Computer Science & Economics

Delivery optimisation through data mining for micro-sized restaurants in the SME sector

T.Y. Cheng

Supervisors:

Dr. G.J. Ramackers & Dr. A. J. Knobbe

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

Abstract

Technological innovation is at the forefront of unprecedented changes in the business landscape of various industries. Numerous businesses use data mining algorithms to optimise their processes and gain a competitive advantage on the market. This study aims to explore how data mining can support restaurants to adapt to the challenges they currently face, namely the rise in demand for delivery services. Other studies have mainly focused on predicting restaurant sales in general but overlooked delivery and the micro restaurant firms in the SME sector. For this purpose, prediction models with seven different algorithms for deliveries have been developed and we have tried to explore how a restaurant can utilise a delivery prediction model to reduce costs and maximise potential revenue. The algorithms used are gradient boosting regression, linear regression, ridge regression, lasso regression, decision tree regression, random forest regression and k-nearest neighbours regression. The delivery data used in this model belongs to an Asian restaurant that falls in the SME category located in a town in the province of South-Holland in the Netherlands.This restaurant has characteristics that are typical of a micro-sized firm based on the criteria of the annual turnover, the number of employees and the total value of assets. The data was gathered from the website of online food delivery platform Thuisbezorgd, which is a partner of the restaurant. Additionally, public weather data from The Royal Netherlands Meteorological Institute (KNMI Datacentrum) and in-house sales data of the restaurant was used throughout the study. As a result of the experiments conducted in this research, we have concluded that the best predictors are the days of the week which is in accordance with the findings of the other studies that tried to predict restaurant sales and that some weather features can bring minor improvements to the model.

Table of contents

1. Introduction

1.1 Background

The restaurant industry faces many challenges in the modern economy. Although the prognosis of growth in the sector is positive and higher on average than other sectors of the Dutch economy [1], restaurant owners are confronted with new challenges such as increased competition, stagnant revenue [2] and other emerging threats on the market. According to the Koninklijke Horeca Nederland (Royal Catering Netherlands), the largest Dutch organisation for the hospitality industry [3], revenue growth in 2019 diminished to half of what it was in the previous year. Revenue per business has barely increased in the past years, despite the overall growth of revenue generated by the hospitality industry. A potential reason for this could be that the growth in revenue is offset by the large increase in the number of businesses on the market [4], as the selection of restaurants in the Netherlands has increased by more than 40% during the last decade [4]. Other possible reasons include the VAT increase from 6% to 9% on food items which caused an increase in prices, as well as shortages in staff which are expected to persist in the upcoming years [5]. The rapid advancement of technology is also affecting the restaurant sector. In particular, food delivery has recently become a major demand for consumers and is currently one of the fastest growing markets in the food industry [6].

The use of technology has sparked various disruptions in the restaurant industry. One of the most powerful current trends occurring in the industry is the high consumer demand for online food delivery services [7]. As specified by dealroom [8], convenience is one of the main drivers for this trend. As reported by Muller [7], delivery might threaten the current eat-in restaurants, rendering them obsolete. Amidst these changes, online delivery platforms such as Thuisbezorgd or UberEats have made a tremendous surge in popularity and are growing at a fast pace [9]. To illustrate the growth, in 2018, Takeaway, the parent company of Thuisbezorgd, reported 4 million active users in the Netherlands and processed 33 million orders which accounted for 674 million euros [10]. Takeaway is active in multiple countries and the number of deliveries went up with 38% in comparison to the previous year [10]. In 2019, this number has risen to 38 million in the Netherlands, an increase of 16% in the number of deliveries [11].

Online delivery presents both opportunities as well as challenges for restaurant business owners, who need to ensure they can adapt to the changes on the market. On one hand, delivery allows restaurants to operate and do business on a larger market, thereby expanding their customer reach and potential revenue streams [12]. On the other hand, there are still many uncertainties surrounding the business sustainability of online food delivery platforms. For instance, UberEats makes losses of a few billion euros each year and is expected to do so in the next few years [13]. Another issue is that restaurants are at risk of going bankrupt due to the way that business models of online delivery platforms are constructed. In these constructions, the revenue is split and the platform takes a commission out of every order. Given that deliveries yield lower profits, it is difficult for restaurants to sustain themselves if some platforms charge up to 30% in commissions for every sale [7]. All in all, it is imperative that restaurants seek methods to efficiently meet the new demands of the market and implement delivery in a way that does not lead to bankruptcy and closure.

In the most vulnerable category lie SME restaurants which are, in many cases, not as resilient as larger companies, especially in a saturated market. As such, it is of heightened importance that SMEs allocate resources efficiently and monitor their business closely [14]. In addition to being more vulnerable, SMEs also lag behind in the adoption of new technology unlike their larger counterparts [15]. Within the SME category, this effect might be exacerbated for micro-sized restaurants as they have the least available resources. Even though the restaurant sector is generally known as a low-tech industry [16], larger restaurant brands have already begun to make investments in data analytics [17]. Because the digital maturity is low in the industry, there are many opportunities for businesses that act swiftly. Regardless of this, the threat to slow technology adopters may be a more urgent matter. According to Boston Consulting Group, data and analytics programs can help business operations to build a competitive advantage. Consequently, it is expected that digital leaders will outpace the others and slow-moving restaurants will have difficulties catching up. At the same time, customers' expectations are shifting towards a better digital experience [17]. Although most restaurants are still passive in this matter, it is clear that restaurants in the SME sector, and in particular smaller restaurants need to embrace the capabilities of data if they want to stay ahead of the curve.

2. Aims and objectives

The use of data mining in the business world is rapidly growing. The ability to explore vast amounts of data in a short amount of time offers an immense advantage for businesses who want to obtain insightful information. For this reason, data mining is often used to assist with decision-making in various industries such as health care, insurance and banking [14]. Applying data mining can help businesses to strengthen their position in an increasingly competitive environment. Given this, restaurants should be no exception to this and should also seek to reap the benefits of this practice.

This research aims to explore how micro-sized restaurant firms within the SME sector can benefit from analysing their own delivery data, and how data mining can help them to achieve better results through a data-driven approach. The exact definition and criteria for micro enterprises is given in chapter 3.1 on page 8. To help with carrying out this research, the following research question has been formulated:

Research question:

“How can data mining and analysis of delivery data contribute to improving the business processes of micro-sized restaurants in the SME sector?”

To answer the main research question we will introduce another set of underlying research questions to guide the process.

1. To what extent can micro-sized restaurants predict the number of deliveries per day?
2. What are the most important features for the delivery prediction?
3. What impact does the weather have on the number of deliveries?
4. In what step of the delivery process can the prediction model be used?

By answering these research questions, valuable new insight will be obtained on how to optimise the delivery. The crux of this research is constructing a delivery prediction model to improve the delivery business process. The hypothesis is that the efficiency of deliveries can be improved in theory when a prediction model is deployed.

3. Context and definitions

3.1 Small and medium-sized enterprises (SME)

The acronym SME stands for small and medium-sized enterprises. More than 99 per cent of all Dutch enterprises belong to this category and are responsible for more 70 per cent of all employment [18]. The European Commission reported in 2018 that on average an SME in the Netherlands has 3.2 employees [19]. The definition of SME given by the OECD is as follows: “SMEs are non-subsidiary, independent firms which employ fewer than a given number of employees” [20]. Generally, two criteria are used to determine whether an enterprise is considered an SME. The first one is based on the number of employees and the second is based on financial assets. These two criteria for SMEs may vary across countries. For this project, we will be taking the criteria of the Netherlands Chamber of Commerce [21]:

- An upper limit of 250 employees

And either of the two criteria below:

- Equal to or less than 40 million turnover
- Equal to or less than 20 million total value of assets on the balance sheet

The restaurant that we have taken as a sample fulfills all three criteria and therefore falls in the SME category. Within this category, the business size can be further distinguished in micro, small or medium-sized. The characteristics to determine the business size according to the Chamber of Commerce is shown in table 1:

	Number of employees	Annual turnover (in € x1.000)	Total value of assets (in € x1.000)
Micro	Less than 10	Less than 700	Less than 350
Small	Between 10 and 50	Between 700 and 12.000	Between 350 and 6.000
Medium	Between 50 and 250	Between 12.000 and 40.000	Between 6.000 to 20.000

Table 1: SME business size categories as specified by the Netherlands Chamber of Commerce

Based on the number of employees, total value of the assets, and annual turnover, we can further classify this particular restaurant into the micro enterprise category which is a size typical for regular Asian restaurants.

3.2 Knowledge Discovery in Databases & Data Mining

In this paper, data mining techniques will be deployed to explore and analyse data. The field of data mining is broadly defined and is, in essence, the process of discovering previously unknown patterns in a large amount of data [22]. Data mining is an essential step in Knowledge Discovery in Databases (KDD), which refers to the overall process of finding meaningful and useful knowledge in a collection of data [23][24]. In figure 1, the process of KDD is shown:

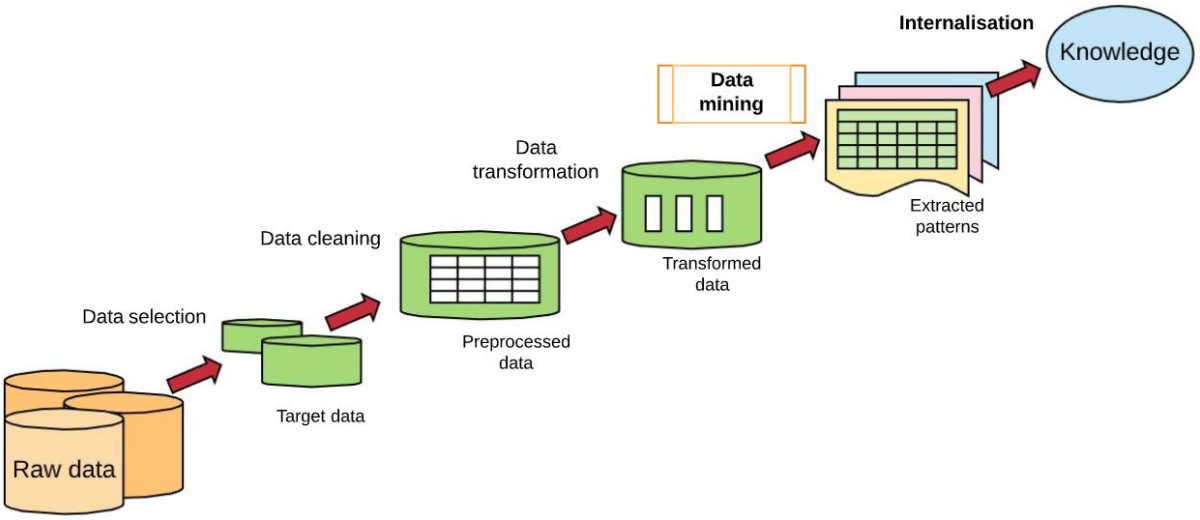

Figure 1: The steps in a KDD process

However, the information gained from data mining only has value if it is incorporated in the decision-making, which is illustrated in the last step in the KDD model. Data mining is widely applied in different domains, such as insurance, banking, telecommunication, finance and retail [25]. In the context of business operations, data mining is used frequently to find patterns in the data which managers use to chart a better course for the business. For example, a business can use the data mining results to develop more effective strategies to increase profitability or reduce costs, such as improved fraud detection, marketing and customer segmentation.

3.3 Machine learning

Machine learning is a subfield in computer science that focuses on programming computers to learn how to independently adapt to new data [26]. Although the field of machine learning has some overlap with data mining, the latter mainly concerns itself with the process of discovering unknown properties of the data [27]. Machine learning, on the other hand, focuses on the development of algorithms to allow computers to learn by itself [28]. Since data mining is an interdisciplinary field, it can take on a machine learning approach in order to uncover new patterns [27]. In figure 2, the different relationships between disciplines are illustrated.

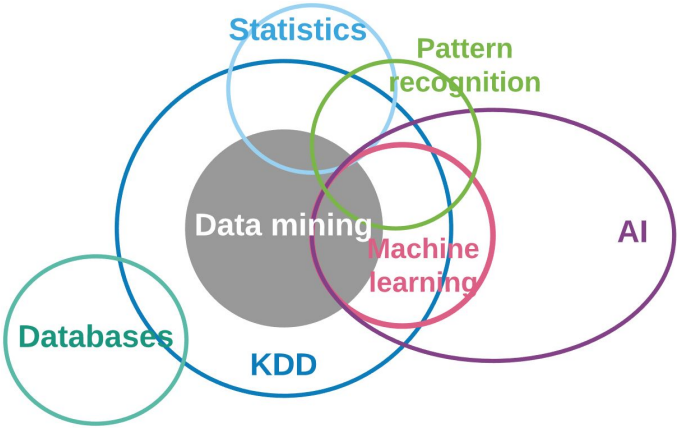

Figure 2: Venn diagrams of the interdisciplinary fields. Adapted from “An Overview of Machine Learning with SAS Enterprise Miner” by Hall, P., Dean, J., Kabul, I. K., and Silva, J, 2014, p.3.

Often two main categories in machine learning are distinguished: supervised learning and unsupervised learning. In the first category, a predictive model is developed based on a labelled data set, meaning that the output is known. The algorithm learns with the input and the corresponding output (e.g. the correct number of deliveries). In contrast, an unsupervised learning uses an unlabelled data set and tries to interpret or search for patterns only based on the input. An example of a common unsupervised algorithm is clustering, a technique that groups similar data points by assigning them to clusters.

There is no best machine learning model, all algorithms have their own benefits and drawbacks depending on the data set and purpose. For this reason, choosing machine learning algorithms is not a trivial process. Nevertheless, there are broad guidelines for choosing the most suitable model.

There are several supervised techniques used within this research. The main one is linear regression, an approach to model linear relationships between the target variable and the given explanatory variable. When a single variable is used, it is called a univariate linear regression. In the presence of multiple explanatory variables, we speak of multivariate linear regression. In linear regression the model parameters are estimated by training the model on a data set with already known labels. The fitted model can then be used to make predictions with the explanatory variables.

When using linear regression, five key assumptions are made:

- **Linearity**: There is a linear relationship between the target variable and explanatory variables
- **Independence**: The observations are independent from each other
- **No multicollinearity**: The explanatory variables are not (highly) correlated with each other
- **Homoscedasticity**: There is a constant variance for the error terms
- **Normality**: The distribution of the errors (residuals) is normal

Since linear regression captures linearity, other algorithms were used to find non-linear patterns. These other algorithms include the following:

- K-Nearest Neighbours (KNN) regression: interpolates the target value based on how similar the target is with its associated neighbours.
- Decision tree regression: builds a tree structure by splitting the data set in increasingly smaller subsets and predicts the target value with the constructed tree.
- Random forest: Builds decision trees independently and trains them by using a random sample of the data. Then the model combines multiple decision trees and then averages the results for the prediction.
- Gradient boosting regression: based on weak learners; builds one tree at a time and uses the new tree to correct errors made by the previous tree and then combines the results.

4. Related work

The body of literature that is concerned with the application of data mining techniques in the micro-sized restaurant sector and online delivery services is very limited. In the following paragraphs, an overview of studies with respect to technological disruptions in the restaurant sector, demand forecasting and impact of the rise of delivery on the restaurant industry will be presented.

A recent study done by Holmberg and Hallden has attempted to develop a sales forecast model and to investigate if weather has any influence on the overall sales of restaurants [29]. Their research was constructed with three datasets from three different restaurants which originate from similarly populated cities in the southern part of Sweden. Certain useful features have been created and selected for the purpose of training the data sets to ensure the results are relevant and help to predict the amount of sales. The algorithms used in the project were XGBoost, LSTM and Uplift. This study indicates that the day of the week is the best feature for the prediction of sales and weather impacts the prediction the least. However, weather features are still beneficial to use on the machine learning models considering minor improvements to the models can be seen.

In one study by Takashi et al [30], researchers tried to use machine learning and statistical analysis to forecast the number of customers of restaurants. For this purpose, a model was developed by combining data retrieved internally from the restaurant. Primarily used was point of sale (POS) data, which are records of the retail transaction stored by the till system as well as external data such as national holidays or weather data. On the basis of this model, three types of regression models, namely Bayesian Linear Regression, Boosted Decision Tree Regression, Decision Forecast Regression and the Stepwise statistical analysis method were applied to 5 different restaurants from Japan. The results showed minute differences between the different models, and forecasted values of expected customers were almost identical to actual values.

Other papers, such as the one by Khan [31] focus on the impact of disruptive technologies and innovations on the restaurant industry, such as the current delivery trend. This study proposes that the increase in the use of technology will occur gradually in 6 stages. Stage 1 is characterised by the limited use of technology and is based mainly on traditional systems such as POS, while stage 6 represents a highly technological dependent service such as drone delivery that would eliminate personal contact between restaurants and consumers. Some important findings of this study are that delivery services, and especially those that involve third-party services, may eventually lead to losses rather than profits. Another important aspect is the vulnerability of smaller players in the market, who may not be able to cope with the disruptions. Changes in food quality and the separation formed between consumer and provider also need to be considered. It is clear that restaurant owners need to carefully think about how to respond to these alterations.

Similarly, Bujisic, Bogicevic and Parsa also conducted research on restaurant revenue and its relationship with the weather [32]. The main focus of this work was the effect of weather factors on the sales via stepwise regression. Bujisic et al. considered temperature the most important weather factor when predicting sales of menu items in a restaurant. Bujisic et al. argued that precipitation, on the other hand, was not significant in any of their models. The results of their analysis indicate that weather factors have a strong influence on both total sales and on the sales of specific menu

items. Considering that temperature is the best predictor for sales, Bujisic et al. advises management to include temperature indicators in the forecast of their sales. Another key point is that the data set was obtained from a restaurant located within a hotel complex. Hence the sales might have been affected by the hotel's occupancy rate which limits the generalisability.

In all the studies reviewed here, various data mining techniques have been applied to the restaurant industry. However, these studies show that scant attention has been paid to delivery business process improvements and smaller restaurants in general. In this paper, the use of data mining to optimise delivery for specifically a micro-sized restaurant is investigated. The contribution of this paper is theoretical as well practical. By expanding the implementation of data mining to business operations in a micro restaurant firm, this study contributes to the existing literature of restaurant data mining research and fills this gap of knowledge. To the best of our knowledge, no study has attempted to predict deliveries of micro-sized restaurants within the SME sector. In addition, the proposed research approach can be used as a practical application in the real world.

This paper also tries to address the importance of small businesses for society. It is relevant because it aims to discover ways in which smaller sized restaurants can remain competitive on the market. As stated by the European Commission, SMEs are a crucial component of the EU economy and they account for more than 99% of all businesses and 75% of total private sector employment in the EU [33]. The absence of SMEs would be detrimental to economic growth, innovation, job creation or social integration [34]. Among this, the food and drink industry is the most important sector in the EU in terms of jobs and value added [35]. For the previously specified reasons, it is essential to ensure the well-being and survival of small businesses and to maintain the diversity of market provision. Data mining is an increasingly important and useful tool that can assist managers of restaurants to gain access to valuable insights which can help them remain competitive against bigger and more resourceful enterprises.

5. Methodology

To carry out this research, the datasets of a restaurant (typical of SME and specifically micro-sized) will be utilised. The paper will primarily focus on the delivery sales data of the restaurant. The general approach for this research is as follows. Firstly, a description of the data sets will be given. Thereafter, the data cleaning process and the steps within executed are described. Consequently, we will describe the process of the data preprocessing including the data transformation and feature selection. Lastly, we outline the approach of the experiments.

5.1 Tools

The experiments were performed in the programming language Python 3.7.3 with the use of Jupyter Notebook (version 5.7.8) including multiple additional packages that were required to carry out specific parts of the experiment. Modules from NumPy (1.18.1) and Pandas (1.0.3) were used to provide the data structures, Scikit-learn (version 0.22.1) and Statsmodels (version 0.11.0) for the machine learning algorithms and metrics and, Seaborn (versions 0.10.1) and Matplotlib (version 3.1.3) for the visualisations.

5.2 Description of the data

To carry out this research, we have collected three datasets from two different sources: (1) actual delivery data from the micro restaurant firm extracted from the Thuisbezorgd software as well as the in-house sales data from the POS software, and (2) publicly available meteorological data obtained from The Royal Netherlands Meteorological Institute (KNMI Datacentrum).

5.2.1 Delivery sales data

The restaurant is an Asian cuisine restaurant located in a town in South-Holland in the Netherlands that offers services in takeaway, eat-in and delivery. For the delivery, the restaurant is affiliated with the third-party online food delivery platform Thuisbezorgd. This collaboration entails that the restaurant is listed as an option where customers can order from on the online platform of Thuisbezorgd, which acts as an intermediary between the customer and the restaurant. After an order has been placed, it is communicated to the restaurant via the Thuisbezorgd software so it can be prepared and delivered. The restaurant carries out all deliveries on its own and Thuisbezorgd takes a commission of 13% out of the value of any order. Next to that, Thuisbezorgd also stores the delivery data for the previous 12 months on its software, which is where the data used in this research was obtained from. This dataset contains 5033 records and it covers slightly less than the span of one year from May 2018 until May 2019.

5.2.2 Restaurant in-house sales data

The restaurant computer runs on the POS software Posbill, which is a German commercial off-the-shelf (COTS) till system for the catering industry. The restaurant in-house sales data covers around 4 years' worth of data which has been collected over the period of 5 May 2015 up until 3 May 2019. It contains the sales record of every item in the aforementioned period. This database has been extracted from the main computer of the restaurant and has around 290.000 instances from the takeaway and eat-in section combined.

5.2.3 Meteorological data

The last data set is acquired from the public information centre of The Royal Netherlands Meteorological Institute (KNMI Datacentrum), which concerns itself with monitoring weather, meteorology, seismic activity and climate [36]. For this project, we have chosen to use the meteorological station Rotterdam, station number 344, seeing that it covers roughly 90% of the delivery area. The original data set included 40 different metrics of the weather per day for the period between October 1, 1956, and present.

6. Data cleaning

As noted by Zhang [37], data cleaning is a fundamental process of data mining. Due to the nature of data collection, real-world data tends to be ‘dirty’, which means that they are incomplete, noisy and/or inconsistent. Errors in data usually appear during the collection and acquisition phase. Some examples of data errors are typos and duplicated or wrongly submitted entries [38]. Incomplete data refers to missing values or attributes. Noisy data is when the data contains erroneous values or anomalies. Additionally, inconsistent data occurs when the data set contains discrepancies.

Zhang, therefore, states that preprocessing data sets is required to ensure the quality and soundness of the data to ultimately improve the data mining results [37]. As machine learning algorithms are highly dependent on the quality of the input, working with impure data will yield low-quality results or even incorrect conclusions. Coined by George Fuechsel, this concept is often referred to as ‘Garbage in, garbage out’ (GIGO) [39]. If not avoided, this may lead businesses to make poor financial decisions due to the inaccuracies of the models. Thus, data cleaning is a crucial step to increase the reliability of the data. Unfortunately, Wickham argues that cleaning the data is not a linear process as new issues come to light during the data manipulation. As a result, data cleaning is often an iterative process over the course of analysis. Hence it is a common trope that 80% of the time of data analysis is spent on data cleaning [40].

To clean the data sets, we have applied the following data cleaning techniques:

- Conversion of values to their correct data types
- Removal of missing values and empty columns
- Filling in missing values
- Removal of “deleted entries” from sales data
- Removal of instances with odd or incorrect values in the delivery data

6.1 Cleaning of the data

During the data cleaning step, the delivery data undergoes multiple phases. For the sales data only the first phase of data cleaning is applicable. The purpose of the first phase is to ensure that the data type is correct and to fill in any blank values such as NaN. For the columns paid online (online betaald) and pickup, we have filled in the missing values NaN with 0. Another operation the data has to undergo is the replacement of all commas with a dot, because Python requires a dot as a decimal separator. Initially, all of the data were a non-null object or non-null float64 type. An example of a conversion is converting a date from a non-null object to a DateTime object or the price to a float. This is necessary to ensure that the data can be manipulated in a later phase. Here below a table of the raw delivery data imported in pandas is shown:

	Bestelling	Datum	Postcode	Totaalbedrag	Online betaald	Pickup
0	IEV3NS	01-05-2018 18:53	2672DM	33,50	1.0	NaN
1	DA1XEX	01-05-2018 19:24	2681PC	32,00	1.0	NaN
2	CP7F0J	02-05-2018 14:49	2671LT	58,50	NaN	NaN
3	GEG6HJ	02-05-2018 17:13	2672EA	20,50	NaN	NaN
4	2U93X2	02-05-2018 18:41	2681BS	24,50	1.0	NaN
5	LVRGD3	02-05-2018 18:50	2681LC	21,50	1.0	NaN
6	98PQ7P	02-05-2018 18:59	2673CR	22,50	NaN	NaN

Table 2: Sample of the delivery dataframe imported in pandas

After exploring the data set, it was discovered that the attribute postal code contained some odd values which did not conform to the standard Dutch postal code format. Since Dutch postal codes start with four numerals followed by two letters (1234AB), it was clear that these postal codes were invalid and therefore removed from the data set. Seven instances were identified where this was the case, and they probably originate due to people not filling their postcodes in correctly when ordering. Those instances with odd values were temporarily replaced with the value 0.

To determine the location, only the numerical part of the postal code will suffice, so only the first 4 characters are taken into account. Based on the postal codes, the corresponding place is assigned to their respective instances. When a postal code is not located in the delivery area, the place ‘Unknown’ is assigned to it. All of the instances with an unknown place are deleted, as these are incorrect postal codes that are outside the designated delivery area.

Aside from delivery, Thuisbezorgd also offers pickup as an option which was rarely used. Along with deleting the entries with incorrect postal codes, a couple of instances that were meant for pickup were deleted.

Before the cleaning of the delivery dataset, there were 5033 instances. Seven of the 5033 instances have been removed from the data set due to the invalidity of postal codes, and another 23 instances were removed because they were intended for pickup. Altogether 30 instances were removed, and thus 5003 instances remained after the data cleaning.

7. Data preprocessing

7.1 Feature creation

Given the fact that the number of predictors is very limited, the first step will be to construct new features that can potentially be used by the algorithm. Using too few features may lead the model to be underfitted. Domain knowledge, understanding of what can affect the target variable, can aid feature creation. The following five features were constructed:

- Nr_Deliveries: number of deliveries per day.
- Weekdays: categorical variable to indicate which day of the week it is
- Season: categorical variable for the season
- Payday: from domain expertise, there is more demand at the end of the month. The restaurant owner attributed this to the fact that wages are usually paid at the end of the month. Payday is a dummy variable if the given day likely was a common payday.
- Holiday: dummy variable to determine if the given date was a holiday
- School_holidays: dummy variable to determine if the given day was an official school holiday in South-Holland

Next, we incorporated the five most prominent weather features from the public meteorological data set of the KNMI that we presume could affect the number of deliveries. Data that was in a period not covered by the delivery data was discarded. The chosen features with their respective description from the KNMI are shown in table 3 below:

Feature name	**Original name by KNMI**	**The description of the feature by KNMI**
Daily wind speed	FG	Daily mean wind speed (in 0.1 m/s)
Daily temperature	TG	Daily mean temperature (in 0.1 degrees Celsius)
Sunshine duration	SQ	Sunshine duration (in 0.1 hours) calculated from global radiation (-1 for <0.05 hour)
Daily precipitation	RH	Daily precipitation amount (in 0.1 mm) (-1 for <0.05 mm)
Daily cloud coverage	NG	Mean daily cloud cover (in octants, 9=sky invisible)

Table 3: Selected features from the KNMI weather database

7.2 One-hot encoding

It is important to realise that machine learning algorithms generally have more difficulty interpreting text. A common practice is to label binarize or one-hot encode the categorical variables to circumvent this issue. The first is a common practice for ordinal variables, categorical variables that can be ordered (e.g. good, moderate, bad), and the latter for nominal variables (non-ordered) [41]. Because the variables are non-ordinal, the categorical variables were one-hot encoded. That is,

the data is converted in a form that is represented by integers instead of strings. For each distinct category in a categorical variable, a new dummy variable is added and given the value 1 for the instances that belong to the category and 0 in the other case [42]. In other words, the categorical variable itself is removed and mapped to a binary array that is added to the data set.

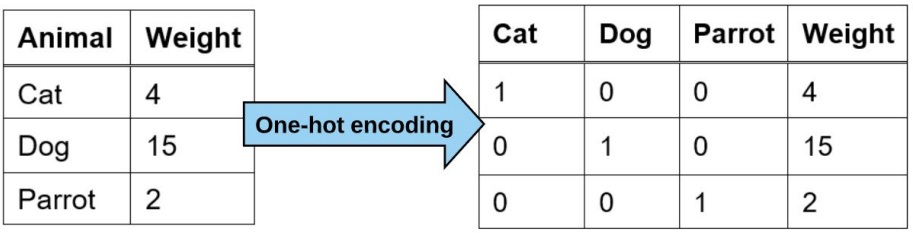

Figure 3: Categorical feature being one-hot encoded

The downside of this is that it increases the dimensionality of the data quickly as each feature increases the dimensionality exponentially, particularly if the one-hot encoded feature has a high-cardinality [43]. Since the same number of data points are spread out over a larger space (higher dimension), it quickly leads to data scarcity [44]. Data scarcity is undesirable because it can lead to poorer performance of the machine learning algorithm [45]. This phenomenon is commonly referred to as the curse of dimensionality. Another matter is that one-hot encoding induces perfect multicollinearity which is discussed on page 18.

7.3 Standardisation

Variables with larger values can intrinsically influence certain machine learning algorithms, because they may attribute larger weights to features of magnitude [46]. This is a problem for machine learning algorithms that for example make use of the Euclidean distance (e.g. K-Nearest Neighbours) [47]. For algorithms such as Ridge or Lasso regression, which places a penalty on the feature based on their magnitude, standardisation ensures that all variables are penalised equally. Since features with widely varying ranges can lead to a bias in the results, it is necessary to standardise the values of the numeric variables to prevent this. Standardisation, also known as Z-score normalisation, is a widely used technique to rescale the quantitative variables to center around the mean (mean becomes zero) with a standard deviation of 1. By standardising, each feature will contribute proportionately to the result, preventing that large scale features outweigh other features heavily. The formula of standardisation is given as:

$$x_{standardised} = \frac{x - \mu}{\sigma}$$ [48]

where $X_{standardised}$ is the standardised value of the original value x, μ for the mean of x and σ for the standard deviation.

7.4 Multicollinearity

One of the key assumptions of linear regression is that variables are independent of each other. If this is not the case, it can cause the estimated coefficients of the regression to be unstable, making it difficult to interpret the coefficients [49]. This phenomena occurs when predictors have a high correlation with each other and is called multicollinearity. Since the one-hot encoded columns of the categorical features are mutually exclusive, this leads to perfect multicollinearity when the

features are combined linearly [50]. An example of a perfect linear relationship of the days of the week is given below:

$$weekday_{Monday} + weekday_{Tuesday} + \ ... \ + weekday_{Saturday} + weekday_{Sunday} = 1$$

The relationship of any of these variables can be expressed as a set of the other independent variables:

$$weekday_{Monday} = 1 - weekday_{Tuesday} - weekday_{Wednesday} - \ ... \ - weekday_{Saturday} - weekday_{Sunday}$$

Since it is possible to substitute any of the categorical variables by rearranging the formula, one of the variables supplies redundant information. In other words, when the values of the 6 other categorical variables are known, the value of the last variable can be inferred (e.g. if the values for Tuesday till Sunday are all 0, then the value for Monday has to be 1). Therefore the 'last' variable only provides additional weight without giving any extra information and one of the one-hot encoded variables should be removed to prevent dependencies between the categorical variables.

Thus, to prevent multicollinearity one column of each hot-encoded categorical variable is dropped. Dropping these features does not lead to a change in the R-square score since the dropped variable is redundant either way. Because it does not affect the outcome of the model, the dropped columns were arbitrarily chosen. In this specific case the columns "weekday_Friday" and "season_Fall" have been dropped from the categorical variables "weekday" and "season".

	Features	**VIF**
1	Sunshine duration	3.397
2	season_winter	3.144
3	season_summer	2.790
4	Daily temperature	2.743
5	season_spring	2.615
6	Daily cloud coverage	2.610
7	School_holiday	2.085
8	weekday_Thursday	1.740
9	weekday_Monday	1.740
10	weekday_Tuesday	1.734
11	weekday_Sunday	1.732
12	weekday_Saturday	1.731
13	weekday_Wednesday	1.726
14	Daily wind speed	1.309
15	Daily precipitation	1.661
16	holidays	1.064
17	payday	1.021

Table 4: Variance inflation factors of all features

One method to quantify multicollinearity is by calculating the variance inflation factor (VIF) for each predictor. VIF measures to what extent the variance of a coefficient is inflated as a result of multicollinearity [51]. A commonly cut-off point for severe multicollinearity is a VIF of five [52]. The VIF for the data set is shown in table 4. Since all VIF were below 5, no other features have been removed from the data set in this step.

7.5 Feature selection

Feature selection according to Wikipedia is "*... the process of selecting a subset of relevant features for use in model construction*" [53]. It is a crucial process since it can impact the outcome of the model greatly [54]. Using fewer features with an equal or better accuracy for a model is more desirable for multiple reasons. First of all, it helps to keep the computational time low and reduces the training time. Secondly, according to Occam's Razor, a principle in the data science field, it is better to keep the model simple to have more explainability. And thirdly, as explained prior with the curse of dimensionality it is important to reduce the number of features to prevent overfitting so that the model can generalise well [55]. Irrelevant features which do not contribute or contribute little to the accuracy can disrupt the model in a similar manner as noise [56] and therefore negatively impact the performance [54]. There are many ways to perform feature selection, however feature selection methods are typically classified in three categories, namely filter methods, wrapper methods and embedded methods, each with their own benefits and drawbacks [48].

In the first step of the feature selection, a Pearson correlation matrix, which is a filter method, is used to assess the bivariate linear relationship between the variables. The correlation coefficient, denoted with r, lies between the values -1 to 1.

- A r of 0 implies no correlation between the variables
- A r of 1 implies a perfect positive correlation
- A r of -1 implies a perfect negative correlation

There are no hard rules for describing the strength of the correlation, but generally, a |r| below 0.4 is considered a weak correlation and a |r| between 0.5 and 0.7 is considered a moderate correlation. Any |r| greater than 0.7 can be seen as a strong correlation. Features may have to be removed if they exhibit high correlation, because it indicates multicollinearity. There is no hard standard for the coefficient value with regards to removing features, but for this paper, we have taken a |r| of 0.7 as threshold.

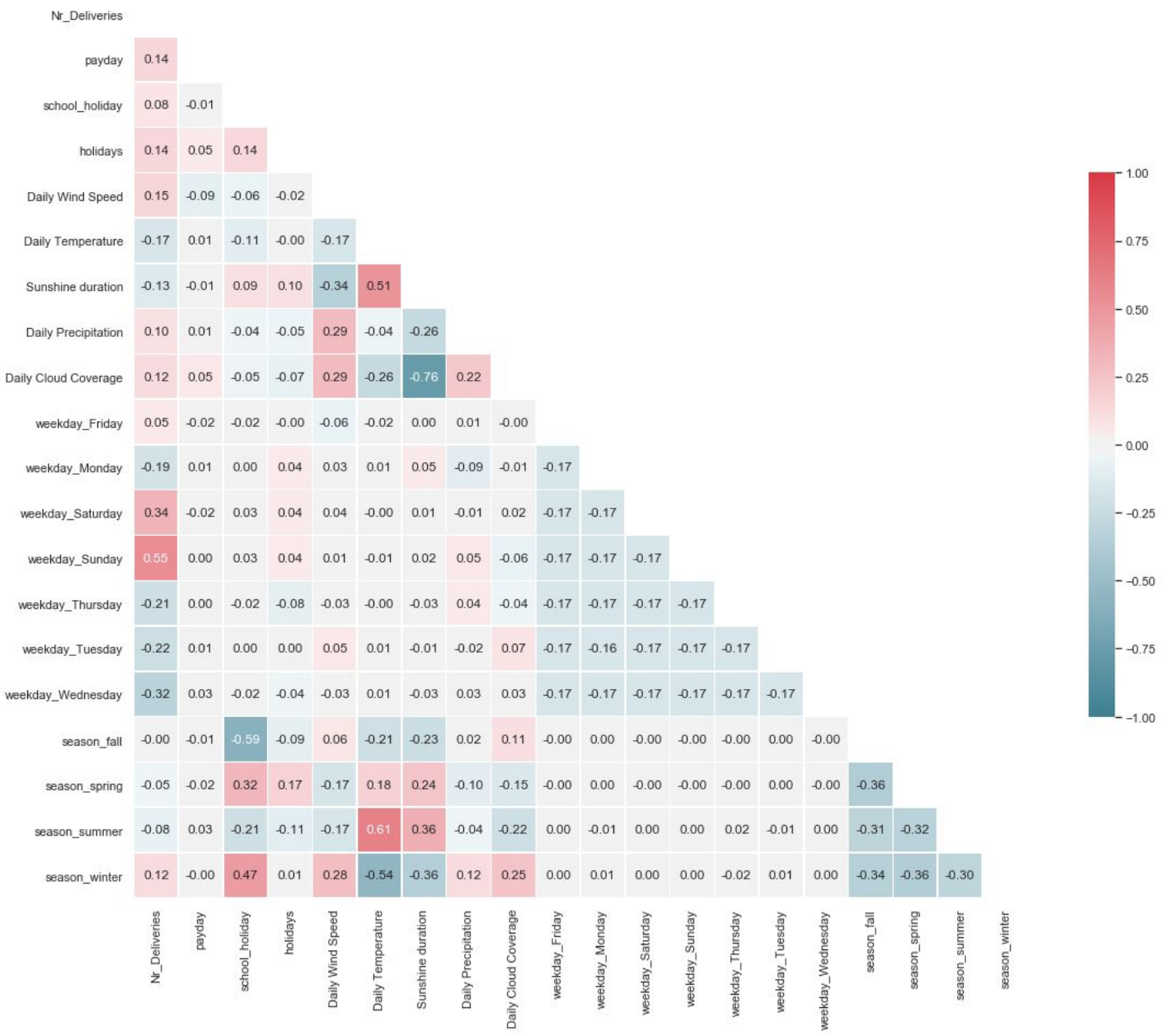

Figure 4: Pearson correlation matrix of the features

Figure 4 presents the intercorrelations among the features. Most of the correlations between the weather features are as expected. For example, the moderate correlation between Daily Temperature with the seasons summer and winter is a logical observation considering temperature has a seasonal relationship. The same goes for sunshine duration and daily cloud coverage given that there is less sunshine when it is cloudy. Because the correlation between these features exceeds the threshold, the feature daily cloud coverage is removed from the data set. For linear regression it would be optimal if there is a high correlation between the predictors and the target variable, which is in this case the number of deliveries. Overall, the majority of the predictors have a weak correlation with the number of deliveries meaning. Another notable observation is that the days of the week have the highest correlation with the number of deliveries. So it is likely that these will be the most important features in a linear regression. In appendix A.2, the Spearman correlation for the features are shown. Spearman can capture non-linear relationships as it

assesses the monotonicity between variables. Overall, there is no big difference between the Spearman and Pearson correlation matrix.

7.6 Transformation of predictors

In some cases it can be useful to investigate whether predictors can be transformed to have a linear relationship with the target variable in order to improve the performance of the linear regression model. To investigate the dependency of the target variable on the predictors, a pairplot of the quantitative variables has been made.

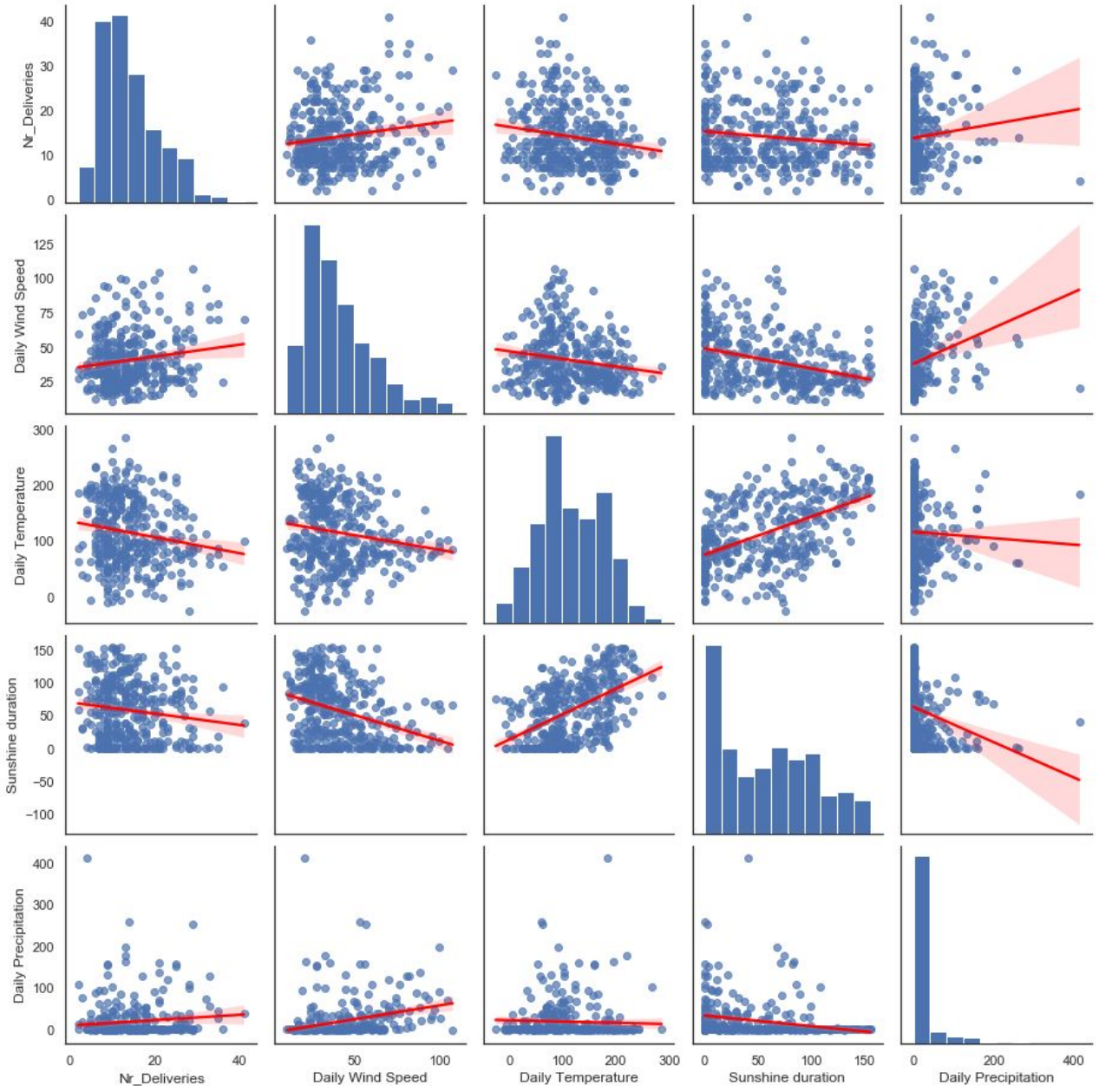

Figure 5: Pairplot of the quantitative features

In the pairplot, if we fit a line through the points (see figure 5), we can observe that there is little to no linearity in the scatter plots which was also indicated by the weak correlations in the Pearson correlation matrix. Furthermore, there are no other clear non-linear relationships to be seen in the

dependencies that we could transform to a linear relationship with the target variable. Additionally, another version of the pairplot has been made of the quantitative features coloured by weekday which is shown in appendix A.1.

7.7 Homoscedasticity

In order to confirm whether the findings of a linear regression are valid, the assumption of homoscedasticity must sufficiently hold. This assumption holds when the residuals, the difference between the observed value and the predicted value, have a constant variance. In other words, the error term of the linear regression is roughly the same for every value of any predictor and the mean of the residuals should be zero. A quick scatterplot of the residuals against the fitted values of a linear regression is made to verify whether the assumption of homoscedasticity holds. The points on the scatter plot (see figure 6) should be uniformly distributed around the grey dotted line. A smooth red line is fitted through the points for visual inspection which ideally should be completely horizontal. In this case, the red line is relatively flat and because no definite pattern can be discerned from the cloud of points, it is adequately evenly distributed and thus the assumption is not violated.

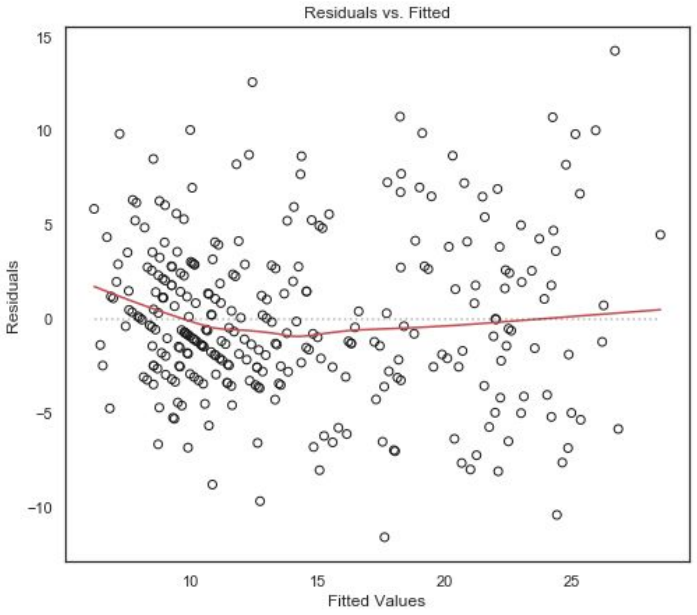

Figure 6: Residuals vs fitted plot of linear regression

7.8 Normality of errors

The normality of errors assumption is necessary to ensure the p-values for the linear regression are valid. Because backward feature elimination makes use of p-values, it is necessary to confirm whether this assumption holds. Central limit theorem states normality is implied when there is a sufficiently large sample size. A sample size equal or greater than 30 is generally considered large enough for the central limit theorem to hold.In this research, there are 355 data points in total which should approximate to a normal distribution according to the central limit theorem. Nonetheless, we can evaluate whether this is indeed the case by constructing a normal

quantile-quantile (Q-Q) plot. The majority of the data points are properly on the red line so the distribution is indeed normal. On the left side, the distribution is somewhat skewed.

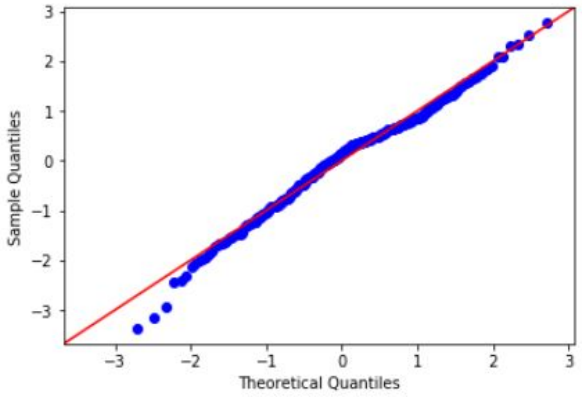

Figure 7: Quantile-quantile plot of residuals

7.9 Backward feature elimination

To select the most relevant features for the linear regression, backward feature elimination (BFE) wrapper is used. BFE is an iterative process that eliminates one feature at the time until all features are below a certain p-value. The p-value tests the null hypothesis whether the feature has no correlation with the target variable. If the p-value for a feature is below the chosen threshold, then the null hypothesis is rejected and there is enough evidence that the feature is associated with the target variable. The process of BFE is described in figure 8.
In the first step of backward feature selection a significance level is chosen. In this case, a significance level of 0.05 is selected. Backward feature selection starts with all features of the data set, then fits the model and calculates the corresponding p-values for each variable. After that, BFE compares the variable with the highest p-value with the threshold. If the p-value is above the threshold, then the feature is eliminated and the BFE process is repeated with the subset. The iterations stop when there are no more features with a p-value higher than the significance level. The BFE eliminated in total seven features in the following order: season_spring, daily precipitation, season_winter, sunshine duration, season_summer, school holiday and holidays.

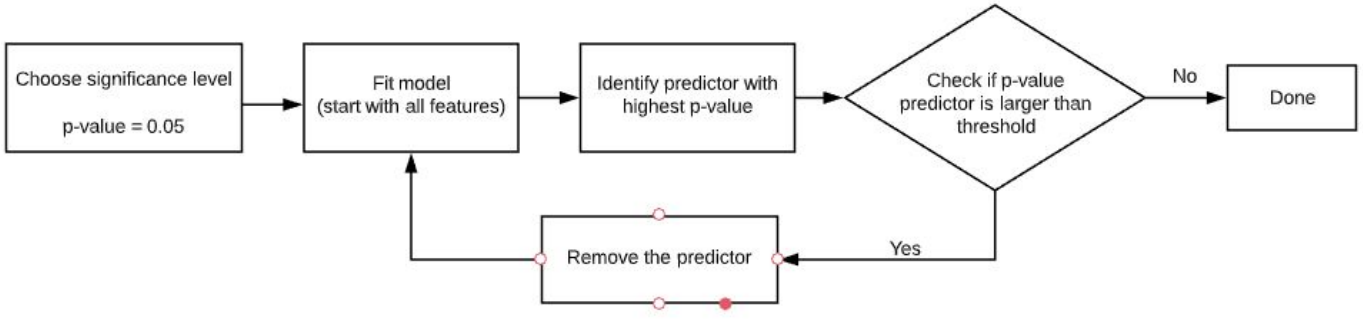

Figure 8: Backward feature elimination process

A manual approach of BFE based on the feature importance of the linear regression is attempted to determine the optimal number of columns which is 8 or 9. This finding is consistent with the results of BFE based on p-value. Description of the manual BFE can be found in appendix A.3 to A.7.

7.10 Final data set

During the preprocessing, nine features of the 18 features were dropped. Thus the subset of features consists of nine features. In total there are 355 data points in this final data set. A discussion of the potential explanations of these eliminations can be found in chapter 10.1.

Subset of features
Payday
Daily wind speed
Daily temperature
weekday_Monday
weekday_Tuesday
weekday_Wednesday
weekday_Thursday
weekday_Saturday
weekday_Sunday

Table 5: The remaining subset of features after feature selection

Dropped features	Reason
weekday_Friday	One-hot encoding, multicollinearity
season_fall	One-hot encoding, multicollinearity
season_spring	BFE, exceeds p-value
Daily precipitation	BFE, exceeds p-value
season_winter	BFE, exceeds p-value
Sunshine duration	BFE, exceeds p-value
season_summer	BFE, exceeds p-value
School_holiday	BFE, exceeds p-value
holidays	BFE, exceeds p-value

Table 6: The dropped features

8. Experiments

In the first part of the experiment, some exploratory analysis with the delivery and in-house sales data is done to gain more insight regarding the delivery domain in general. In the second part, predictive experiments were held to construct a delivery prediction model with seven different algorithms and the subset of features after the feature selection as well as with the set containing all features.

All predictive experiments were performed with a k-fold cross validation. K-folds cross validation is a technique to evaluate how the performance of the model generalises. This implies that the data set has been split into k equal subsets, called folds. For each iteration, one fold is held back as a validation set and the other remaining k-1 folds are used as the training set. During the iterations the folds are rotated in such a way that each fold is used once as a validation set. In this way, each data point is tested once and is used in the training four times. When choosing k, the size of the dataset and the variance- bias tradeoff must be considered. Choosing a large k means splitting the dataset in many folds which leads to a higher variance, but less bias. Vice versa happens when choosing a small k. Through experimentation, values of 5 or 10 for k have shown to have the best compromise between bias and variance in general. For our case, we have taken a k of 5 since the delivery dataset only contains 355 instances which is not too large.

Before doing the k-fold cross validation, the hold-out method is used. With the hold-out method, the dataset is split into a training set and test set in a certain ratio. It is difficult to determine what the

best ratio to split a data set is, but commonly a ratio of 80:20 is used. This ratio gives 284 instances in the training set and 71 instances in the test set.

K-fold cross validation is then performed with the training data to roughly approximate adequate hyperparameters for the algorithms. Not all possible combinations of hyperparameters have been tested, because finding the optimal hyperparameters is too costly in terms of computation time and resources and therefore limited hyperparameters have been tested. The goal of tuning the hyperparameters is so that we can get a reasonable idea of how well each model can perform. In order to find the best hyperparameters, 'GridSearchCV' is used to test the performance of each set of hyperparameters. After tuning them, the best model is fitted on the remaining test set to assess the performance. The performance on the test set is then taken as the final performance of the model. This approach is schematically represented by figure 9.

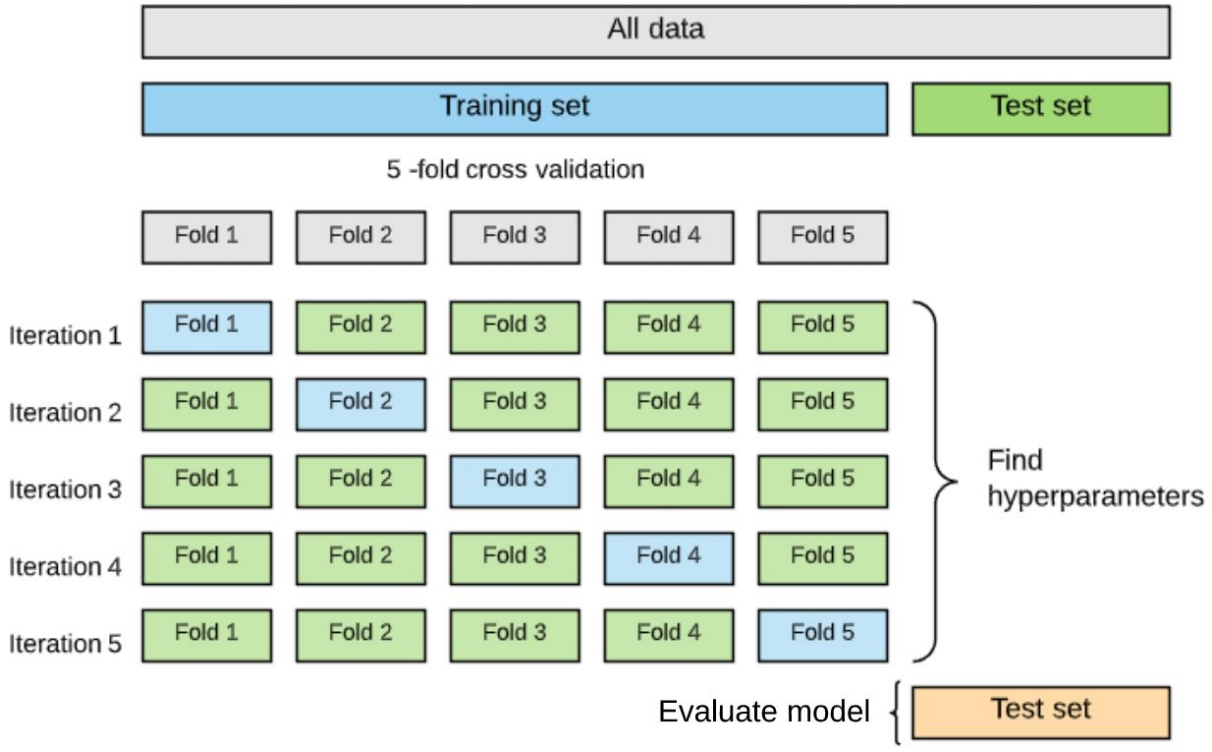

Figure 9: Schematic approach of the experiments

Interestingly, one of the algorithms had an unusual performance and was subjected to scrutiny. Therefore, at the end of the experiments, one more 5-fold cross validation was done over all data to investigate whether the R^2 score in the hold-out method reflected the overall performance well.

9. Results

9.1 Exploratory results

Figure 10: Amount of euros delivery sales vs amount of euros restaurant sales

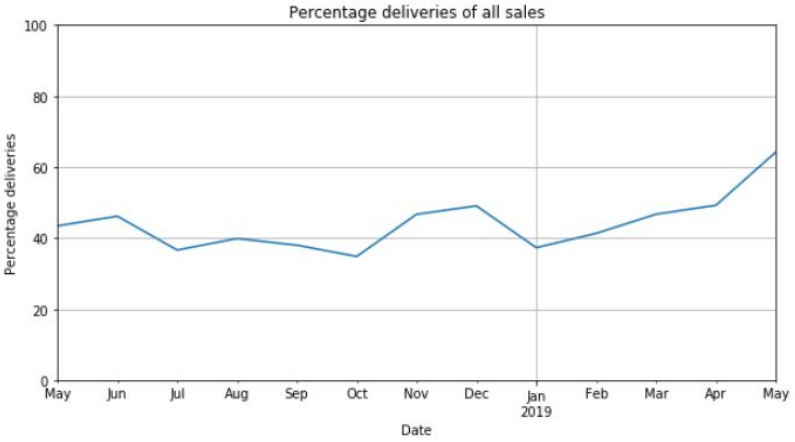

Figure 11**:** The increasing number of deliveries as percentage of total restaurant sales

What stands out in figure 10 is that in the year 2018 delivery sales and restaurant sales do not necessarily move together, apart from the month August when the restaurant was closed due to holidays. However, restaurant sales seem to be more volatile, meanwhile delivery sales increase on a more consistent basis. This consistency is also shown in figure 11. If this trend continues, delivery may soon overtake in-house restaurant sales as the primary source income for the business.

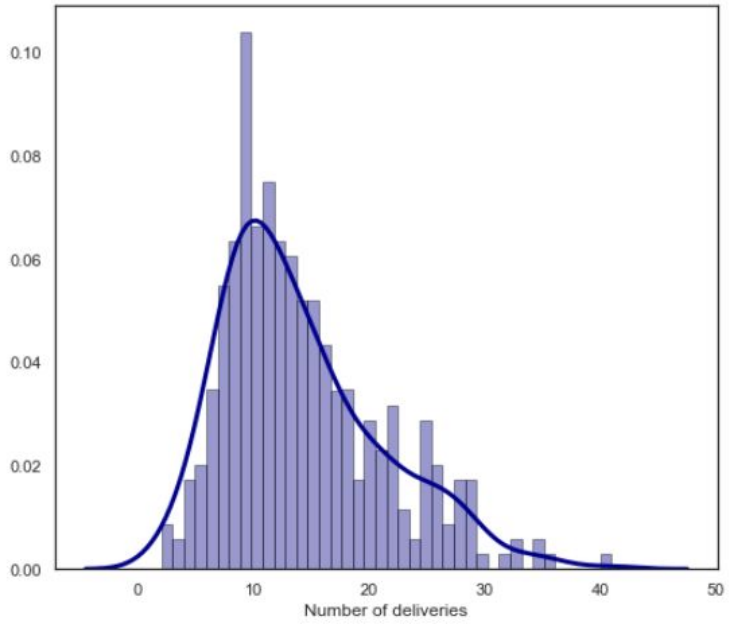

Figure 12: Density plot of the number of deliveries

In figure 12, a density plot of the distribution of the number of deliveries is shown. The maximum number of deliveries on a single day is 41 and the minimum is 2 deliveries. The distribution is

approximately normal and slightly skewed to the right by a few larger instances. It is interesting that there is a much larger peak around the 9~10 deliveries compared to the neighbouring numbers of deliveries.

In table 7, different descriptive statistics are shown to indicate the amount of variability in the data set. These statistics are visualised by using a boxplot.

Weekday	**Count**	**Mean**	**Std.**	**Min.**	Q_1	Q_2	Q_3	**Max.**
Monday	50	10.86	3.88	3	9	10	12.75	25
Tuesday	50	10.34	4.19	2	8	10	12	27
Wednesday	51	8.75	3.41	2	7	8	11	17
Thursday	51	10.47	3.64	3	8	10	13	20
Friday	51	15.00	4.61	7	11.5	15	18	26
Saturday	51	19.78	6.67	6	15	18	23.5	41
Sunday	51	23.35	5.55	13	19.5	23	26.5	36

Table 7: Descriptive statistics of the distribution of the number of deliveries

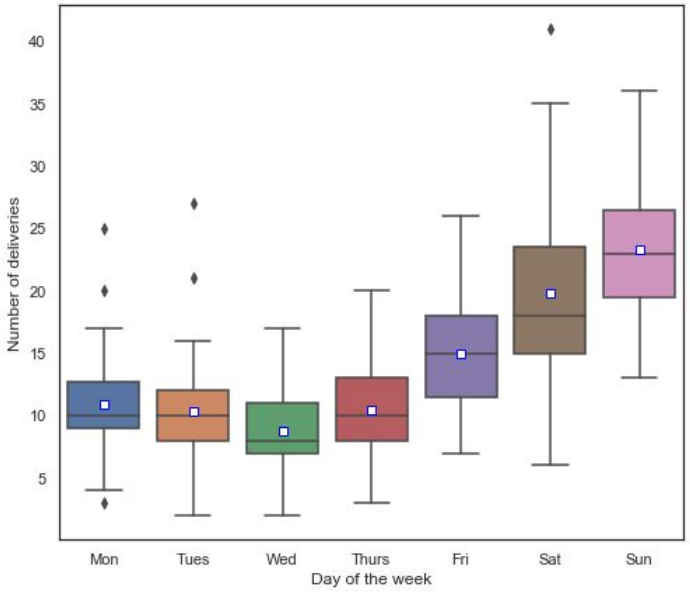

Figure 13: Boxplot of the number of deliveries by weekday
(white square indicates the mean)

In figure 13, a boxplot of the number of deliveries is plotted against the day of the week. In the figure it is shown that the demand for deliveries is reasonably constant from Monday through Thursday with Wednesday as the least busy day. For these days, the number of deliveries do not vary too much. However, it is unexpected that Monday and Tuesday have the most fliers (outliers in a boxplot), considering the whiskers have the smallest range. Starting from Friday, the demand for deliveries noticeably goes up until it reaches its peak on Sunday. Another insight is that

Saturday has the largest variance in the number of deliveries. A discussion of the fliers is presented on page 36.

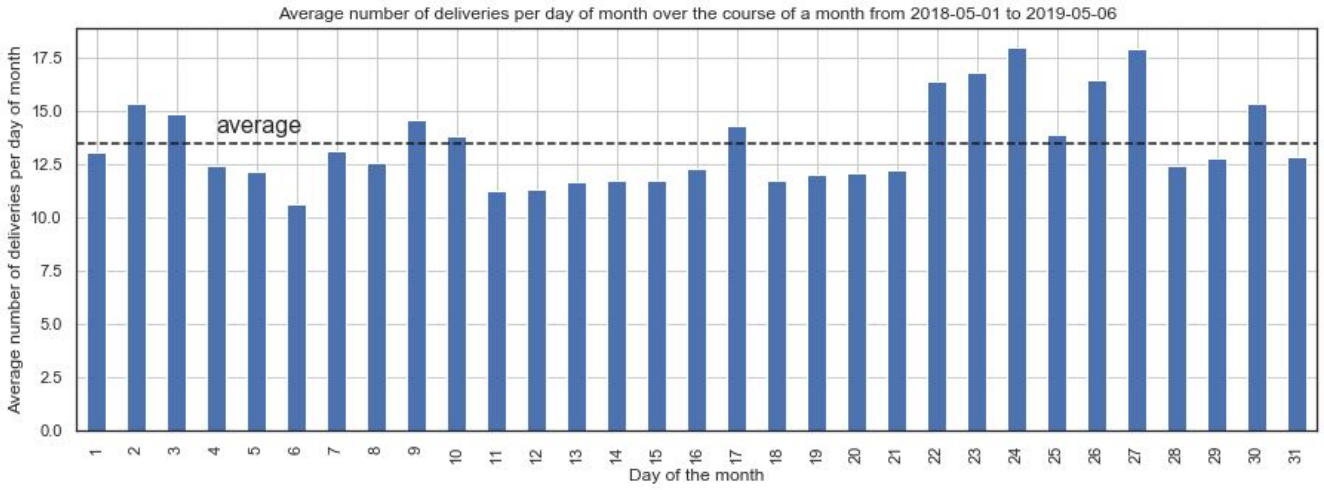

Figure 14: Average number of deliveries per day of the month

As shown in the graph above (figure 14), the demand for delivery is at its highest between the 22th and 27th day of the month. A potential explanation for this is that most people receive their wages during that time. The effect of this is reflected in the experimental feature payday. When investigating the feature importance in tables 11 to 16, it can be seen that payday is consistently the most important predictor after the days of the week. In figure A.4 in the appendix, removing the payday feature results in a lower R^2 score which decreased from 0.47 to 0.45.

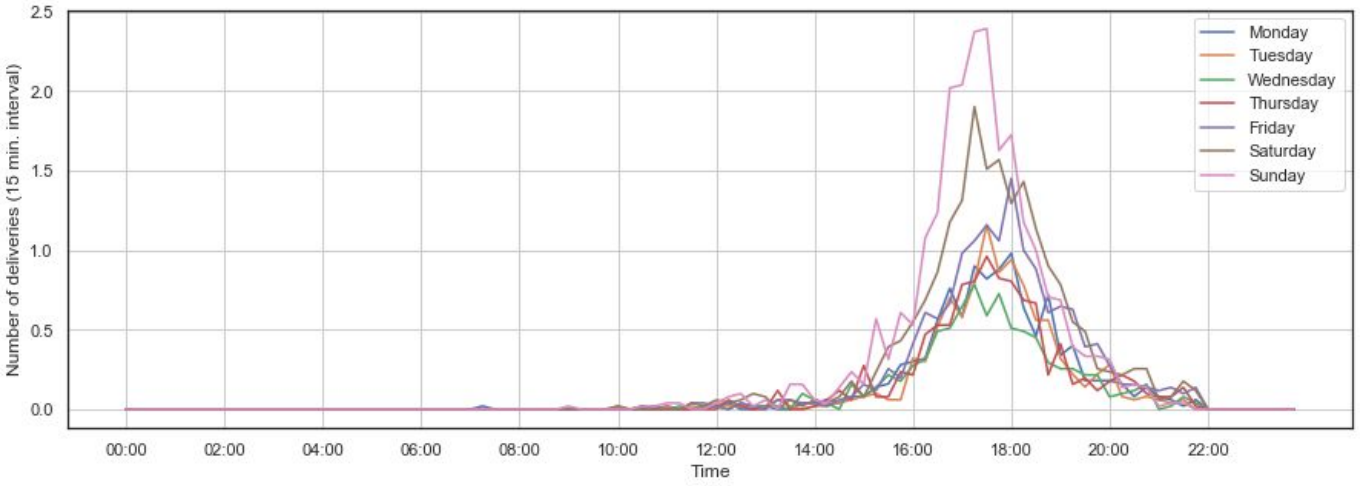

Figure 15: Average time of delivery orders received per 15 min. interval on the days of the week

In figure 15 it can be seen that the majority of customers order food somewhere between 17:00 and 18:00 hours. This comes as no surprise considering this is a common dinnertime for most households. On Friday, there is a second larger peak at around 18:00 hours. This could be caused by the fact people tend to meet for a Friday afternoon drink at the end of the workweek and therefore dine slightly later.

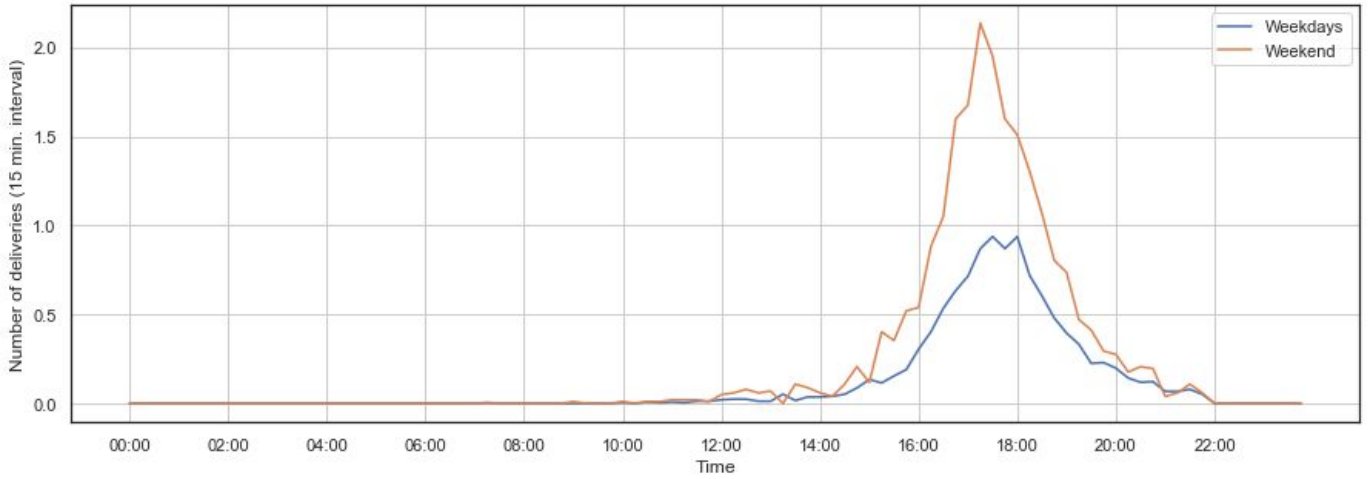

Figure 16: Average number of deliveries per 15 min. interval on the weekend and weekdays

As shown in figure 16, on the weekend the highest peak of deliveries occurs slightly earlier than during the weekdays. A possible reason for this is that people have to get off work first to order food during the week. We can observe that normally the weekdays have roughly slightly less than four orders per hour during the peak. Given the fact that the average maximum productivity is four orders per hour delivered per employee, one deliverer is sufficient on weekdays. Further, we can observe that the influx of orders is more immediate on the weekend, but also decreases much faster after the peak (the slope is steeper). During the busiest time, the peak is slightly more than double the usual number of orders (between 8 and 9) than during weekdays. This implies that the restaurant must be better prepared in advance on the weekend opposed to during the week since many orders have to be handled at once, thus two to three extra deliverers are needed on those days based on the labour productivity.

9.2 Predictive results

In total seven different algorithms have been tested with two different sets of features. One time with the subset of features after feature selection and another time with all features. The results of the experiments are shown in tables 8 and 9. For both sets of features, gradient boosting regression achieved the highest score with a score of 0.853 for the subset and a R^2 score of 0.913 with all features. The R^2 score of lasso and linear regression on the subset are the same, because the alpha parameter was near 0, thus there was little to nearly no regularisation. For most algorithms, there is no significant difference in performance between the subset and set of all features. There is a big discrepancy in the performance of the decision tree depending on the set of features used. The decision tree did not do well with all the features. Further investigation with cross validation was done afterwards (see table 10). It is notable that all models have a similar performance when cross validating with R^2 scores around the 0.50. Another interesting observation is that the R^2 scores for every model is higher on the subset of features than with all features.

	Model	R^2	MSE
1	Gradient boosting regression	0.853	7.506
2	Decision tree regression	0.633	18.743
3	KNN regression	0.569	22.002
4	Linear regression	0.557	22.621
5	Lasso regression	0.557	22.624
6	Ridge regression	0.556	22.690
7	Random forest regression	0.518	24.631

Table 8: Performance of the algorithms with the subset of features

	Model	R^2	MSE
1	Gradient boosting regression	0.913	4.428
2	Linear regression	0.573	21.796
3	Ridge regression	0.563	22.296
4	Lasso regression	0.556	22.659
5	KNN regression	0.541	23.431
6	Random forest regression	0.487	26.183
7	Decision tree regression	0.201	40.799

Table 9: Performance of the algorithms with all features

	Model	R^2 on subset	R^2 with all features
1	Gradient boosting regression	0.553	0.547
2	Linear regression	0.539	0.523
3	Ridge regression	0.535	0.523
4	Lasso regression	0.533	0.515
5	KNN regression	0.530	0.412
6	Random forest regression	0.508	0.488
7	Decision tree regression	0.474	0.445

Table 10: 5-fold cross validation with the best estimator of each model

Below in table 11, 12 and 13 the coefficient of the ridge, lasso and linear regression taken as the feature importance. The features are ranked from most important to least important by sorting them on the absolute value of the coefficients. Since all three algorithms have a similar performance it is not unexpected that the feature importance ranking is more or less the same. For all cases, the weekdays are on the top of the feature importance ranking with Sunday and Wednesday as the most important features consistently. After the days of the week, paydays are considered somewhat substantial by the models. As displayed in table 14, 15 and 16, the least important features from the whole set are the seasons. Sunshine duration, precipitation, school holidays and public holidays have small coefficients below the 0.6 and were not considered very significant either.

Feature	Importance
weekday_Sunday	2.943
weekday_Wednesday	-1.980
weekday_Thursday	-1.553
weekday_Tuesday	-1.465
weekday_Saturday	1.402
weekday_Monday	-1.259
Payday	1.099
Daily wind speed	0.970
Daily temperature	-0.948

Table 11: Feature importance of ridge regression with subset

Feature	Importance
weekday_Sunday	2.982
weekday_Wednesday	-2.140
weekday_Thursday	-1.695
weekday_Tuesday	-1.604
weekday_Monday	-1.389
weekday_Saturday	1.384
Payday	1.165
Daily wind speed	1.009
Daily temperature	-0.983

Table 12: Feature importance of lasso regression with subset

Feature	Importance
weekday_Sunday	2.981
weekday_Wednesday	-2.143
weekday_Thursday	-1.698
weekday_Tuesday	-1.607
weekday_Monday	-1.392
weekday_Saturday	1.383
Payday	1.166
Daily wind speed	1.010
Daily temperature	-0.984

Table 13: Feature importance of linear regression with subset

Feature	Importance
weekday_Sunday	2.885
weekday_Wednesday	-1.872
weekday_Thursday	-1.484
weekday_Saturday	1.382
weekday_Tuesday	-1.368
weekday_Monday	-1.160
Payday	1.010
Daily wind speed	0.802
Daily temperature	-0.784
Sunshine duration	-0.459
School_holidays	0.429
Holidays	0.412
Daily precipitation	0.244
season_summer	0.217
season_spring	-0.152
season_winter	0.038

Table 14: Feature importance of ridge regression with all features

Feature	Importance
weekday_Sunday	3.112
weekday_Wednesday	-1.787
weekday_Saturday	1.489
weekday_Thursday	-1.349
weekday_Tuesday	-1.245
weekday_Monday	-1.016
Payday	0.949
Daily wind speed	0.736
Daily temperature	-0.636
Sunshine duration	-0.363
Holidays	0.262
School_holidays	0.243
Daily precipitation	0.155
season_winter	0.093
season_spring	-0.063
season_summer	0.000

Table 15: Feature importance of lasso regression with all features

Feature	Importance
weekday_Sunday	2.959
weekday_Wednesday	-2.132
weekday_Thursday	-1.721
weekday_Tuesday	-1.584
weekday_Monday	-1.364
weekday_Saturday	1.358
Payday	1.117
Daily temperature	-0.970
Daily wind speed	0.911
School_holiday	0.556
Sunshine duration	-0.505
holidays	0.433
season_summer	0.354
Daily precipitation	0.243
season_spring	-0.141
season_winter	-0.133

Table 16: Feature importance of linear regression with all features

9.2.1 Visualisation linear regression

To gain more insight in the linear regression model, the performance of the model is visualised graphically.

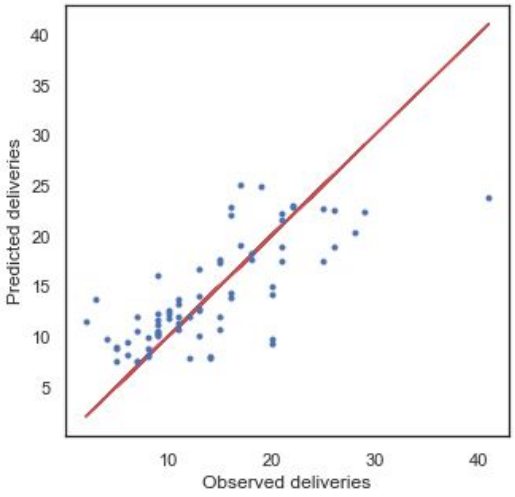

Figure 17: Predicted vs observed deliveries of test set (71 data points)

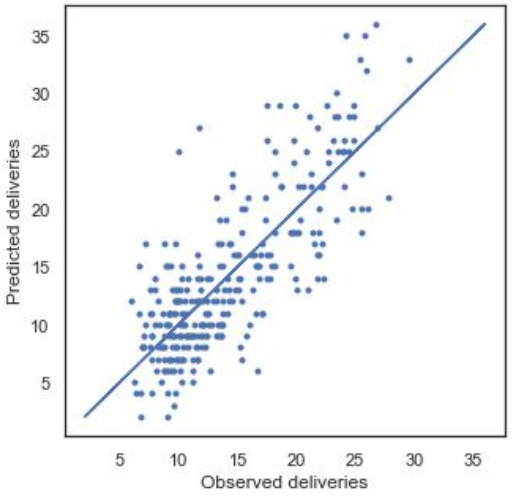

Figure 18: Predicted vs observed deliveries of training set (284 data points)

The predicted deliveries are plotted against the observed deliveries for the test and training set to get a sense of the model's accuracy (see figure 17 and 18). Although there are no strict guidelines in scientific literature regarding the interpretation of R^2 scores, it is generally accepted that a R^2 score between 0.5 and 0.7 is considered to be moderate. Given that the Pearson correlation of the variables and target variable were reasonably weak, it is surprising that the correlation between the prediction and actual deliveries is still moderate with R^2 scores around the 0.5. A few outliers can be spotted on the graph as well with the one of 41 deliveries being the most notable. A more in-depth discussion about the outliers can be found in the discussion in chapter 10.1 and a plot of the residuals can be found in chapter 7.7, where homoscedasticity is discussed.

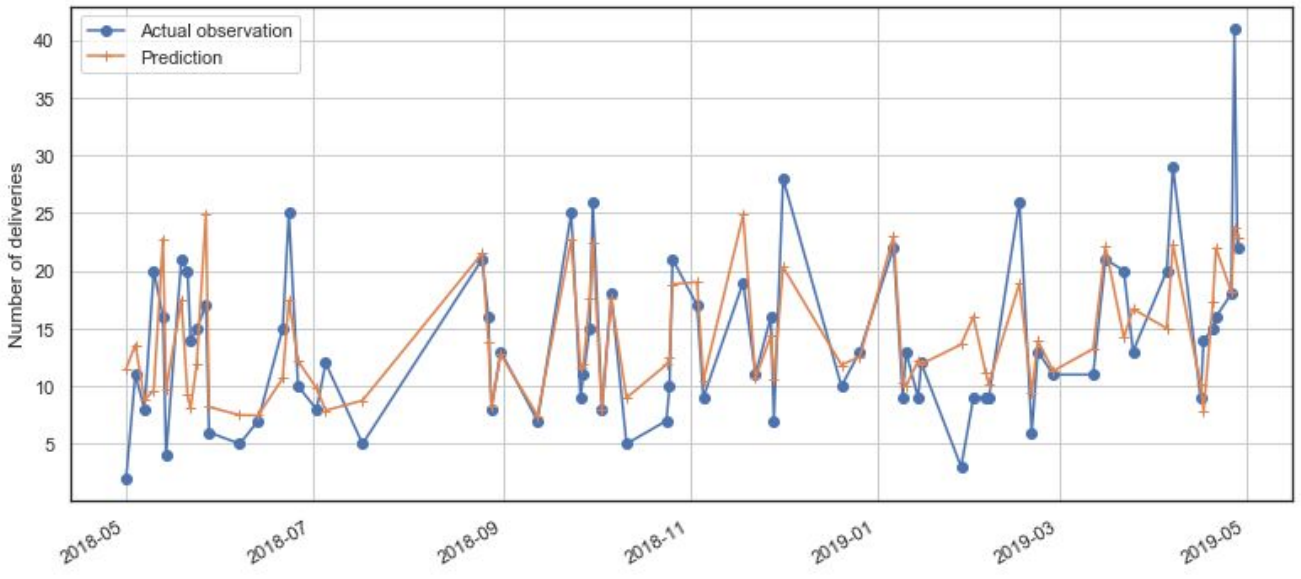

Figure 19: Predictions on test set (blue is observation, orange is prediction)

In the graph above (figure 19), it can be discerned that the linear regression does not work well for when the number of deliveries is low. We can visually approximate that the prediction did not go below 7 despite various actual observations being lower than that. This is because the intercept is set at 14.118 and the negative coefficients are small (see table feature importance).

9.2.2 Visualisation decision tree

The optimal hyperparameters found by 'GridSearchCV' for the decision tree with the subset of features was a depth of 3 and the criterion 'mse'. In order to get a better understanding of how the model predicts the number of deliveries, the decision tree is displayed below (see figure 20). In contrast to the linear models, the decision tree only takes two weekdays, Saturday and Sunday, into consideration. Furthermore, the feature 'daily wind speed' is eliminated.

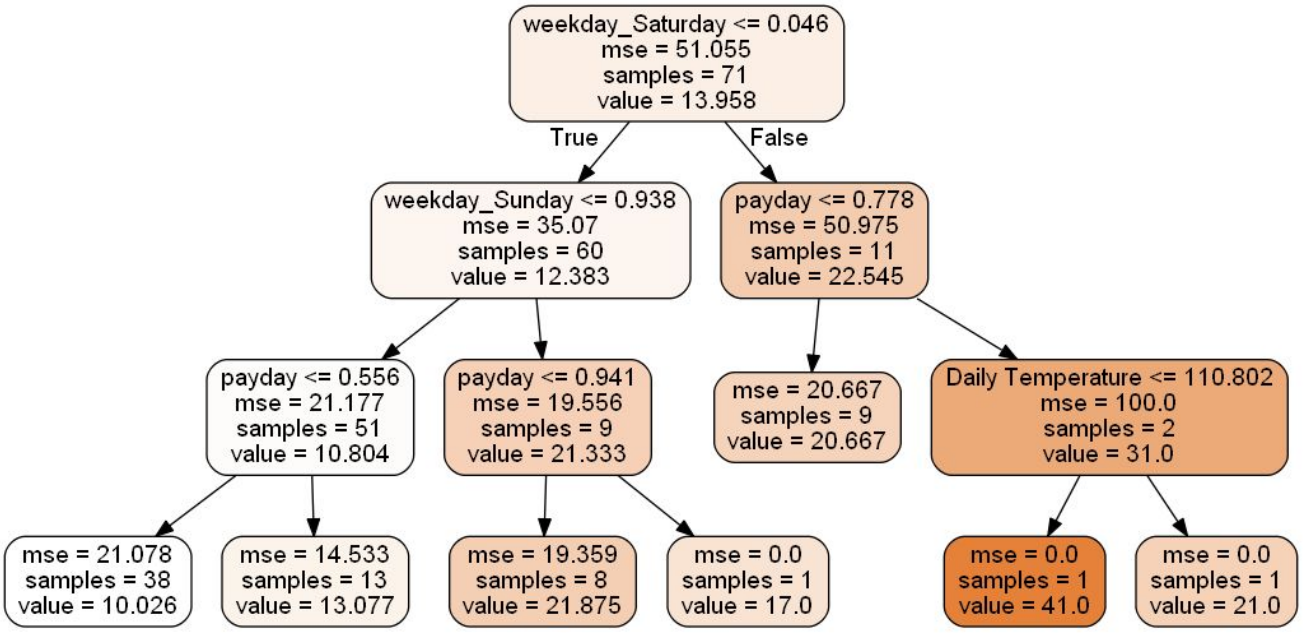

Figure 20: Decision tree with the subset of features

10. Discussion

The results indicate there is moderate predictability for the number of deliveries with the given predictors for the majority of the algorithms. However, given the business context, the usefulness of this result is mediocre which is further discussed in chapter 10.2. The ridge, lasso and linear regression performed similarly as anticipated given all three models capture linearity and the shrinkage factors were low. The performance of these models may be influenced negatively by the fact that the observations may not be completely independent and autocorrelation is present (suggested by the finding at outliers). With the highest R^2 score of 0.913, gradient boosting regression achieved the best performance and the decision tree the worst performance with a R-square score of 0.201 both on the set with all features. However, these data must be interpreted with caution because the R^2 training score of the gradient boosting tree is only 0.389, which is a high bias, and must be treated with suspicion. This could be attributed to mere coincidence; the easiest cases to predict ended up in the test set and the harder cases in the training set. Depending on how the data set is split in a training and test set with the hold-out method, the algorithms may perform differently and achieve other performance results. Although it is unlikely considering the data set is split randomly and the probability is low that this occurs. To circumvent this issue, we attempted to cross validate the best estimators of each model to get a better estimation of the overall performance (see table 10). When using 5-fold cross validation with the best estimator of the gradient boosting tree, the average score is only 0.547 with all features so the previous R^2 score with the hold-out method is debatable. The R^2 score for the subset was slightly higher, a score of 0.553. Surprisingly, the variance during the cross validation was not that high (std. of 0.094 between the R^2 scores) despite the large discrepancy of the R^2 score between the hold-out method and cross validation. Comparing all performances of the cross validation in table 10**,** there is only a small difference between all the estimators. Seeing these results, we cannot conclude that one model is better than the other since the difference in performance is so insignificant (on average a R^2 score of 0.525 on the subset). Also, based on the consistent performance of the KNN and random forest, it seems likely there are some non-linear effects in the data as well.

In figure A.1 in the appendix, it can be seen that the number of deliveries varies monthly. This may suggest seasonality, however, upon further investigation with the BFE it seems that seasonality is not of importance for the prediction model. There may be seasonality present when all deliveries are aggregated per season, however, on a daily basis, the season does not seem to influence the number of deliveries. Alternatively, the differences in the number of deliveries per season are more likely to be just irregularities. It is difficult to discern the actual significance of seasonality on sales, especially given that the data only spans one year in length. There could be other reasons that are particular to that year which might have caused the fluctuations seen in figure 10, such as a specific short-term increase in demand or some special economic events. Next to that, the BFE also eliminated two weather features, namely, daily precipitation and sunshine duration. This is an unexpected result considering precipitation is often thought to be correlated positively with the number of deliveries, because customers are more likely to stay home and order instead of going out. One reason could be that the set of customers that order food and the set of customers that dine in are disjoint from each other. Another potential explanation is that the restaurant has a loyal customer base that orders regularly regardless of the rain or sunshine. In addition to that, school holidays and public holidays were eliminated also by the BFE. This is in accordance with the study

of Holmberg and Hallden (2018) that concluded that the feature holidays did not capture the irregularities in sales and therefore had no effect. The most important features to determine the number of deliveries are the days of the week, in particular Sunday and Wednesday were the most consistent predictors according to the feature importance rankings.

10.1 Outlier detection

In the data set it was found that the number of deliveries usually does not exceed 33-35 orders on a day. A possible explanation for this is that the restaurant owner is known to close the restaurant for deliveries on Thuisbezorgd manually when it is too busy. Any daily deliveries larger than 35 is greater than 3 standard deviations (std). In this case, it could have been that the restaurant manager was too late or forgot to close the restaurant for online ordering timely and too many orders came through. One may argue that it is better to remove these outliers as they are rare occurrences and it would improve the generalisation of the model. Despite that, the outliers have not been removed for the following two reasons:

- These data points are not erroneous and actual valid observations;
- With a total of 355 data points, the effect of removing/keeping the outliers is modest.

The largest outlier found in the data set was dated on 27th of April in 2019, with 41 deliveries. This number may be explained by the fact that it was King's Day, a public holiday in the Netherlands. In the tables below the top outliers and fliers from the boxplot, data points outside the whisker range, are shown and possible explanations for it. Based on these plausible reasons, it seems likely that certain yearly events can lead to a change in the delivery demand. Unfortunately, these annual patterns will be difficult to capture due to the limited data set spanning only over 355 days. The flier on Monday the 28th of January, 2019, suggests that the observations may not be independent and some autocorrelation may be present. On this day, there were only 3 deliveries, an unusual small amount for which we do not have a direct explanation. However, prior to that day, there was a much larger demand for deliveries than usual on three previous days; Friday, 25th of January with 26 deliveries (normally 15 on average), Saturday, 26th of January with 35 deliveries (19.78 on average) and Sunday, 27th of January with 33 deliveries (23.35 on average).

Date	**Nr. of deliveries**	**Possible explanation**
27-04-2019	41	King's Day (public holiday)
24-03-2019	36	Payday
09-12-2018	35	Early payday including 13th month allowance
26-01-2019	35	Payday and annual raise salary

Table 17: Top 4 outliers

Date	**Weekday**	**Nr. of deliveries**	**Possible explanation**
28-01-2019	Monday	3	Autocorrelation, too much spending on previous days, little demand
21-05-2018	Monday	20	Whit Monday (public holiday)

22-04-2019	Monday	25	Easter Monday (public holiday)
20-11-2018	Tuesday	21	Unknown - perhaps competitor(s) closed
01-01-2019	Tuesday	27	New Year's Day (public holiday)
27-04-2019	Saturday	41	King's Day (public holiday)

Table 18: Fliers from the boxplot on page 28

10.2 Managerial implications

In the last couple of years, in-house restaurant sales have declined considerably (see appendix B.1). While this decline cannot entirely be attributed to an increase in deliveries due to lack of data, it may signal a changing trend in the market. In the two graphs below it can be seen that in the period May 2018 to May 2019, the sales from deliveries have increased overall. Interestingly, by the end of period, deliveries were the majority of the sales. If the trend continues, the role of delivery will become increasingly important for restaurants. Delivery prediction can help the restaurant manager to estimate the correct number of delivery employees needed. A BPMN model of the usual delivery process can be found in appendix B.2.

In figure 19, the business process of hiring extra deliverers is illustrated. The prediction model can be used in the hiring process of on-call deliverers which usually takes place maximal one week up to one day in advance of the workday. Since the prediction model is dependent on weather forecasts and its predictability decreases with the increase of the prediction time horizon, it is most optimal to use the model one day in advance (the minimal required notice). This way, the weather forecasts would be the most accurate while allowing the management to have the opportunity to hire more delivery employees when needed. Subsequent to the prediction, the restaurant can determine what the demand for deliveries will be the following day and hire an appropriate number of employees to carry them out. Thereby, this is one of the more direct ways in which the restaurant can minimise costs and maximise revenue by employing the exact amount of delivery personnel that is required.

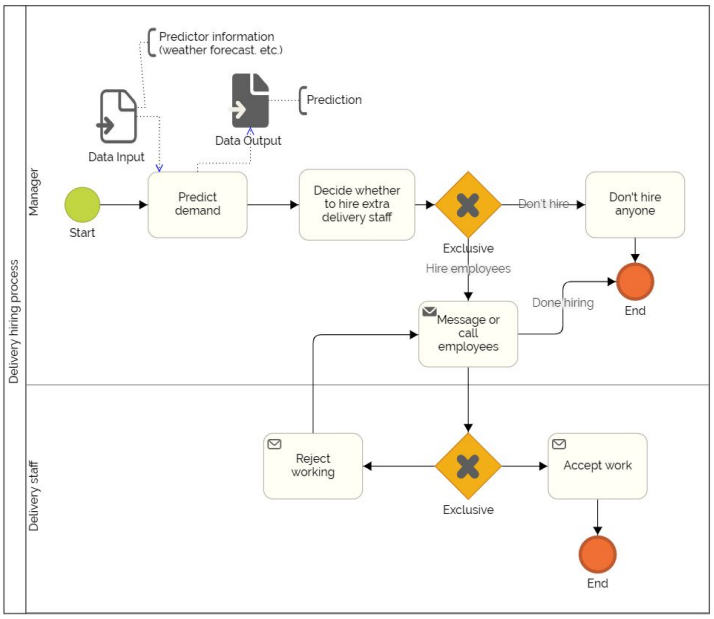

Figure 21: BPMN of the delivery hiring process

10.2.1 Cost-benefit analysis

From the managerial perspective, this study does not seem to offer any meaningful insights which managers can use to improve their business processes in a manner that would lead to significant increases in revenue, customer satisfaction or cost-savings for the restaurant. The results of this paper indicate that the use of machine learning for the purpose of forecasting future deliveries currently does not offer much business value to micro-sized restaurants. The MSE of the models were more or less around 20 on average. Taking the square root of this, this would roughly imply that the prediction of the number of deliveries would be 4.5 off on average. For the restaurant on which the study was performed 4.5 is a considerable error in comparison to the average of 14 daily deliveries, because the error amounts to more than 30% of the average daily deliveries. Therefore, the prediction is not reliable enough and we can conclude that most likely, it is scarcely beneficial for restaurants who only use Thuisbezorgd delivery data and weather data to try to predict the delivery demand. The usability of the results in an actual business/management setting is limited at best and may also be too costly to apply, given that restaurants require a data specialist to analyse their data. A further problem is that often restaurants gather their data in a primitive way that does not facilitate its usage for data mining to draw insights from it. For it to be usable, the data must be transformed which is costly both timewise and financially.

If we assume that the prediction model is hypothetically accurate, then it may have benefits in the delivery process in terms of cost saving and extra revenue. Although this is hard to quantify without the financial data of the restaurant. Further, it also depends on many factors of a restaurant's business model and is therefore different for each restaurant. Nonetheless, for these cost calculations in this paragraph we will give some rough estimates for the oriental restaurant with the estimates given by the restaurateur. Based on these estimates, we explore one method to optimise the delivery process of the restaurant. In accordance with the figures (see appendix C.3), it always pays off to hire one extra deliverer to do one extra delivery because the operating profit per delivery is positive (operating income is 11.32 euros for the average delivery and the cost of hiring one delivery employee is 7.50 euros, which is a minimal profit of 3.82 euros). From a microeconomic perspective, hiring more workers is subject to the law of diminishing returns due to limited capital such as kitchen equipment. One deliverer is able to fulfill on average 4 deliveries per hour so the maximum output would be 4 per unit of labour. If the marginal product of labour, the number of deliveries that one extra deliverer can realise, would be maximised, this would yield 37.78 euros operating income per hour (11.32 * 4 - 7,50). Furthermore, presuming that the prediction is made one day in advance, only the number of delivery staff is flexible and can be adjusted. Because of this, the maximum amount of orders that the restaurant can handle on a day is predetermined (e.g. it is not possible to hire one extra cook or procure another delivery vehicle one day in advance) and the restaurant will close for orders once its maximum capacity is reached. Restaurants can save costs with the prediction model with the knowledge that allows them to determine precisely how many delivery staff they will need the following day. In this case, each staff not hired saves the restaurant 7.50 euros per hour. Next to that, maximum potential revenue from delivery can be realised by hiring extra staff to deliver until the maximum capacity of the kitchen is reached. For example, if the expected deliveries require 2 deliverers, the restaurant needs to hire a second deliverer. Even though the output for the second deliverer is not maximised, the restaurant will still generate extra profit no matter what since the operating profit is always positive for any positive MPL. Over time, these profits and cost savings may add up to a considerable amount especially if deliveries will become more popular in the future.

Following the logic from the previous paragraph, extra profits and cost savings on an annual basis have been crudely calculated to give more business context. For these calculations, the observations of the delivery data have been used. Based on figure 15 and 16, we assume that the number of orders during peak hour is roughly 35% of the total number of deliveries. With this number, we can determine the exact number of employees needed to fulfill all the orders during peak hour which is the number of employees that the manager will hire to not miss any revenue. Normally this specific restaurant hires 1 deliverer on Monday till Thursday, 2 deliverers on Friday and Saturday, and 3 deliverers on Sunday. We then compare the difference in labour output of how many deliverers the restaurant normally would have hired against what they should have hired if they hypothetically could predict the demand accurately. With this difference the potential profit or cost saving is then calculated for every day in the data set. Example calculations are shown in appendix C.4. The estimated sum of all the hypothetical savings and extra profit add up to 1010.17 euros for 355 days, the number of days in the delivery data set. On a yearly basis, the restaurant would earn/save slightly more than one thousand euros if they were to be able to predict the number of deliveries per day. Comparing this to the annual turnover of thirty five thousand euros, this is a minor profit increase of 2.9%. Overall, the benefit of trying to maximise delivery efficiency through prediction is minimal for similar smaller restaurants.

We have found that restaurants can get better value by simply observing the historical trends in their data, without needing to do any data mining, which gives limited benefits. It is important to observe general trends in order to better understand their position in the market and external environment. In this manner, restaurants can notice if any changes occur to their sales and try to pinpoint what may be the cause of this (e.g. higher sales because of an increase in wages, improving mobile delivery app, successful marketing campaign or lower sales because recession, drop in food quality, longer waiting times and cold food on delivery). This seems to be a more reliable and cost-efficient method as opposed to using machine learning based on data collected by Thuisbezorgd. Because Thuisbezorgd is a large and popular platform that is used by many restaurants in the Netherlands and throughout other countries (under different brand names owned by the dot-com company Takeaway.com N.V.), the generalisability of the approach to analyse delivery data that originates from the platform is high and should be applicable to other micro-sized restaurants.

11. Limitations

The analysis of this paper has several limitations that are listed below. No attempt has been made to actually implement the prediction model and put them into practice.

11.1 Sample size

- **Small sample size:** We only used data from one restaurant in one region with a small variety of Asian cuisines. The location, cuisine and the menu selection of this restaurant will certainly have a specific influence on the sales. This limits the potential generalisability of the analysis to apply the results to other SME restaurants, especially if the aforementioned aspects differ greatly from the oriental restaurant that we have taken as a sample.

11.2 Data set

Due to the nature of data collection, real-world data tends to be ‘dirty’, which means that they often contain erroneous values or noise. Therefore, data cleaning is required to ensure the quality and soundness of the data. Depending on how thorough the data preprocessing has been, the quality of the data set may vary.

- **Delivery data set:** The data set obtained from Thuisbezorgd for delivery is considered far from ideal for machine learning due to the small amount of available predictors and its relatively small size (only roughly 5000 instances and 355 if transformed per day). For that reason, the selected features in the data analysis may not be optimal for its intended purpose. Consequently, the machine learning models’ performance may be hindered since they are highly dependent on the quality of the data input. In addition, yearly patterns and long-term trends cannot be analysed since the data spans less than one year only. It is also important to realise that a small number of customers do not order via Thuisbezorgd, but via an alternative method such as phone call or email. Unfortunately, it was not possible to discern the sales records that were deliveries since deliveries were not explicitly categorised in the POS system. As a result, the number of deliveries may be slightly understated on some days and in-house sales may slightly be overstated.
- **Hyperparameters**: No attempt has been made to find the complete optimal set of hyperparameters for each machine learning model for this problem. Only a limited number of hyperparameters have been tuned to allow each machine learning model to have a chance to perform well and to get a reasonable assessment.

11.3 Weather data set

- **Unhomogenised data:** The Royal Netherlands Meteorological Institute (KNMI Datacentrum) stated that the acquired meteorological data set is unhomogenised. This may imply that the weather metrics in the data set are not completely accurate (e.g. due to relocation of the weather station, temperature measurements under different air pressure, etc.).
- **Prediction with weather forecasts**: The final model uses the weather features ‘daily wind speed’ and ‘daily temperature’ to predict the number of deliveries. To train the model, the actual measurements of the weather variables have been taken from the KNMI. However, for the prediction, a restaurant manager would only have the weather forecast information available and not the actual weather on the day itself. Considering there will be some discrepancy between the weather forecast and the actual weather, the performance of the models are not completely accurate and overstated. Nevertheless, the difference between the forecast and actual weather should be modest, because weather forecasts are generally known to be very reliable. Noodweer Benelux, a Belgian organisation specialised in extreme weather, stated a 98% reliability for the European Centre for Medium-Range Weather Forecasts (ECMWF) models for 3-days forecasts in 2016 [57]. The National Oceanic and Atmospheric Administration in the United States, approximated that a 5-day forecast would be around 90% accurate [58]. Next to that, the coefficients of the weather features in the prediction model are relatively small so we believe the influence of the weather is limited regardless.

12. Conclusions & future work

12.1 Conclusion

The main research question for this project was: “How can data mining and analysis of delivery data contribute to improving the business processes of micro-sized restaurants in the SME sector?”. To answer the main research question, another set of underlying research questions were introduced:

1. ***To what extent can micro-sized restaurants predict the number of deliveries per day?***
   The highest score achieved by gradient boosting regression is 0.913, however, the stability of this model is highly debatable. When tested with cross validation, the gradient boosting regression is not significantly better than the other models and achieves a worse performance than with the hold-out method. Thus, in reality, the actual score might be closer to 0.553. As discussed in the limitations, we must also take into account that the model used actual weather metrics to predict the number of deliveries. When deployed, only weather forecasts will be available since the actual weather on the day itself is unknown. This may slightly decrease the performance of the model due to the minor inaccuracies of the weather forecasts. Nonetheless, these influences should be modest for the given reasons in the limitations. So with a R^2 score of around 0.553, we can assume there is moderate predictability of the number of deliveries.
2. ***What are the most important features for the delivery prediction?***
   The most important features for delivery prediction given the available predictors in the data set are the days of the week. Sunday, Wednesday and Thursday are frequently the three largest coefficients, followed by the payday variable. After that, daily wind speed and daily temperature are roughly considered equally important.
1. ***What impact does the weather have on the number of deliveries?***
   Surprisingly, weather features daily precipitation and sunshine duration are not considered significant at all for linear regression. Daily wind speed and daily temperature have a minor influence on the number of deliveries according to the feature importance. Overall, we can conclude that the weather features do not have much impact on the prediction.
3. ***In what step of the delivery process can the prediction model be used and how can it help to improve the delivery process?***
   According to these data, we can infer that prediction for the number of deliveries should be done one day in advance for the most optimal predictability. With the assistance of an accurate delivery prediction model, the restaurant can obtain maximal revenue (by fulfilling all demand) and minimise costs (by avoiding standby workers). Additionally, if the number of expected deliveries is known, the opportunity cost (losing out on potential revenue) of not employing delivery personnel may be avoided.

In spite of its limitations, this study adds to our understanding of the usefulness of data mining in the delivery process and lays the groundwork for future research into micro restaurant firms and delivery prediction. The findings of this study on delivery prediction are similar to what is known from other papers about restaurant sales prediction. These results reflect those of Holmberg and Hallden (2018) who also found that the day of the week is the best feature for prediction of sales and that the weather features only improved the model slightly. However, the finding is contrary to

the previous study of Bujisic, Bogicevic and Parsa who found that temperature was the most important weather feature for restaurant sales. This discrepancy could be attributed to the fact that the restaurant from Bujisic et al. was located in a hotel complex and consequently their results are not as generalisable.

12.2 Future work

For future work, we make a few recommendations. A natural progression of this work is to include data sets of multiple micro restaurant firms with a larger period spanning over more than one year to capture annual patterns. The feature importance ranking and correlation has shown that there is a lack of good predictors and therefore it would be advisable to explore different predictors in further research. Considerably more work will need to be done to determine appropriate predictors for delivery prediction to reach a satisfactory accuracy. Another question that remains unanswered is whether there is an autocorrelation between the number of deliveries. Further investigation and experimentation into time-series models is required to determine whether deliveries can be forecasted based on the previous values if autocorrelation is present.

12.3 Outlook

As technology continues to advance and becomes more integrated into the provision of restaurant services, it is vital for restaurants to try to capitalise on the potential benefits. A good practice that restaurant managers who seek to improve the business processes of their restaurants can take advantage of is the adoption of better data storage and management techniques. This can serve them well for the future, as the use of data analytics continues to grow. On the same note, restaurants stand to gain tremendously from the emergence of IoT technology. It is expected that many restaurants will adopt this technology as it should improve customer service. Once integrated, this new concept will make the gathering of data far more convenient. This new data can show, for example, how fast deliveries are being done, customer satisfaction and retention rate and better cost overview. It is clear that restaurants need to prepare to deal with the disruptions that will occur in the market. Currently, a declining trend of in-house restaurant sales is noticeable, while delivery has already grown to 44% of total yearly sales. As third-party delivery platforms grow larger, they may charge bigger commissions, and this may be a financial threat for restaurants because the profit margin on deliveries is already low [59]. Micro-sized restaurants are small and have no bargaining power, thus their survival is more uncertain. Furthermore, it is fairly unclear what the future of the industry will be like and how much of a role delivery will play over a decade or more. But in the meantime, restaurant owners need to find ways to cut costs and stay relevant in the market considering that consumer preferences are shifting. Such reasons amplify the necessity of a better incorporation of data utilization for operational and strategic decision making, which data mining can support. In conclusion, even though the gains may be marginal, micro restaurant firms must find superior ways to incorporate data mining into their delivery process if they wish to thrive in this highly technological and changing market.

Bibliography

[1] Koninklijke Horeca Nederland. (2019, October 31). Economisch belang horeca groeit met 60 procent in 10 jaar. Retrieved from https://www.khn.nl/nieuws/economisch-belang-horeca-groeit-met-60-procent-in-10-jaar

[2] Koninklijke Horeca Nederland. (2020, January 13). Groei horeca vlakt af in 2019. Wet Arbeidsmarkt in Balans grootste bedreiging 2020. Retrieved from https://www.khn.nl/nieuws/groei-horeca-vlakt-af-in-2019-wet-arbeidsmarkt-in-balans-grootste-bedreiging-2020

[3] Koninklijke Horeca Nederland. (n.d.). Koninklijke Horeca Nederland in English. Retrieved from https://aanmelden.khn.nl/about/koninklijke-horeca-nederland-in-english

[4] Geijer, Thijs. “Groei Horeca Loopt Tegen Grenzen Aan.” *ING Website*, ING, 28 Feb. 2019, www.ing.nl/zakelijk/kennis-over-de-economie/uw-sector/outlook/horeca.html.

[5] Van der Reijden, Demian. “Personeelstekorten Remmen Groei Horeca Komende Jaren Af.” *Misset Horeca*, Vakmedianet, 14 Dec. 2018, www.missethoreca.nl/restaurant/nieuws/2018/12/personeelstekorten-remmen-groei-horeca-komende-jaren-af-101313563.

[6] Misset Horeca. (2019, November 29). Groei thuisbezorging eten en drinken zet zich onverminderd voort. Retrieved June 21, 2020, from https://www.missethoreca.nl/horeca/nieuws/2018/11/groei-maaltijdbezorging-zet-zich-onverminderd-voort-101312847

[7] Muller, C. (2018). Restaurant Delivery: Are the "ODP" the Industry's "OTA"? Part II. *Boston Hospitality Review, 6*.

[8] Wijngaarde, Y., & De Miguel, S. (2017, March). *Food Delivery Tech: Battle for the European Consumer* [PPT]. Amsterdam: Dealroom.co.

[9] ANP Producties. (2020, February 06). Meer groei voor Uber, maar nog wel verlies. Retrieved June 22, 2020, from https://www.telegraaf.nl/financieel/1070863437/meer-groei-voor-uber-maar-nog-wel-verlies

[10] Van Asselt, D. (2019, March 14). Thuisbezorgd.nl: 4 miljoen gebruikers, 33 miljoen bestellingen in 2018. Retrieved June 21, 2020, from https://www.snackkoerier.nl/bedrijfsvoering/nieuws/2019/03/thuisbezorgd-nl-4-miljoen-gebruikers-33-miljoen-bestellingen-in-2018-101300820?_ga=2.166090043.461625302.1589703021-553172251.1581773496

[11] Stil, H. (2020, January 14). Thuisbezorgd leverde in 2019 38 miljoen maaltijden aan de deur. Retrieved June 21, 2020, from

https://www.parool.nl/nederland/thuisbezorgd-leverde-in-2019-38-miljoen-maaltijden-aan-de-deur~b530c87d1/?referer=https%3A%2F%2Fwww.google.com%2F

[12] Khan, M. A. (2020). Technological Disruptions in Restaurant Services: Impact of Innovations and Delivery Services. Journal of Hospitality & Tourism Research, 44(5), 715–732. https://doi.org/10.1177/1096348020908636

[13] Griswold, A. (2019, August 23). Analysts expect Uber Eats to lose money on every order for at least the next five years. Retrieved from https://qz.com/1693843/uber-eats-will-lose-money-until-at-least-2024-say-cowen-analysts/

[14] Kim, S. Y., & Upneja, A. (2014). *Predicting restaurant financial distress using decision tree and AdaBoosted decision tree models. Economic Modelling, 36, 354–362.* doi:10.1016/j.econmod.2013.10.005

[15] OECD. (2018). *Strengthening SMEs and entrepreneurship for productivity and inclusive growth: Key issues paper*. OECD Ministerial Conference, Mexico City, 22-23 February 2018, page 13. Retrieved from OECD: https://www.oecd.org/cfe/smes/ministerial/documents/2018-SME-Ministerial-Conference-Key-Issues.pdf

[16] Quinn, B., McKitterick, L., McAdam, R., & Dunn, A. (2014). *Introduction. The International Journal of Entrepreneurship and Innovation, 15(3), 143–145.* doi:10.5367/ijei.2014.0158

[17] Bolden, D., Martin, M., Luther, A., & Hadlock, P. (2018, November 9). *Feeding the Algorithm: How Restaurants Use Data to Capture Competitive Advantage*. https://www.bcg.com/en-nl/publications/2018/feeding-algorithm-restaurants-use-data-capture-competitive-advantage.aspx.

[18] MKB-Nederland. (2020, March 25). Informatie over het mkb (midden- en kleinbedrijf) in Nederland: Mkb cijfers, definities en organisaties belangrijk voor marktonderzoek. Retrieved from https://www.mkbservicedesk.nl/569/informatie-over-midden-kleinbedrijf-nederland.htm

[19] European Commission Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs. (2018). *2018 SBA Fact Sheet - Netherlands.* Retrieved from https://ec.europa.eu/docsroom/documents/38662/attachments/21/translations/en/renditions/native

[20] OECD Glossary of Statistical Terms. (2005, December 2). Small and Medium-sized Enterprises (SMEs). Retrieved April 12, 2020, from https://stats.oecd.org/glossary/detail.asp?ID=3123

[21] Netherlands Enterprise Agency, RVO. (n.d.). What is an SME? Retrieved April 8, 2020, from https://business.gov.nl/starting-your-business/first-steps-for-setting-up-your-business/what-is-an-sme/

[22] Unnisabegum, Ahmed & Hussain, Mohammed & Shaik, Mubeena. (2019). Data Mining Techniques For Big Data, Vol. 6, Special Issue ,. 10.13140/RG.2.2.25408.07686.

[23] Tutorialspoint. (n.d.). Data Mining - Knowledge Discovery. Retrieved April 14, 2020, from https://www.tutorialspoint.com/data_mining/dm_knowledge_discovery.htm

[24] Techopedia. (2017, August 18). What is Knowledge Discovery in Databases (KDD)? - Definition from Techopedia. Retrieved April 14, 2020, from https://www.techopedia.com/definition/25827/knowledge-discovery-in-databases-kdd

[25] Bharati, M. & Ramageri, Bharati. (2010). Data mining techniques and applications. Indian Journal of Computer Science and Engineering. 1.

[26] DeepAI. (2019, May 17). Machine Learning. Retrieved April 14, 2020, from https://deepai.org/machine-learning-glossary-and-terms/machine-learning

[27] Heiler, L. (2017, March 20). Difference of Data Science, Machine Learning and Data Mining. Retrieved from https://www.datasciencecentral.com/profiles/blogs/difference-of-data-science-machine-learning-and-data-mining

[28] Expert System Team. (2017, March 7). What is Machine Learning? A definition. Retrieved from https://expertsystem.com/machine-learning-definition/

[29] Holmberg, M., & Halldén, P. (2018). Abstract Machine Learning for Restaurant Sales Forecast. Retrieved June 06, 2020 from http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-353225

[30] Tanizaki, T., Hoshino, T., Shimmura, T., & Takenaka, T. (2019). Demand forecasting in restaurants using machine learning and statistical analysis. In Procedia CIRP (Vol. 79, pp. 679–683). Elsevier B.V. https://doi.org/10.1016/j.procir.2019.02.042

[31] Khan, M. A. (2020). Technological Disruptions in Restaurant Services: Impact of Innovations and Delivery Services. Journal of Hospitality & Tourism Research, 44(5), 715–732. https://doi.org/10.1177/1096348020908636

[32] Bujisic, Milos & Bogicevic, Vanja & Parsa, H.. (2016). The effect of weather factors on restaurant sales. Journal of Foodservice Business Research. 20. 1-21. 10.1080/15378020.2016.1209723.

[33] European Commission. (n.d.). SME competitiveness. Retrieved May 08, 2020, from https://ec.europa.eu/regional_policy/en/policy/themes/sme-competitiveness/

[34] European Commission. (2017, June 28). Entrepreneurship and Small and medium-sized enterprises (SMEs). Retrieved May 08, 2020, from https://ec.europa.eu/growth/smes_en

[35] European Commission. (2017, August 30). Food and drink industry. Retrieved May 08, 2020, from https://ec.europa.eu/growth/sectors/food_en

[36] KNMI. (n.d.). Wij zijn het KNMI. Retrieved June 21, 2020, from https://www.knmi.nl/over-het-knmi/over

[37] Zhang, Shichao & Zhang, Chengqi & Yang, Qiang. (2003). Data Preparation for Data Mining.. Applied Artificial Intelligence. 17. 375-381. 10.1080/713827180.

[38] Ilyas, I.F. (2016). Effective Data Cleaning with Continuous Evaluation. IEEE Data Eng. Bull., 39, 38-46.

[39] Techopedia. (2017, January 4). What is Garbage In, Garbage Out (GIGO)? - Definition from Techopedia. Retrieved May 4, 2020, from https://www.techopedia.com/definition/3801/garbage-in-garbage-out-gigo

[40] Wickham, H. (2014). Tidy Data. *Journal of Statistical Software, 59*(10), 1 - 23. doi:http://dx.doi.org/10.18637/jss.v059.i10

[41] Srinidhi, S. (2018, July 30). Label Encoder vs. One Hot Encoder in Machine Learning. Retrieved from https://medium.com/@contactsunny/label-encoder-vs-one-hot-encoder-in-machine-learning-3fc273365621

[42] Brownlee, J. (2020, April 27). Why One-Hot Encode Data in Machine Learning? Retrieved May 3, 2020, from https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/

[43] Ambielli, B. (2018, February 11). When to Use One Hot Encoding. Retrieved May 3, 2020, from https://bambielli.com/til/2018-02-11-one-hot-encoding/

[44] DeepAI. (2019, May 17). Curse of Dimensionality. Retrieved May 3, 2020, from https://deepai.org/machine-learning-glossary-and-terms/curse-of-dimensionality

[45] Al-Janabi, Samaher & al-bakry, Abbas. (2010). Genetic Programming Data Construction Method to Handle Data Scarcity Problem. International Journal of Advancements in Computing Technology (IJACT)..

[46] Lakshmanan, S. (2019, May 17). How, When and Why Should You Normalize/Standardize/Rescale Your Data? Retrieved May 4, 2020, from https://medium.com/@swethalakshmanan14/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff

[47] Asaithambi, S. (2017, December 4). Why, How and When to Scale your Features. Retrieved May 4, 2020, from https://medium.com/greyatom/why-how-and-when-to-scale-your-features-4b30ab09db5e

[48] Wikipedia. (2020, January 31). Feature scaling. Retrieved June 03, 2020, from https://en.wikipedia.org/wiki/Feature_scaling

[49] Allison, P. (2012, September 10). When Can You Safely Ignore Multicollinearity? Retrieved June 05, 2020, from https://statisticalhorizons.com/multicollinearity

[50] Mahto, K. (2019, July 20). One-Hot-Encoding, Multicollinearity and the Dummy Variable Trap. Retrieved June 05, 2020, from https://towardsdatascience.com/one-hot-encoding-multicollinearity-and-the-dummy-variable-trap-b5840be3c41a

[51] Wikipedia. (2020, May 29). Variance inflation factor. Retrieved June 03, 2020, from https://en.wikipedia.org/wiki/Variance_inflation_factor

[52] Rogerson, P. A. (2001). Statistical methods for geography. London: Sage

[53] Wikipedia. (2020, May 10). Feature selection. Retrieved June 01, 2020, from https://en.wikipedia.org/wiki/Feature_selection

[54] Shaikh, R. (2018, October 28). Feature Selection Techniques in Machine Learning with Python. Retrieved June 01, 2020, from https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e

[55] Agarwal, R. (2019, July 27). The 5 Feature Selection Algorithms every Data Scientist should know. Retrieved June 01, 2020, from https://towardsdatascience.com/the-5-feature-selection-algorithms-every-data-scientist-need-to-know-3a6b566efd2

[56] Paul, S. (2020, January 02). Beginner's Guide to Feature Selection in Python. Retrieved June 01, 2020, from https://www.datacamp.com/community/tutorials/feature-selection-python

[57] Noodweer Benelux. (2016, April 06). Betrouwbaarheid van weersvoorspelling. Retrieved June 22, 2020, from https://www.noodweer.be/betrouwbaarheid-weersvoorspelling-10-dagen/

[58] SciJinks. (n.d.). How Reliable Are Weather Forecasts? Retrieved June 22, 2020, from https://scijinks.gov/forecast-reliability/

[59] Maze, J. (2019, February 08). As delivery grows, debate rages over its profitability. Retrieved from https://www.restaurantbusinessonline.com/financing/delivery-grows-debate-rages-over-its-profitability

Appendices

Appendix A: Data transformation

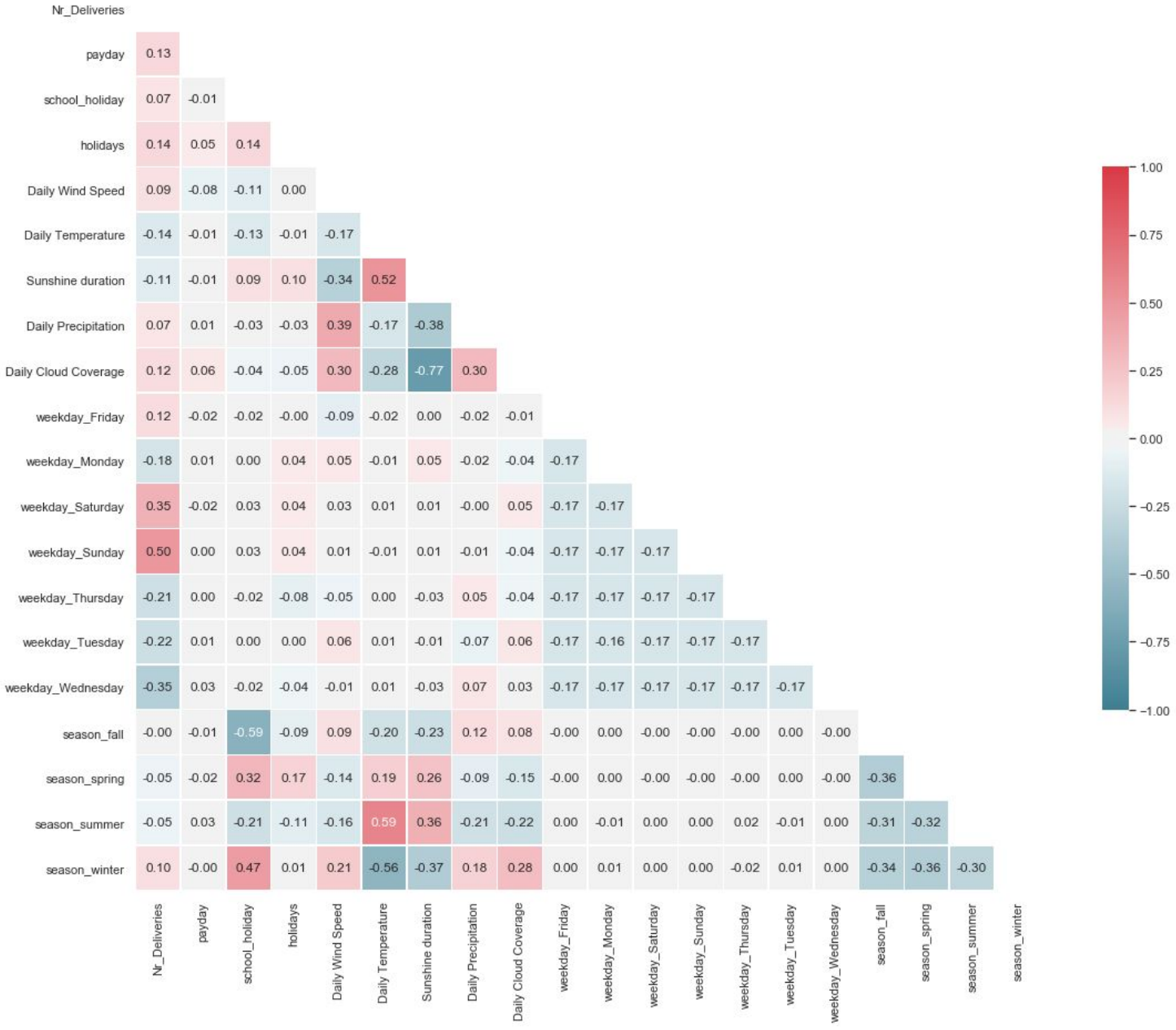

Figure A.1: Spearman correlation matrix of the features

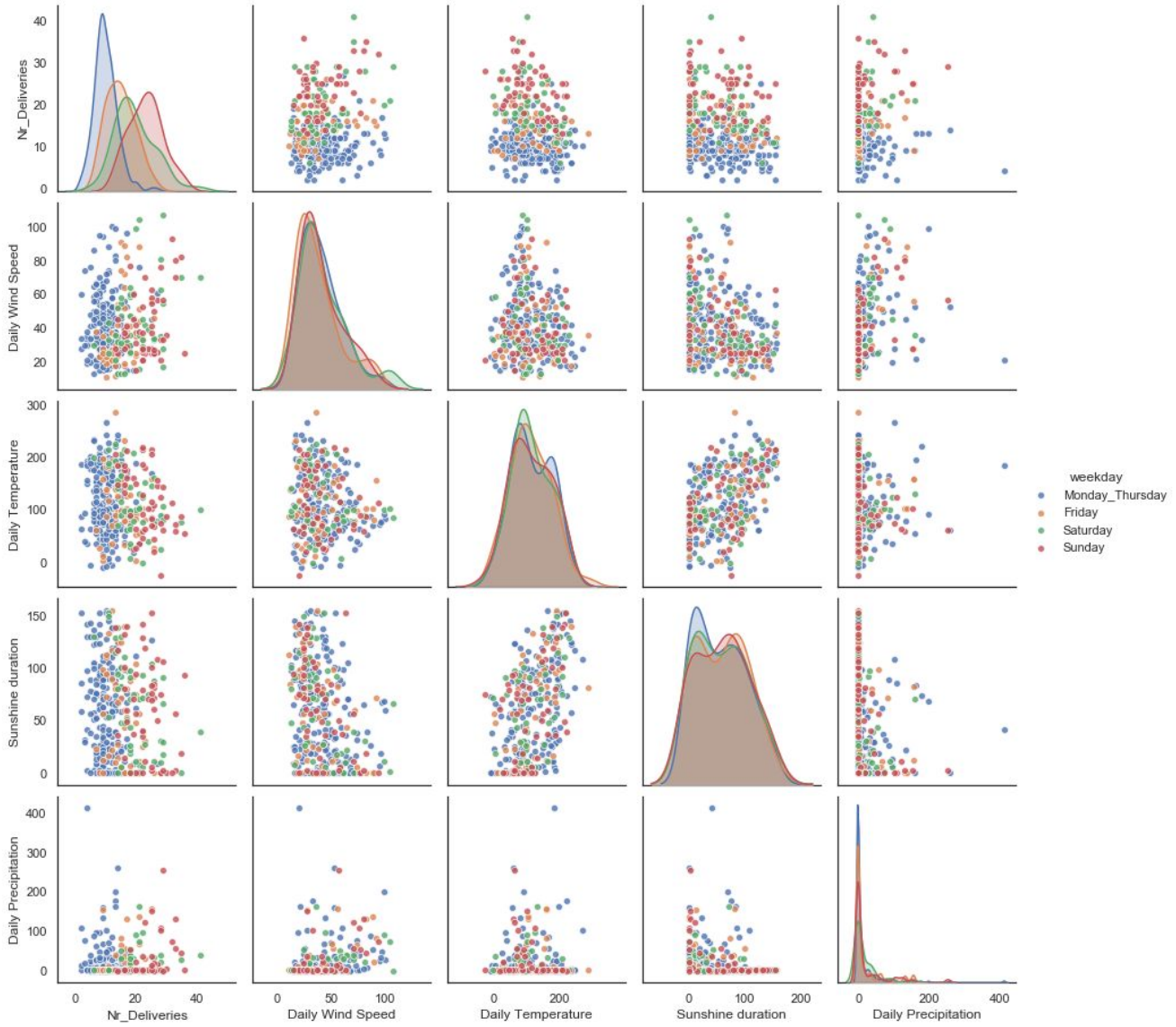

Figure A.2: pairplot of quantitative variables by weekday

```
# Manual backward feature selection based on feature importance
Copy columns of dataset to cols
While length of cols are larger than 0
      Cross validate with cols and linear regression
      Print scores and coefficients
      Append scores (R^2 and std) to list
      Determine column with smallest coefficient
      Remove this column
Plot scores
```

Figure A.3: pseudocode of manual backward feature elimination based on feature importance

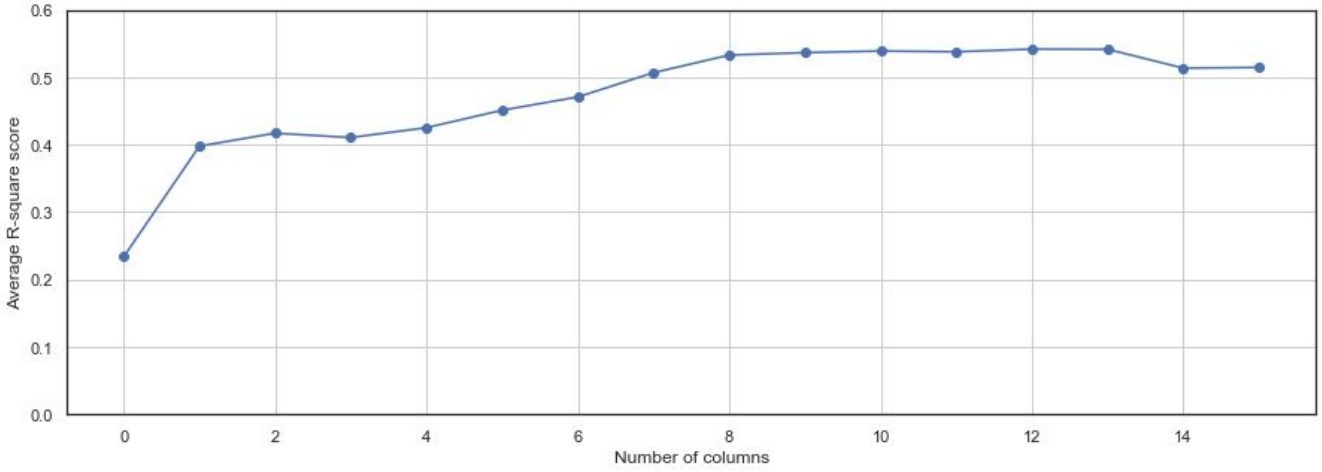

Figure A.4: Average R^2 score of the cross validation against the number of columns

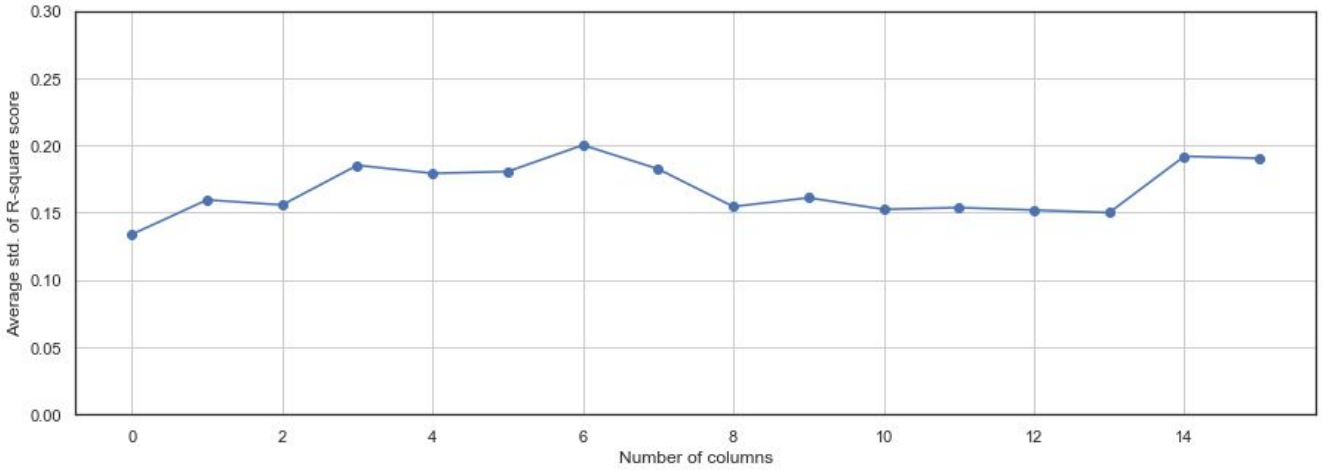

Figure A.5: Average std. of R^2 score of the cross validation against the number of columns

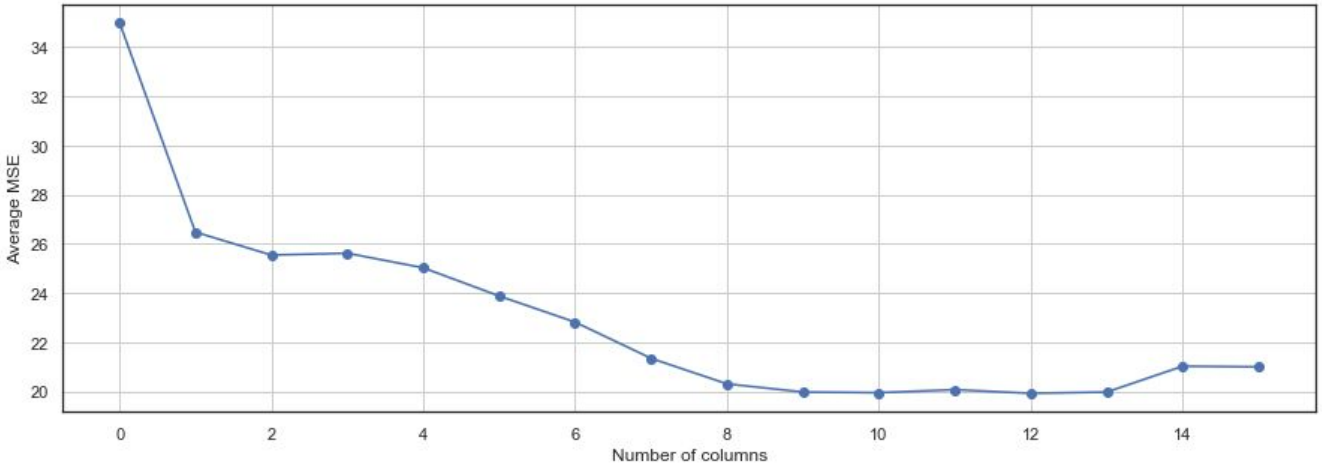

Figure A.6: Average MSE of the cross validation against the number of columns

Order	Features
1	Daily precipitation
2	season_spring
3	season_winter
4	School_holidays
5	season_summer
6	Sunshine duration
7	Holidays
8	Daily temperature
9	Daily wind speed
10	Payday
11	Weekday_Monday
12	Weekday_Tuesday
13	Weekday_Thursday
14	Weekday_Wednesay
15	Weekday_Saturday
16	Weekday_Sunday
17	Daily temperature

Table A.7: Order of removal of columns in manual BFE

Appendix B: Results

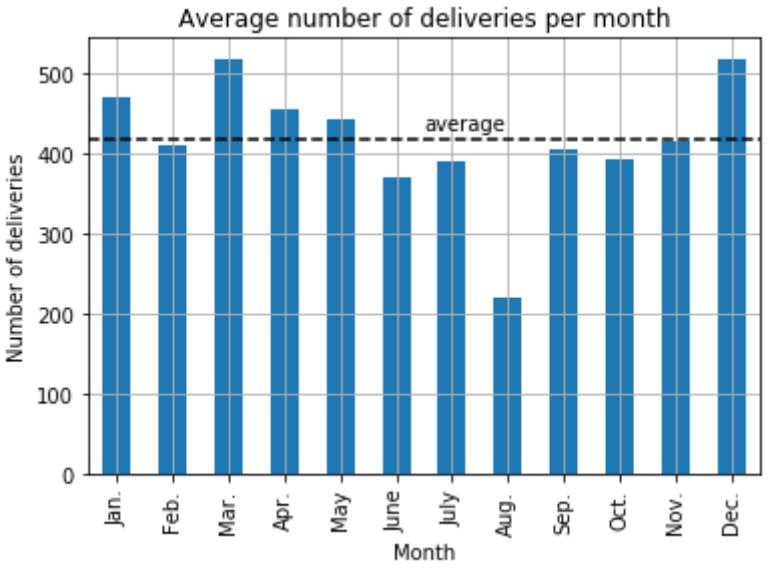

Figure A.1: Average number of deliveries per month

Appendix C: Discussion

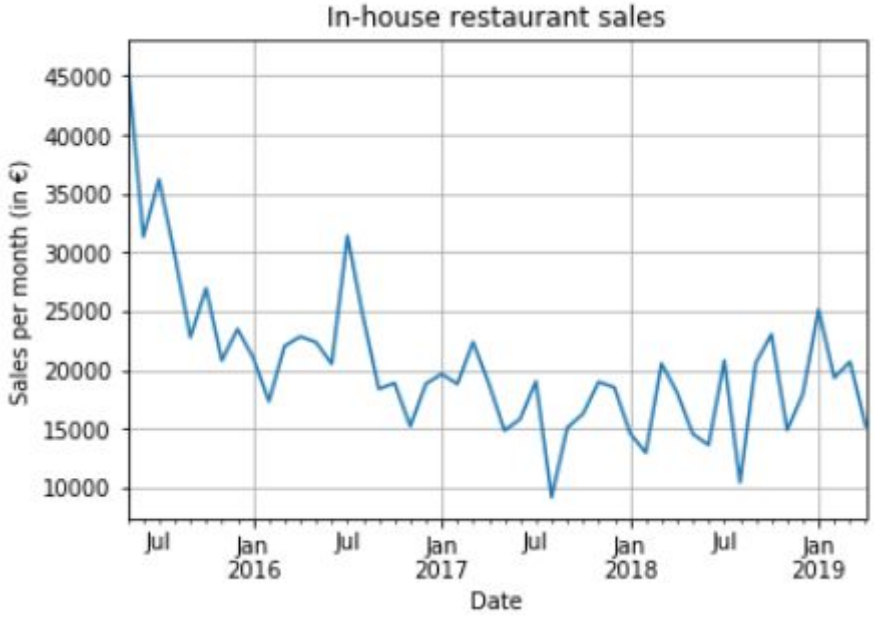

Figure C.1: Decreasing number of in-house restaurant sales over the last 5 years

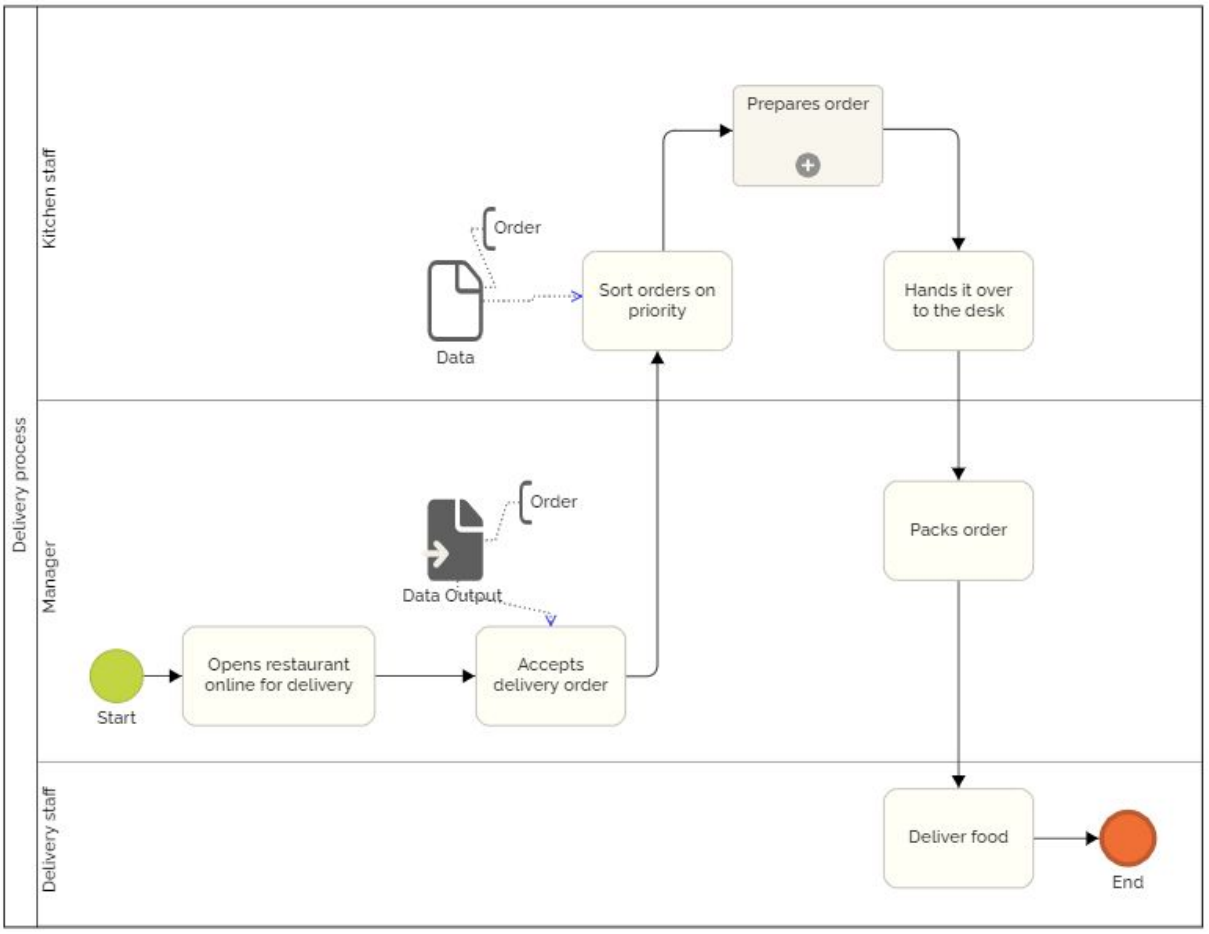

Figure C.2: BPMN model of the delivery process

Economic calculations using estimated values

Assumptions:

- Cooks cannot be hired one day in advance.
- Fixed amount of cooks in the kitchen
- Fixed costs are not taken into account: kitchen personnel, delivery vehicles, overhead costs

Note: numbers are estimates and may not reflect the actual costs of the market.

Delivery per staff:	±4 orders maximum per hour
Wage/hour:	7.50 euros/hour
Average sales per delivery:	32.34 per delivery
Cost of goods sold (in %):	35%
Commission Thuisbezorgd:	13% (fixed fee included)
Operating income:	11.32 per delivery (35% of 32.34 euros)

Table C.3: Economic calculations on cost savings and potential revenue

Break-even point calculation:
Revenue - cost of hiring one extra driver = 0
number of deliveries * operating income per delivery - cost of hiring one extra driver = 0
x * 11.32 - 7.50 = 0
x * 11.32 = 7.50
x = (11.32 / 7.50) ≈ **0.66 deliveries**
Hiring one extra deliverer for 0.66 deliveries does not generate any extra revenue nor loss for the restaurant.

Example calculation 1 - Potential revenue
Day of the week: Wednesday
Actual/predicted deliveries = 20
Maximum orders per hour (peak hour) = 18 * 0,35 = 7

Normally:
Number of deliverers = 1
Maximum output per hour for deliverer(s) = 1 * 4 = 4 orders per hour
Revenue missed: 7 - 4 = 3 orders could not be delivered

With accurate prediction
For those 3 orders, one extra deliverer would have been hired.
Potential revenue = 3 * 11.32 = 33.96
Potential profit = 33.96 - 7.50 = 26.46 euros extra

Example calculation 2 - Cost savings
Day of the week: Sunday
Actual/predicted deliveries = 14
Maximum orders per hour (peak hour) = 18 * 0,35 = 4.9

Normally:
Number of deliverers = 3
Maximum output per hour for deliverer(s) = 1 * 4 = 12 orders per hour
Hired too many deliverers: 4.9 - 12 = -7.1
The deliverers were able to deliver 7.1 orders per hour more than needed.

With accurate prediction
With 2 deliverers, the same revenue could have been achieved.
Potential cost saving of hiring one driver less = 7.50

Example calculation 3 - No cost savings nor extra revenue
Day of the week: Monday
Actual/predicted deliveries = 13
Maximum orders per hour (peak hour) = 13 * 0.35 = 4.55

Normally:
Number of deliverers = 1
Maximum output per hour for deliverer(s) = 1 * 4 = 4 orders per hour
Orders missed: 4.55 - 4 = 0.55 orders could not be delivered

With accurate prediction

No change in profit or cost savings.

Considering the break-even point is at 0.66, it is not worth hiring one extra employee to deliver 0.55 deliveries.

```
For each index and row in cost calculations:
    # difference in output = maximum orders per hour – maximum output per hour for deliverers
    If difference in output > 0:
        Extra drivers needed = difference in output / 4
        # 0.66 is the break-even point to hire one more deliverer or not
        If (difference in output % 4) < 0.66:
            Extra drivers needed rounded down
        Else:
            Extra drivers needed rounded up
        Hiring costs extra drivers = extra drivers needed * 7.5
        If extra drivers needed > 0:
            Earnings extra deliveries = difference in output * 11.32
        Profit = earnings extra deliveries – hiring costs extra drivers
    Elif difference in output < 0:
        # 3.34 is the break-even point for when you have hired too many deliverers (4 – 0.66)
        # abs value stands for absolute value
        If abs value of difference in output < 3.34:
            Drivers too many = 0
        Elif abs value of difference in output >= 3.34 and abs value of difference in output < 7.34:
            Drivers too many = 1
        Elif abs value of difference in output >= 7.34 and abs value of difference in output < 11.34:
            Drivers too many = 2
        Else:
            Drivers too many = 3
        Drivers cost savings = drivers too many * 7.5
        Profit = drivers cost saving
    Update cost calculation at index with profit (cost savings)
Return the sum of all the profits (cost savings)
```

Figure C.4: Semi-pseudocode cost savings and potential revenue for accurate prediction model