''

Universiteit Leiden
The Netherlands

Opleiding Informatica

Personalized modeling of training load and physical capacity of an elite rower

Victor Neuteboom

Supervisors:
A. Knobbe, R. Meerhoff, A.-W. de Leeuw & S. van der Zwaard

BACHELOR THESIS

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

09/07/2019

Abstract

There are many opportunities for using data science within the sports domain. One such opportunity is modeling the relationship between training load and an athlete's training response. As this response is athlete specific, it is interesting to investigate this relationship on an individual level. Understanding this response may provide opportunities to optimize an athlete's training plan and further improve athletic performance. Additionally, understanding this relationship may also aid in injury prevention. The idea of personalized models has been investigated in medicine, but is quite new in the sports domain.

This thesis aims to investigate the application of personalized models within the sports domain. Specifically, we aim to build a personalized regression model, capable of predicting a specific athlete's physical capacity based on historical training data.
We constructed our model from the historical training data and physical capacity tests of an elite rower. This historical data was supplied in the form of a logbook, where several measures were logged daily. These measures described internal training load, external training, recovery and well-being. From this logbook, we derived features describing several aspects of training, with a focus on training load. Furthermore, we defined *peak* VO_2 as the target measure of physical capacity.

We used a feature-based approach, in which we translated daily logbook data into features corresponding to specific time windows. From this approach, we were able to obtain aggregated features and overcome the problem of unequal and inconsistent sampling rates. Hereafter, we applied the Least Absolute Shrinkage and Selection Operator (or LASSO) regression to select a subset of important features and fit a model. The resulting model was validated within a nested cross-validation procedure.

The best cross-validated models only reached an R^2 value of 0.30. Given the low amount and type of data available, the personalized models were not reliable enough for meaningful predictions. Future work should further explore under which conditions (such as more controlled data collection, higher frequency or more rigorous target variables) our modeling approach can yield insightful results.

Contents

1 Introduction

In the recent years, the field of data science and data analytics has grown explosively. Data science and machine learning enable us to recognize patterns in big datasets and even predict future events. In a special issue of *Machine Learning*, several papers concerning practical applications of machine learning are showcased [1]. This issue demonstrates the practical value of machine learning applied to many domains in modern society and science.

One of the many domains where data science techniques show promise is the sports domain. Within the sports domain there are many different opportunities for collection and analysis of data. To some extent, sport has always been about data: goals scored, distance covered, time raced, even sports that are difficult to quantify are scored by judges (e.g., gymnastics).

By collecting data from both training and performance tests, it's possible to gain insights into how athletes respond to training. For example, one aspect of training that is often monitored is training load [2]. Training load is a description of volume, frequency and intensity of performed training. Changes in training load have been found to influence athletic performance and may help us understand the quantitative relationship between training and performance of athletes [3].

Understanding this relationship offers opportunities to optimize training strategies and further improve athlete performance. In sports where an athlete's physical capacities are very important in match performance, understanding the training-performance relationship can prove very useful. An example of such a sport is rowing, which is a very physically demanding sport. Furthermore, rowing performance is found to be highly correlated to measures of an athlete's physical capacities. For example, in a recent paper by van der Zwaard et al., maximal oxygen consumption and peak power together were found to explain 98% of variance in rowing ergometer performance [4].

However, understanding the relationship between training and physical capacity or performance has proven to be difficult in practice. In a paper by J. Borresen and M.I. Lambert, several limitations of modelling the training-performance relationship are mentioned [5]. First of all, small differences in athletic performance can have a considerable effect on match outcomes in elite competitions. Therefore, models that predict performance need a very narrow error margin in order to be practically viable. Furthermore, the lack of an individual measure of athlete training response has been identified as a possible explanation for poor model performance. Changes in performance caused by training can heavily depend on the specific sport and even the specific athlete, as response to training can be highly individual [6].

Therefore, it would be very interesting to narrow down our view and focus on the training response of an individual athlete. With machine learning techniques, it may be possible to develop a personalized model, tailored to the specific training response of an individual athlete. Thus far, such models haven't been applied to athlete data.
However, the development of personalized models is being studied extensively in another domain: medicine. Personalized medicine could help optimizing medical care for each individual patient [7]. With personalized medical models, there is a major challenge to overcome: in most settings, there are many more possible predictors (for example genes) than observations (clinical outcomes) [8].

This obstacle may explain why personalized models aren't widespread in different domains, as the problem of a disbalance between predictors and observations is not unique to medicine. There are several techniques that try to deal with such disbalanced data problems, where there are many more predictors than observations [9]. Many of these methods apply a form of regularization in order to deal with the high dimensionality of the data. In this thesis, we will focus on one such technique: the Least Absolute Shrinkage and Selection Operator (or LASSO).

Using the LASSO technique, we will try to build a personalized model of the relationship between training and an athlete's physical capacity. We will specifically focus on the data of a single elite rower. We will base our model on this single athlete's historical training data. Based on personal historical training data, we will seek to predict this athlete's physical capacity.

1.1 Research question

The aim of this thesis is to model the relationship between historical training data and physical capacities. However, constructing a model on its own is insufficient to be valuable to athletes and coaches. We need a measure of how reliable and accurate our model is. Furthermore, we are interested in the amount of data needed to build a reliable predictive model.

Considering the aspects mentioned above, we formulated our main research question:

Can a regression model based on historical training data of a single athlete reliably predict physical capacity and identify important training aspects?

1.2 Thesis overview

Firstly, in order to understand the features that will be used as inputs for our final model, in Chapter 2.1 we will explore relevant sport science theory concerning training load and performance. Secondly, in Chapter 2.3 we will look into data science theory and address how a regression model in general works, providing base knowledge of how a prediction is made using a model. Furthermore, we will look at a specific linear regression technique: the Least Absolute Shrinkage and Selection Operator (or LASSO).
In Chapter 2.4, after we have established background knowledge of our features and the workings of our regression model, we will focus on ways of measuring the accuracy and reliability of regression models.
With a basic understanding of our data and regression models, we will demonstrate the method we used in order to build our regression model in Chapter 3.
Afterwards, in in Chapter 4, we will provide the results we obtained by building a model according to our method.
Finally, in Chapters 5 and 6, we will discuss our findings, drawbacks and conclusions.

2 Theory & related work

In this Chapter, we will address relevant theory regarding training load, training structure and rowing performance. Furthermore, we will look at the theory behind regression models. Finally, we will touch upon model evaluation techniques.

2.1 Training load monitoring

In order to model the relationship between training effects and rowing performance, we need measures that quantify characteristics of training. In high-performance sports programs, a commonly monitored dimension of training characteristics is training load. Training load describes the frequency, volume and intensity of training. Monitoring training load of athletes is often considered a valuable tool for assessing adaptation to the training program and minimizing over-training or injury risks [6].

As technology improves, more advanced ways of measuring training load become available. Positional tracking technology allows monitoring and quantification of specific accelerations, speed and individual distances covered within a training session. Furthermore, sensors attached to equipment such as bicycles are able to track athlete power output. These measurements are more expensive as they require equipment and expertise to use effectively, but can provide a more detailed view on training load [10].
Training load can be described by many different measures. These measures are commonly split in two groups: measures of external training load and measures of internal training load [6].

2.1.1 External training load

Measures of external training load objectively describe the physical effort exerted by an athlete, as prescribed by a training plan. External training load is subject to the structure, quantity and quality of training [10].
These measures do not consider the physiological or psychological response of an athlete to training. They give an objective summary of prescribed training. Simple measures of external training load are descriptions of training characteristics and are specific to the kind of training that is performed. Examples of such measures of training load for rowing are: duration, distance covered, stroke rate and (ergometer) stroke power curves.
One of the major advantages of measuring external training load is that these measures are easy to track and directly result from the training program. In some cases, there are no suitable measures of internal training load available, whereas measures of external training load are always available.

2.1.2 Internal training load

However, training load as experienced by an athlete can differ with a same external training load. For example, two athletes may respond differently to the same training based on factors such as health and genetics [10]. Furthermore, the way a specific athlete responds to the same external training load may differ over time due to changed circumstances such as an increase in fitness [10]. As a result, it is important to not only monitor external training load, but to also monitor individual indicators of internal load of an athlete [10].

In contrast to external training load, internal training load is based on the effort and stress perceived by the athlete. Measurements of internal training load are often subjective and related to physical or mental stress experienced by athletes [6].

Several markers of internal training load can be measured objectively using sensors. A common occurrence is the use of heart rate tracking during training. A more expensive example is testing blood lactate concentrations during exercise, which gives more details on the physiological response to training but is much more intrusive and expensive to measure. Furthermore, there are cases where more intrusive measures such as blood lactate concentrations are hard to gather. For example, it is very impractical to perform blood lactate measurements during a boat training.

It is also possible to measure internal training load in a subjective way. Self-reported measures are commonly used in monitoring internal training load, as they are generally less intrusive and less expensive. A commonly used self-reported subjective measure of internal training load is the Rating of Perceived Exertion (or RPE) [2]. The premise of RPE is that directly after a training, an athlete rates the intensity of the workout on a scale of 0 to 10, where 0 is equivalent to rest and 10 is equivalent to maximal effort. In the case of rowing, an RPE score of 10 would be comparable to an all-out 2-km race. Table 1 describes how RPE scores should be classified.

Rating	**Description**
0	Rest
1	Really easy
2	Easy
3	Moderate
4	Sort of hard
5	Hard
6	-
7	Really hard
8	-
9	Really, really hard
10	Just like my hardest race

Table 1: Borg CR10 RPE scale, modified by Foster (2001)

A drawback of RPE scores is that they only describe the intensity of a training, whereas training load is considered a combination of intensity and volume. Furthermore, it is a highly subjective

measure as it depends on how the athlete feels. Even though RPE is a highly subjective measure, it has been proven valuable in many situations due to its relatively high validity and particularly due to the low costs of monitoring RPE [11, 12, 13].

2.1.3 Derived features

Measures of both internal and external load can be combined with each other to derive features that describe training characteristics in a different perspective or within a different scope. For example, RPE is often used as a measure of internal training load. However, RPE only describes the intensity of a training. By combining RPE with a measure of training time, we get a feature that describes the total volume of training.
RPE scores of a training are often combined with the duration of that training in minutes, resulting in a measure of training load regarding that specific training. This measure is called the Session Rating of Perceived Effort (or sRPE) [14].

Training load expressed in sRPE is calculated as follows:

$$sRPE = duration * RPE,$$

where duration is measured in minutes and sRPE is measured in arbitrary units (AU).

The session RPE is often used as a basis for other derived features.
One of the features that can be derived from sRPE over time is the Acute:Chronic Workload Ratio (or ACWR). The ACWR describes the proportion of acute workload to chronic workload [15].
Acute workload is defined as the sum of all most recent sRPE scores, usually those of the last 7 days. Chronic workload is defined as the sum of all sRPE scores over a longer period, usually of the last 28 days.

In most cases, this ratio is calculated either as a rolling average or as an exponentially weighted rolling average. The rolling average is calculated like so:

$$ACWR = \frac{sRPE_{acute}}{sRPE_{chronic}},$$

where $sRPE_{acute}$ is the average daily training load of the acute period (usually 7 days) and $sRPE_{chronic}$ is the average daily training load of the chronic period (usually 28 days).

The exponentially weighted ACWR places more emphasis on more recent training loads compared to older training loads by assigning them higher weights [16].
Usually, the values of the ACWR range between 0 and 2. A value of 0 indicates a period of rest, the acute load has to be equivalent to 0. A value of 1 means the acute and chronic training loads are equal. Generally, an ACWR value between 0.80 and 1.30 is considered optimal. A value below 0.80 indicates the acute load is low in comparison to the chronic load, indicating undertraining. A value of above 1.50 is associated with a higher relative injury risk in team sports (Australian football, cricket and rugby) [17].

Another feature derived from the session RPE is Training Monotony (or TM) [14]. Training Monotony is an index that describes the variation of training load over a period, usually a week. It is calculated by taking the daily mean training load over a period and dividing it by the standard deviation over that period. An increase in training monotony score may pose a risk for overtraining and illness [14].

2.2 Rowing performance

The international standard distance for rowing races is 2-km, with a race typically lasting between 5,8 and 7,4 minutes [18]. Race times may vary quite significantly, mainly due to weather conditions and different types of boats. In order to objectively test race performance, rowers typically perform a 2-km time trial on a rowing ergometer.
To perform well on this distance, rowers need a combination of endurance capacity and sprint capacity [18].

For endurance capacity, three essential aspects have been identified [19]:

1. Maximal oxygen consumption ($VO_{2_{max}}$): the maximum amount of oxygen an athlete can use during intense exercise
2. Lactate threshold: the point at which there is a steep increase in blood lactate levels during incremental exercise
3. Efficiency: the amount of oxygen consumed in order to reach a certain speed or power output

In previous studies concerning rowing, maximal oxygen uptake ($VO_{2_{max}}$), power output at 4mmol^{-1} lactate ($W_{4mmol^{-1}}$) and oxygen uptake at the lactate threshold ($VO_{2_{LT}}$) were found as possible determinants of 2000-m rowing ergometer performance [20].

For sprint capacity, the maximum amount of power an athlete can generate is crucial. The maximum amount of power is often defined as peak power (P_{peak}), which is measured during a short high intensity effort. Average power output during a short high intensity effort can also be used as an indicator for sprint capacity [21]. Peak power output was found as possible determinant of 2000-m rowing ergometer performance. Furthermore, peak power output was found to explain 84.6% of 2000-m ergometer performance in elite male rowers [20, 22].

Using a combination of determinants of both sprint and endurance capacity, it is possible to accurately predict 2000-m rowing ergometer performance. $W_{VO_{2max}}$, $W_{4mmol^{-1}}$, W_{max} and $VO_{2_{LT}}$ together were found to explain 98.3% variance in 2000m rowing ergometer times. Moreover, P_{peak} and $VO_{2_{max}}$ together were able to explain 98% of variance in 2000-m rowing ergometer performance [20, 4].

An all-out 2000m rowing ergometer trial is very mentally taxing for athletes, because a 2000-m rowing ergometer time trial closely emulates a real race. Therefore, different tests are often used to

assess athlete performance. In this thesis, we will focus on two types of tests: the Wingate test and the Maximal incremental step test.

2.2.1 Wingate test

The Wingate test consists of a maximal effort of 30 seconds. Such a short maximal effort stresses the anaerobic system of an athlete and allows measurement of anaerobic power and capacity. The test is most commonly performed using a cycling ergometer, but can be performed on a rowing ergometer as well [23].

Before the start of the test, the athlete performs a low-resistance warm-up of at least 5 minutes. During this warm-up, several short sprints are performed to get used to the fast movement required during the test. Five seconds before the test, the athlete starts pedalling at maximum speed. At the start of the test, the workload is set to a predetermined value of resistance, this value is often based on the athlete's bodyweight. During 30 seconds, the athlete delivers an all-out effort, pedaling as fast as possible. From the data resulting from the 30 second effort, several values are commonly derived: peak power output, lowest power output and a fatigue index (the difference between peak and lowest power output) [24].

The Wingate test is very suitable for the prediction of medium distance sports performance, as is the case with rowing. With this test, a very high predictive power was obtained in several studies. In elite rowing, P_{peak} and $VO_{2_{max}}$ derived from a Wingate test were able to explain 98% of variance in ergometer performance [4]. In another medium distance sport, speed-skating, studies have shown that a Wingate test performed in the summer has predictive value for performance during the following winter in elite athletes [25]. Additionally, the test's validity has been extensively studied [26, 27, 24].

2.2.2 Maximal incremental step test

A maximal incremental step test may be performed to assess the athlete's endurance capacity. The test is performed on a rowing ergometer. It consists of several steps of increasing intensity.
Each step lasts 4 minutes. Between each step, there is a one minute rest period. During the first 6 steps, the target workload (in Watts) is predetermined by a protocol and increases with a constant increment every step.
During the final step, the athlete performs a maximal effort for 4 minutes, without a set predetermined target workload. During rest periods, blood samples are collected to determine lactate levels. From the test results, measures describing power output, oxygen uptake and blood lactate levels are derived. A complete description of the test protocol can be found in [28].

2.3 Regression model

From the tests discussed in Chapter 2.2, several measures of an athlete's physical capacity can be derived. Several of these measures are highly reliable predictors of 2-km rowing performance.
In this thesis, we aim to predict these measures of physical capacity based on historical training data. In order to make a prediction, we need to construct some sort of model. As both measures of training load, derived from historical training data, and measures of physical capacity are continuous data, it is possible to construct a (linear) regression model.

2.3.1 Linear regression

A linear regression model [9] attempts to fit a linear regression line to the data. In the case of a single predictor and n observations, simple linear regression, the equation of the regression line is of the form:

$$Y_i = a + b * X_i + \epsilon,$$

where Y is the value of the target in sample i, X is the value of the predictor in sample i , a is the intercept, b is the slope of the line with i ranging from 1 to n and ϵ is the residual error.

In practice, datasets often have more than one predictor, so a simple linear regression is not appropriate. In order to fit a regression line using two or more predictors, we use multiple linear regression. The formula of a multiple linear regression model [9] with p predictors and n observations is of the form:

$$Y_i = a + b_1 X_{i1} + b_2 X_{i2} + b_p X_{ip} + \epsilon_i,$$

where Y is the value of the target in sample i, X_{ij} is the value of predictor j in sample i , a is the intercept, b_j is the coefficient of predictor j, ϵ_i is the residual error between predicted and actual value of the target in sample i, with i ranging from 1 to n and j ranging from 1 to p.

In order to fit a multiple linear regression model, we have to calculate the coefficients of our predictors that minimize our model error. A commonly used method is to use the Ordinary Least Squares (OLS) procedure. OLS minimizes the sum of squared residuals (RSS). [9]

2.3.2 LASSO model

In practical applications, data often has more potential predictors than test samples. When working with data with many predictors compared to test samples, we generally want to avoid using too many predictors in a linear regression model, as this will lead to a high chance of overfitting. Instead, we want to use a small subset of the best predictors.

In order to select a subset of important predictors it is possible to use a technique called the Least Absolute Shrinkage and Selection Operator (or LASSO)[29]. In this method, coefficients of predictors are shrunk. It is possible to shrink the coefficients of some predictors to zero, effectively removing them from the resulting model.

Within this method, shrinkage is achieved by applying a penalty based on the so-called L1 vector norm [30]. The L1 norm measures the length of a vector by calculating the sum of all absolute values of the vector [31]. In the case of a regression model, this would be the sum of all absolute coefficients of potential predictors. A regularization term containing the L1 norm is added to the ordinary least squares loss function, resulting in the following definition of the LASSO estimate [9].

$$LASSO_{estimate} = argmin_{\beta}\Big\{\frac{1}{2}\sum_{i=1}^{n}((y_i - \beta_0 - \sum_{j=1}^{p}(x_{ij}\beta_j))^2) + \alpha\sum_{j=1}^{p}|\beta_j|\Big\},$$

where n is the total number of test samples, y_i is the true value at sample i, β_j is the coefficient of predictor j, α is a predetermined constant, p is the total amount of predictors and x_{ij} is the value of predictor j at sample i.

The amount of shrinkage introduced by LASSO is controlled by a hyperparameter called α. The value of α directly influences the number of predictors used in the resulting linear regression model. As the value of α increases more shrinkage is introduced and more coefficients are set to 0, with all coefficients being zero at $\alpha = \infty$. When the coefficient of a predictor is set to 0, the predictor is effectively eliminated from the regression model. When the value of α is decreased, less coefficients are set to 0. When α is equal to zero, all predictors are included in the model. The solution computed by using the LASSO method at $\alpha = 0$ is equal to the ordinary least squares regression solution as the penalty term in the estimate computed with LASSO is equal to zero, resulting in no shrinkage.

Finding the right fit for a model is finding the right balance between bias and variance [9]. In this case, some bias is purposefully introduced to the model by using a regularization term, effectively decreasing model complexity. This decrease in model complexity may introduce additional error in modelling the dataset, but improves the stability of the model estimates and thereby the generalizability of the model [32]. As a result, the model will generally perform worse on the initial dataset, but may perform better on unseen data.

2.3.3 LARS algorithm

There are several ways to fit a LASSO regression model. One of them is the computationally efficient algorithm named Least Angle Regression (or LARS) [33]. LARS is capable of efficiently calculating the solution path used by LASSO and finding values of α. It begins at an infinitely large α, where the solution computed by LASSO is equivalent to the ordinary least squares solution. Step by step, it decreases the value of α. As α decreases, a piecewise linear solution path is calculated. For every decrease in α, the active set of predictors changes. It continues until all predictors have entered the model. See [9] for more details on fitting LASSO using the LARS algorithm.

2.4 Model validation

To judge the reliability of a model, it is important to understand the prediction accuracy and generalizability of a predictive model. In order to gain insight in model performance, model validation methods are used. Model validation methods guide the choice of model and quantify model quality [9].

Within model validation methods, the dataset is typically split in a training set and a test set [9]. The regression model is fit using only the training data, leaving a set of unseen test data. After fitting the model, the performance of the model is tested using the unseen test set.

2.4.1 Cross-validation

The same principle of splitting the data in a train and test set is used in cross-validation [9]. However, cross-validation uses multiple iterations (or folds) with different splits of the data. For each iteration a scoring metric for the model can be computed. This scoring metric can be averaged over all folds, resulting in an indication of how well the model performs on average.

Commonly used scoring metrics include: mean absolute error (MAE)[34], mean squared error (MSE)[35], root mean squared error (RMSE)[34] and coefficient of determination (R^2)[35].

There are several methods of splitting the dataset in training and test tests. One of them is leave-one-out cross-validation (LOOCV)[36]. On a dataset with n samples, LOOCV performs n iterations of validation. In each iteration, one sample is used as the test set and the remaining samples are used as the training set.

Cross-validation fold	**Training set**	**Test set**
1	2, 3, ..., n-1, n	1
2	1, 3, ..., n-1, n	2
...	...	...
n-1	1, 2, ..., n-2, n	n-1
n	1, 2, ..., n-2, n-1	n

Table 2: Visualisation of leave-one-out cross-validation of a dataset with n samples

2.4.2 Nested cross-validation

Cross-validation may have biased results when fitting hyperparameter(s) and assessing model performance within the same cross-validation procedure, especially in small sample scenarios. This possible bias arises from the fact that regular cross-validation both fits hyperparameters and evaluates model score on the same dataset. This may lead to an overly optimistic estimation of model performance. For more details on this possible introduction of bias, see [37]. In order to obtain

an unbiased measure of performance, fitting hyperparameters and evaluating model performance need to be separated. There exists a modified cross-validation solution that accomplishes this separation by using two levels of cross-validation. This solution is called nested cross-validation [38].

To simplify the explanation of nested cross-validation, let us consider fitting a model with one hyperparameter α. It uses an outer layer of cross-validation in order to validate model performance. For every fold of this outer cross-validation, we choose an optimal value of α using another layer of cross-validation. This inner cross-validation is performed only on the training set of the outer cross-validation, so when choosing our α the model never sees our outer validation set. From our inner cross-validation, we obtain an optimal α value. Within every out cross-validation fold, we fit a model based on the best α found for that fold. The resulting model is evaluated with the outer validation set. In summary, the inner layer of cross-validation is used for hyperparameter tuning, the outer layer is used for evaluation of model performance on unseen data. Figure 1 shows a visual example of one outer fold of this nested cross-validation method.

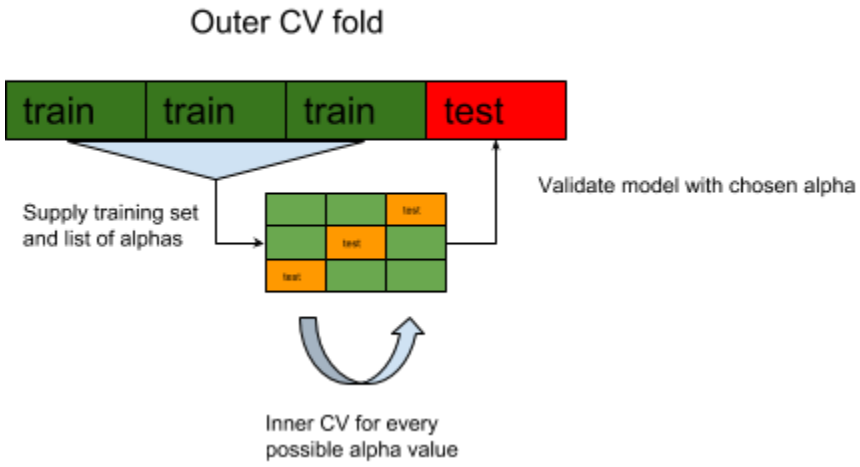

Figure 1: One outer fold of nested cross-validation, where hyperparameters are fitted in the inner cross-validation.

2.4.3 Implications of cross-validation scores

From the scores obtained from cross-validation, we can gain several insights. First and foremost, cross-validation scores give a good estimate of how well our model performs on held-out, unseen, data. The RMSE can be compared to the standard deviation of our target variable, giving an indication whether our model performs better than a baseline. Our aim is to construct a model with a minimal RMSE value. The R^2 score gives an easily interpretable indication of model performance by describing the proportion of variance explained by the predictive model [35]. Ideally, the R^2 score is close to 1, meaning that the model explains close to 100% of the variance.

Furthermore, we can compare our training errors with our testing errors obtained by cross-validation. By plotting training and testing errors against model complexity, we can observe possible over- or underfitting [9]. A model that suffers from underfitting will have both high training error and high testing error, where both will decrease with increased model complexity. Such a model won't be able to make accurate predictions on either seen or unseen data. On the other hand, a model that is affected by overfitting has a significant difference between training error and testing error. This model will make accurate predictions on training data, but perform much worse on unseen data as it fails to generalize well. Usually, cross-validation selects a model without overfitting or underfitting.

3 Method

The aim of this thesis is to predict physical capacity of an athlete based on historical training data. In order to make this prediction, a model is constructed. In this chapter, we will describe all steps we performed in order to build this model, from data collection and feature engineering to fitting and validating the resulting model. For each step, all software was written in Python [39], using the pandas [40] and scikit-learn [41] modules.

3.1 Data collection & cleaning

Within our data collection process, we looked at several aspects of rowing training and physical capacity of athletes. These aspects were monitored using two different sources of data: logbook entries and performance test results.

From the logbook entries, several measures were monitored. These measures describe internal training load, external training load, recovery & well-being and other aspects. Within the logbook data, the following measures were monitored:

Internal training load

1. Resting heart rate (beats/min)
2. RPE score for 1st, 2nd and 3rd training per day (on a scale of 1-10)

External training load

3. Distance for 1st, 2nd and 3rd training per day (km)
4. Duration for for 1st, 2nd and 3rd training per day (minutes)
5. Training frequency (count per day, between 0 and 3)

Recovery & well-being

6. Recovery (on a scale of 1-10)
7. Sleep quality (on a scale of 1-10)
8. Sleep duration (minutes)
9. Mood (on a scale of 1-10)
10. Eaten well (yes/no)

11. Slept well (yes/no)
12. Max strain (yes/no)

Other

13. Bodyweight (kg)
14. Comments (text)

Besides tracking logbook data, the athlete performed several performance tests. The athlete performed two types of tests: the Wingate tests and Maximal incremental step tests. For more information, see Chapter 2.
From the Wingate test, we identified the following variable as potential target:

1. *Peak power*

From the Maximal incremental step test, we identified the following variables as potential targets:

1. *Peak* VO_2
2. *Peak power*

In this thesis we focused on the results from the step tests, because considerably more Maximal incremental step tests (14) were performed than Wingate tests (5). From the step test, we chose Peak VO_2 as our main target. Peak VO_2 was chosen as it was found to have a very high correlation with 2000-m rowing ergometer performance, as described in Chapter 2.2.

3.1.1 Cleaning

Before we were able to use our data to build a model, we had to clean the data. In order to clean our data, we first looked at missing values. There were some variables that were missing for the majority of logbook entries, with gaps of missing values of several months. We dropped the following variables for having too many missing values:

1. Resting heart rate
2. Weight
3. Max strain

Furthermore, we restricted our data to the period between november 2013 and december 2017. After december 2017 measures regarding training load were no longer tracked. As these measures are important for building our model, we decided to not use the data collected after december 2017. This resulted in three step tests being excluded from analysis. Furthermore, another step test was excluded for not having data for the final step, resulting in 11 usable step tests in total.

3.2 Feature engineering

Our goal is to build a regression model with a high predictive power. The features we used as inputs for our model will have a large influence on the resulting prediction accuracy. Without good features, it is difficult to obtain a good model. The goal of our feature engineering process is two-fold:

1. To obtain a better description of the underlying structure of our data.
2. To transform the data into a format that allows us to fit a regression model.

3.2.1 Feature construction

In an attempt to better describe our data we construct additional features. These features are constructed without adding new data. Instead, they are based on existing features or combinations of features. New features can be manually constructed based on domain knowledge (see Chapter 2). These features may better describe the underlying structure of the data than the original variables. In training load data, there is also an important temporal component. We will discuss this temporal component in Section 3.2.2. Furthermore, better features allow you to build a good model using a smaller number of features as these better features may be able to provide more information on the underlying structure of the data.

In this process we constructed the features listed below. We expect that these features capture the most important characteristics of the training load. For further information on these features, see Chapter 2.

Daily training load: calculated by multiplying the RPE score by the duration of every training, summed per day (there are a maximum of 3 trainings per day).

Weighted ACWR: Exponentially weighted acute:chronic workload ratio, where acute workload is last 7 days and chronic workload is last 28 days.

Unweighted ACWR : Rolling average acute:chronic workload ratio, where acute workload is last 7 days and chronic workload is last 28 days.

Weekly monotony: Training monotony, calculated over the last 7 days.

Monthly monotony: Training monotony, calculated over the last 28 days.

Weekly total load: Sum of all training loads of last 7 days.

Monthly total load: Sum of all training loads of last 28 days.

Change in training load: Training load of last day - training load of current day.

Load per training type: Sum of training load, grouped by training type (rowing, ergometer, weights, bike, other).

Time per RPE zone: Sum of training durations, grouped by RPE scores (where RPE scores of 1-2, 3-4, 5-6, 7-8, 9-10 are grouped together).

Load per RPE zone: Sum of training loads, grouped by RPE scores (where RPE scores of 1-2, 3-4, 5-6, 7-8, 9-10 are grouped together).

3.2.2 Feature aggregation

After cleaning our data and constructing additional features, we should have some informative features to use in our model. However, at this point our features can't directly be used to predict our target variable, the athlete's $Peak\ VO_2$ (as measured in a Maximal incremental step test). In order to be useful for our model, our features (or predictors) must occur at the same frequency as our target variable.

Our target data has a lower sampling rate than our predictors. Our predictors are measured daily, whereas our target data is measured infrequently. Normally, you would be able to downsample

the predictors to the sampling rate of the target or even upsample the target to the rate of the predictors. However, this will not work well due to the infrequent nature of the target data.

In order to solve this problem, we chose to only consider data within a specific time window before every performed step test. For example, we can consider all data in a window of 24 weeks from a specific step test up to the day before this specific step test. The values of every original feature within this window are aggregated to new features specific to that window. A drawback of this method is the potential loss of information, as data outside of the chosen window is not considered in the resulting features. We try to minimize this loss of information by using a set of windows of different sizes. The use of windowing also allows us to incorporate temporal components in a model.

We experimented with the windows described in 3, with windows denoted as [x, y] where x is the start of window, in weeks (w) or days (d) from the test and y is the end of the window, also in weeks (w) or days (d) from the test.

Start window	**End window**	**Notation**
8 weeks before test	5 days before test	['8w', '5d']
2 weeks before test	3 days before test	['2w', '3d']
1 week before test	1 day before test	['1w', '1d']
2 days before test	day of test	['2d', '0d']

Table 3: Windows used in the feature aggregation process.

These windows were chosen to describe training effects over several different time periods before a test: long term effects, medium term effects, short term effects and acute effects.

For every feature in each window, we computed the following aggregates:

1. Min
2. Max
3. Standard deviation
4. Mean

This process results in 4 new aggregate features for every original feature in a specific window. Using 4 different windows, 4 different aggregation methods and 30 original features, we would end up with $4 * 4 * 30 = 480$ aggregate features. The combination of a very high number of predictors and a low number of test samples poses a large risk of overfitting our model. We will address this issue with our choice of model.

3.3 Regression model

Due to the nature of our data and as a result of our feature engineering process, we end up with many more predictors than test samples ($p >> n$ or high dimensionality, low sample size)[9].
Due to the high dimensionality and low sample size of our data, there is a high risk of overfitting a model.

In order to lower the risk of overfitting, we aim to reduce the number of predictors used in our regression model. This leads us to the use of a LASSO regression model (see 2.3 for more information on LASSO).

In order to choose an appropriate value for the hyperparameter α, which determines shrinkage, we use cross-validation. As our sample size is limited, we aim to use as much data for training as possible. As a result, we chose leave-one-out as our cross-validation method.

In order to obtain an unbiased assessment of model performance, we used a nested cross-validation procedure consisting of two layers of leave-one-out cross-validation (as explained in 2.4). Within the inner layer of cross-validation, hyperparameter fitting was done. In the outer layer, the resulting models were assessed and model performance was calculated.

Within our inner cross-validation procedure, we first established a range of predetermined α values. For every fold in the inner cross-validation we fit a model for every value of α in our predetermined list of values. This results in a test error score for every fold for every α. We used mean squared error as our scoring metric.

After our inner cross-validation process, we group our error scores by α value and take the mean error over all folds. This results in a mean test error score for every value of α. Initially, we chose the value of α with the lowest average test error as our best α, which we used to fit our final model. However, we also wanted to have more control over how many predictors could be chosen. We decided to add a constraint to the value of α based on the resulting amount of predictors. Only values of α that resulted in a number of predictors lower or equal to our manually selected boundary were considered in finding the best α value.
After choosing a value for α, we assessed model performance in the outer layer the nested cross-validation procedure. Every fold of outer cross-validation may have a different value for α, as the entire inner cross-validation procedure is done for every single fold of outer cross-validation.

3.4 Experiment on limited data

Datasets describing training load and physical capacity test results of a single athlete will generally be limited in size, due to the impact of performing physical capacity tests. Therefore, we were interested in seeing if we could construct a reliable model with smaller datasets and if we can find a pattern in how fast model accuracy improves when expanding the dataset with more test samples. Furthermore, it could be interesting for athletes if they are able to iteratively improve performance of a predictive model by logging new data.

To gain insight in how much data is needed and how fast model performance improves when adding data, we estimate model performance on smaller parts of our original dataset. We repeated our main experiment several times using only a selected part of our total dataset. We tried portions of our datasets ranging from a size 4 test samples to 11 of test samples. To get a more accurate measure of model performance we used different distributions of the dataset for every possible number of samples. In order to account for possible chronological effects, we used both chronological and random distributions, as can be seen in the first split in Figure 2. The chronological method only chooses samples which are directly adjacent to each other, using a sliding window. With this method, for a dataset of length n and a windowsize of w, there will be $n + 1 - w$ possible distributions. The random method also chose $n + 1 - w$ distributions, without the limitation that samples have to be adjacent to each other.
For every possible number of samples, the mean score is computed separately for chronological and random distributions. This workflow results in a list of cross-validated model performance scores for every possible n, for both chronological and random distributions.

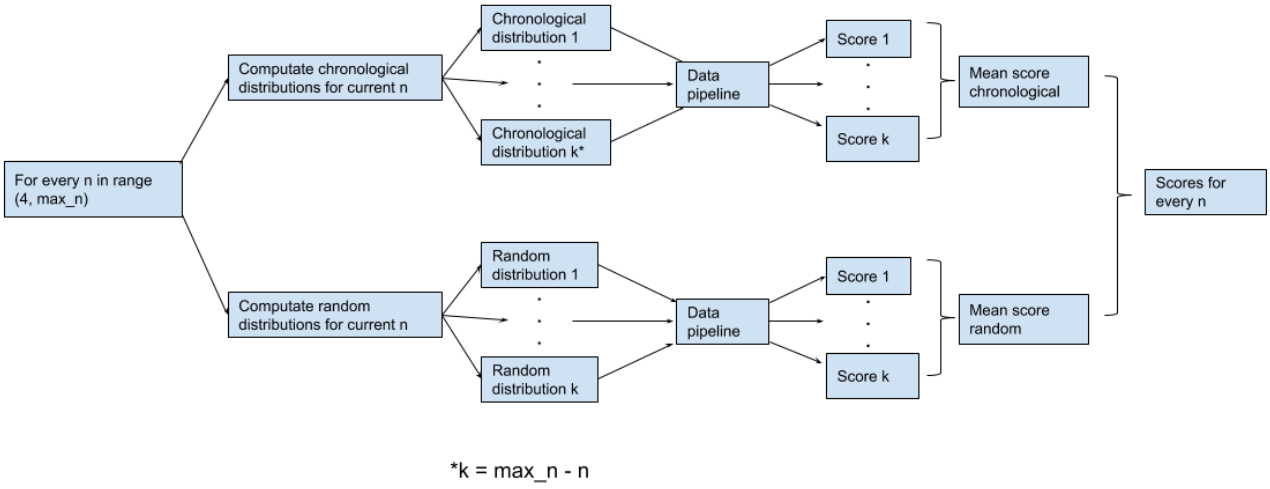

Figure 2: General workflow of experiment with limited samples.

4 Results

Our main goal was to construct a regression model based on historical training data to predict physical capacity of an athlete. We defined *peak* VO_2 (from maximal incremental step tests) as our target measure of physical capacity. The final model will use features derived from logbook data to predict an athlete's *peak* VO_2.

In this Chapter, we will elaborate on several important aspects related to our main goal. First and foremost, we are interested in the final model, fitted on the entire dataset. Secondly, we are interested in which features are used in models resulting from cross-validation and how stable the feature selection process is. Thirdly, we will discuss the model's cross-validated quality and fit. Next, we will consider the statistical significance of the predictions. Finally, we will investigate the required dataset size to fit a useful, reliable model.

Firstly, we will look at the regression model constructed by following the procedures described in Chapter 3. For our final model, we fitted hyperparameters using regular leave-one-out cross-validation. We chose 3 predictors as our maximum amount of predictors. This resulted in the following model:

$$VO_{2_{peak}} = 105.04 * p_1 + b$$

where p_1 is the minimal weighted ACWR in window [8w, 5d], and b is the intercept.

This model was fit with an α of 98.07. Even though the model was allowed to choose at most three predictors, it only picked one. Apparently, using only one predictor led to a higher cross-validated score in the hyperparameter fitting procedure.
Using a nested cross-validation procedure, an R^2 of 0.30 was obtained. We will further discuss this cross-validated model performance later in this chapter.

Secondly, we will look at model stability and chosen features when using a nested cross-validation. In our nested cross-validation procedure, every fold of the outer layer of cross-validation results in a separate model. To gain insight in the important features, we counted how often each of the significant predictors was included within different models obtained by this method of cross-validation. These counts can be found in Table 4. Even though we set a boundary of three predictors for our model, sometimes a model with less predictors has a better cross-validated accuracy and was therefore chosen as final model by the cross-validation procedure.

Feature	**Aggregate**	**Window**	**Times included in model**
Weighted ACWR	Min	8 weeks - 5 days	11
sRPE first training	Std	8 weeks - 5 days	3
Weighted ACWR	Mean	8 weeks - 5 days	1
Distance first training	Min	2 days - 0 days	1
sRPE first training	Max	8 weeks - 5 days	1
Sleep quality	Min	1 week - 1 day	1

Table 4: Features included in models built in outer cross-validation folds with all samples (n=11), using a maximum of 3 predictors (p=3)

Next we will look at the stability of our model. We base the stability of our model on the consistency of chosen features. If a model doesn't consistently select a similar subset of features, it is not robust. However, it has proven difficult to quantify this stability. In recent research, a measurement quantifying stability was proposed [42]. Using the method proposed in this paper, we calculated a measure of stability (Φ) based on the method described in [42]. With this method, a value for Φ of < 0.40 is considered poor, 0.40 to 0.75 is considered intermediate to good and > 0.75 is considered excellent.

Number of predictors	**Stability measure (Φ)**
1	0.00
3	**0.64**
4	**0.56**
6	0.51
8	0.39

Table 5: Stability of selected features from nested cross-validation using all test samples (n=11), differing number of predictors

Thirdly, we will deal with model quality and fit. Furthermore, we will examine statistical significance of the regression model. The quality of our resulting regression model can be assessed based on three attributes: Accuracy, model fit and statistical significance. Firstly, we will look at the accuracy of our model. We are mostly interested in the performance of our model on unseen data. So in order to measure accuracy, we will look at cross-validated measures of model performance. Specifically, we look at the scores calculated by the outer layer of cross-validation in our nested cross-validation procedure, as these give an unbiased view of model performance on unseen data. We evaluated several models, differing in the amount of predictors chosen by LASSO. These models were all fitted using the method described in Chapter 3, using all available test samples. The resulting performance metrics can be found in the table below. From this table, it is clear that models with 3 or 4 predictors obtained the best accuracy on unseen data.

Number of predictors	R^2	**RMSE**	**std(target)**
1	-0.05	446.64	436.87
3	**0.30**	**365.10**	**436.87**
4	**0.30**	**365.69**	**436.87**
6	0.20	389.60	436.87
8	0.12	409.68	436.87

Table 6: Scores from nested cross-validation using all test samples (n=11), differing number of predictors (p)

In Chapter 2.4.3, we mentioned the possibility of assessing model fit by comparing training and testing errors. In our nested cross-validation procedure, we can consider the cross-validated scores from the outer layer as our testing error. The cross-validated scores of our inner cross-validation can be considered as our training error. Plotting the results from our nested procedure results in Figure 3.

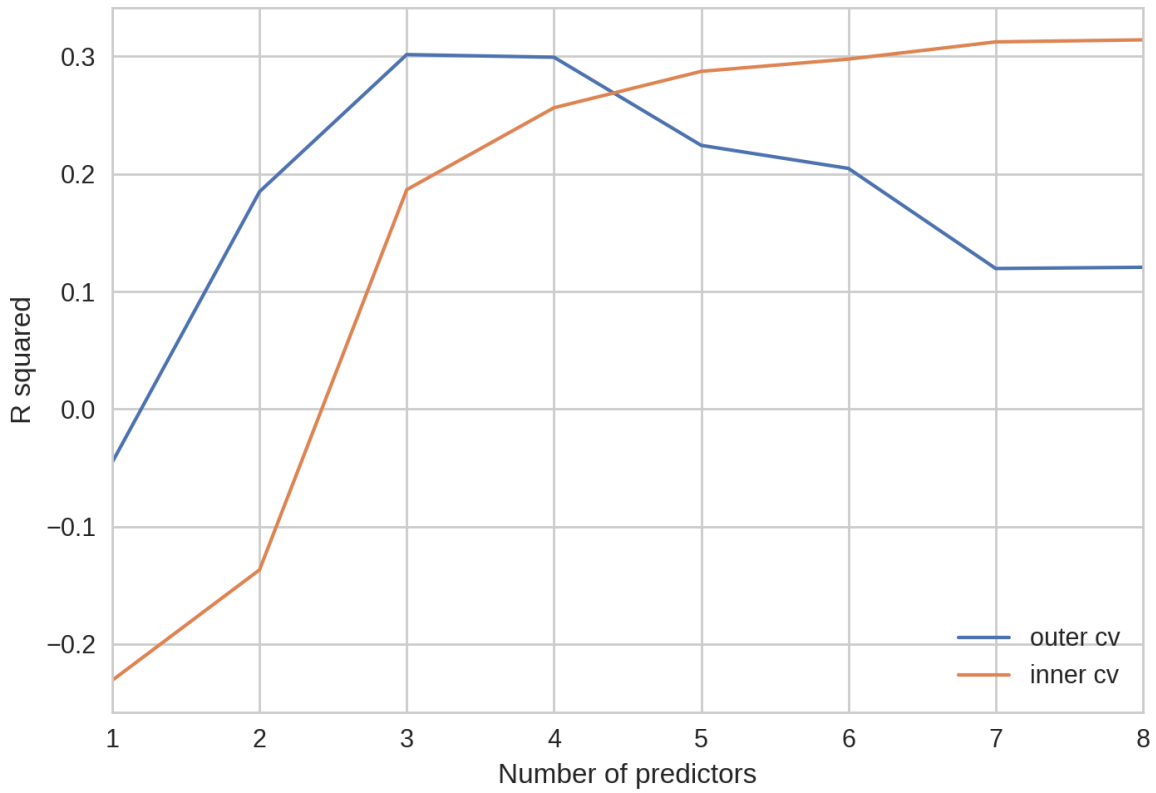

Figure 3: Comparison of training and testing errors, where outer cv can be considered as testing error and inner cv as training error.

As expected, training error decreases when model complexity increases (or more predictors are added). However, testing error reaches a minimum at a certain point of complexity and increases with further rises in model complexity. From this plot, it is clear that the best fit for our model is at 3 or 4 predictors. Models with less predictors are possibly underfitted and models with more predictors are possibly overfitted.

Next, we will look at the statistical significance of our model. We will derive an F value from the R^2 and degrees of freedom resulting from our model using the formula listed below.

$$F = \frac{R^2}{1 - R^2} * \frac{p}{n - (p + 1)},$$

where p is the number of predictors and n is the number of observations

The resulting value of F is 1.00 whereas the critical F-value is 4.25. The corresponding p-value is 0.45. As our $F < F_{critical}$, we fail to reject the null-hypothesis, indicating our model is not significantly different from an intercept-only model.

Finally, we are interested in the size of the dataset that is needed to fit a reliable model. In order to examine the effects of dataset size on prediction accuracy, we performed an additional experiment using limited portions of our dataset. From this experiment we obtained cross-validated scores for models with different amounts of predictors and which were trained and tested on limited parts of our dataset.
In Figure 4, we plotted the cross-validated R^2 scores of models based on increasing sizes of data available to these models. At every value of n, we plotted the best model performance, determined by highest average cross-validated R^2.

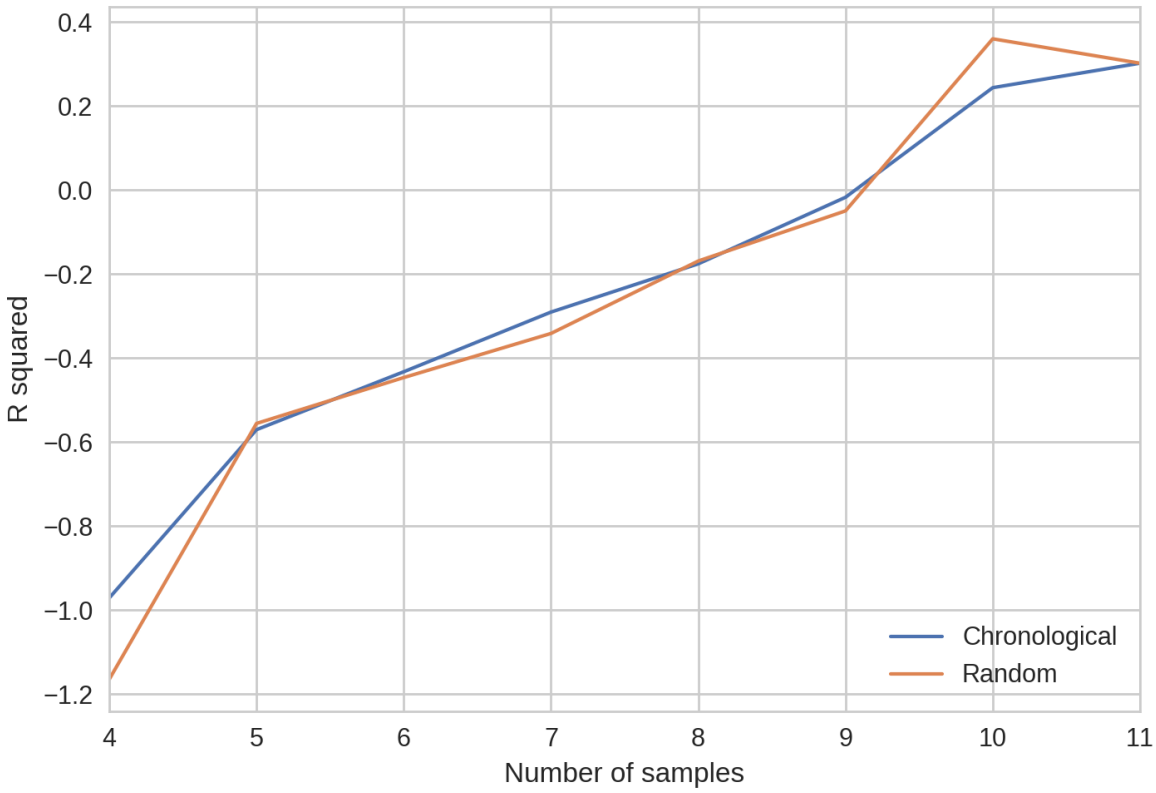

Figure 4: Model performance on different dataset sizes, using chronological and random distributions of samples

In this plot, there are quite some negative R^2 values. This may occur when the fitted model makes worse predictions than fitting just a straight line as a baseline (the intercept). It indicates that the model is completely unable to make accurate predictions.

5 Discussion

The main goal of this thesis was to provide a model based on the relationship between historical training data and physical capacities of an elite athlete.

Using a feature construction and aggregation procedure, we constructed features describing training load. We used these features to fit a model with LASSO regression. We assessed the quality of our models based on four aspects: accuracy, stability, important features and required dataset size.

In order to evaluate model accuracy, we used two metrics: cross-validation accuracy and statistical significance. Our best model, fit with three predictors and an intercept, had a cross-validated R^2 score of 0.30. Furthermore, we derived an F-score and critical F-score. As our F-score was smaller than the critical F-score, our model was found to be insignificantly different from a baseline model consisting of only a fitted intercept.
All-in-all, we consider our model accuracy and significance to be too low to be of practical use.

We assessed our model stability based on consistency of chosen features by applying the method described in a recent paper by S. Nogueira, K. Sechidis, and G. Brown [42]. For our best model, fitted using 3 predictors, we obtained a stability value of 0.64, which is considered moderate to good. Further inspection of individual cross-validation folds shows that our model consistently chooses features from a small subset. We consider our model to be stable, implying the model is able to choose important features.

From the small subset of variables often chosen by our model, one feature was chosen every single fold: the minimal acute:chronic workload ratio within the window ranging from 8 weeks to 5 days to our test. The value of this feature is mainly influenced by periods of rest or low training loads, which will result in a low value.

Additionally, we investigated the effect of sample size on model performance in order to gain insight into how much data is needed to fit a reliable model. Models fitted with the minimum amount of samples (4) performed terrible, achieving R^2 scores far below zero. Negative scores imply that the model performed worse than a baseline model (fitting just the intercept). Plotting model score alongside sample size showed a clear upwards trend in model performance, with an approximately linear relationship between model performance and sample size. Based on these results, we expect model accuracy to improve further when supplied with more data.

In conclusion, considering our model accuracy was both low and statistically insignificant, we deduce that our model in its current state is unable to accurately and reliably make personalized predictions of performance based on historical data.

There are several possible explanations why our model is unable to accurately predict the physical capacity of an athlete. Firstly, the relationship between our features and our target may be too complex to model using a linear regression model such as a LASSO model [43]. This possibility may be investigated by fitting different types of models on our dataset.
Secondly, we may have not succeeded in capturing the effects of training within the right features.

There are opportunities to vary chosen windows or use different aggregate functions. Furthermore, different features may be constructed from the raw logbook data.
Thirdly, the dataset may have simply been too small. Within machine learning, a dataset can easily have thousands of observations. A dataset consisting of only 11 samples is undoubtedly a very small dataset, making it harder to find statistically significant relationships.

On a positive note, the experiment with varying dataset sizes shows a promising upwards trend in model performance over increasing amounts of samples. This may indicate that adding more samples will improve our regression model's performance. This could possibly lead to obtaining a statistically significant and practically useful model performance, provided that the sample size is sufficiently large. It is hard to conclude how much data exactly is needed to fit a reliable model, but the current sample size is most likely too small [43].

Due to different frequencies that training and physical capacity tests occur, datasets of a similar nature to the dataset examined in this thesis will very likely be unbalanced. Athletes training daily or multiple times a day will quickly build up a dataset describing training load. However, testing physical capacities can be intrusive and is done on a much lower frequency than regular training, leading to a low target sample size. In order to obtain a more balanced dataset, physical capacity tests may simply need to be performed more often. But due to the intrusiveness of such tests, this may not be desirable. Instead, it may be interesting to identify less intrusive measures of physical capacity or markers that heavily correlate with physical capacity or performance, which can be used as a surrogate for performance tests. Such a measure could then be used as the target for fitting a model. Until a less intrusive measure or a good marker for physical capacities is found, it will be very hard to build up a sufficiently large dataset for a single athlete. Even if physical capacity tests are performed more often (for example monthly), training and capacity have to be monitored for a very long time (likely multiple years) before a dataset reaches a sufficient size to fit a useful model.

The concept of a personalized model of training load and physical capacity remains an interesting opportunity. Being able to optimize training programs on a personal level would provide a valuable tool for both athletes and coaches. In previous studies, several models of this relationship have been proposed, often with some trade-off between fitness and fatigue. However, these models were found to lack a measure of individualism [5]. To realize a useful personalized model, it is essential to find the right metric that quantifies physical capacity or performance. For such a metric to be suitable, it should be possible to regularly measure, so that the imbalance between the frequency of predictors from training load data and this metric as a target remains limited. Finding such a metric and a way of frequently measuring it remains a challenge that warrants further studies.

6 Conclusion

The primary aim of this thesis was to build a personalized regression model based on historical data from a single elite rower, in order to predict physical capacity.
With this historical data, we performed a feature engineering process with a large focus on metrics describing training load and feature aggregation in several windows. We used LASSO regression to fit a model on the resulting features. Using a nested cross-validation procedure, the best model was only able to explain 30% of the total variance ($R^2 = 0.3$). Furthermore, this model was found to be statistically insignificantly different from an intercept-only model.
Based on these findings, we conclude that our model was unable to accurately and reliably predict physical capacities based on historical data.

In conclusion, we were able to translate raw logbook data to aggregated features in several windows and thereby overcoming the problem of unequal and inconsistent sampling rates. However, our final model (fitted with 11 test samples) does not perform well enough to be considered reliable or useful.

Our method does show promise in identifying useful predictors. Furthermore, based on our experiment with different sample sizes, our model performance may quite possibly improve with increased sample sizes. For future research, it would be very interesting to find a way to gather more target data and retry our method with a larger dataset. Moreover, other features, time windows or aggregation methods could be considered.

References

[1] C. Rudin and K.L. Wagstaff. Machine learning for science and society. *Mach learn*, 2014(95):1–9, 2014.

[2] C. Foster, J.A. Florhaug, J. Franklin, L. Gottschall, L.A. Hrovatin, S. Parker, and C. Dodge P. Doleshal. A new approach to monitoring exercise training. *J Strength Cond Res.*, 15(1):109–115, 2001.

[3] C. Foster, E. Daines, L. Hector, and A.C. Snyder. Athletic performance in relation to training load. *Wisconsin medical journal*, 95(6):370–374, 1996.

[4] S. van der Zwaard, G. Weide, K. Levels, M.R.I. Eikelboom, D.A. Noordhof, M.J. Hofmijster, W.J. van der Laarse, J.J. de Koning, C.J. de Ruiter, and R.T. Jasper. Muscle morphology of the vastus lateralis is strongly related to ergometer performance, sprint capacity and endurance capacity in olympic rowers. *J Sports Sci.*, 36(18):2111–2120, 2018.

[5] J. Borresen and M.I. Lambert. The quantification of training load, the training response and the effect on performance. *Sports Medicine*, 39(9):779–795, 2009.

[6] S.L. Halson. Monitoring training load to understand fatigue in athletes. *Sports Med*, 44 Suppl 2:S139–S147, 2014.

[7] G.S. Ginsburg and H.F. Willard. Genomic and personalized medicine: foundations and applications. *Translational research*, 154:277–287, 2009.

[8] R. Spang. Diagnostic signatures from microarrays: a bioinformatics concept for personalized medicine. *Biosilico*, 1:64–68, 2003.

[9] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, second edition, 2009.

[10] F.M. Impellizzeri, S.M. Marcora, and A.J. Coutts. Internal and external training load: 15 years on. *International Journal of Sports Physiology and Performance*, 14(2):270–273, 2019.

[11] L.K. Wallace, K.M. Slattery, and A.J. Coutts. The ecological validity and application of the session-rpe method for quantifying training loads in swimming. *Journal of Strength and Conditioning Research*, 23:33–38, 2009.

[12] T.J. Scott, C.R. Black, J. Quinn, and A.J. Coutts. Validity and reliability of the session-rpe method for quantifying training in australian football: A comparison of the cr10 and cr100 scales. *Journal of Strength and Conditioning Research*, 27:270–276, 2013.

[13] M. Haddad, G. Stylianides, L. Djaoui, A. Dellal, and K. Chamari. Session-rpe method for training load monitoring: Validity, ecological usefulness, and influencing factors. *Front. Neurosci.*, 11:612, 2017.

[14] C. Foster. Monitoring training in athletes with reference to overtraining syndrome. *Med Sci Sports Exerc.*, 30(7):1164–1168, 1998.

[15] J. Gabbett, B.T. Hulin, P. Blanch, and R. Whiteley. High training workloads alone do not cause sports injuries: how you get there is the real issue. *Br J Sports Med*, 50:444–445, 2016.

[16] S. Williams, S. West, M.J. Cross, and K.A. Stokes. Better way to determine acute:chronic workload ratio? *Br J Sports Med*, 51:209–210, 2017.

[17] T. Soligard, M. Schwellnus, and J. Alonso et al. How much is too much? (part 1) international olympic committee consensus statement on load in sport and risk of injury. *Br J Sports Med 2016*, 50:1030–1041, 2016.

[18] N.H. Secher. The physiology of rowing. *Journal of Sports Sciences*, 1:23–53, 1983.

[19] M.J. Joyner and E.F. Coyle. Endurance exercise performance: the physiology of champions. *J Physiol*, 586:35–44, 2008.

[20] S.A. Ingham, G.P. Whyte, K. Jones, and A.M. Nevill. Determinants of 2,000m rowing ergometer performance in elite rowers. *Eur J Appl Physiol*, 88(3):243–246, 2002.

[21] A.W.J. Stevens, T.T. Olver, and P.W.R. Lemon. Incorporating sprint training with endurance training improves anaerobic capacity and 2,000-m erg performance in trained oarsmen. *Journal of Strength and Conditioning Research*, 29(1):22–28, 2015.

[22] M. Bourdin, L. Messonnier, J.-P. Hager, and J.-R. Lacour. Peak power output predicts rowing ergometer performance in elite male rowers. *Int J Sports Med*, 25(5):368–373, 2004.

[23] S.E. Riechman, R.F. Zoeller, G. Balasekaran, F.L. Goss, and R.J. Robertson. Prediction of 2000m indoor rowing performance using a 30s sprint and maximal oxygen uptake. *J Sports Sci*, 20(9):681–687, 2002.

[24] O. Bar-Or. The wingate anaerobic test: An update on methodology, reliability and validity. *Sports Medicine*, 4:381–394, 1987.

[25] N. Hofman, J. Orie, M. Hoozemans, C. Foster, and J. de Koning. Wingate test is a strong predictor of 1500m performance in elite speed skaters. *International Journal of Sports Physiology and Performance*, 12:1–17, 03 2017.

[26] A. Legaz-Arrese, D. Munguia-Izquierdo, E.L. Carranza-Garcia, and G.C. Torres-Davila. Validity of the wingate anaerobic test for the evaluation of elite runners. *Journal of Strength and Conditioning Research*, 25(3):819–824, 2011.

[27] Y. Hachana, A. Attia, S. Nassib, R.J. Shephard, and M.S. Chelly. Test-retest reliability, criterion-related validity, and minimal detectable change of score on an abbreviated wingate test for field sport participants. *Journal of Strength and Conditioning Research*, 26(5):1324–1330, 2012.

[28] T Rice. *7 x 4 min step test protocol*. Australian Institute of Sport, 2008.

[29] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–288, 1996.

[30] C.F. van Loan G.H. Golub. *Matrix Computations*, chapter 2.2. Johns Hopkins University Press Baltimore, 3rd edition edition, 1996.

[31] E.W. Weisstein. L1-norm. From Mathworld–A Wolfram Web Resource. http://mathworld.wolfram.com/L1-Norm.html. Accessed: 15-03-2019.

[32] D.M. McNeish. Using lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences. *Multivariate Behavioral Research*, 50:471–484, 2015.

[33] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann Statist*, 32(2):407–499, 2004.

[34] C.J. Willmott and K. Matsuura. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *CR*, 30(1), 2005.

[35] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning*. Springer, first edition, 2013.

[36] P. Refaeilzadeh, L. Tang, and H. Liu. *Cross-Validation*, pages 532–538. Springer US, Boston, MA, 2009.

[37] I. Tsamardinos, A. Rakhshani, and V. Lagani. Performance-estimation properties of cross-validation-based protocols with simultaneous hyper-parameter optimization. *International Journal on Artificial Intelligence Tools*, 24(5), 2015.

[38] S. Varma and R. Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 91, 2006.

[39] About python. https://www.python.org/about/. Accessed: 15-03-2019.

[40] The pandas project. https://pandas.pydata.org/about.html. Accessed: 15-03-2019.

[41] scikit-learn: machine learning in python. https://scikit-learn.org. Accessed: 15-03-2019.

[42] S. Nogueira, K. Sechidis, and G. Brown. On the stability of feature selection algorithms. *Journal of Machine Learning Research*, 18:1–54, 2018.

[43] A. Schneider, G. Hommel, and M. Blettner. Linear regression analysis. *Dtsch Arztebl Int*, 107(44):776–782, 2010.