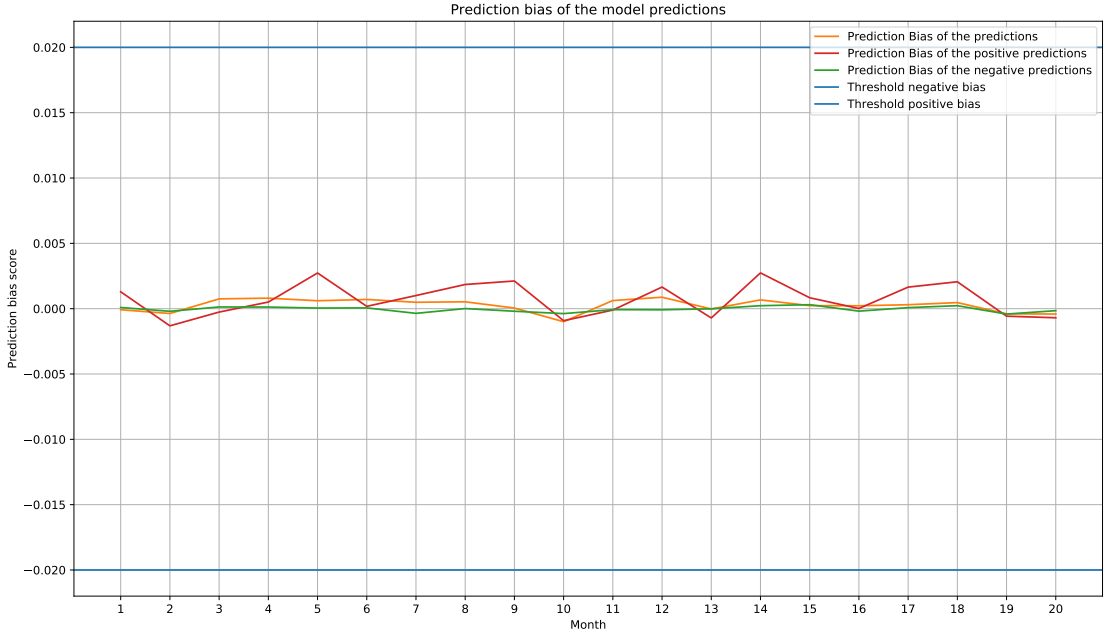


Figure 5.34: Scenario Slow Changes: Prediction bias of the model predictions

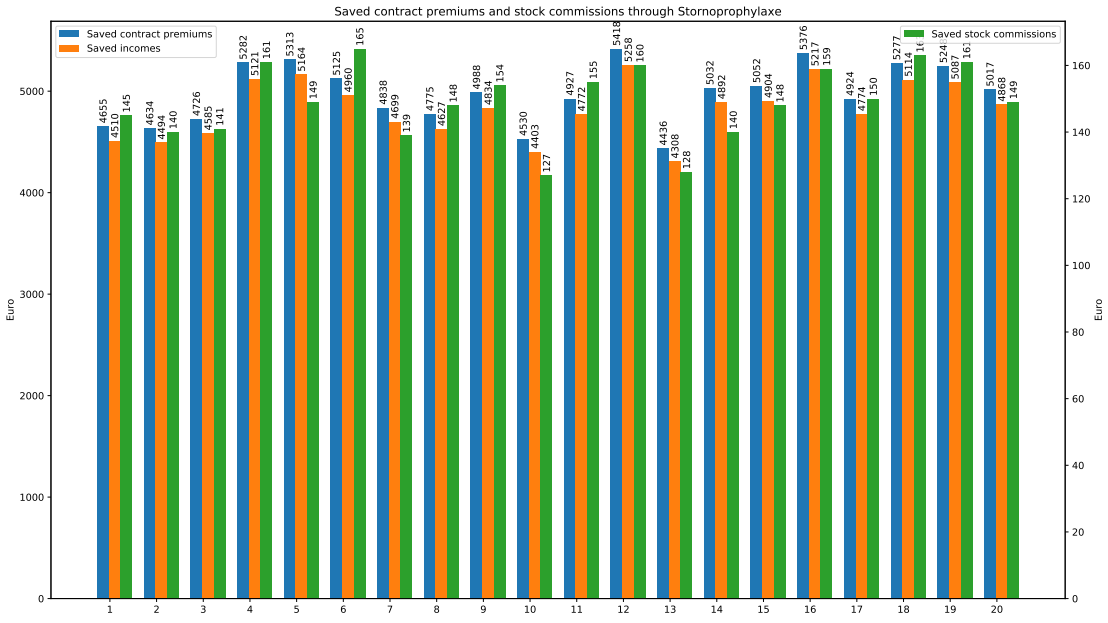


Figure 5.35: Scenario Slow Changes: Saved contract premiums and stock commissions through Stornoprophylaxe

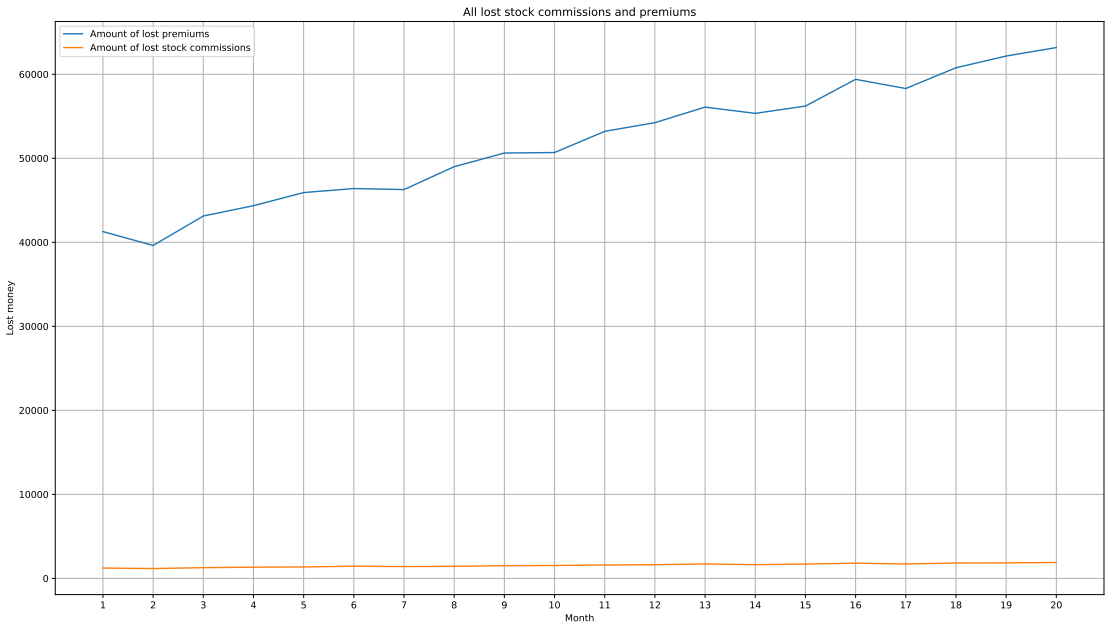


Figure 5.36: Scenario Slow Changes: All lost stock commissions and premiums

A trend is rather less to be recognized here, even though a minimal increase would be recognizable via the calculation of averages over time.

Question 8: Unlike the previous question, where we could not see any major changes in Figure 5.35, we can see a clear trend in the plot in Figure 5.36. This is the result of adjusting the data set for our "Slow Changes" scenario so that more people will quit over the months due to the two adjusted data features. Thus, this behavior of the metric was expected. Consequently, it is a confirmation for us to use this metric because it is now proven to work and we can see that it adds value. It is clearly visible that more contracts are lost from month to month, which cannot be reported about the saved contracts from the previous plot.

From this it can be concluded that both metrics should be included, but especially the metric of lost stock commissions and premiums should have a stronger weight for slow changes. Here it is clearly evident that the model works worse with the new changing data. Since the trend is so clearly visible, it is a worth consideration to include a threshold for this metric, although it was intended to be informative, as explained in Experiment 1.

Only the scale of lost stock commissions should be chosen differently. Due to the different sizes of the values between the two data lines, the trend of the lost stock commissions is not easy to recognize. This can be solved for example by a secondary y-axis.

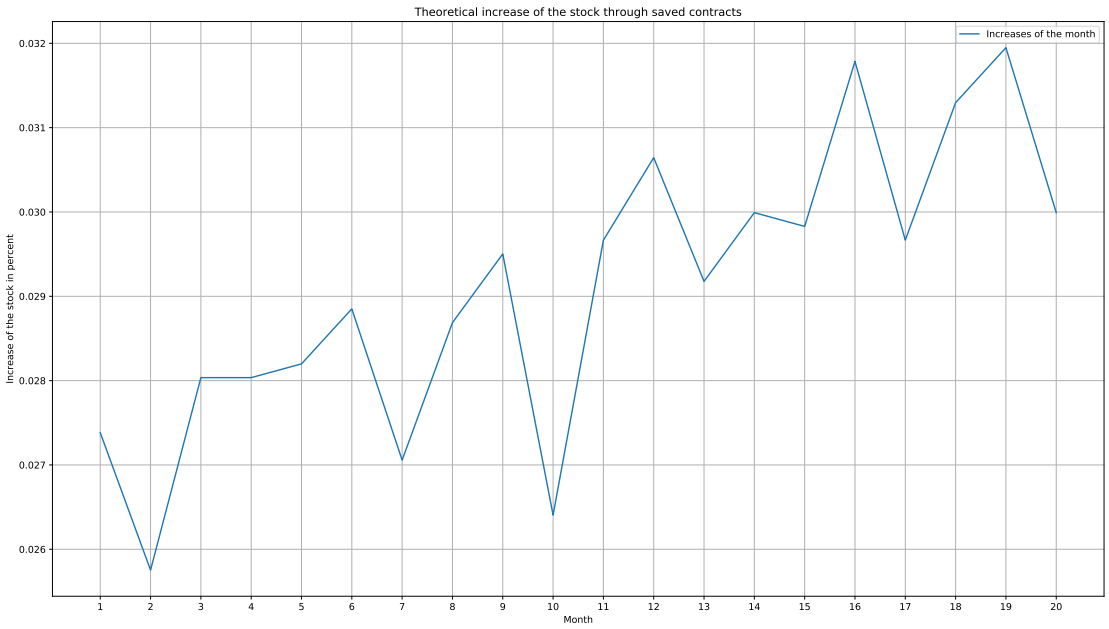


Figure 5.37: Scenario Slow Changes: Theoretical increase of the stock through saved contracts

Question 9: In the following plot, Figure 5.37, one can see the theoretical increase of the insurance stock. The plot was already described in detail in experiment 1.

The most important difference to the Figure 5.16 is that the theoretical enlargement is much smaller and, above all, fluctuates much more. In Question 6 & 7 we have already stated that no major changes can be seen in the saved contracts, but a small positive trend could be seen if one would take the average values. This is again confirmed here, since one can clearly see a positive trend. However, this trend is so marginal that it can almost be neglected. It amounts to only 0.0026 percentage points in the comparison of month 1 and month 20. This observation is also supported when comparing the best month with the worst month. Hence, month 2 had added only 0.0257 percent and month 19 0.0319 percent contracts. The difference with 0.0062 percentage points is larger than when comparing month 1 and month 20, but still insignificant. For this reason, we would like to point out that one should always try to keep the same scale zoom at all times, otherwise wrong assumptions could be made if one does not pay close attention to the axes of the plots.

In conclusion, there is both, a strong fluctuation and a positive trend, although it is so small that it was nearly not visible in the previous figures such as Figure 5.35. However, this trend has become much clearer in this metric.

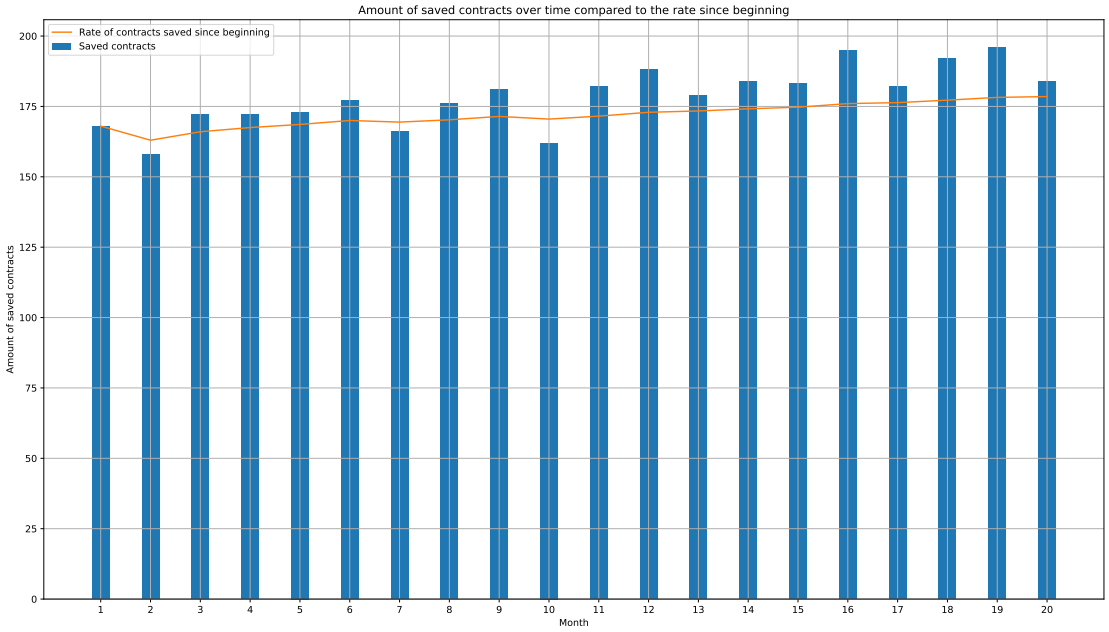


Figure 5.38: Scenario Slow Changes: All cancellations (TP and FN) and therefore all lost contracts

Question 10: The metrics presented in Figure 5.38 show the number of contracts saved each month and the average of this number over the entire life of the current model. In the first experiment, we criticized that the average loses significance at some point and that a rolling average should be chosen. This cannot be supported here so far, since no generalization of the average is yet visible in this figure. To what extent this behavior might still occur can unfortunately not be considered here after 20 months.

Otherwise, the behavior of the plot course is similar to that of Figure 5.37. But in contrast to the previous figure, you can read the exact number of the saved contracts here, which might be a better illustration than presenting only very small percentages. Therefore, we consider this metric to be more suitable to be used for presentations on, for instance, business targets. Furthermore, it is much more concrete, since it directly refers to the number of contracts and does not use a percentage, which appears to be very abstract when one does not know the relation between total inventory and forecasts.

All in all, we consider this metric to be more appropriate than the previous metrics which dealt with saved contracts.

Question 11: Since Question 11 only refers to the metric Recall, but with different terminology, we will not describe the plot or the benefit again here. It was a

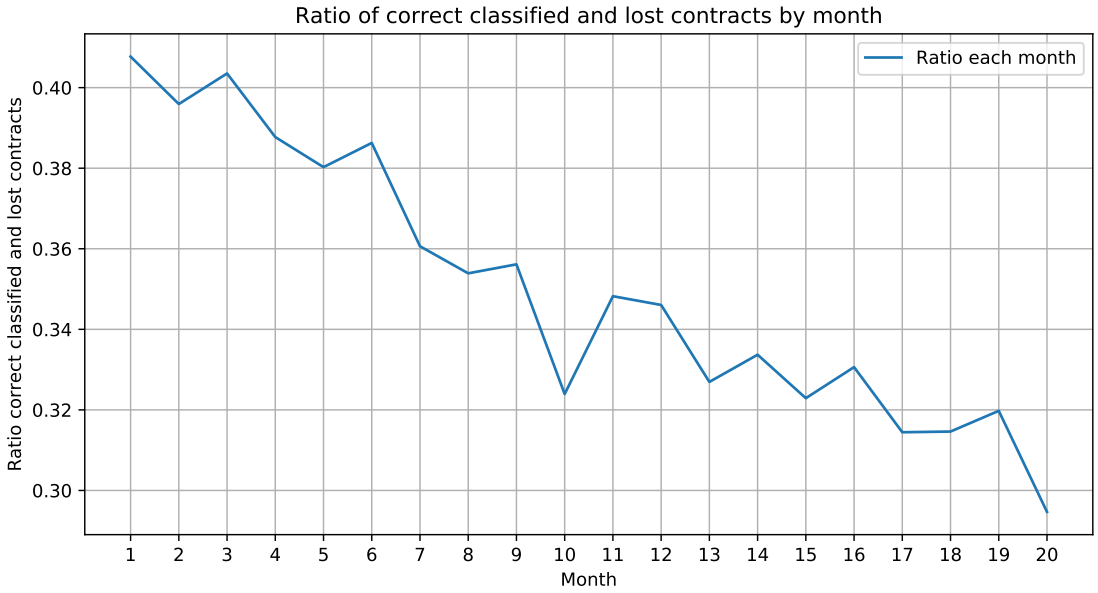


Figure 5.39: Scenario Pandemic: Ratio of correct classified and lost contracts by month

business goal to keep or reduce the value of Recall as low as possible. Accordingly, this would also be the goal here.

The metric is displayed in Figure 5.39 with the adjusted terminology.

Question 12: Question 12 is not presented here as we explained beforehand in Section 4.4 Question 12. A manual evaluation by a data owner and a creation and maintenance of the permission database is beyond the scope of this thesis.

Question 13: As already described in the first experiment, it is very difficult to detect discrimination just by using metrics. Nevertheless, we try to simplify the process of detection by metrics. For the intention of these metrics we refer to the description in Section 5.2 Question 13.

It can be seen that in the first two metric plots the most important data features are all the same as in Experiment 1, except for one. The average values remained similar as well. Only the monthly values differ, while remaining basically within a similar overall range. We therefore conclude that the machine learning model and the weighting of the data features remain constant for both abrupt and slow changes.

From this we derive that, once the model has been classified as non-discriminatory, continuous monitoring is not required. Nevertheless, it should be checked from

time to time whether the courses or the ranking of the data features have changed and thus a different behavior has occurred.

This assumption is supported if the two additional metric plots are added. These show similar behavior as the first two plots. Only with the negative contributing data features a change in the ranking of the data features can be seen when comparing them with experiment 1. We believe that the constancy of the values and the ranking of the individual data features is a good starting point for the evaluation of non-discrimination as we suggested in the last paragraph.

Question 14: In order not to repeat ourselves in the last question, we would like to refer to the explanation and elaboration in Section 5.2 Question 14. In addition, we have already compared the metrics in the previous question to Experiment 1 and have presented the results in the figures 5.10, 5.19, 5.20, 5.21 & 5.22.

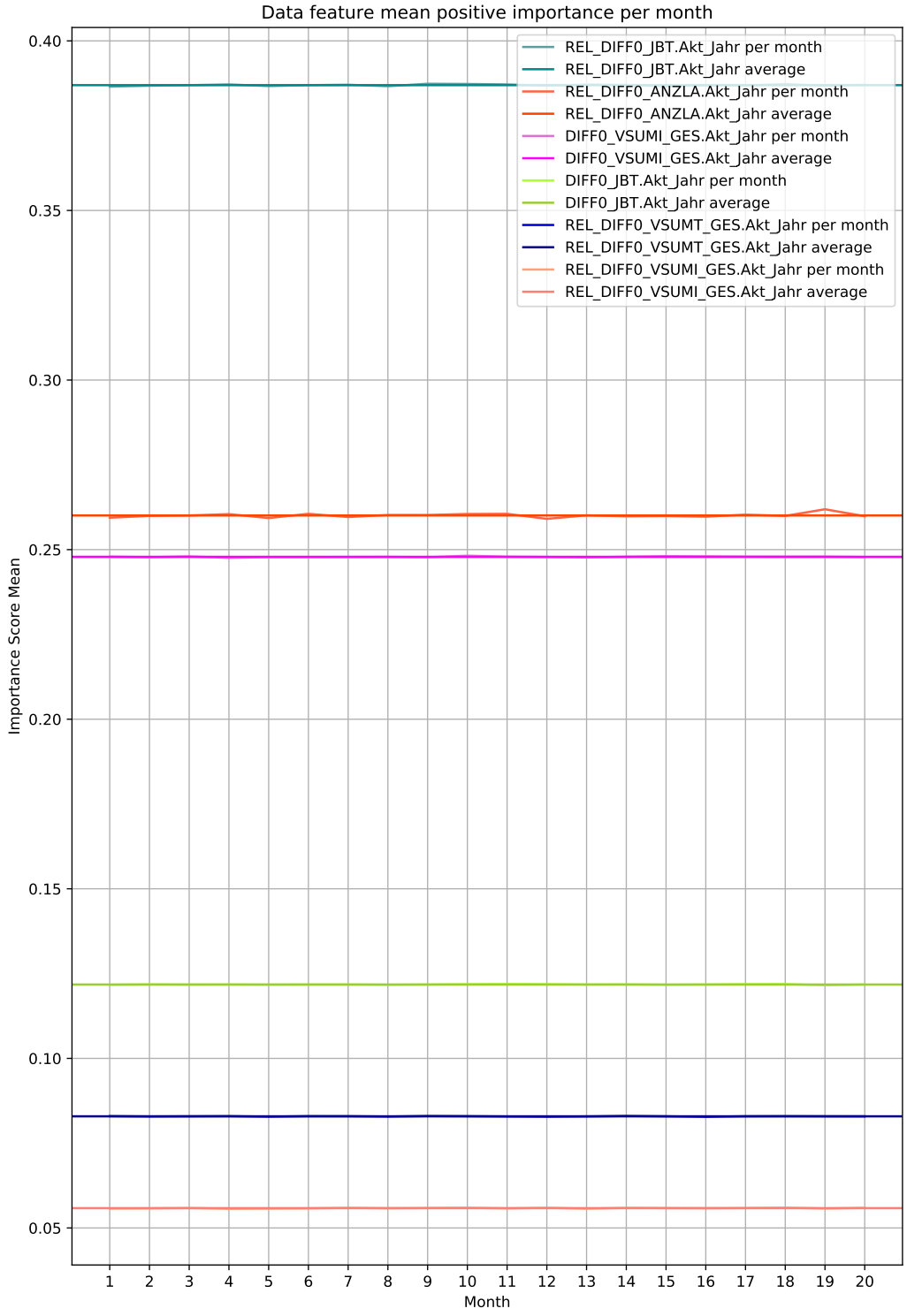


Figure 5.40: Scenario Slow Changes: Data feature mean positive importance per month

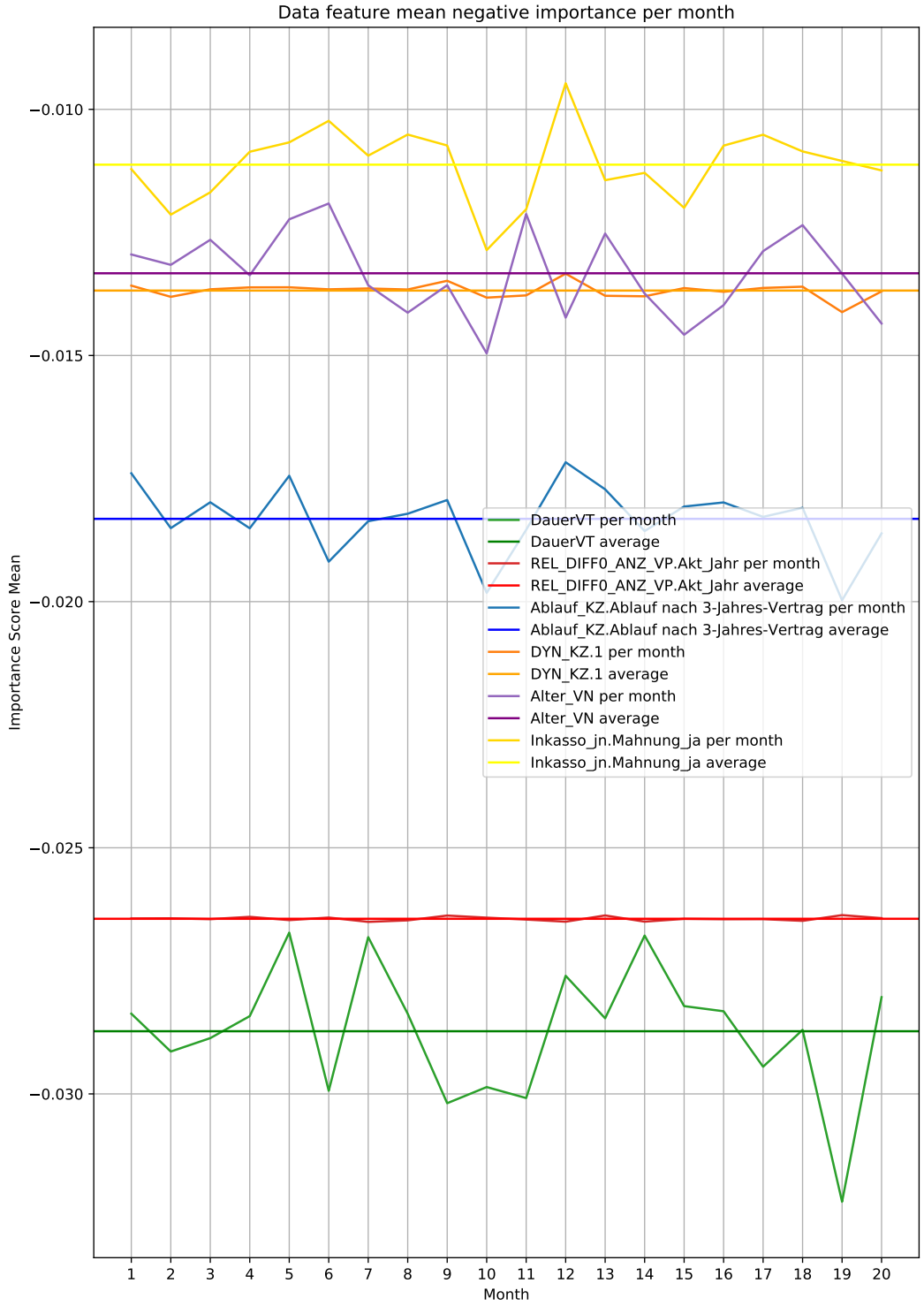


Figure 5.41: Scenario Slow Changes: Data feature mean negative importance per month

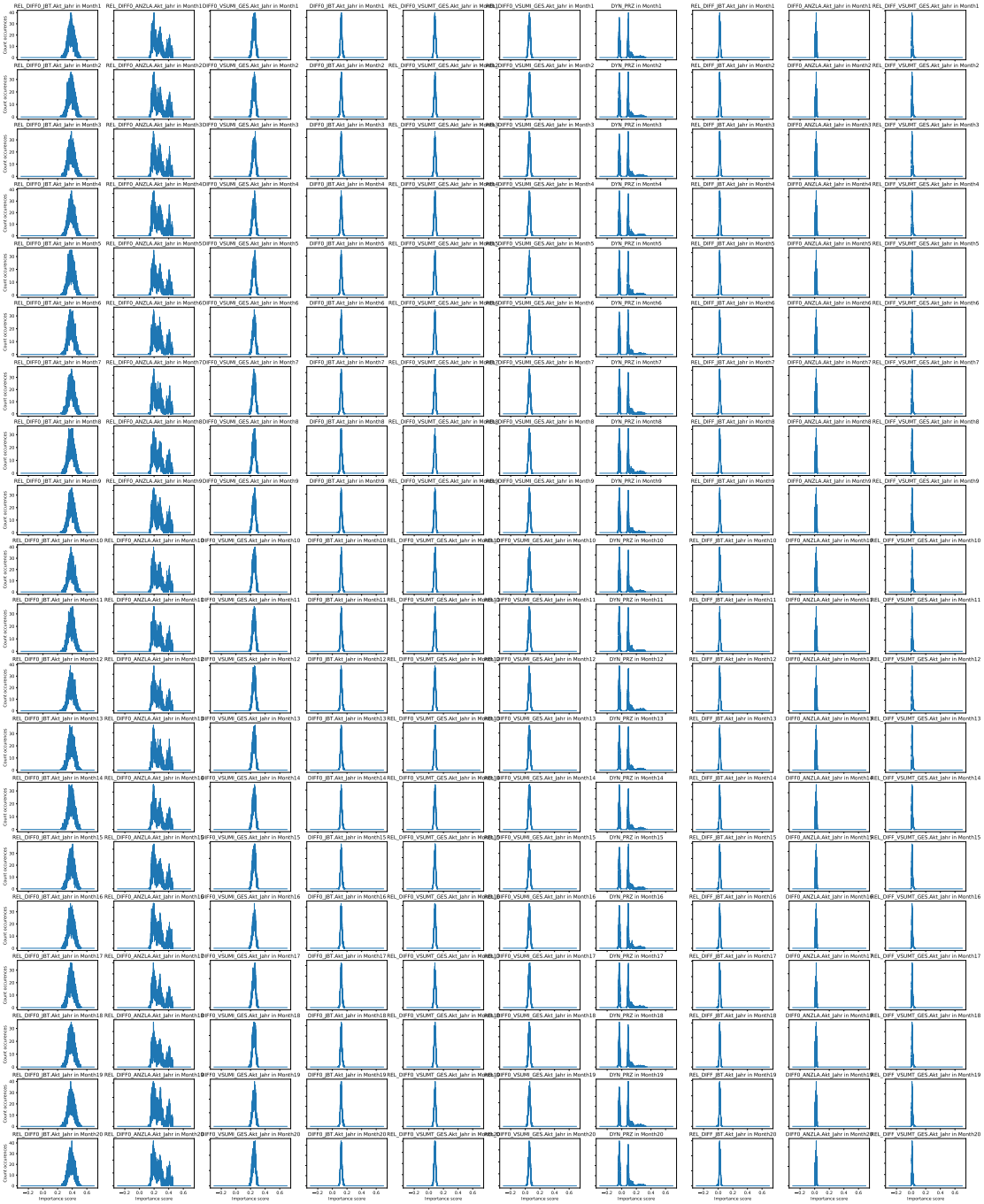


Figure 5.42: Scenario Slow Changes: Data feature over time - positive

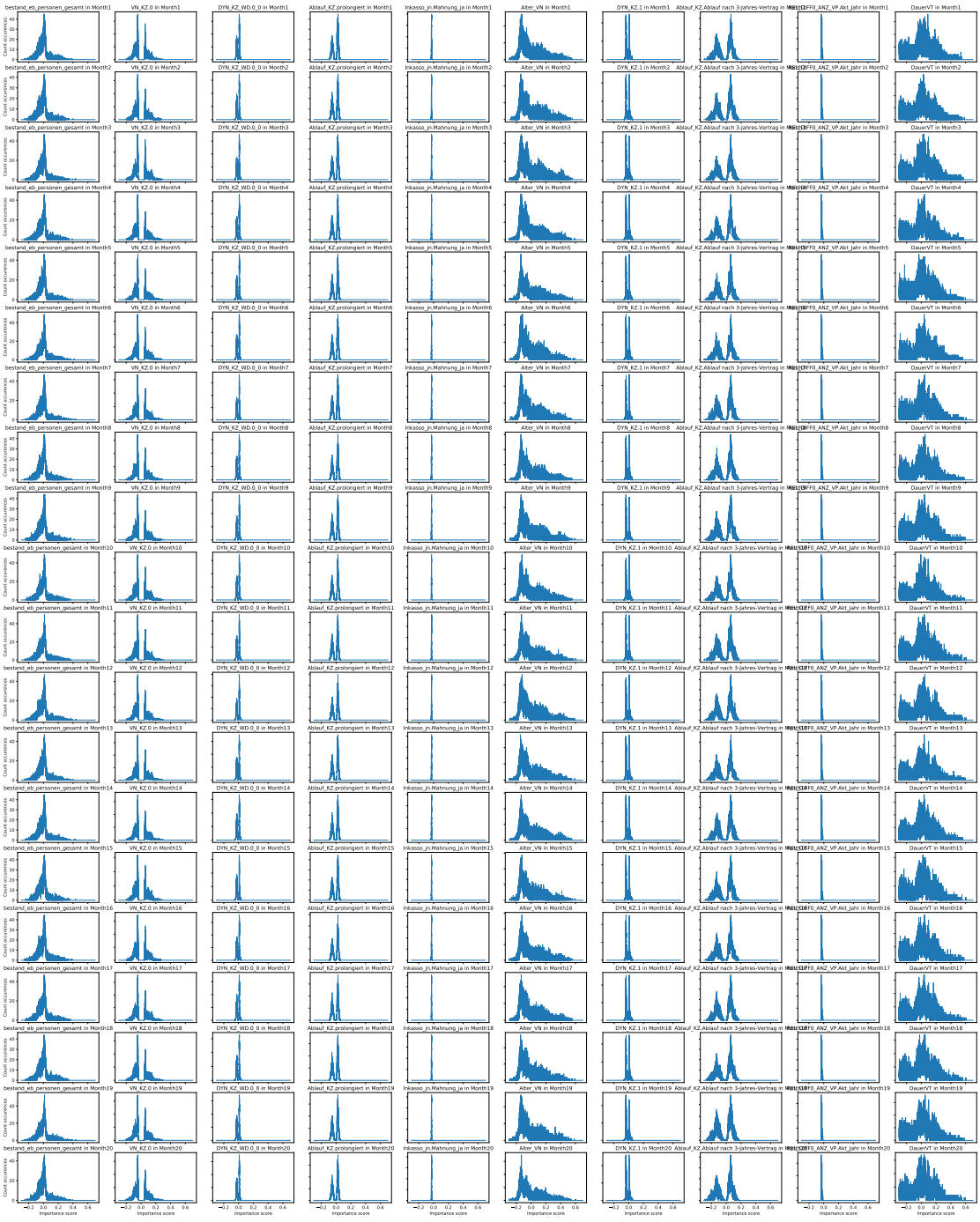


Figure 5.43: Scenario Slow Changes: Data feature over time - negative

Chapter 6

Evaluation

In the evaluation chapter we will take a closer look at the figures shown in Chapter 5 and evaluate how well the metrics have performed in each scenario. We want to show and evaluate how effective and efficient our metrics work, meaning how well the created artifact contributes to solving our problem.

6.1 Effectiveness and efficiency of the metrics

This section is divided into three subsections, since we have also concentrated on three areas in the metric design. We will therefore deal with the respective metrics separately for each of these areas in order to be able to consider separate evaluation criteria.

6.1.1 ML specific metrics

Standard metrics

The results of the standard metrics Precision, Recall, F1-Score, FPR, Accuracy and Specificity are very different. This behavior can be explained, but still surprised us. While Precision tended to develop positively, Recall and Accuracy in particular fell more sharply towards the threshold.

It turned out that Recall and Accuracy were effective in both scenarios, although Recall performed better than Accuracy. Thus, both seem to be valuable metrics to ensure the quality of a machine learning model.

Specificity and the FPR, on the other hand, only hit the pandemic scenario and remained unaffected by the slow changes scenario. Given that specificity is in principle FPR reversed, with one having the count of the true negative predictions and one having the count of the true positive predictions above the fraction line,

it should be decided on a case-by-case basis which metric should be used in which use case. Nevertheless, they only seem to be suitable for fast abrupt changes.

For our second scenario, the F1-Score and Precision turned out to be unsuitable. While Precision even improved in the Scenario Slow Changes, it fluctuated upwards in the other scenario, meaning away from the threshold. Although the F1 score briefly exceeded the threshold in the Scenario Pandemic, it remained far above it in the following months. In the Slow Changes scenario, it merely fluctuated. It is a good sign that Precision and Recall are mostly balanced, but this is not interesting for our use case. Hence, both metrics are not suitable for monitoring the quality of the model predictions.

Rolling averages

Contrary to our expectations, the rolling averages of the standard metrics turned out to be performing worse than using them without rolling averages.

The basic assessment and explanation of the metrics remains the same. The only difference is that there is significantly less noise.

Nevertheless, a major limitation is evident. On average, if the thresholds have been passed at all, the rolling averages over five months always needed one month longer.

It is important to mention again that due to our use case the standard metrics have an evaluation period of three months, which is not the case for the rolling averages. For this reason, we have to add three months to the standard metrics to be able to compare those two metrics. Nevertheless, the indicated one month more time needed compared to the rolling average metrics remains.

Consequently, we cannot recommend the rolling averages based on this analysis.

Distribution of metrics

Neither can we recommend the "model certainty over time". However, this may be related to our special use case. While the majority of our predictions are negative and the threshold is very low at 0.21, there is little room for the predictions to spread. Therefore we have a $\cap$-shape rather than a $\cup$-shape. Additionally, it is not possible to distinguish between the compared months. It simply does not change enough throughout them. We can't say whether the metrics are better reflected in a different presentation and thus could provide a good contribution. It also remains open whether the metric is more helpful in a different use case. Nevertheless, we cannot recommend this metric because it has not contributed to the evaluation of the model quality in either of the two scenarios.

In contrast, the "average model predictions over time" and the "standard deviation of pos and neg predictions" performed very well. Especially in the pandemic

scenario, both metrics hit directly at the critical month 7 and raised an alarm that the model was behaving abnormally. This metric is therefore very well suited for abrupt changes in the external world. For slow transitions, which we tested in the Scenario Slow Changes, the metrics did not trigger an alarm. However, the two metrics differ in the way information is displayed. The first one considers the average of the prediction values, the second metric considers the distribution of the prediction values. Preferably we recommend the metric of "average model predictions over time" and the second metric, the "standard deviation of pos and neg predictions", as an additional metric to the first.

Prediction bias

As the "prediction bias" is a similar metric to the two metrics described before, it is not very surprising that the courses are similar. Accordingly, the metric assessment is the same. The core difference, however, is that the "prediction bias" has a direct score, which in its presentation is independent of the actual prediction values. The major advantage is that even with models that have average prediction values around 0.1 and 0.9, there is no distortion of the representation. Another advantage is that the "prediction bias" is also comparable between models because it does not refer to the direct prediction values. We consider it a matter of taste whether one ultimately uses the "prediction bias" or the "average model predictions over time". Both are excellent and comparable metrics and are therefore very well suited.

6.1.2 Business metrics

Saved contract premiums and stock commissions

As could be seen in the survey, this metric was considered the best and most important of the business metrics. This assessment can indeed be found in the first experiment, where a positive trend in saved contracts can be seen during the abrupt changes. In the second experiment, however, this cannot be observed as clearly, although an increase could have been expected. We already described this phenomena in the previous chapter.

However, since we still have the issue of traceability in this metric since we are dependent on external information sources that are not equally available, we consider this metric to be rather problematic. Unless the requirements for a seamless information feedback loop are ensured beforehand, we advise against using this metric. However, if it is available, it should be integrated.

All lost stock commissions and premiums

In contrast to that, we can definitely recommend the metric "all lost stock commissions and premiums". This metric tends to signal negative behavior, as you want to hold as many contracts as possible, but we have no dependencies on external sources. This metric can be evaluated automatically at any time. Furthermore, it shows a clear trend in scenario 2, which is caused by changes in the outside world, as well as an outbreak of the metric in scenario 1.

These are all very good preconditions for this metric, so we consider it to be very good.

Theoretical increase of the stock through saved contracts

Next, we look at the metric that deals with the theroetic increase in the insurance stock through contract saves. In principle, it can be positively mentioned that the expected behavior of the metric occurred in both the Pandemic and the Slow Changes scenarios. However, the magnitudes of these values are so small that they lose their significance. A theoretical magnification of 0.0475% maximum in the Scenario Pandemic and 0.032% maximum in the Scenario Slow Changes are simply too insignificant to consider this metric important and relevant.

This behavior occurs because the ratio of saved contracts to the total stock is too small, which is also a consequence of the few cancellations per month compared to the total stock. Thus, this metric is indeed interesting because it has shown a good response to scenarios, but is less important than the other metrics created.

Number of saved contracts over time compared to the rate since beginning

Logically, the course of this metric is similar to the metric "Saved contract premiums and stock commissions". This is of course because in both cases these are metrics, which correlate directly to the saved contract premiums and stock commissions. The big difference with this metric is that it reflects the number of contracts saved and does not represent the direct monetary value. Thus, this metric can be used to draw conclusions about market dominance on the basis of the number of contracts and thus the corresponding market size.

Since sometimes the size of the stock is more important than the monetary equivalent, especially if one wants to grow, we consider this metric to be perfectly legitimate. However, we would like to state that we think it is more a matter of taste which of the two metrics is used.

Ratio of correct classified and lost contracts by month

Since this last business metric is a different representation of the metric recall, we refer to the evaluation of this metric in this chapter.

Finally, we would like to state our conclusion. We consider the metrics dealing with saved contracts to be generally more suited than those dealing with lost contracts. However, since the latter can be evaluated without external dependencies, at least in our use case, we would currently favor them.

6.1.3 Regulatory metrics

Our last part of this section deals with regulatory metrics. Here we have presented two different metrics in total. The first one is the "model certainty", which we have already evaluated above. We will not go into this again. The second one is the distribution of data features importance for positive and negative predictions in two representations each.

Data feature importance over time - positive and negative

The first representation represents the data feature importance in a compressed way. This has the big advantage that the most important data features can be compared easily throughout all months. However, this leads to a pooling of information, so that for example the values of the distribution of the data feature drop out. We consider this representation to be slightly unsuitable because the compression and the low fluctuation in both scenarios do not create a high information value. Nevertheless, it is still suitable for a rough overview.

This is different with the second representation variant. Here there is much more depth of information by displaying the data feature importances divided into individual plots for each month. The direct comparability is reduced a bit, because they are no longer located above or next to each other. Nevertheless, the distribution of the values in the respective plots provide much more information from the metrics. For example the distribution of importance can be seen here much better.

In principle, we think that the second plot should be integrated in order to support the manual evaluation of whether discrimination prevails or whether the model decides transparently. We consider this metric to be highly recommendable. The second metric, on the other hand, we are critical of the fact that the information content is shrinking due to the merging of the months. We would recommend it as a complementary metric representation and therefore only to be used if the second representation is also used.

6.2 Key-Metrics

In the last section we described which metrics turned out to be valuable and useful and which ones turned out to be dispensable or unnecessary. From this evaluation of the metrics we have now formed an outline, which we would define as the Key-Metrics. We intentionally limited them to six in order to have a special focus on them. Four are from the area of evaluating the machine learning model from a technical perspective and one metric each is from the area of Business Metrics and Regulatory Metrics.

It is important to mention that metrics that were not temporally, technically or, due to missing requirements, not feasible for us do not appear in this enumeration. Next to that we have selected the metrics based on our use case. They are generally transferable, but the use case may influence the selection.

1. Recall - Ratio of correctly classified and lost contracts by month (technical and business metric)
2. False Positive Rate
3. Accuracy - Ratio of correctly classified predictions
4. Averages of positive and negative predictions
5. Sum of lost stock commissions and premiums
6. Data feature importance over time - positive and negative

For our Key-Metrics we have chosen the technical machine learning metrics Recall, False Positive Rate, Accuracy and the averages of positive and negative predictions.

In our use case, Recall has achieved excellent results for both experiments and has thus been defined as Key-Metric by us.

FPR, on the other hand, seems to be particularly useful for fast, abrupt scenarios where suddenly many positive predictions are made, which then result in false positives. For our pandemic scenario, we think the metric is so powerful that it should also become a Key-Metric. For slow transitions, however, it seemed to be rather secondary.

In comparison, Accuracy seems to be out of place here, because although it showed a trend, it performed worse than the two previous metrics. However, since this is due to the unbalanced data set, where we have an average of only 5% cancellations on average and only those 5% are increased in percentage terms in the scenarios, this behavior was to be expected. We are convinced that with

balanced data sets this metric will perform much better. For this reason it should also be mentioned in the Key-Metrics.

For the last technical machine learning metric we have chosen the averages of positive and negative predictions. Compared to the prediction bias it has the advantage that it shows the result values of the predictions in the metric at the same time while giving information about the current month and not only a ratio like the prediction bias metric. In addition, we also preferred it over the standard deviation, as we believe that the standard deviation should mostly be an extension of the averages of positive and negative predictions. In general, all three metrics performed similarly well, especially in our first experiment. Only because of the above remarks we finally decided to use averages of positive and negative predictions as Key-Metric.

In business metrics, we consider "all lost stock commissions and premiums" to be the Key-Metric. This is mainly due to the fact that we always receive correct values for this metric. Neither by saving a contract is the value manipulated, nor do we depend on the feedback of the Field Service Partners. Furthermore, we think that the monetary representation is a good way for the management to see the results.

In addition to the business Key-Metric just presented, we also consider the "ratio of correct classified and lost contracts by month" to be of immense importance. However, since this is just another representation of the metric Recall and this is already included, it does not need to be introduced here a second time.

Our final Key-Metric is the metric of the data feature distribution in the form presented in the figures 5.21, 5.22, 5.42 and 5.43. We think it is essential to understand how the model works to support the decision whether discrimination is present and whether the model is deciding transparently. Thereby this metric is extremely helpful, especially because of the clear representation.

6.3 Assessment against design objectives

In the last section of this chapter we return to the design objectives and compare if our artifact, thus our set of metrics, has achieved them or not. If not, we also discuss the reasons for this.

As in Chapter 3 we divide this section into four subsections.

6.3.1 Business objectives

We have defined three objectives in the area of business objectives:

1. Maximize insurance premiums

2. Maximize stock commission

3. Maximize contract stock size

For all three objectives, we believe that we have provided appropriate metrics. With the metrics "saved contract premiums and stock commissions" and "theoretical increase of the stock through saved contracts" we have already covered all three objectives. In addition, we have also created complementary solutions with the last three metrics to provide additional support for the objectives. This includes the metric "Ratio of correctly classified and lost contracts by month", which as another representation of Recall pursues a direct business goal, which was not defined by us but by the project management of "Stornoprophylaxe". Moreover, we have also defined general metrics that cannot be directly defined as business metrics. These would be metrics like Precision or FPR. Such metrics of course also add value to the achievement of business goals, since they directly influence them by maximizing their value.

Therefore it is only fair to say that we have achieved all business objectives, even if we cannot apply the effect of the metrics to the actual stock data.

6.3.2 Data protection objectives

As described in Section 3.2.2 the topic around the regulations for using artificial intelligence, especially in insurance companies, is very complex and huge. Therefore, we limited the objective to make as much information as possible available to support the compliance managers in their work as well as possible. This is unfortunately of subjective nature to what extend our metrics support those people and when the support is sufficient enough to call it good.

However, we defined the main objectives:

1. Revealing of information

2. Support of the compliance work

3. Adherence to all regulations

Under the item "revealing of information" we define to reveal as much information about the workflow of the model and in the model as possible. For this we have considered several metrics. We consider the "data feature over time - positive and negative", defined by us as Key-Metric, the most important one for this objective. Furthermore, one can also draw conclusions from metrics like "model certainty", regarding what the courses of the metrics mean in relation to the regulations. Nevertheless, at the beginning of the thesis we expected more of these metrics in relation to the objective. Therefore, we think that we have already

created a basis with our metrics and provide a lot of information, but that this is not enough to support the work of the compliance managers in an optimal way.

The second item of the data protection objectives is based on the first one and thus describes a similar objective. The objective here was to make as much information as possible available about the processes surrounding the model in order to facilitate the work of compliance managers. For this reason, this item also follows the conclusion from the first item, as our assessment could not achieve any additional results.

However, the situation is different for the last point. The objective is to check the regulations regarding artificial intelligence partly automatically and partly semi-automatically. Unfortunately, we must clearly summarize that our metrics cannot provide this. A much more sophisticated concept would be necessary, for example by using a second separate machine learning model to check the behavior of machine learning models and then make an assessment. Such a objective is not feasible as part of a thesis project. It could be a future project on a much larger scale.

6.3.3 Technical objectives

As we defined the technical objectives in the beginning of our thesis project we also tried to focus on a complete implementation of a monitoring tool in our scope, next to the creation of sophisticated metrics. Later on, we decided to not include the implementation of this tool in the scope of this thesis as it is not generally transferable and the thesis would have gotten too big. Therefore, we defined objectives in the beginning of the project, which we will drop here in the assessment as they are no longer part of this thesis anymore. The items 2, are therefore dropped.

At the beginning we defined the following objectives:

1. Ensuring the model quality
2. Current status
3. Detailed information
4. Ease of use
5. Standard views
6. Metric help
7. Notifications
8. Anomaly detection

9. Metric presentation
10. Time offset for metrics

The first objective was indeed achieved. Through our variety of both technical and business metrics, we have been able to ensure that despite our time offset, both scenarios recognized the changes in the external world in time. We have described this in detail in Chapter 5 and 6.

However, an overall status of the model was not integrated in the form of metrics. We believe that this is part of the monitoring tool to observe the different metrics and derive an overall status from them. For this reason, it is no longer part of this assessment.

Nevertheless, we have integrated the subsequent item and are of the opinion that we have achieved this goal. While we are able to assess the current status of the model through various technical and business metrics and thus provide information, we have also integrated metrics that increase the transparency of the model. As previously described in the data protection objectives, we desired more information uncovering, but nonetheless we have already revealed a lot of information.

The following four points, points 4 to 7, again fall under the scope of the monitoring tool and are therefore not described in detail. Still we tried to make the metrics as simple and concrete as possible, so that at least for the individual metrics an "ease of use" is given and a "metric help" is not needed.

Point 8 consists of two parts. On the one hand anomaly detection in the model, on the other hand anomaly detection in the (input) data. The latter could not be implemented by us, because we used a template by the SDA (Service-Dominated Architecture) department, which used a framework that only accepts valid and expected JSON requests. Therefore we are not able to detect any anomalies. Even missing values are simply assigned a default value. We have been able to achieve the former with our metrics. Different types of anomalies, such as an unusually high number of positive predictions or a disproportionate use of a data feature, are reliably detected by our metrics.

Subsequently, we defined the metric presentation as an objective, so that these are displayed on a suitable scale with an additional traffic light if it generates additional benefit. The traffic light representation has not yet been integrated and can therefore not yet be validated. However, since there is already a design for it, it will be implemented in the next version. We have focused on the adherence of thresholds, which require manual intervention if they are passed. Apart from that, our metrics are mostly displayed in an appropriate scale, so that they can be evaluated in an optimal way. Only for the metric "data feature mean positive importance per month" the large distance between the data features and the small fluctuation causes poor readability.

The last and probably the most important point in our use case describes the goal to provide metrics that can guarantee an accurate assessment of the current quality level despite the late evaluation. This we have achieved. The metrics we have defined can determine whether the model is still working as expected for the current month of prediction. However, we still recognize room for improvement, because the metrics can detect if something is wrong, but not if this is a positive or negative change. This would be another interesting challenge that could be dealt with in the future.

6.3.4 SIGNAL IDUNA specific objectives

The area of SIGNAL IDUNA specific objectives is a special area of the thesis, since it contains objectives that are specifically designed for SIGNAL IDUNA and do not claim to be transferable. Here we have defined four objectives:

1. Number of saved contracts
2. Maximize the recall
3. Adjustments causing data drifts
4. Demographic changes

The first item refers to the SIGNAL IDUNA business goal to determine and prevent 10% of the annually cancellations. Our technical and business metrics have created a good basis to support this goal. Therefore, even though it has not been tested during production, we are convinced that we have achieved this goal.

Maximize the recall, on the other hand, is ultimately more dependent on how the metrics recall is handled. Other metrics can also provide additional input to support this objective. We consider this to be achieved as well, because we can support this process very well.

The last two items, however, were not integrated in our metrics and are therefore not achieved. We have not integrated a metric that detects data drifts based on new contract designs or adjustments of insurance conditions, nor a metric that detects demographic changes before they affect the model. Especially the last one would go beyond the scope of a metric. This would have to be covered in a separate program.

Chapter 7

Conclusion

Our last chapter is intended to conclude this thesis and to summarize our experiences, insights and achievements.

We first address the four previously defined research questions and answer them with regard to their feasibility and their outcomes.

Thereafter, we summarize the contributions of the thesis on the topic of multi-aspect monitoring machine learning models in operation and also want to address the objectives presented in the introduction, such as general transferability or reusability of training metrics. In addition, we summarize for which stakeholders the results are of particular value.

At last we present work that unfortunately could not be realized or researched within the scope of this thesis project. This includes work that was beyond the limits of this thesis, but also concepts that could be further investigated in future projects.

7.1 Answer to the research questions

As described in the introduction, we would like to go back to the previously defined Research Questions in detail at the end of the thesis. In the course of the thesis project it turned out that we can answer some of the questions in more detail than others. Thereby the answers to Research Question 1 and 2 are fairly short, whereas the answer to Research Question 3 is relatively long. The answer to Research Question 4 is provided at a reasonable length corresponding to the result.

RQ.1 What features are necessary for a monitoring tool for machine learning models in operation?
During the development of the metrics and the integration into a monitoring tool it turned out that many of the features mentioned in the first survey brought a

direct increase in utility to the tool and therefore became a necessity for the efficient and effective use of a monitoring tool for machine learning models in operation. Especially important were:

1. Simple install and go solution. - immediately operational
2. Straightforward usability
3. An overall status to summarize the models quality.
4. Display metrics as a scale representation and add a traffic light representation where useful.
5. Button or link as interface for explaining the displayed metric.
6. Notification to the correct group of people based on the type of metric.

After evaluating the first survey, but more importantly during the development of the project, these were considered most important and essential to provide a monitoring tool for machine learning models in operation. It was especially important that users who are neither experienced in machine learning nor technically versed will also use this tool. Accordingly, the explanatory button or link and the straightforward usability particularly stood out. Still, the points of an overall status and notification of the correct groups of people are equally important.

To summarize, these six conditions are not very technical and not very focused on machine learning, nevertheless they have to be considered and should be implemented if one is providing a monitoring tool for machine learning models in operation.

RQ.2 What defines a good metric for assessing the quality of machine learning models in operation?

Throughout the thesis project we have gained different insights into metric development. First of all, it should be stated that the development of metrics is a huge challenge, which we were not aware of at the beginning.

Our findings for answering this research question were initially drawn from our first survey. Then, as we reviewed and improved our metrics, we drew further conclusions. Finally, our findings were partially confirmed and partially expanded by our second survey.

Nevertheless, we can state the following two lists. The first one describes three key goals that need to be achieved by the used metrics. Whereas the second listing states how to achieve those goals. It can be seen that these lists are quite abstract as they have to be applied differently from use case to use case. Nevertheless we tried to be as concrete as possible.

Metrics must have the following goals:

- Easy to comprehend
- Helpful and not overwhelming
- Represent meaningful information

For any metric developed or applied, the metric should effect the monitoring in all three ways listed above. As previously described, they reflect the goals that should be achieved when utilizing metrics.

Therefore, it must be checked whether the metrics used to monitor machine learning models in operation are simple to understand and whether the status can be evaluated and understood when viewed directly.

Furthermore, the metric must serve a purpose and monitor an aspect of the model. Once a metric monitors an aspect of the model and this adds measurable value to the monitoring, the metric is considered helpful. Nevertheless, it must be ensured that the metric does not overwhelm the monitors, so the metric must not be too complex. This would again directly affect the usefulness of the metric.

Subsequently, it must also be ensured that this metric reflects meaningful and useful information. In so far as the monitored aspect is represented in a proper way and best reflects the status of the model.

Metrics must meet the following five requirements:

- Statuses for metrics must be defined.
- Applicable in day to day business.
- Automatically evaluable.
- Metrics must be linked to a group of people.
- Compatible to the model.

To achieve these goals, we have identified five requirements that must be met.

Statuses for metrics must be defined in order to directly assess whether the metric implies action or not. We have chosen three statuses: “Normal”, “Attention required” and “Severe Impairment”. Others and even more are of course possible as well.

Furthermore, each metric must be applicable at any time during normal operation. There must not be any restrictions on evaluability, for example, no availability of dependent data or systems. A continuous applicability of metrics must be ensured.

Moreover, a metric must be automatically evaluable. The goal is to achieve a degree of flexibility and speed to be able to react to suddenly occurring events. This cannot be achieved by manual evaluations. Sometimes it cannot be avoided that humans fill metrics with data, as we also explained in chapter 4.4. However, this should not be the norm, but rather the exception.

Besides the automatic evaluability of metrics, it is also important to notify the right people as soon as a threshold is exceeded by a metric. Therefore there must be at least one group of people for each metric, who is responsible for this metric or what it represents.

Finally, it must also be ensured that the metric used is compatible with the machine learning model. Metrics which are not compatible should not or cannot be used. An example would be our prediction certainty, which only works for models that return a scale value as return value. A model which returns only a true or false value as result is not suitable.

RQ.3 Which of those good metrics are best suited to evaluate a machine learning model in operation?
A lot of metrics contributed somewhat to the result. However, this differed depending on the scenario. In our scenario we did not limit ourselves to the technical evaluation of the model, but also considered the business impact and regulatory requirements, whereas the last two might be completely irrelevant for other use cases. Furthermore, with our three months time offset, we have a special requirement that does not apply to many scenarios. But there might as well be scenarios where the model's predictions cannot be evaluated automatically, where our metrics become interesting again.

In Chapter 6.2 we have already selected the six most important metrics. These Key-Metrics deal with all three areas and are from our point of view the most suitable metrics to evaluate machine learning models. These are the following metrics:

1. Recall - Ratio of correctly classified and lost contracts by month (technical and business metric)
2. False Positive Rate
3. Accuracy - Ratio of correctly classified predictions
4. Averages of positive and negative predictions
5. Sum of lost stock commissions and premiums
6. Data feature importance over time - positive and negative

In addition, there are other metrics that make an excellent contribution to the evaluation of models in operation. We have presented these in Chapter 5 and evaluated them in Chapter 6. Also not to be neglected are the metrics we provided that did not make it into the thesis, namely those that could not be implemented or were simply not included. These were marked yellow in the tables or added as appendix.

RQ.4 What are possible thresholds where metrics indicate a necessary change of a machine learning model?

Our last research question is the only one we cannot answer well. We have found good thresholds for us, but since they are designed for our use case, they can only be applied to others to a very limited extent. Especially with metrics like accuracy it depends on the use case how far the metric can fall before the quality losses are no longer acceptable. While for our use case we have decided that a drop of 0.02 points is the limit, for others it may be for example 0.10 points.

Besides that, certain metrics can get a higher weighting and thus the thresholds for these can be chosen more strictly. For instance, in our case the metric recall is defined as very important, but we have decided against a stricter threshold-value here.

Moreover, different metrics such as prediction bias and lost contracts are simply too different to define general thresholds.

Ultimately, however, it remains clear that, depending on the particular use case, the thresholds should be set close enough to the training average so that one can react to changes as quickly as possible. The allocation of the thresholds must therefore be carried out with care and caution and should not be carried out in a rash manner. That is crucial and should be followed at all times.

7.2 Contributions

Our thesis project has made several different contributions to the current development around multi-aspect monitoring of machine learning models in operation, but also to machine learning in general.

We have found a lot of insights through our intensive work with regulations around artificial intelligence. This includes data protection, data security, how to handle personal data, prevent discrimination and respect for privacy or transparency in decision-making. We have revealed how delicate it is to use machine learning in Europe, but especially in the insurance sector.

Furthermore, our survey revealed that the topic of monitoring of machine learning models in operation has a large interest base, but that most of the interested

parties do not yet have a concrete idea of how to realize it. This is also reflected in the scientific literature, where there is plenty of material about machine learning techniques, pipelines etc., but almost no material about monitoring those models in operation.

Subsequently, we showed with the help of our concept how monitoring could look like in operation. For this purpose we provided a sophisticated technical architecture design, which we consider to be generally transferable.

Additionally we uncovered, that monitoring machine learning models is not always about flawless quality, performance or working properly. Especially for certain groups of people, such as directors, managers or project leaders, other aspects can be of greater importance. For businesses for instance the generated economic performance is more important and for those no thresholds can or should be defined.

The most important contributions are of course our metrics. We have shown that metrics, which are also used in model training, also serve an excellent utility in operation. However, we have also shown that, depending on the respective use case, other additional metrics are necessary as the training metrics cannot provide sufficient depth of information. For this reason we have developed suitable metrics based on our use case, which complement the monitoring in operation and allow it to be carried out successfully. We have shown this in Chapter 5 and Chapter 6. In addition, we developed metrics that were not examined in the context of the thesis, but which must not be neglected. These are also excellent metrics, which may be even better than those reviewed. They were simply not implemented in the thesis and will be realized in future projects.

In summary, we have done a lot of research in the area of multi-aspect monitoring of machine learning models in operation and have gained various insights that can be used by other independent researchers as a basis, but can also be implemented by interested individuals.

7.3 Future work

As it is often the case with scientific projects, there are also parts of the project that could not be implemented for various reasons and were therefore categorized as future work. Among others, such reasons are the extent of the implementation, the lack of necessary requirements such as authorizations or automated information or even approaches that did not fit into the project at this stage.

Especially the SIGNAL IDUNA specific metrics belong to this category. These can give SIGNAL IDUNA a little more assurance about the quality of its application. Nevertheless, the implementation of the metrics was either too complex and costly or promised too little added value to address them in detail. We have

described this issue in detail in the last paragraph of Chapter 4.4.

Furthermore, we were not able to collect the information about "unknown data features". This was due to the fact that the machine learning model service was originally prepared as a template by the SDA (Service-Dominated Architecture) department, which used a framework that only accepts valid and expected JSON requests. Thus unknown data features are simply dropped or not accepted. This would be another aspect which can be optimized in the future.

Moreover, it was not possible within the scope of the thesis to provide a partly automated and partly semi-automated control of the regulations with regard to Artificial Intelligence. Simple metrics, however, cannot achieve this goal. A much more sophisticated concept is needed, for instance a second separate machine learning model trained for this purpose, which checks the behavior of machine learning models and makes an assessment. It is likely that individual models will be necessary for the respective aspects, such as discrimination or transparent predictions. This would certainly be an interesting topic for a project on a larger scale, as it has to cover several regulations and laws at the same time.

One last future work item we would like to point out, as this is most likely a topic where we are going to work on in the near future. Despite the late time of evaluation of our model predictions we are able to analyze and evaluate the current quality state. Still, we face the issue that it is possible to see if something has changed in quality, but not if it has changed for the better or the worse. So an increased number of positive predictions could have both positive and negative effects. Identifying the direction of the effects is a challenge that we want to realize in a further project.

Finally we would like to motivate everyone again to deal more with the topic of "multi-aspect monitoring of machine learning models in operation". More variations of metrics will be possible to extend our portfolio of metrics. Here we would like to encourage everyone to take our work as a basis for further development and research.

Bibliography

[1] Prometheus Authors. *OVERVIEW*. 2020. URL: https://prometheus.io/docs/introduction/overview/ (visited on 07/23/2020).

[2] Victor R. Basili, Gianluigi Caldiera, and H. Dieter Rombach. "The Goal Question Metric Approach". In: 1994.

[3] Denis Baylor et al. "TFX: A TensorFlow-Based Production-Scale Machine Learning Platform". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '17. Halifax, NS, Canada: Association for Computing Machinery, 2017, pp. 1387–1395. ISBN: 9781450348874. DOI: 10.1145/3097983.3098021. URL: https://doi.org/10.1145/3097983.3098021.

[4] Bundesministerium für Bildung und Forschung. *Bekanntmachung*. Mar. 2019. URL: https://www.bmbf.de/foerderungen/bekanntmachung-2392.html (visited on 07/15/2020).

[5] Eric Breck et al. "The ML Test Score: A Rubric for ML Production Readiness and Technical Debt Reduction". In: *Proceedings of IEEE Big Data*. 2017.

[6] Deutscher Bundestag. *Allgemeines Gleichbehandlungsgesetz (AGG)*. https://www.antidiskriminierungsstelle.de/SharedDocs/Downloads/DE/publikationen/AGG/agg_gleichbehandlungsgesetz.pdf;jsessionid=EC1A0952847BC834ABDD6912DB069A30.2_cid369?__blob=publicationFile&v=22. Aug. 2006.

[7] Ori Cohen. *Monitor! Stop Being A Blind Data-Scientist*. Oct. 2019. URL: https://towardsdatascience.com/monitor-stop-being-a-blind-data-scientist-ac915286075f (visited on 09/26/2020).

[8] Konferenz der unabhängigen Datenschutzbehörden des Bundes und der Länder (Datenschutzkonferenz). *Hambacher Erklärung zur Künstlichen Intelligenz*. Nov. 2019.

[9] Konferenz der unabhängigen Datenschutzbehörden des Bundes und der Länder (Datenschutzkonferenz). *Positionspapier der Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder*. Nov. 2019.

[10] Gesamtverband der Deutschen Versicherungswirtschaft e.V. der deutschen Versicherer. *Künstliche Intelligenz in der Versicherungswirtschaft*. Oct. 2019.

[11] Gesamtverband der Deutschen Versicherungswirtschaft e.V. der deutschen Versicherer. *Zusammenfassung: Positionspapier der DSK zu empfohlenen TOMs bei Entwicklung und Betrieb von KI-Systemen*. Oct. 2019.

[12] Google Developers. *Testing Pipelines in Production*. 2019. URL: https://developers.google.com/machine-learning/testing-debugging/pipeline/production (visited on 09/25/2020).

[13] Google Developers. *Testing Pipelines in Production*. 2019. URL: https://developers.google.com/machine-learning/crash-course/classification/prediction-bias (visited on 10/10/2020).

[14] European Union. *Charter of Fundamental Rights of the European Union*. Vol. 53. Brussels: European Union, 2010, p. 380.

[15] Council of the European Union. *REGULATION (EU) 2016/679*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679. Apr. 2016.

[16] High-Level Expert Group on AI. *Ethics guidelines for trustworthy AI*. eng. Report. Brussels: European Commission, Apr. 2019. URL: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[17] Waldemar Hummer et al. "ModelOps: Cloud-Based Lifecycle Management for Reliable and Trusted AI". In: June 2019, pp. 113–120. DOI: 10.1109/IC2E.2019.00025.

[18] Grafana Labs. *Grafana Features*. 2020. URL: https://grafana.com/grafana/ (visited on 07/23/2020).

[19] Li Erran Li et al. "Scaling Machine Learning as a Service". In: *Proceedings of The 3rd International Conference on Predictive Applications and APIs*. Ed. by Claire Hardgrove et al. Vol. 67. Proceedings of Machine Learning Research. Microsoft NERD, Boston, USA: PMLR, Nov. 2017, pp. 14–29. URL: http://proceedings.mlr.press/v67/li17a.html.

[20] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[21] Ken Peffers et al. “The design science research process: A model for producing and presenting information systems research”. In: *Proceedings of First International Conference on Design Science Research in Information Systems and Technology DESRIST* (Feb. 2006).

[22] Qualtrics. *Qualtrics Home Page*. 2021. URL: https://www.qualtrics.com/ (visited on 01/20/2021).

[23] Cynthia Rudin. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. In: *Nature Machine Intelligence* 1 (May 2019), pp. 206–215. DOI: 10.1038/s42256-019-0048-x.

[24] D Sculley et al. “Hidden Technical Debt in Machine Learning Systems”. In: *NIPS* (Jan. 2015), pp. 2494–2502.

[25] IDG Research Services. *Studie Machine Learning / Deep Learning 2018*. Tech. rep. 2018.

[26] IDG Research Services. *Studie Machine Learning / Deep Learning 2019*. Tech. rep. 2019.

[27] IDG Research Services. *Studie Machine Learning / Deep Learning 2020*. Tech. rep. 2020.

[28] European Data Protection Supervisor Spanish Data Protection Agency. “INTRODUCTION TO THE HASH FUNCTION AS A PERSONAL DATA PSEUDONYMISATION TECHNIQUE”. In: (Oct. 2019).

[29] statista.com. *AI Startup Funding Reaches Record High*. 2019. URL: https://www.statista.com/chart/18878/artificial-intelligence-startup-funding/ (visited on 08/09/2020).

[30] OMDIA | TRACTICA statista.com. *Revenues from the artificial intelligence for enterprise applications market worldwide, from 2016 to 2025*. 2016. URL: https://www.statista.com/statistics/607612/worldwide-artificial-intelligence-for-enterprise-applications/ (visited on 08/09/2020).

List of Figures

List of Tables

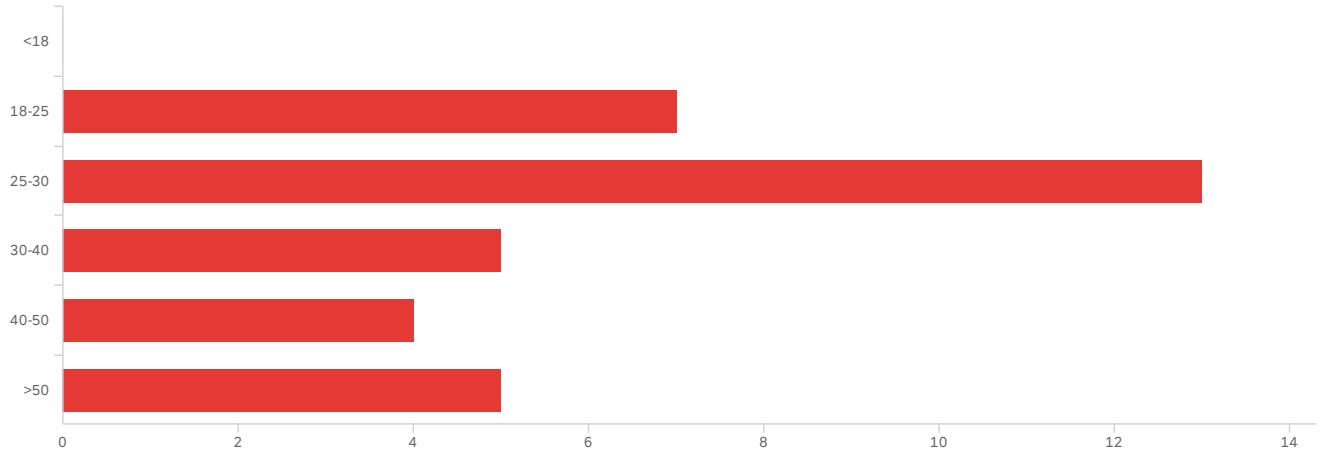
Appendix

- Survey I
- Metrics who did not make it into the demonstration
- Survey II

Default Report

Monitoring tools for ML-Models in operation
April 6, 2020 5:19 AM MDT

Q1 - What age are you?

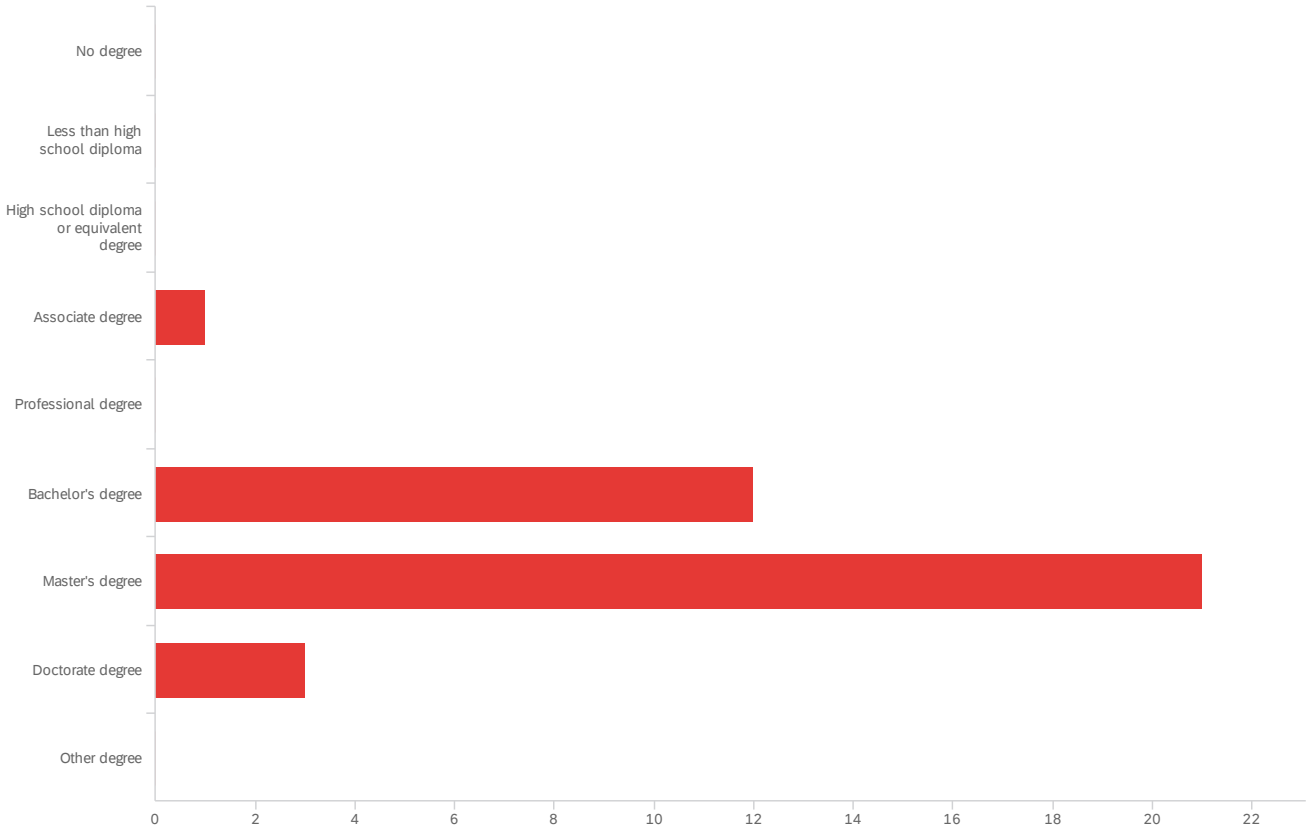


#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
1	What age are you?	2.00	6.00	3.62	1.33	1.77	34

#	Field		Choice Count
1	<18	0.00%	0
2	18-25	20.59%	7
3	25-30	38.24%	13
4	30-40	14.71%	5
5	40-50	11.76%	4
6	>50	14.71%	5
			34

Showing rows 1 - 7 of 7

Q2 - What is your educational background?



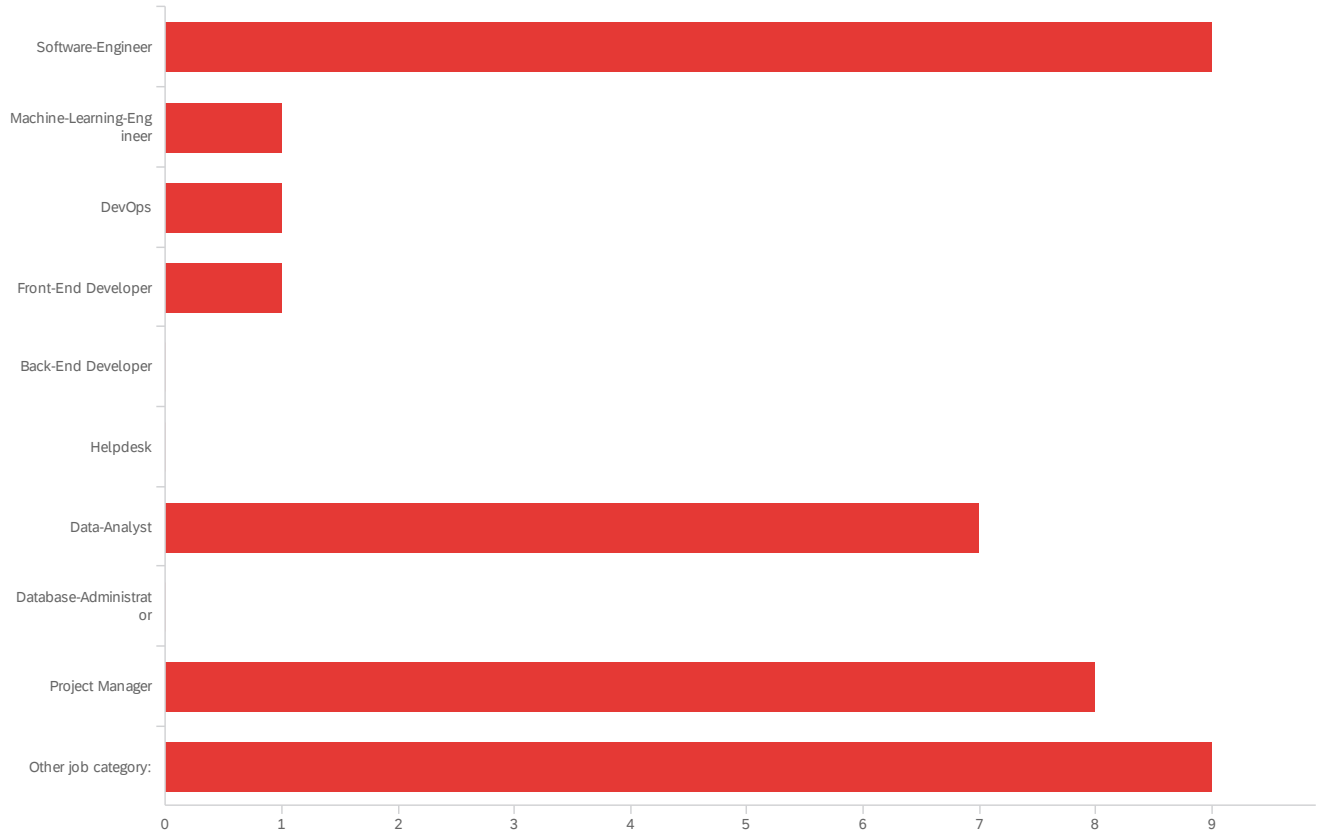
#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
1	What is your educational background?	4.00	8.00	6.68	0.74	0.54	37

#	Field	Choice Count	
1	No degree	0.00%	0
2	Less than high school diploma	0.00%	0
3	High school diploma or equivalent degree	0.00%	0
4	Associate degree	2.70%	1
5	Professional degree	0.00%	0
6	Bachelor's degree	32.43%	12
7	Master's degree	56.76%	21

#	Field		Choice Count
8	Doctorate degree	8.11%	3
9	Other degree	0.00%	0
			37

Showing rows 1 - 10 of 10

Q3 - Which job category do you belong to?



#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
1	Which job category do you belong to? - Selected Choice	1.00	10.00	6.36	3.62	13.12	36

#	Field		Choice Count
1	Software-Engineer	25.00%	9
2	Machine-Learning-Engineer	2.78%	1
3	DevOps	2.78%	1
4	Front-End Developer	2.78%	1
5	Back-End Developer	0.00%	0
6	Helpdesk	0.00%	0
7	Data-Analyst	19.44%	7

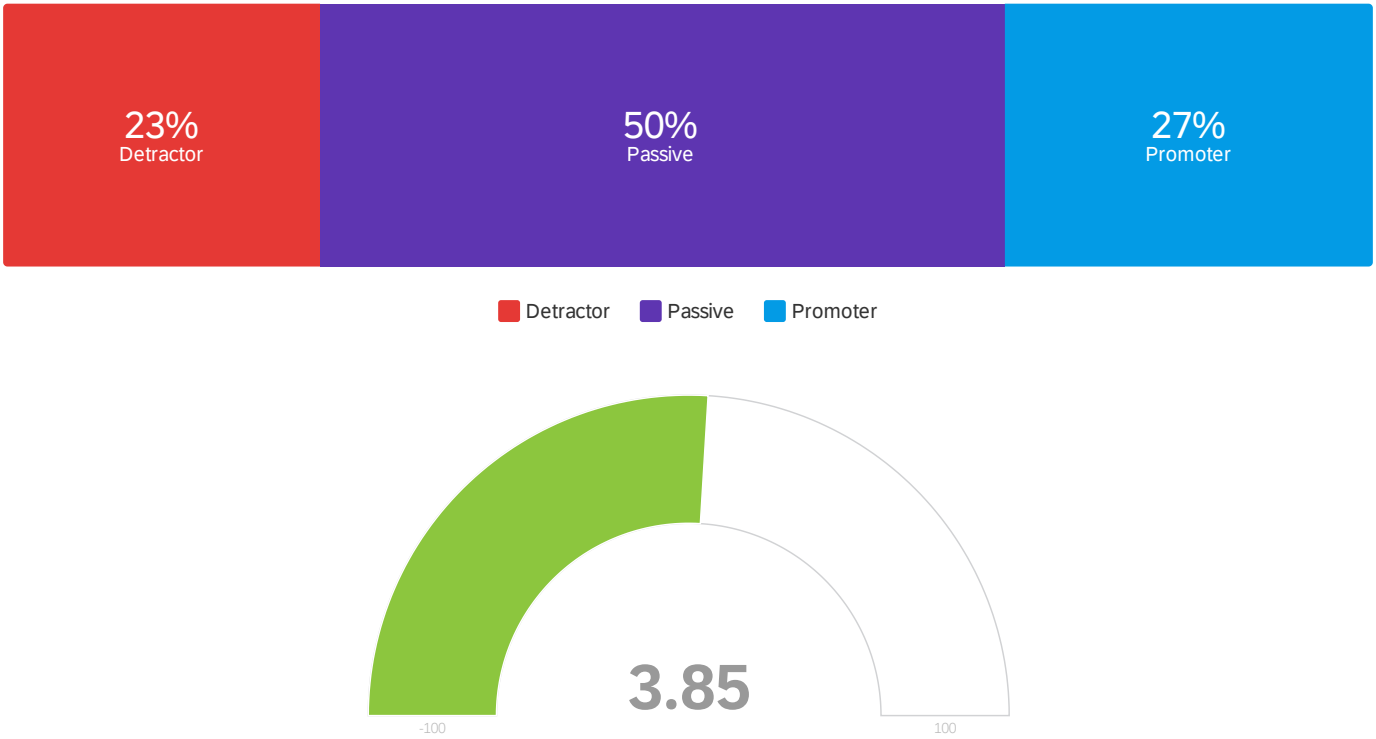
#	Field		Choice Count
8	Database-Administrator	0.00%	0
9	Project Manager	22.22%	8
10	Other job category:	25.00%	9
			36

Showing rows 1 - 11 of 11

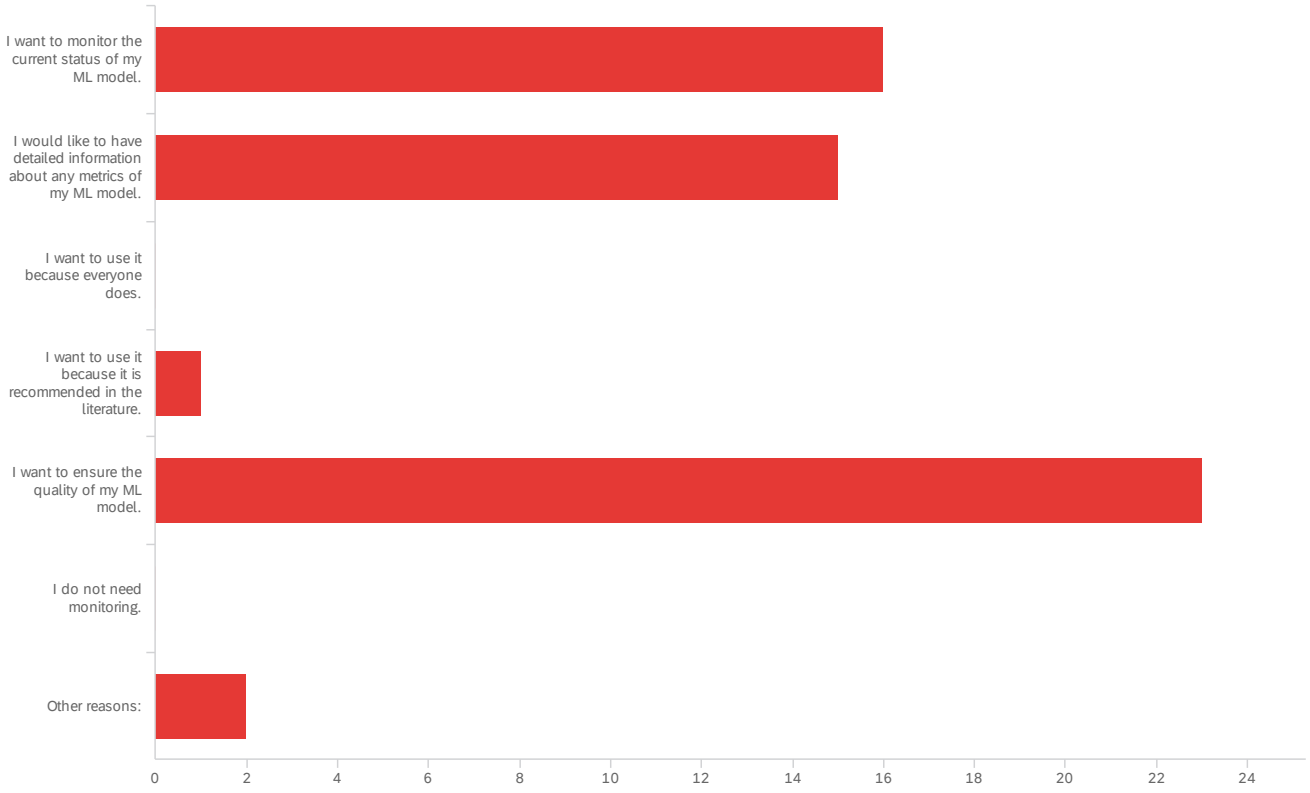
Q3_10_TEXT - Andere Berufsgruppe:

Andere Berufsgruppe:
Astronomy MSc researcher
Mathematiker
Fullstack
Compliance-Officer
Aktuar
Aktuar
Aktuar
Controller

Q4 - How important do you consider monitoring of ML models in operation?



Q5 - What are the main reasons for using monitoring of ML models in operation? (multiple choice)



#	Field		Choice Count
1	I want to monitor the current status of my ML model.	28.07%	16
2	I would like to have detailed information about any metrics of my ML model.	26.32%	15
3	I want to use it because everyone does.	0.00%	0
4	I want to use it because it is recommended in the literature.	1.75%	1
5	I want to ensure the quality of my ML model.	40.35%	23
6	I do not need monitoring.	0.00%	0
7	Other reasons:	3.51%	2
			57

Showing rows 1 - 8 of 8

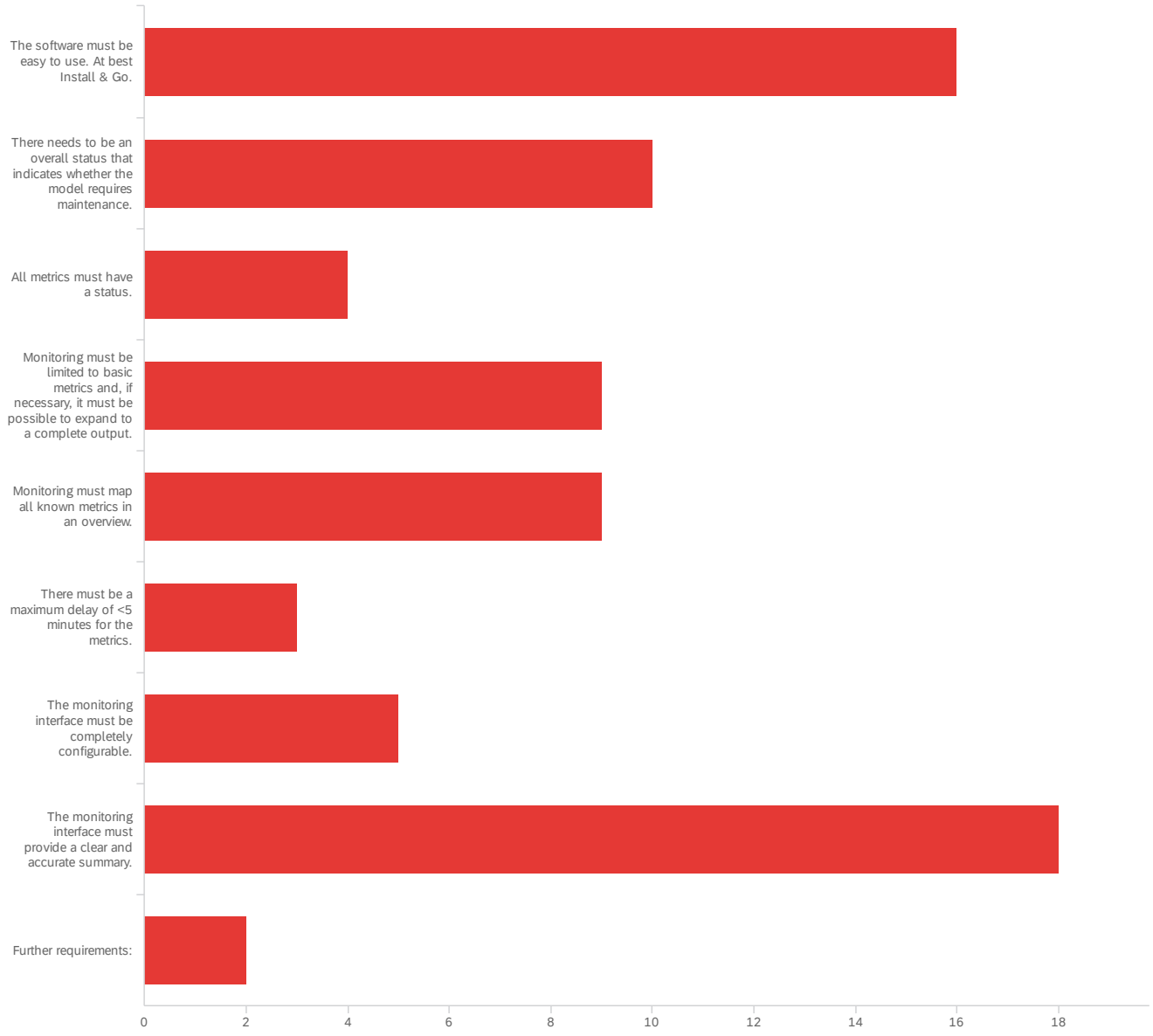
Q5_7_TEXT - Weitere Gründe:

Weitere Gründe:

Es muss neben der fachlichen Qualität und Aussagefähigkeit sichergestellt sein, dass u.a. die Ergebnisse nicht zur ungewollten Diskriminierung von Personengruppen führen, was Schadensersatzforderung und Haftungstatbestände erzeugen kann und dass Modell und seine Ergebnisse muss grundsätzlich erklärbar und vermittelbar gegenüber Dritten sein.

Nachweißbarkeit beziehungsweise Regulatorik -> BaFin

Q6 - What conditions must be met before you would use monitoring? (multiple choice)



#	Field		Choice Count
1	The software must be easy to use. At best Install & Go.	21.05%	16
2	There needs to be an overall status that indicates whether the model requires maintenance.	13.16%	10
3	All metrics must have a status.	5.26%	4
4	Monitoring must be limited to basic metrics and, if necessary, it must be possible to expand to a complete output.	11.84%	9
5	Monitoring must map all known metrics in an overview.	11.84%	9
6	There must be a maximum delay of <5 minutes for the metrics.	3.95%	3

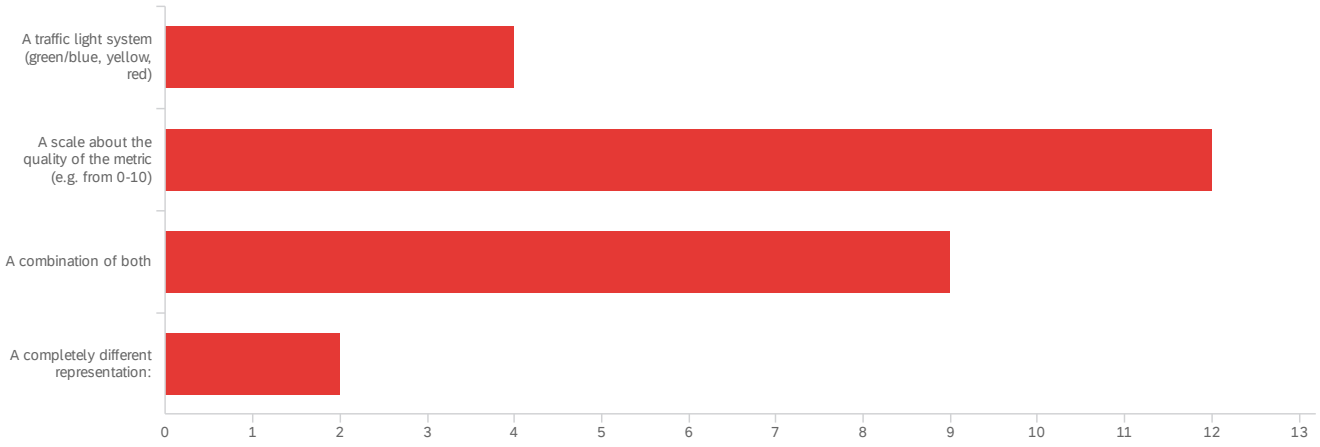
#	Field		Choice Count
7	The monitoring interface must be completely configurable.	6.58%	5
8	The monitoring interface must provide a clear and accurate summary.	23.68%	18
9	Further requirements:	2.63%	2
			76

Showing rows 1 - 10 of 10

Q6_9_TEXT - Weitere Voraussetzungen:

Weitere Voraussetzungen:
Transparenz der dargestellten Informationen mit Drill-Down-Optionen. Ggf. viel erklärender Hilfstext auf Wunsch um fachliche und technische Zusammenhänge entnehmen zu können. Das Tool muss Schulung / Fachliche Information und Monitoring zu gleich sein. Schulung / Fachliche Information muss gleichfalls regelmäßig an technischen Fortschritt angepasst werden. Evergreen-Ansatz: Optimalerweise wird die Software kontinuierlich weiterentwickelt und automatisch aktualisiert. Adapter / Schnittstelle zu meinen eigenen Systemen muss einfach wartbar sein.
Es muss in der Lage sein, sinnvolle Metriken zeitnah bereitzustellen

Q7 - How should the status of the metrics be presented?



#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
1	How should the status of the metrics be presented? - Selected Choice	1.00	4.00	2.33	0.82	0.67	27

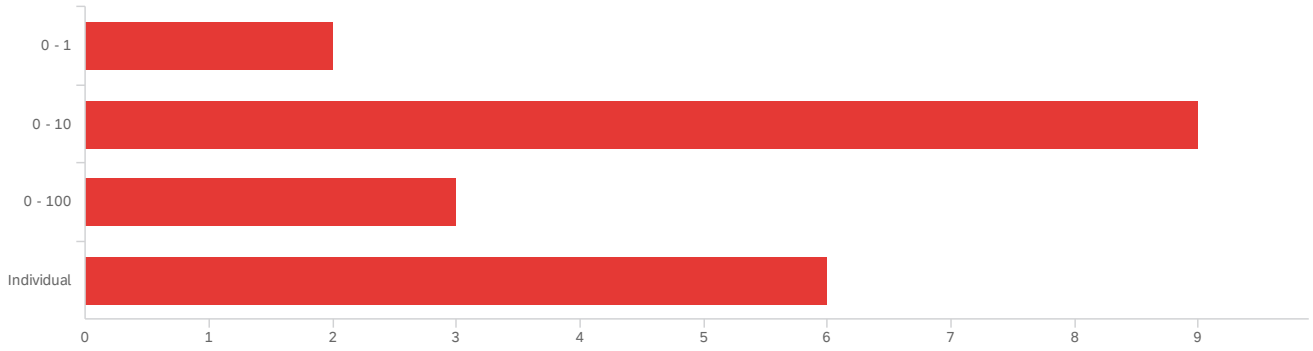
#	Field		Choice Count
1	A traffic light system (green/blue, yellow, red)	14.81%	4
2	A scale about the quality of the metric (e.g. from 0-10)	44.44%	12
3	A combination of both	33.33%	9
4	A completely different representation:	7.41%	2
			27

Showing rows 1 - 5 of 5

Q7_4_TEXT - Eine ganz andere Representation:

Eine ganz andere Representation:
freie Auswahl und gerne auch eigene Definitionen sollten möglich sein
Ausgabe Metrik und Historie der Metrik

Q7.1 - Which scale is most suitable?



#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
1	Which scale is most suitable?	1.00	4.00	2.65	1.01	1.03	20

#	Field		Choice Count
1	0 - 1	10.00%	2
2	0 - 10	45.00%	9
3	0 - 100	15.00%	3
4	Individual	30.00%	6
			20

Showing rows 1 - 5 of 5

Q8 - Which metrics are necessary to assess the status of an ML model in operation?

Please specify:

Welche Metriken zur Einschätzung des Status eines ML-Modells im Betrieb sin...
The accuracy on the test set The number of test sets completed
Spread in prediction performance: often, outliers in predictions are most useful to find issues in ML models.
Anomalies and Performance
outliers that pushes model in wrong direction
Efficiency metrics like memory, processing time I'm not sure how to assess the validity
Accuracy, specificity, classification report, error, R^2 goodness of fit
verschiedene denkbar s.o.
ich denke, das hängt vom jeweiligen Modell ab
Klassische ML Metriken wie Accuracy, RMSE etc. operative Metriken zum Performance
ROC,accuracy, Konfusionsmatrix, AUC,precision, recall, Vielleicht möchte ich mir für ein bestimmtes Modell auch selbstdefinierte Metriken anschauen, d.h. dafür sollte es auch Möglichkeiten geben.
fortlaufende Backtestingperformance des Modelles Metrik, die aussagt, inwieweit sich die aktuellen Prognosedaten strukturell von den Trainingsdaten unterschieden. In diesem Zusammenhang ist es auch sinnvoll, Metriken bzgl. der Robustheit des Modelles auszuweisen. Nicht nur Punktschätzer, sondern auch Aussagen zur Schätzunsicherheit (Wahrscheinlichkeitsverteilung anstatt "nur" Punktschätzer der Zielgröße)
Durchsatz Ansprechbar Reaktionszeit

Q9 - What do you expect from a monitoring tool for ML models in operation in general.

(Which has not been questioned before.)Please specify:

Was erwarten Sie von einem Monitoring-Tool für ML-Modelle im Betrieb im All...
The time passed Expected finish time
To provide guidelines through metrics on what parts should be re-evaluated.
Detect the anomalies of the ML model, monitor the performance and suggest improvements
The spike of change when model has changed
Is the model likely to fail? Is the hardware supporting well the model?
How specific parameters relate to optimum, which parameters need tweaking
Hohe Transparenz des ML-Prozesses an sich und Bewertung zur Qualität der Ergebnisse.
Hübsche Grafiken Blueprint für den Standardfall, allerdings für jedes Modell selber gestaltbar
Vor Allem eine Aussage, wie gut das Modell funktioniert und (wie oben schon erwähnt) Kennzahlen zur Robustheit des Modelles mit Blick auf eine geänderte Datenlage. Auch oben schon erwähnt: Information zur Unsicherheit der Modellprognose in Form einer Wahrscheinlichkeitsverteilung der Zielgröße
Einfach zu bedienen Leicht einzusehen Leicht verständlich Oberflächliche Zusammenfassung aber auch Anhaltspunkte um detailliert in Log-Analyse einzusteigen

End of Report

Goal	Purpose	Ensure
	Issue	the computational performance
	Object (process)	of the model
	Viewpoint	from the customers perspective.
Question	Q1	What is percentage of my resources are gone from my service?
Metrics	M1	Current CPU usage
	M2	Current GPU usage
	M3	Current network traffic
	M4	Current RAM usage
	M5	Current Client load
Question	Q2	Is the performance usage increasing over time?
Metrics	M6	Average CPU usage over time
	M7	Standard deviation of the CPU usage
	M8	% cases at maximum utilization of the CPU
	MX	M6-8 for all metrics M1-5
Question	Q3	Do we have enough spare resources for a sudden increase of requests?
Metrics	M9	Current usage of resources in use
	M10	Is the current usage underneath a certain threshold
	M11	M1-5 in percentage compared to the max utilization possible.

Goal	Purpose	Ensure
	Issue	the model information is always up to date
	Object (process)	to recognize issues after new releases immediately
	Viewpoint	from the developers perspective.
Question	Q4	Which model version is in use?
Metrics	M12	Current version
	M13	Timeline of all versions
Question	Q5	Which model size has the model?
Metrics	M14	Current size
	M15	Size alteration of releases over time.
Question	Q6	Which ML-Algorithm was used?
Metrics	M16	Algorithm used
Question	Q7	Which parameter were used while training the model?
Metrics	M17	Set of parameter

Goal	Purpose	Ensure
	Issue	the model accuracy
	Object (process)	of the model
	Viewpoint	from the supporters perspective.
Question	Q8	How accurate is the model over the last 1000 predictions? (1000 seems reasonable for our use case.)
Metrics	M18	Confusion Matrix
	M19	AUC
	M20	Deviation between TP, TN & FP, FN
	M21	Training-Serving Skew - Accuracy skew
Question	Q9	Is the distribution of preditctions still in a plausible range?
Metrics	M22	Prediction bias
	M23	Prediction bias compared by training-serving

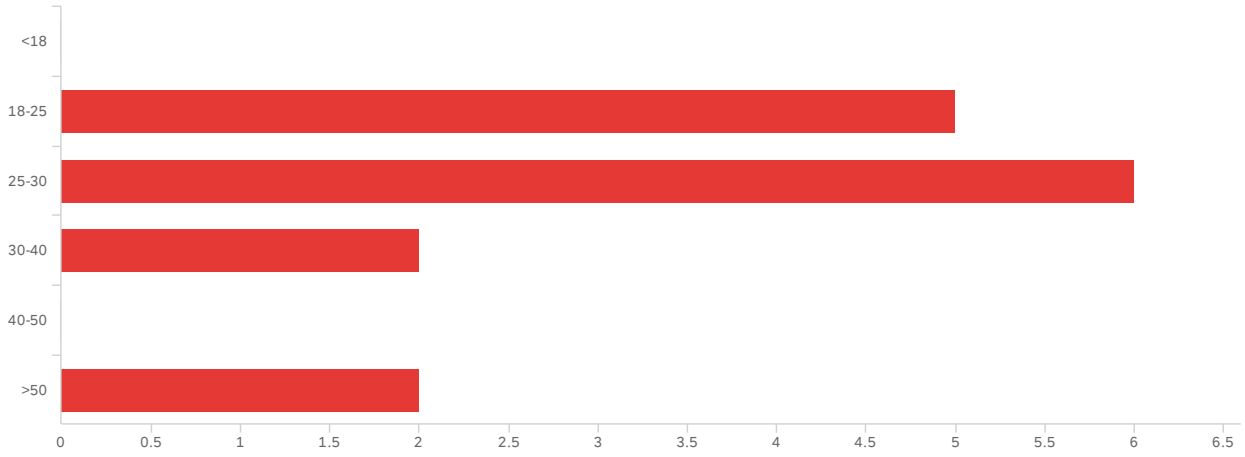
Goal	Purpose	Ensure
	Issue	a solid data basis
	Object (process)	for the model to not influence the predictions
	Viewpoint	from the supporting perspective.
Question	Q9	Is the data schema still similar to the one which the model was trained with?
Metrics	M24	Training-Serving Skew - Schema skew --> Distance measure
Question	Q10	Are still all features served while handling an input request?
Metrics	M25	Training-Serving Skew - Feature skew
	M26	Average count of missing features
	M27	Missing features with a count and a percentage in a table.
Question	Q11	Are there unexpected features appearing?
Metrics	M28	Count of unexpected features.
	M29	Extra features with a count and a percentage in a table.
Question	Q12	Are the incoming values the expected type?
Metrics	M30	Count of NAN, infinities, NULLS and N/As
	M31	Table with count and percentage of those values
	M32	True/False distribution of the expected datatype of each feature.
Question	Q13	Are there anomalies inside the input data? (Like N/A or NULL or infinity)
Metrics	M33	Count of NAN, infinities, NULLS and N/As
	M34	Table with count and percentage of those values

Report

Metrics for monitoring ML-Models in operation

November 29, 2020 4:52 AM MST

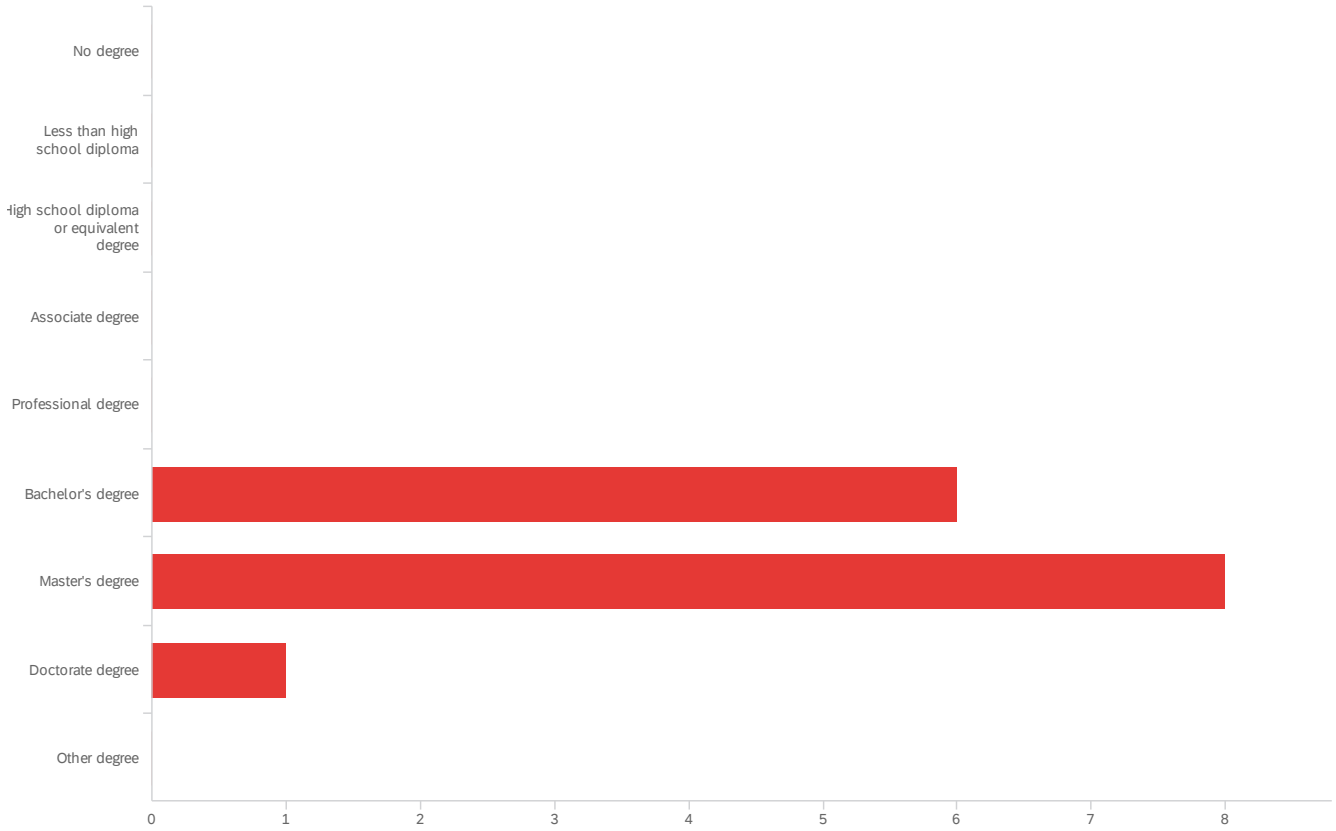
Q1.2 - What age are you?



#	Field		Choice Count
1	<18	0.00%	0
2	18-25	33.33%	5
3	25-30	40.00%	6
4	30-40	13.33%	2
5	40-50	0.00%	0
6	>50	13.33%	2
			15

Showing rows 1 - 7 of 7

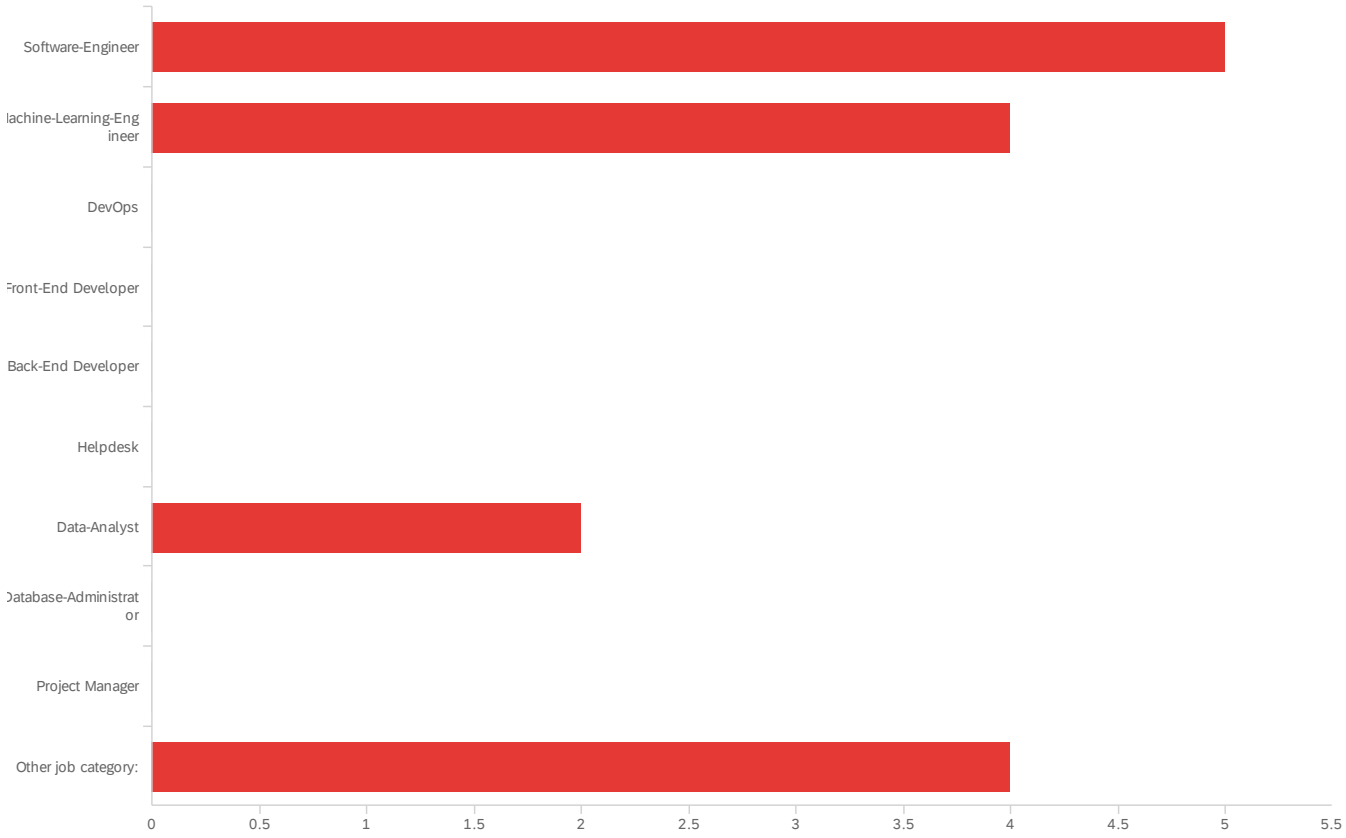
Q1.3 - What is your educational background?



#	Field	Choice Count
1	No degree	0.00% 0
2	Less than high school diploma	0.00% 0
3	High school diploma or equivalent degree	0.00% 0
4	Associate degree	0.00% 0
5	Professional degree	0.00% 0
6	Bachelor's degree	40.00% 6
7	Master's degree	53.33% 8
8	Doctorate degree	6.67% 1
9	Other degree	0.00% 0
		15

Showing rows 1 - 10 of 10

Q1.4 - Which job category do you belong to?



#	Field	Choice Count	
1	Software-Engineer	33.33%	5
2	Machine-Learning-Engineer	26.67%	4
3	DevOps	0.00%	0
4	Front-End Developer	0.00%	0
5	Back-End Developer	0.00%	0
6	Helpdesk	0.00%	0
7	Data-Analyst	13.33%	2
8	Database-Administrator	0.00%	0
9	Project Manager	0.00%	0
10	Other job category:	26.67%	4

15

Showing rows 1 - 11 of 11

Q3_10_TEXT - Andere Berufsgruppe:

Andere Berufsgruppe:
Student
Aktuar
Optical Science / Data science

Q2.3 - Which aspects of the software "Stornoprophylaxe" should be monitored?

(Example: ML aspects, regulatory aspects etc.)

Welche Aspekte von der Software "Stornoprophylaxe" sollten überwacht werden...

Demographics and distribution of characteristics.

Some aspects which considers the difference of time until the prediction is evaluable. Of course also standard metrics such as CPU usage, Accuracy, Precision etc.

Regulatorik und Management Ziele sind leider meistens besonders wichtig.

Definitiv allgmeine ML-Aspekte wie prediction bias, accuracy und Ähnliches, aber auch Auslastung und Gesetzeskonformität.

Age, date-time

Welche Kriterien haben einen hohen Einfluss auf die Entscheidung?

Data Drift

False positives vs false negatives, because it matters how the prediction fails. Data from more than one insurance provider.

Schadenaspekte, sollen die Kunden nicht besser stornieren?

- Modellaspekte: Welches waren bei den einzelnen Kunden die ausschlaggebenden Merkmale? - Datenqualität: Haben die Daten noch die für das Modell erforderliche Datenqualität? (Auch im Zusammenhang mit dem ersten Punkt zu sehen, z.B. Auswahl auf Basis falsch ermittelter Kundenmerkmale) - Modellstabilität und Modellgüte (Modellgüte dabei hinsichtlich der tatsächlichen Stornos) - Modellgüte über die Zeit (z.B. nachlassende Güte aufgrund neuer der Modellbasis noch unbekannter Zusatzinformationen)

Wirkung der vorbeugenden Maßnahmen, Datenschutz (Profiling)

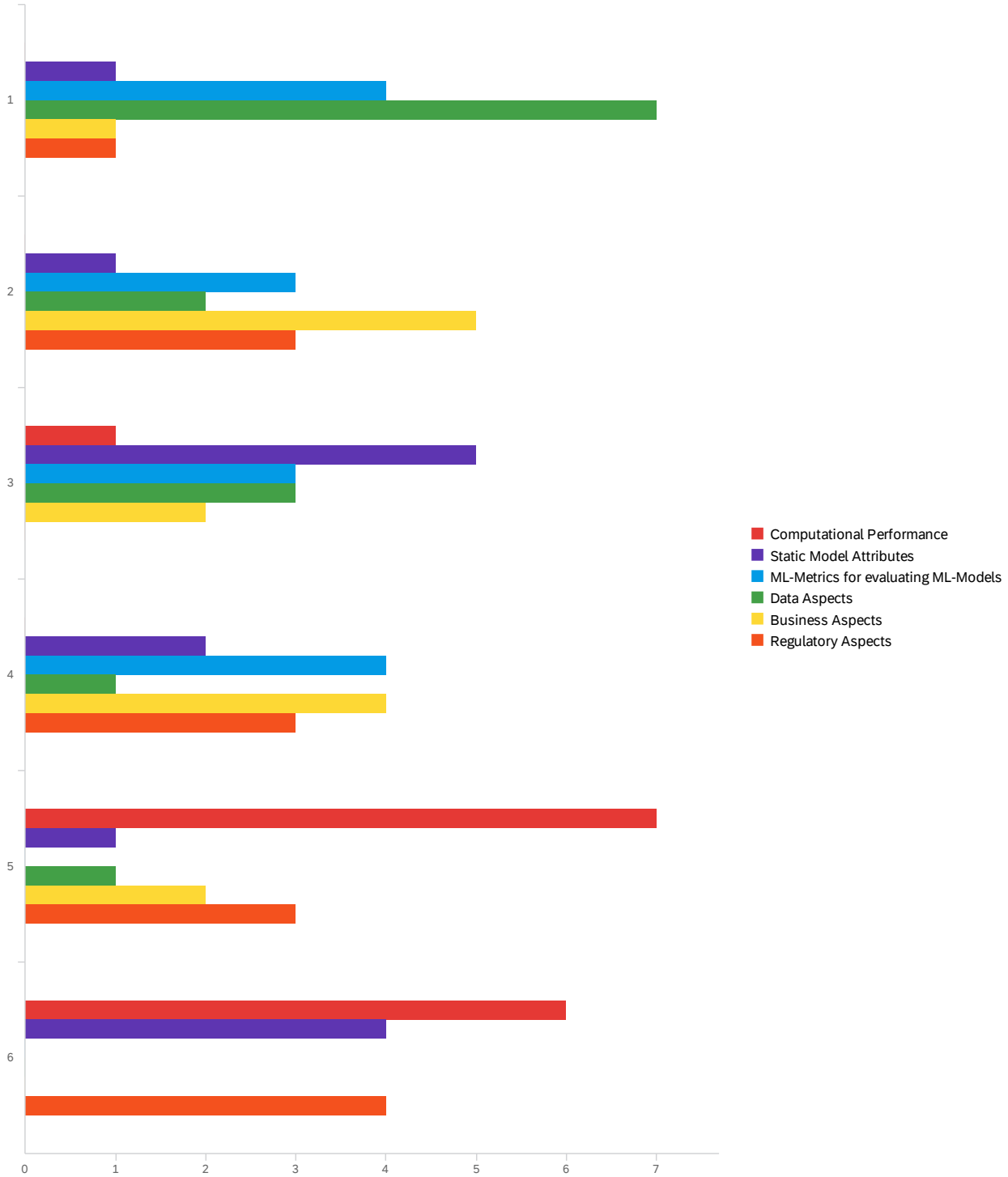
It might be interesting to somehow use the social or economic class of the customers as input to the model.

Nur ML-Metrics

All seem important.

ML-aspects, regulatory aspects, monetary aspects, container aspects like CPU etc.

Q2.4 - Which aspects do you consider important and how important? (Ranking by Drag & Drop)



#	Field	1	2	3	4	5	6	Total

#	Field	1		2		3		4		5		6		Total
1	Computational Performance	0.00%	0	0.00%	0	7.14%	1	0.00%	0	50.00%	7	42.86%	6	14
2	Static Model Attributes	7.14%	1	7.14%	1	35.71%	5	14.29%	2	7.14%	1	28.57%	4	14
3	ML-Metrics for evaluating ML-Models	28.57%	4	21.43%	3	21.43%	3	28.57%	4	0.00%	0	0.00%	0	14
4	Data Aspects	50.00%	7	14.29%	2	21.43%	3	7.14%	1	7.14%	1	0.00%	0	14
5	Business Aspects	7.14%	1	35.71%	5	14.29%	2	28.57%	4	14.29%	2	0.00%	0	14
6	Regulatory Aspects	7.14%	1	21.43%	3	0.00%	0	21.43%	3	21.43%	3	28.57%	4	14

Showing rows 1 - 6 of 6

Q3.3 - Answers question: What is the ratio between confirmed positive predictions and all positive predictions?

Field		Choice Count
10	26.67%	4
9	6.67%	1
8	13.33%	2
7	40.00%	6
6	0.00%	0
5	13.33%	2
4	0.00%	0
3	0.00%	0
2	0.00%	0
1	0.00%	0
0	0.00%	0
		15

Showing rows 1 - 12 of 12

Q3.4 - Answers question: What is the ratio between confirmed positive predictions and positive results?

Field	Total
10	6
9	1
8	0
7	3
6	3
5	1
4	0
3	0
2	0
1	1
0	0

Showing rows 1 - 11 of 11

Q3.5 - Answers question: What is the ratio between the confirmed predictions and all predictions?

Field	Total
10	3
9	2
8	0
7	5
6	2
5	0
4	0
3	1
2	0
1	0
0	2

Showing rows 1 - 11 of 11

Q3.6 - Answers question: What is the ratio between confirmed negative predictions and negative results?

Field	Total
10	1
9	0
8	2
7	2
6	4
5	2
4	1
3	2
2	0
1	0
0	1

Showing rows 1 - 11 of 11

Q3.7 - Answers question: What is the harmonic mean between precision and recall?

Field	Total
10	2
9	1
8	4
7	3
6	0
5	1
4	2
3	1
2	0
1	0
0	1

Showing rows 1 - 11 of 11

Q3.8 - Answers question: What is the ratio between positive predictions that did not come true and negative results?

Field	Total
10	2
9	1
8	4
7	3
6	0
5	1
4	2
3	1
2	0
1	0
0	1

Showing rows 1 - 11 of 11

Q3.9 - Remarks to above questions:

Anmerkungen zu den obigen Fragen:

Im Rahmen der Software sind die oberen Metriken bedingt gut. Teilweise eignen sie sich besser, als andere. So scheint mir der F1-Score für den Fall der Stornoprophylaxe eher als uninteressant.

Es fiel mehr schwer, o.a. Kenngrößen ohne ein konkretes Modell zu bewerten. In der Regel bewerte ich mit solchen Kenngrößen konkrete Modellergebnisse und versuche gleichzeitig das konkrete Modell besser zu verstehen. Entscheidend ist aber letztlich der ökonomische Nutzen eines Modells. Ein Modell kann sehr gute Voraussagen machen. Wenn es mir aber keine neuen Erkenntnisse oder verbesserten Geschäftsprozesse liefert, ist es trotzdem wertlos.

It would be nice to have what true and false positives and negatives mean in this context. E.g. "a true positive would mean that the customer..."

Q3.11 - Metric:Trend of general metrics (accuracy, specificity, etc.) over five months. How suitable is this metric?

Field	Total
10	3
9	2
8	2
7	6
6	1
5	1
4	0
3	0
2	0
1	0
0	0

Showing rows 1 - 11 of 11

Q3.12 - Remarks to above questions:

Anmerkungen zu den obigen Fragen:

Man muss allerdings auf das Problem der Unbalanciertheit Rücksicht nehmen. Die Accuracy ist also eher problematisch, die TPR dagegen besser.

Q3.14 - Metric:Comparison of the distribution of model predictions over the prediction months. How suitable is this metric?

Field	Total
10	3
9	1
8	0
7	1
6	5
5	3
4	2
3	0
2	0
1	0
0	0

Showing rows 1 - 11 of 11

Q3.15 - Metric:Comparison of the average positive and negative predictions over the prediction months. How suitable is this metric?

Field	Total
10	4
9	0
8	4
7	2
6	3
5	1
4	1
3	0
2	0
1	0
0	0

Showing rows 1 - 11 of 11

Q3.16 - Remarks to above questions:

Anmerkungen zu den obigen Fragen:

Die Beschreibung der Metriken scheinen sehr ähnlich. Der Unterschied kommt nicht klar zum Vorschein.

Q3.18 - Metric:Bias of model predictions X = Avg test predictions - Avg predictions of month X How suitable is this metric?

Field	Total
10	4
9	1
8	2
7	3
6	2
5	2
4	0
3	1
2	0
1	0
0	0

Showing rows 1 - 11 of 11

Q3.19 - Remarks to above questions:

Anmerkungen zu den obigen Fragen:

Why comparing with the test data set?

Q4.2 - Metric:Saved contract premiums per month. How suitable is this metric?

Field	Total
10	5
9	2
8	4
7	1
6	0
5	2
4	0
3	0
2	0
1	0
0	1

Showing rows 1 - 11 of 11

Q4.3 - Metric:Lost contract premiums per month. How suitable is this metric?

Field	Total
10	1
9	2
8	1
7	3
6	0
5	2
4	0
3	2
2	3
1	0
0	1

Showing rows 1 - 11 of 11

Q4.4 - Metric:Theoretical percentual increase of the contract stock by the number of saved contracts. How suitable is this metric?

Field	Total
10	2
9	0
8	5
7	1
6	2
5	3
4	0
3	0
2	1
1	0
0	1

Showing rows 1 - 11 of 11

Q4.5 - Metric:Average monthly rescue rate over a moving average of five months. How suitable is this metric?

Field	Total
10	2
9	2
8	2
7	0
6	1
5	5
4	0
3	1
2	0
1	1
0	1

Showing rows 1 - 11 of 11

Q4.6 - Remarks to above questions:

Anmerkungen zu den obigen Fragen:

Die monatliche Rettungsrate wird maßgeblich von den kommunikativen Fähigkeiten des Außendienstes und dem Marktumfeld, u.a. auch dem Verhalten der Konkurrenz bestimmt.

Q5.2 - Metric: Distribution of the weighting of the used data-features for the predictions per month. Topic: To support the monitoring of discrimination. How suitable is this metric?

Field	Total
10	2
9	1
8	2
7	1
6	2
5	4
4	0
3	1
2	1
1	0
0	1

Showing rows 1 - 11 of 11

Q5.3 - Metric:Comparison of the data features used for predictions with the respective entry in the permissions database regarding data sensitivity. Topic:Supporting the monitoring of the use of the minimum amount of personal data that is actually required to fulfill the agreed contractual service. How suitable is this metric?

Field	Total
10	6
9	3
8	2
7	3
6	0
5	0
4	0
3	0
2	1
1	0
0	0

Showing rows 1 - 11 of 11

Q5.4 - Metric:Subjective evaluation of the predictions by the data owner or the person responsible for the data. Topic:Discrimination and data use How suitable is this metric?

Field	Total
10	2
9	1
8	3
7	2
6	3
5	3
4	0
3	0
2	1
1	0
0	0

Showing rows 1 - 11 of 11

Q5.5 - Remarks to above questions:

Anmerkungen zu den obigen Fragen:

Letzteres echt eine Metrik?!

Is question 3 actually a reasonable metric? Seems to complex. It needs a lot of requirements.

End of Report