Predicting bicycle facility usage through check-in and check-out data

Casper Antoon Slenders

ANR: 358550

Thesis submitted in partial fulfillment

of the requirements for the degree of

Master of Science in Communication and Information Sciences,

Master Track Data Science Business & Governance,

at the School of Humanities and Digital Sciences

of Tilburg University

Thesis committee:

dr. G.A. Chrupała

dr.ir. P.H.M. Spronck

Tilburg University

School of Humanities and Digital Sciences

Department of Cognitive Science & Artificial Intelligence

Tilburg, The Netherlands

December 2018

Preface

This thesis is written for the fulfillment of the master's program Data Science: Business and Governance at Tilburg University. In August 2018, the search for a thesis topic brought me to the company Movares, a consultancy and engineering company based in Utrecht. Together with the Dutch railway company, ProRail, the municipality of Utrecht and the Province of Utrecht, an interesting topic was set up for my master thesis project. I would like to express special gratitude towards my supervisors Simone Jorink and Marco van der Linden from Movares for providing domain knowledge and support during this project. Another thanks goes to Lovisa Lundgren who carried out observations that provided insight for this work. I also would like to thank Jorrit Visser and Caspar Muijsert from the NS. They provided data for this research and put forth interesting ideas. A special thank you to dr. Grzegorz A. Chrupała for his supervision and advice. Additionally, I would like to thank my friends, roommates and fellow students for their support during this hectic period. Finally, a big thanks to my family for their unconditional love and (financial) support.

Casper Slenders

Utrecht, December 2018

Abstract

Utilization of bicycles has grown rapidly over the past few decades, causing the demand of bicycle facilities to grow as well. Hence, it becomes increasingly important to study the patterns of the usage of the existing facilities in order to manage them more efficiently. The current study focuses on the prediction of bicycle facility usage for two storages around Utrecht central station and aims to enhance efficient management. Previous studies focused on the prediction of bicycle sharing systems and included several algorithms for forecasting their usage. Moreover, results of existing literature stressed the importance of weather features for predicting bicycle usage. The work presented in this thesis did not only focus on bicycle sharing systems, but also on the usage of common bicycle storages. Datasets obtained by the NS were used to train algorithms together with weather data of KNMI. A Multilayer Perceptron, Support Vector Machine and Random Forest were employed for this task. Results showed that a Random Forest model yielded best performance. Surprisingly, weather features did not yield high predictive power, whereas inventory features prove to be most important predictors. Understanding the different user types of the storages generated more insight in facility usage and led to suggestions for future research.

Table of contents

1. Introduction

1.1 Context

The Netherlands is famous for its bicycle culture. Bicycles are broadly used as a common means of transportation when commuting to work or visit relatives. The utilization of bicycles has grown rapidly over the past few decades, which has caused the demand for bicycle facilities to grow as well. The high demand for these facilities does not always correspond to the available space in the Netherlands, especially around railway stations. The size of bicycle facilities around railway stations is increasing, but due to limited space, expansion is reaching its boundaries while the demand is still growing. In addition, a latent demand exists. This means that a reasonable amount of people does not store their bicycles in a storage near a railway station, simply because they have experienced that it is difficult to routinely store them. (Jonkeren et al., 2018). Utrecht, one of the fastest growing urban areas in the Netherlands is currently building the largest bicycle storage in the world, called the Stationsplein storage. This storage is scheduled to be finished by 2019 and will have a capacity of 12.500 places. Nonetheless, upon completion it is expected that it still lacks enough space to accommodate the growing demand. For this reason, it seems crucial to investigate alternative ways to optimize the usage of bicycle facilities in such way that further increase in capacity is needed to a lesser extent.

The storages investigated in this study are multifunctional structures that offer storage places for thousands of bicycles. These storages offer a variety of services such as a bike sharing system (BSS) and a service point for the maintenance and repair of bicycles. The realization of these bicycle storages is a complex task, especially in crowded areas. Therefore, storage capacity expansion is expensive and takes several years to realize. Using the existing storages in a more efficient way can offer a solution for this issue and may solve the over-demand of these facilities. Over-demand could ultimately obstruct people in using bicycles and cause them to use other forms of transportation (Singla et al., 2015). A decrease in the usage of bicycles could not only affect this means of climate friendly transport, but also the general health of the population. According to the study of Fishman, Schepers, and Kamphuis (2015) the use of bicycles in the Netherlands prevents 6500 deaths annually.

A considerable amount of research has been done on predicting the demand in BSS and can be found in section 2. Previous studies found that weather conditions and hour of the day have shown to be the most influential factors to bicycle facility usage. Taking these influential factors into account, this current study uses regression models to predict the usage of bicycle storages for optimization purposes. The prediction of bicycle facility usage is executed by combining data from various sources (Weather data, Check-in and Check-out data, OV-fiets data).

Utrecht central station is the largest railway station in the Netherlands and processes more than 180,000 daily travelers (NS, n.d.). This number is growing every year and indicates the need for optimizing the usage of facilities. Multiple parties are involved in this project and share the interest of finding a solution for this ever-growing demand of bike storage space. Over the past few years, newly built facilities are completed near Utrecht central station and are built and maintained by the Dutch railway company (NS), ProRail and the municipality of Utrecht. The bicycle storages in Utrecht keep track of their users in various ways by gathering data about the people who check in and out and make use of the BSS. The generated data gives insight in the usage of different bicycle storages and is useful for data analysis.

1.2 Scientific and practical relevance

With increasing traffic density and the need for more sustainable means of transport, the usage of bicycles as form of transportation has become more popular in recent years. Consequently, a better understanding of bicycle facility demand and the possibility to predict its usage, results in a more efficient management of these facilities. Service providers are able to make an estimation about available shared bikes and parking spots in order to improve the management and coordination of the facilities. Given the large investments made in the construction of bicycle facilities, it is important that they are optimally utilized. For a large portion of travelers, bicycle storages are an important connector to other forms of transportation. Over-demand of these facilities affects user-experience and has negative side effects as they need to look for another bicycle storage. As mentioned above, it can ultimately push people to use other (less sustainable) means of transportation. Providing better regulated bicycle facilities enables a more efficient multi-modal transportation network and encourages people to use bicycles when commuting to work, having the aforementioned substantial health benefits for the Dutch population. Fishman, Schepers, and Kamphuis (2015) also proved that these health benefits could lead to economic advantages. Hence, providing good bicycle facilitites also has economical advantages in the long-term.

This work differs from previous approaches in that it seeks to predict bicycle facility usage for large combined facilities consisting of BSS and common bicycle storages. Different methods have been used to develop models for predicting BSS demand, but not in a combination as seen in Utrecht. Several studies have proven that a Random Forest model is useful for the prediction of bicycle usage and demand for BSS (Lozano, De Paz, Villarrubia González, Iglesia, and Bajo, 2018; Yang et al., 2016; Yin, Lee, and Wong, 2012), whereas Giot and Cherrier (2014) found that a Ridge regression performed best together with an Adaboost regression. Additionally, previous researches used different features for predicting BSS demand. Hence, it is interesting to investigate which models yield the best result for facilities as seen in Utrecht and examine multiple input features to clarify study results.

1.3 Research questions

This research aims to find tools that help optimize the usage of bicycle facilities. In order to optimize bicycle facilities, it must be investigated how well the usage of bicycle storages can be predicted. Therefore, the following main research question is composed:

Main research question:

- *To what degree can future bicycle facility usage be predicted for the secured bicycle storages of Utrecht central station?*

Sub research questions:

The bicycle storages of Utrecht central station that are used in this research consist of a common storage part and a bicycle sharing system (BSS) part. In order to answer the main question, insight is needed in the number of bicycles that enter the storages per day of the week and per different parts of the day. Patterns present in the data can be used for expectations of bicycle facility usage, and therefore provide a baseline model. Consequently, the first question is formulated as follows:

- *What is the average incoming and outgoing number of bicycles and shared bicycles for certain periods of time such as day of the week and part of the day?*

Moreover, exploring the effects of various features (e.g. temperature, precipitation) during peak hours and outside peak hours allows for a better understanding of important determinants. Hence, the second question is formulated as follows:

- *Which features yield the highest influence for the prediction of bicycle facility usage during peak hours and outside peak hours?*

It is expected that a distinction can be made between user types of the bicycle storages at Utrecht central station. These groups (e.g. commuters and recreationists) make use of the facilities at different periods (Faghih-Imani and Eluru, 2015), and are interesting factors to investigate since their impact on the occupancy at specific moments can be great. In order to investigate their impact, the following question is formulated:

- *To what extent does user type influence the usage of bicycle facilities around Utrecht central station?*

Different methods have been used to predict BSS demand. However, until now, these methods were not applied to a combination of a BSS and a common bicycle storage. Hence, in order to add scientific relevance, it is interesting to look at the best performing model for the situation as seen in Utrecht:

- *Which model yields the best results for the prediction of bicycle facility usage?*

1.4 Findings

The task for this study was to make predictions up to 8 hours in the future for the usage of the bicycle facilities. In order to define a baseline, a linear regression model was trained. For the current number of incoming bicycles in the Stationsplein storage, the baseline achieved a R^2 of 0.92, a MAE of 27.17, a RMSE of 47.48 and a RMSLE of 1.25. Moreover, the linear regression showed that, with a 95% confidence, bicycle facility usage depends on whether it is a weekday or not, and whether it is a peak hour or not. Three models were employed: a Multilayer Perceptron (MLP), a Support Vector Regressor (SVR) and a Random Forest Regressor (RFR), which all outperformed the baseline model. Among these algorithms, the RFR yielded the best performance with a RMSLE of 0.35 for the prediction of current incoming bicycles for both bicycle facilities studied in this work. The MAE yielded by the RFR for the prediction of the usage of the common bicycle storages did not exceed 15 for the Jaarbeursplein storage and 21 for the Stationsplein storage on the entire time horizon of 8 hours. For the prediction of BSS usage on the entire horizon, the MAE did not exceed 6 for the Jaarbeursplein storage and 5 for the Stationsplein storage. Inventory features used to integrate increasing and decreasing trends into the time series forecast, showed high predictive power. Surprisingly, weather features did not yield high predictive power for subsets made of peak hours and non-peak hours. A correlation matrix was constructed in order to understand user type influence. Activity side users in a BSS yielded the highest correlation with the total number of rented shared bicycles. For the common bicycle storage part, home side users make up the largest part among the user types.

2. Related work

Large railway stations are often located in urban areas, consisting of commercial and residential buildings, tourist attractions and shopping areas, attracting a large variety of users of Dutch Rail. Moreover, many commuters are travelling by train on a daily basis and thus routinely make use of the facilities being part of railway stations. According to the Dutch railway company, 43% of all rail passengers are using bicycles to get to a railway station. Today, the total number of available parking places for bicycles in the Netherlands is around 460,000 (NS, 2018). This number was 183,000 in 1985 and 279,000 in 1999 (Martens, 2007), demonstrating the enormous growth over the past few decades. In July 2018 the Netherlands Institute for Transport Policy (KiM) did research in order to gain insight in the bicycle parking problem and found a relation between population growth and the occupancy rates of bicycle storages around railway stations (Jonkeren et al., 2018). The growing population of urban areas in the Netherlands explains the persistence of the bicycle parking problem and stresses the importance of finding ways to optimize usage of these storages even more. In addition, Jonkeren et al. (2018) pointed that people preferably make use of large railway stations offering better facilities and connections to other trains instead of small railway stations closer to their point of departure. This behavioral pattern is putting even more pressure on large railway stations and their facilities.

2.1 Prior studies

2.1.1 Bicycle facilities

The bicycle facilities investigated in this study can be divided into two parts: a common bicycle storage part and a BSS part. The common bicycle storing concept is relatively new and requires the user to check-in when entering and check-out when leaving the facility. Analyzing this check-in data could generate insight in the usage of the facilities, which is lacking at the moment. The BSS part of the bicycle facilities consists of public transportation bicycles (OV-fiets), which are introduced in the Netherlands in 2003. Usage has more than doubled from 1.5 million users in 2014 to 3.1 million in 2017 (NS, 2015; NS, 2018), clearly indicating its growing popularity.

The usage of the bicycle facilities around railway stations is expected to vary throughout the day, showing a similar pattern during weekdays and a similar pattern during the weekends. Commuters dominate bicycle facilities during weekdays, demonstrating a large peak in usage in the morning and evening peak hours. Vogel et al. (2011) observed this pattern and found a peak in BSS usage between eight and ten in the morning, and a peak in the late afternoon hours. During weekends they found a distinct peak in the night, whereas the morning peak was absent. The study of O’Brien et al. (2014), Faghih-Imani et al. (2014) and Yin, Lee, and Wong (2012) showed similar patterns, indicating that BSS is commonly used by commuters during weekdays. The common bicycle storages in Utrecht may show differences in usage patterns compared to the BSS. The share of people making use of bicycles at the home side of their trip is much higher than the share of bicycles at the non-home side of their trip, also called the activity-end (Keijer and Rietveld, 2000; Martens, 2004). This can be addressed to the fact that most people store their bicycle at home and use them for access trips to railway stations. However, users in a BSS have a higher share at the activity-end (Jonkeren et al., 2018), suggesting that usage peaks in BSS will occur later during morning rush-hours and earlier during evening rush-hours than in common bicycle storages. The first sub research question: "*What is the average incoming and outgoing number of bicycles and shared bicycles for certain periods of time such as day of the week and part of the day?*" aims to look at the average incoming and outgoing number of bicycles and shared bicycles in the facilities near Utrecht central station, and will be used to investigate if this difference in peak-usage is present. Moreover, gaining insight in daily and weekly cycles of system activity could be useful for facility operators. Studying these cycles could help operators to alter their redistribution strategy in a more efficient way to optimize usage and minimize operational costs (O'Brien et al., 2014; Raviv, Tzur and Forma, 2013).

2.1.2 Prediction of bicycle facility usage

Prediction of bicycle facility usage is useful for multiple reasons and could help facility operators and users to optimize trip experience. Most research on predicting bicycle facility usage focusses on BSS. Giot and Cherrier (2014) aimed to predict BSS usage up to one day ahead. They used a dataset containing information about a BSS in Washington acquired over two years, whereas most other studies validated their results on datasets collected during a few weeks. A modified dataset was used with weather features, number of bikes available and time features such as day of the week and season. The researchers built additional delay features for the number of available bikes up to twenty-four hours before the time they made the prediction for. They proved that it is possible to predict BSS usage up to twenty-four hours in advance with the use of various regression systems. However, they did show an important issue of overfitting for many state-of-the-art regressors. Moreover, since cycling is an outdoor activity, it is prone to different weather conditions. The feature 'temperature' showed a big impact on the prediction in the study of Giot and Cherrier (2014), El-Assi, Mahmoud and Habib (2017) and Saneinejad, Roorda and Kennedy (2012), making this feature interesting for verification. Sub research question two aims to investigate: *"Which features influence the prediction of occupation of the bicycle storages most during peak hours and outside peak hours?"* Yang et al. (2016) proved that weather features have a great influence on the usage of BSS, indicating that people tend to make more use of bicycles when temperatures are higher. However, they did see differences in influence during peak hours, suggesting that users are less influenced by weather conditions when commuting to work. Gebhart and Noland (2014) also showed that unfavorable weather conditions reduces the usage of a BSS. Surprisingly, they found that a trip increase still can be seen when temperatures rise above 32.2 °C, as one would expect these temperatures to be uncomfortable for cycling. Moreover, they state that their results are expected to be applicable to general cycling, making it interesting for verification in this study.

Yang et al. (2016) proposed a bicycle mobility model and devised a prediction for BSS traffic. They used a dataset from a large BSS network consisting of over 103 million trip records. Predicting check-out behavior is done using a Random Forest Model based on historical shifts of bicycles, corresponding time and weather data. In their approach, they evaluated overall performance of check in estimation for the following 30 minutes. For the training-set, the researchers used the first 20 days of each month, the latter was used for the test-set. They showed that less people made use of public bicycles in rainy days and that workdays yielded better prediction results than weekends.

Yin, Lee, and Wong (2012) used historical and meteorological data in order to predict usage of the Washington BSS network for a given hour grounded on the conditions of the hour. They found that the RF method performed best in terms of prediction accuracy and training time. Besides, they found that the predicition problem in a BSS network is higly non-linear. Lozano, De Paz, Villarrubia González, Iglesia, and Bajo (2018) made suggestions for a system that visualizes BSS usage and presented predictive tools. For their study, they also developed a RF model that was used to predict demand for BSS. This regression model outperformed other algorithms such as an Extra Tree Regressor and a Gradient Boosting Regressor. Dias, Bellalta and Oechsner (2015) made predictions of different stations in the Barcelona BSS network. They concluded that a RF model is good for making forecasts up to 48 hours ahead. In their study, they were able to predict the BSS stations' statuses correctly nearly half of the times, up to two days before they occurred. In their recommendations, they state that the use of more observations could potentially help building more powerful models.

In the analysis of Sarkar, Lathia and Mascolo (2015) they used 996 stations included in a dataset for a period of 4.5 months. Analysis was done using different state-of-the art algorithms on a large dataset. For their study, they computed forecasts at different fixed points in the future. The study concluded that predictions for smaller BSS systems yielded better results than bigger systems and that occupancy levels in small BSS systems for 6 minutes in the future can be predicted just as good as occupancy levels 48 minutes ahead. They recommend using additional datasets such as weather data, in order to possibly reveal the impact of favourable/adverse conditions on BSS usage.

In the light of the third sub research question: "*To what extent does user type influence the occupation of bicycle storages around Utrecht central station?"*, it is interesting to look at users of bicycle facilties. Yet, little is known about user types and their motivations for usage. Bachand-Marleau, Lee and El-Geneidy (2012) showed that location is crucial for the encouragement of BSS users. They indicate that a higher number of stations near residential areas result in more users, suggesting that home-side users are an important group. According to Faghih-Imani and Eluru (2015), a clear distinction can be made in the type of user of a BSS by observing the period or time of day of usage.

2.1.3 Algorithms employed in prior studies

Multiple state-of-the-art algorithms are used for the prediction of bicycle usage in bicycle facilities, both regression and classification tasks have been applied. The last sub research question: *"Which model yields the best results for the prediction of bicycle storage occupation?"* aims in looking at the best performing model for the prediction task of this study. Since most previous studies focused solely on the prediction of BSS usage and demand, exploring ways to predict common bicycle storage usage could offer interesting insights. The three models considered in this study are a Random Forest Regressor (RFR), a Support Vector Regressor (SVR) and a Multilayer Perceptron (MLP). These models all have been applied in previous research and shown their effectiveness for the prediction of BSS usage. The problem can be specified as a regression task which takes various input features and constructs a continuous outcome for the number of incoming and outgoing bikes.

The RFR is widely applied in previous research (Dias, Bellalta and Oechsner, 2015; Lozano, De Paz, Villarrubia González, Iglesia, and Bajo, 2018; Yang et al., 2016; Yin, Lee, and Wong, 2012) and proved to perform best amongst all models used in these researches. The SVR is a model that has been applied by Giot & Cherrier (2014) and Yin, Lee, and Wong (2012) for the prediction of BSS usage. Giot and Cherrier (2014) found that the SVR performed well on the training dataset but showed bad results on the validation set due to over-fitting. Moreover, in the study of Yin, Lee, and Wong (2012) an SVR yielded the best results after an RFR with a RMSLE of 0.33, whereas the RFR had a RMSLE of 0.31. The researchers did state that its performance could be improved by optimizing the model parameters. The last model considered in this study is the MLP and is used in previous research for the prediction of BSS usage (Sarkar, Lathia, and Mascolo, 2015; Zhou, Wang, Zhong and Tan, 2018). In the study of Sarkar, Lathia, and Mascolo, (2015), state-of-the-art algorithms did not perform as expected, whereas the MLP did yield acceptable error scores. Taking into account all of the previous studies, the RFR seems to be the best performing model for the prediction of bicycle usage. However, other studies found that an SVR and an MLP are capable of predicting bicycle facility usage as well, making them interesting for employing in this study.

2.2 Current work

Previous research focused solely on the prediction of bicycle usage in BSS. Hence, this study seeks to predict the usage of not only BSS facilities, but will additionally focus on the prediction of common bicycle storage usage. Since data gathering for the storage of common bicycle storages is relatively new, this study will seek to obtain interesting insights from the data provided, in order to compare results with BSS usage prediction. This study will therefore examine if previous used approaches can be applied to common bicycle storages as well. Moreover, by the use of meteorological features, the current work will look at the applicability of these features for the prediction of bicycle usage. Gebhart and Noland (2014) demonstrated that weather features yield a high impact on the usage of bicycles in a BSS and expressed the expectation that these results would be applicable to general usage of bicycles as well. Hence, by using hourly weather data collected by the Royal Netherlands Meteorological Institute (KNMI), verification of this statement will be examined. In addition, Sarkar, Lathia, and Mascolo (2015) employed an MLP for their study but relied only on historical data but did not add weather data. This current study will also employ an MLP and will additionally add meteorological features in order to examine their influence on the prediction of bicycle facility usage. Combining data of BSS and common bicycle storages will additionally allow for interesting insights in usage patterns, since it is expected that patterns vary between usage in BSS and in common bicycle storages. It is assumed that peaks in BSS usage occur at different moments since it is more dominantly used by activity-end users (Jonkeren et al., 2018). The practical relevance obtained from this, is that it allows for redistribution strategies and a better alignment of free bicycle storage space.

The datasets used in this study cover two large bicycle facilities in Utrecht: the Jaarbeursplein storage and the Stationsplein storage. These facilities offer storage for thousands of bicycles and contain 922 and 826 shared bicycles respectively (NS, n.d.). The data concerning the common bicycle storages are gathered by a check-in system. Also, the data concerning the public transportation bicycles (OV-fietsen) are collected by a check-in system. Furthermore, hourly weather data collected by the KNMI of station de Bilt is used. A full description of the datasets used can be found in section 4.1. With the use of these datasets, answers will be given to the research questions.

3. Methods

This chapter describes the models used for answering the research questions. As described in the previous chapter, a Random Forest Regressor, a Support Vector Regressor and a Multilayer Perceptron will be employed and tested for their applicability in this current task.

3.1 Random Forest Regressor

The first model considered is the Random Forest (RF) and can be used for both classification as regression tasks. An RF is a flexible machine learning algorithm that produces great results most of the time. Since this study is dealing with a Regression task, RF becomes RFR, referring to a Random Forest Regressor. An RFR is one of the most effective machine learning models for predictive tasks. The model is formed by building a number of random trees on different bootstraps of the training set, which increases diversity in the forest and leads to more robust predictions. The mean of the numerical outputs of all trees is the regression value (Breiman, 2001). One of the advantages of the Random Forest model is that it can handle many input variables without the need of variable deletion. Moreover, an RFR provides straightforward methods for feature selection, which can be useful for answering the second sub research question. One of the drawbacks of an RFR is its interpretability and its model size, which could make it slow to evaluate.

3.2 Support Vector Regressor

The Support Vector Machine (SVM) is an algorithm used for supervised learning problems which can be employed for classification and regression challenges. An SVM is used to find a hyperplane in an n-dimensional space (where n is the number of features) that classifies the data points distinctly. An SVM used for regression is called a Support Vector Regressor. The SVR attempts to search a subset of samples within the training set in order to compute the regression with them. Furthermore, in order to evenly penalize misestimates, SVR uses a symmetrical loss function for training (Awad and Khanna, 2015). An advantage of the SVR is that it is defined by a convex optimization problem, so no local minima exists. Another advantage is that it scales relatively well to high-dimensional data. However, a serious problem with an SVR is the limitation in speed and size for training and testing.

3.3 Multilayer Perceptron

The last model considered in this study is a Multilayer Perceptron (MLP). An MLP is a deep, artificial neural network that is composed of an input layer, an output layer and an arbitrary number of hidden layers in between which function as the computational engine of the MLP. MLPs are often used for supervised learning problems and are able to model highly non-linear functions. Scaling the variables of your data should always be done since it optimizes the performance results of the MLP, this is also called normalization. One major drawback is that an MLP is only capable of predicting stationary time-series (Koskela, Lehtokangas, Saarinen and Kaski, 1996). while the data of the bicycle facilities in Utrecht may show seasonal differences. In contrast to the SVR, an MLP algorithm can encounter difficulties when having to deal with local minima.

4. Experimental setup

In this section, the experimental setup of this study is described. The first part of this chapter will give a description about the used datasets provided by external parties and the open data used. In subsection 4.2, a description of the exploratory data analysis is given. The next subsection covers the steps taken to pre-process the data. Section 4.4 contains the experimental procedure which explains the algorithms that are used and which parameters are chosen. Subsection 4.5 and 4.6 will cover the implementation of the algorithms and the evaluation criteria used for testing model performance respectively. Codes needed to reproduce this work can be found in a specific GitHub repository[1].

4.1 Datasets

4.1.1 NS - common bicycle storages

The data used for training the algorithms is provided by multiple parties. Four datasets provided by the Dutch railway company (NS) are used. Two of these datasets contain data of users checking in and checking out in two major bicycle storages near the Utrecht central railway station. The first storage is called the Jaarbeursplein storage and covers data of a period from January 2017 to October 2018. The second storage is called the Stationsplein storage and covers data from its opening on August 7, 2017 to August 10, 2018. When a user checks in, no personal information is registered, making this dataset anonymous. Table 1 presents an overview of the two datasets of the common bicycle storages provided by the NS with the period it covers and its features.

Table 1

Datasets concerning the common bicycle storages provided by the NS

Location	Period	Features
Jaarbeursplein	01/01/2017 – 30/09/2018	*Name of storage, starting year, starting month, start date, starting hour, starting time, number of parked bikes, storage time, end year, ending month, end date, end hour and end time.*
Stationsplein	07/08/2017 – 10/08/2018	*Name of storage, starting year, starting month, start date, starting hour, starting time, number of parked bikes, storage time, end year, ending month, end date, end hour and end time.*

[1] https://github.com/u358550/Thesis

The features in the datasets mainly comprise the time of arrival and leaving of the users. The feature 'starting year' means the year of arrival of a user and 'starting month' means the month of arrival and so on. The feature 'ending year' means the year of leaving of a user. The same holds for the features 'ending month', 'ending date' and so on. Moreover, the exact storage time in seconds is included as well as the number of bicycles that entered the facility in the same minute.

4.1.2 NS – shared bicycles

In addition, two datasets were used consisting of shared bicycle (OV-fiets) rentals in the Jaarbeursplein and Stationsplein storage in Utrecht. The Jaarbeursplein storage covered a period from January 2017 to October 2018, while the Stationsplein storage covered a period from its opening in august 2017 to October 2018. The datasets included information about the rental location and return location as well as the exact time of rental and return on minute-level. Furthermore, a feature is present in these datasets concerning the number of rentals in the same minute. Table 2 presents an overview of the two datasets about shared bicycles provided by the NS with the period it covers and its features. The features in the datasets mainly comprise the time of rental and return of a shared bicycle. The feature 'starting date' means the date a shared bicycle was rented. The feature 'end date' means the date a shared bicycle was returned. Furthermore, the number of shared bicycles that were rented in the same minute is included. The location of rental and return are also included in this dataset as they may differ. However, in this dataset, the location of rental always corresponds to the location of return.

Table 2

Datasets concerning the shared bicycles provided by the NS

Location	Date	Features
Jaarbeursplein	01/01/2017 – 30/09/2018	*Location of rental, starting date, starting hour, starting time, number of rentals, return location, end date and end time.*
Stationsplein	07/08/2017 – 30/09/2018	*Location of rental, starting date, starting hour, starting time, number of rentals, return location, end date and end time.*

4.1.3 KNMI

Lastly, hourly weather data collected by the Royal Netherlands Meteorological Institute (KNMI) was used for the extraction of interesting features. These features are used to improve the prediction of BSS usage and common bicycle storage usage. The KNMI collects data at multiple weather stations across the Netherlands and makes it publicly available (KNMI, n.d.). For this work, hourly weather data of station De Bilt from January 2011 to October 2018 was used. Table 3 provides an overview of the dataset. The dataset consists of 24 features that are all collected on an hourly basis. It is important to note that some features, such as average wind speed, represents the average windspeed in that specific timeframe, while other features, such as temperature, represent the temperature at that time.

Table 3

Dataset KNMI station De Bilt

Period	Features
01/01/2011 – 01/10/2018	*Station number, date, hour, wind direction, average wind speed, highest wind gust, temperature, min. temperature, dew point temperature, sunshine duration, global radiation, precipitation duration, precipitation amount, atmospheric pressure, horizontal visibility, cloud cover, Relative atmospheric humidity, weather code, weather code observation, fog, rain, snow, thunder, ice formation.*

4.2 Exploratory Data Analysis

This section provides insight into the datasets used and summarizes its main characteristics. After a combination of datasets, data about the shared bicycles of the NS contained 20111 rows with 90 features for the Stationsplein storage and 30585 rows with 90 features for the Jaarbeursplein storage. The common bicycle storage dataset contains 17649 rows with 90 features for the Stationsplein storage and 30585 rows with 90 features for the Jaarbeursplein storage. No values were missing in both datasets.

In Figure 1, a clear trend can be seen in the usage of the Jaarbeursplein storage. During weekdays, a large peak of incoming bicycles is present during the morning rush hours, while a large peak of outgoing bicycles is present during the evening rush hours. This pattern indicates the storage's usage for commuting purposes during weekdays. During weekend days, a different pattern is present. The number of incoming bicycles is increasing until 12 o 'clock in the afternoon and then slowly decreases as it is reaching its minimum just after midnight. The number of outgoing bicycles during weekends increases until midnight and reaches its minimum in the morning hours. This pattern indicates the storage's usage for recreationists, such as tourist, concertgoers and people who go on a daytrip. An almost identical pattern can be found for the Stationsplein storage.

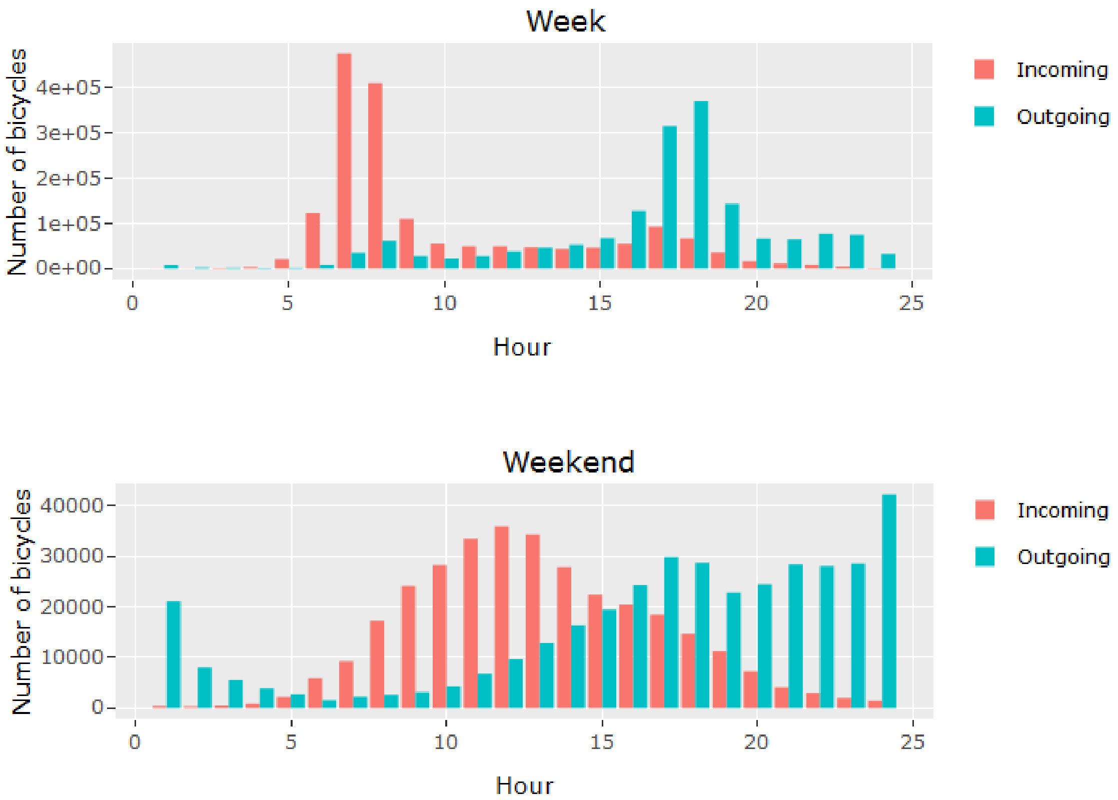

Figure 1. Total number of incoming and outgoing bicycles in the Jaarbeursplein storage during the week and weekend on hourly level

The usage of shared bicycles shows a similar pattern. As can be seen in Figure 2, a large peak in rentals is present during morning rush hours on weekdays, while a large peak in returns is present during evening rush hours. However, when a shared bicycle is rented, it is leaving the bicycle storage, while incoming bicycles are filling the storage. This means that patterns between the common bicycle storages and the BSS are inversely proportional. Hence, the pattern found in the data of shared bicycles indicates that activity-side users contribute largely to BSS usage in these storages while home-side users are a large proportion in the common storage part. Visualizing the patterns demonstrates that usage peaks in BSS occur later during morning rush hours and earlier during evening rush-hours than common bicycle storages, confirming the results of Jonkeren et al. (2018). However, an incoming bicycle cannot be linked to an outgoing bicycle based on visualizations of the data, so further insight is needed and discussed in section 5.4.

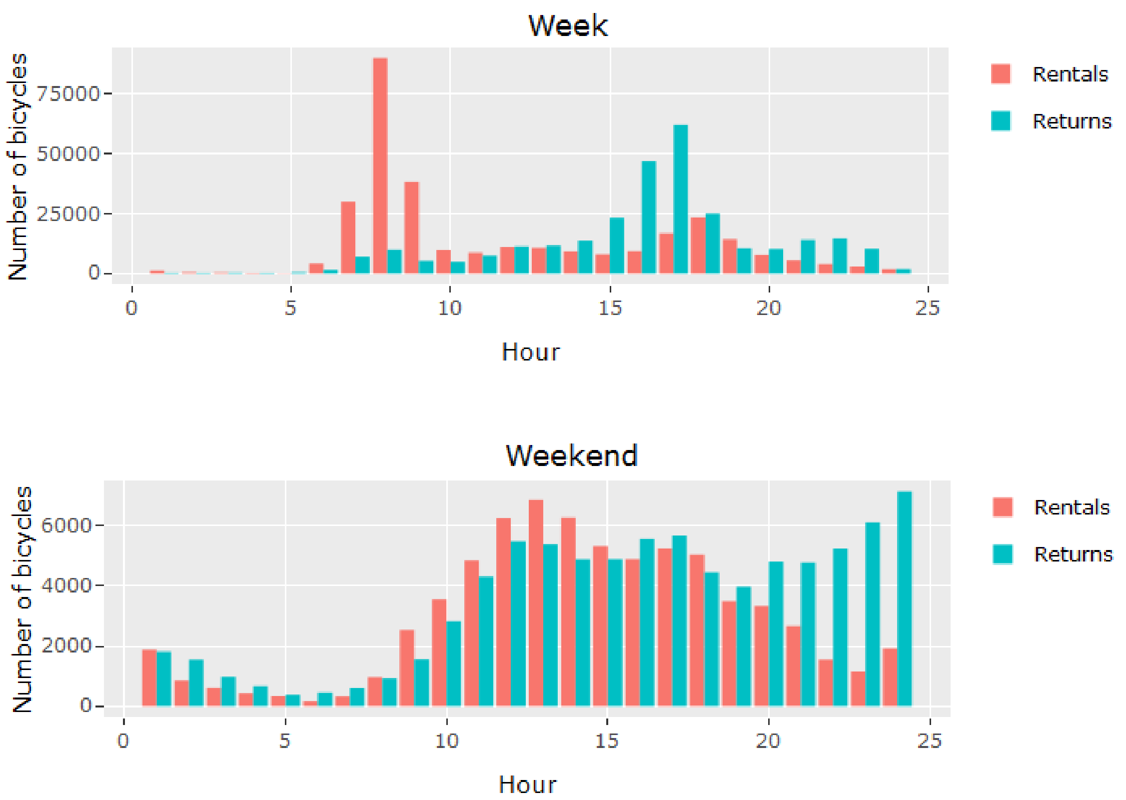

Figure 2. Total number of rentals and returns for shared bicycles in the Jaarbeursplein storage during the week and weekend on hourly level

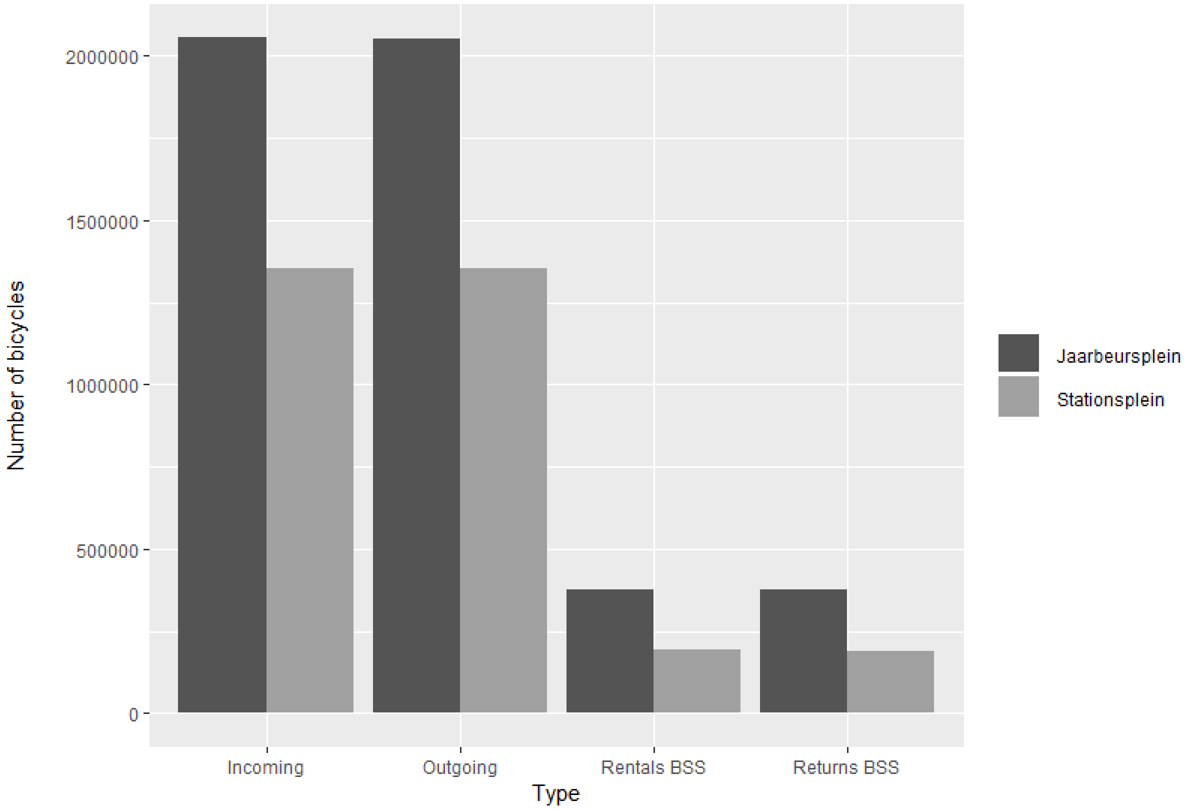

Figure 3. Total number of bicycles and shared bicycles per storage

Figure 3 shows that there is an imbalance between the datasets, meaning that the number of records is not equally represented. Due to a difference in time-coverage of the datasets, this imbalance might cause that predictions results will vary.

4.3 Pre-processing

After data collection from KNMI and NS, all data had to be cleaned and pre-processed, which is done in R. This section describes the steps that were taken and how the final features were created and selected for training the models.

4.3.1 NS

The data provided by NS concerned a total of 4 datasets of two storages: the Jaarbeursplein storage and the Stationsplein storage. Two datasets contained data about the check-ins and check-outs of the common storage part, while the other two datasets contained data about the BSS part of the storages. In order to make predictions for the usage of the storages, the datasets had to be converted into half-hourly data. Before converting the data, interesting features were retrieved from the raw data. Each record in the raw data for the common storage part contained an exact time of storage, which is useful for answering sub research question 3. Six user types were formed by using the time of check-in, time of check-out and storage time. Table 4 provides an overview of these user types. The user types are defined in conjunction with Movares and the criteria are based on observations of the bicycle facilities that demonstrated clear patterns of people entering and leaving at certain moments. A home-side user is using a bicycle to access the railway station at the home-end whereas an activity-side user uses a bicycle at the non-home-end of the trip (Keijer and Rietveld, 2000; Martens, 2004). For the data of the BSS part of the storages, time of rental could be retrieved by calculating the difference between the start time of rental and time of return. After dummy coding these user types per record, the datasets were converted into half-hourly data. These time-series datasets hold information about the incoming and outgoing number of bicycles, as well as the rentals and returns of shared bicycles per 30 minutes. Furthermore, incoming and rented bicycles were added per user type for the time-series data.

No missing values were present in the original datasets of NS. After conversion to the time-series data, missing values occurred as no check-ins or rentals were registered for certain time units, these were set to 0. Moreover, date features were added such as weekday, weekend day and peak hour. These features were transformed into dummies in order to make them dichotomous, where 1 equals an occurrence of a category and 0 otherwise. Furthermore, the history of the incoming and outgoing bicycles and the rentals and returns of the shared bicycles were constructed into features, called inventory features. A total of 16 inventory features were constructed resulting in an inventory of bicycle usage for 8 hours in the past. In this way, the employed models allow for integrating increasing or decreasing trends in usage into the forecasts. In addition, the forecast horizon for this study is 8 hours, so features were included in the datasets for the usage of the storages up to 8 hours ahead, functioning as targets for the prediction.

Table 4

Definitions of user types

User type	Common bicycle storage	Bicycle sharing system
Short-use	ST*:	ST:
	ET*:	ET:
	Storage time: ≤ 5 hours	Rental time: ≤ 5 hours
Home-side	ST: between 5:00 and 10:00	ST: between 13:00 and 23:00
	ET: between 13:00 and 23:00	ET: between 7:00 and 10:00
	Storage time: between 5 and 15 hours	Rental time: between 5 and 15 hours
Activity-side	ST: between 13:00 and 23:00	ST: between 5:00 and 10:00
	ET: between 7:00 and 10:00	ET: between 13:00 and 23:00
	Storage time: between 5 and 15 hours	Rental time: between 5 and 15 hours
Long-term	ST:	ST:
	ET:	ET:
	Storage time: ≥ 15 hours	Rental time: ≥ 15 hours

User type	Common bicycle storage	Bicycle sharing system
Evening-users	ST: between 14:00 and 18:00 ET: between 21:00 and 01:00 Storage time: ≥ 5 hours	ST: between 14:00 and 18:00 ET: between 21:00 and 01:00 Rental time: ≥ 5 hours
Other	ST: Else ET: Else Storage time: Else	ST: Else ET: Else Rental time: Else

* ST = Starting time * ET = Ending time

4.3.1 KNMI

The data retrieved from the KNMI concerned a dataset of hourly measurements at weather station De Bilt in the province of Utrecht. All rows that did not correspond to dates present in the data of NS were omitted in the KNMI dataset. Furthermore, 8 interesting features were retrieved from the dataset: average windspeed, temperature, sunshine duration, duration of precipitation, amount of precipitation, cloud cover, humidity and whether it rained. No missing values were present in the data of the KNMI. Furthermore, temperatures in the KNMI dataset had to be divided by 10 before representing degrees Celsius. Hulot, Aloise and Jena (2018) concluded that using previous hour features for the weather had no real impact on the prediction, therefore, these were not included.

After merging data of the NS and KNMI, the final datasets used for the prediction of bicycle storage usage contained a total of 90 features.

4.4 Experimental procedure

After preprocessing and merging of the datasets, the data was split into a training and test set. Since the aim of this study is predicting most recent bicycle facility usage, data was split by date, where the most recent months functioned as test set. Since the datasets covered different time periods, the last 30% of the datasets were used for testing, the remaining 70% of the datasets were used for training the models. This means that for the BSS dataset of the Jaarbeursplein storage, training set approximately covered the months January 2017 up to and including February 2018, whereas the test set approximately covered the months March 2018 up to and including September 2018. In order to transform this problem into a supervised learning problem, inventory features together with weather features and time features are used as input to predict the observations for the different fixed points in the future time horizon. This approach is used for both training and testing the applied algorithms.

Consequently, a baseline was defined, for which a Multiple linear regression was trained using all features present in the datasets, where the future variables functioned as targets. After employing the baseline model, algorithms for predicting the usage of the bicycle facilities were trained. The regression algorithms that were used in this study are a Random Forest Regressor, a Support Vector Regressor and a Multilayer Perceptron. In order to increase the performance of the regression algorithms, grid search was applied to find the most optimal hyperparameters. 3-fold cross validation was used for optimization purposes. One of these models is an MLP, which is a neural network. Since neural networks perform better with normalized input variables, MinMaxScaler from the scikit-learn library in Python was used. After training the algorithms with the best performing hyper-parameter settings, evaluating was done on the test set. Model performance was evaluated using different error metrics which are discussed in section 4.6. Section 4.4.1 will discuss the optimal hyperparameter settings found by grid search.

4.4.1 Hyperparameter optimization

Grid search was used to find the most optimal hyperparameters for the SVR, MLP and RFR. Table 5 shows the most optimal hyperparameter settings for the selected algorithms.

Table 5

Optimal hyperparameters found by grid search

Algorithm	Optimal hyperparameter-settings
Random Forest Regressor	n_estimators=1000, max_depth = 100, max_features = 20, min_samples_leaf= 3, min_samples_split = 8
Support Vector Regressor	kernel = ‘poly’ , C = 10, gamma = 1
MLP for regression	hidden_layer_sizes= 250, activation = ‘relu', solver = 'lbfgs', learning_rate = “constant”

The following hyperparameters are considered for the RFR: n_estimators (100, 500, 1000), max_depth (80, 100), max_features (20, 30, 35), min_samples_leaf (3, 4, 5) and min_samples_split (8, 10, 12). For the SVR, the following hyperparameters are considered: kernel (linear, poly, rbf, sigmoid), C (0.001, 0.01, 0.1, 1, 10) and gamma (0.001, 0.01, 0.1, 1). Finally, the considered hyperparameters for the MLP are: hidden_layer_sizes (50, 150, 250), activation (identity, relu), solver (lbfgs, sgd) and learning_rate (constant, adaptive).

4.4.2 Feature importance

One of the goals of this study is to find features with the biggest influence on bicycle storage usage during peak hours and outside peak hours. In order to test feature importance for these different periods, subsets of the data are made for 30-minute periods within peak hours and 30-minute periods outside peak hours. Selected algorithms for this study provide straightforward methods for looking at feature importance. The RFR model has an attribute “feature_importance” derived from the scikit-learn library in python that allows for feature importance testing. With applying this attribute, a list will be returned where high values reflect the most important features, which can be used to answer sub research question 2.

4.5 Implementation

Exploratory data analysis and pre-processing was performed using the programming language R, version 3.5.1 with RStudio (Version 1.1.456). Modelling was done using programming language Python, version 3.7.0. Table 6 provides an overview of the used libraries and packages for R and Python.

Table 6

Used libraries and packages in R and Python

Language	Library/package
R	DataExplorer, data.table, dplyr, forcats, ggplot2, Graphics, lubridate, plotly, plyr, reshape2, scales, tidyr, timeDate, timeSeries,
Python	math, matplotlib, numpy, pandas, scipy, seaborn, sklearn

4.6 Evaluation Method

To evaluate performance of the models, different metrics are applied. The R-squared (R^2), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Root Mean Squared Logarithmic Error (RMSLE) are used to evaluate performance for each of the chosen models. MAE is the average of all absolute errors, RMSE is the root of the means of the errors squared and RMSLE is the root mean square of the logarithm of the ratio between the predicted values and the actual values. The metrics that are applied all show the error of the prediction. MAE gives a number that can be used to determine the number of incorrectly predicted bicycle storage visits on average. RMSE penalizes predictions more that are far from the true value and RMSLE gives more penalties to under-prediction than it does to over-prediction, which can be useful in the sense that it is better to over-prepare for usage peaks in the bicycle storages.

5. Results

In this section, the performance of the employed models is presented. The aim of this study is to provide clarity in the degree to which bicycle storage usage can be predicted. With the use of three regression models, predictions are measured on a time-horizon of 8 hours. This prediction task was executed for two bicycle storages in Utrecht for the common bicycle storage part and the BSS part of the storages. Furthermore, grid search is applied to find the most optimal hyper-parameter settings for the algorithms and feature contribution is tested.

5.1 Baseline

In order to define a baseline for this study, a linear regression model is used. All features are trained in this baseline model except for the features that reflect the future time-horizon and the incoming user type features. Table 7 shows the performance of the baseline model for the prediction of the incoming and outgoing number of bicycles as well as the rentals and returns of the shared bicycles in the Stationsplein storage.

Table 7

Performance of baseline model for the Stationsplein storage

Type of usage	Forecast horizon	R^2	MAE	RMSE	RMSLE
Rental	t = 0	0.74	4.90	9.20	0.76
	t + 240	0.24	8.34	15.83	1.05
	t + 480	0.53	6.95	12.39	1.00
Return	t = 0	0.82	4.11	6.19	0.68
	t + 240	0.74	5.40	7.41	0.89
	t + 480	0.73	5.42	7.50	0.88
Incoming	t = 0	0.92	27.17	47.48	1.25
	t + 240	0.35	79.54	138.31	1.91
	t + 480	0.69	44.49	95.09	1.50
Outgoing	t = 0	0.88	25.67	48.77	0.93
	t + 240	0.94	22.75	32.88	1.00
	t + 480	0.94	23.30	33.24	0.93

The results of the baseline show that there is a difference in prediction accuracy among the features that were tested for prediction. For the current number of incoming bicycles in the

Stationsplein storage, the baseline achieved a R^2 of 0.92, a MAE of 27.17, a RMSE of 47.48 and a RMSLE of 1.25. Moreover, as can be seen in the results of the baseline model, t + 480 generally has lower error scores than t + 240 and deviates from expectations.

The linear regression that is constructed as baseline can be used to answer the first sub research question. As seen in the exploratory data analysis, usage peaks occur during the morning and the evening rush hours in the bicycle facilities around Utrecht central station for both the common storage part and the BSS part. To test significance per feature for the incoming and outgoing number of bicycles and the rentals and returns of shared bicycles, results of the regression model are evaluated. The results are shown in appendix A. The table shows that, with a 95% confidence, the number of incoming bicycles in the Stationsplein storage depends on whether it is a weekday or not and whether it is a peak hour or not. This is also applicable for the number of outgoing bicycles and the rentals and returns of shared bicycles in the same storage. Again, similar results were achieved for the Jaarbeursplein storage, indicating that usage of both storages is related to rush hours and whether it is a weekday or not.

5.2 Performance of the models

The Random Forest Regressor is widely applied for the prediction of BSS usage in previous research and proved to perform best amongst all models employed in these researches. This study seeks clarity in the application of the RFR for BSS usage and if it is effectively applicable to common bicycle storage usage as well. Moreover, a Multilayer Perceptron and a Support Vector Regressor are trained and evaluated. Figure 4 shows the performance of the different models on the set forecast horizon of 8 hours when evaluated on the test set. The RMSLE of the baseline deteriorates immediately but improves when the forecast horizon increases. The RFR clearly shows the best results for all graphs, demonstrating its applicability for the prediction of both BSS usage as well as common bicycle storage usage. When predicting the current number of incoming bicycles in the Stationsplein storage, the RFR has a RMSLE of 0.35, while an MLP and an SVR yield 0.64 and 0.67 respectively. The RMSLE of the RFR slightly increases over the time horizon and reaches 0.46 for the incoming number of bicycles at t + 480, whereas the MLP and the SVR yield a RMSLE of 0.78 and 0.84. A similar trend is present for the rentals and returns of shared bicycles in the Jaarbeursplein storage. The RFR again yields the lowest RMSLE of 0.48 for the current number of rentals, whereas the MLP and SVR produce a RMSLE of 0.65 and 0.69 respectively. When looking to the other metrics, the overall performance of the RFR is yielding the best results, proving its applicability for the prediction of bicycle facility usage. Moreover, the course of the error over the time horizon shows a stable pattern for the RFR, demonstrating its robustness, contrary to the baseline model that showed deviating results. An overview of all error metrics for the incoming and outgoing number of bicycles as well as the rentals and returns of shared bicycles for both storages can be found in appendix B.

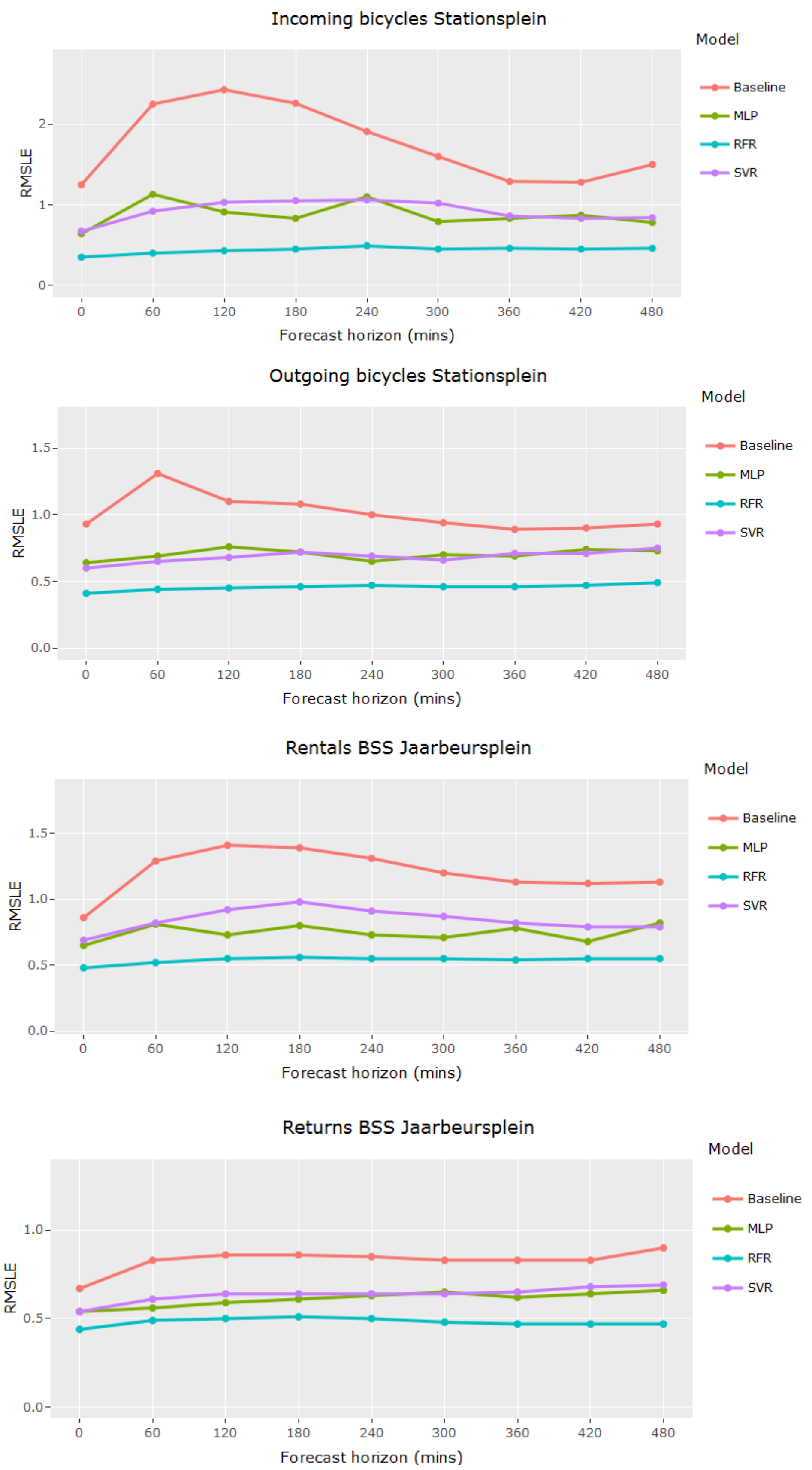

Figure 4. Performance of models on set time horizon

5.3 Feature importance

The Random Forest model provides a straightforward way to examine feature importance for the prediction of bicycle facility usage. Using the attribute ‘feature_importance’ derived from the scikit-learn library in python, a list will be returned with importances denoted on a 0 to 1 scale, where higher feature importances are closer to 1. Table 8.1 and table 8.2 provide the top 10 features with the highest importances for the prediction of the incoming number of bicycles and number of rented shared bicycles during peak hours and outside peak hours for the Jaarbeursplein storage. Feature importances for the Stationsplein storage yielded similar results. As was to be expected, inventory features are important in predicting bicycle facility usage in both peak hours and outside peak hours. An interesting observation is that weather features are almost not present, indicating that usage of the Jaarbeursplein storage and Stationsplein storage is scarcely dependent to the weather. This applies to both peak hours and outside peak hours. The RFR also shows that time features such as ‘HH’ and ‘Halfuurnum’ (Hour and Half-hour) are important factors for the prediction of bicycle facility usage.

Table 8.1

Feature importance for the prediction of incoming bicycles in and outside peak hours for the Jaarbeursplein storage derived from the RFR

Peak		**Outside peak**	
feature	importance	feature	importance
t_min30_Incoming	0.27	t_min30_Incoming	0.43
Halfuurnum	0.23	t_min60_Incoming	0.18
HH	0.19	t_min330_Outgoing	0.08
t_min60_Incoming	0.06	t_min300_Outgoing	0.05
t_min240_Incoming	0.06	t_min360_Outgoing	0.03
t_min210_Incoming	0.04	Halfuurnum	0.03
t_min90_Incoming	0.03	t_min90_Incoming	0.03
t_min150_Incoming	0.03	HH	0.02
t_min120_Incoming	0.02	t_min210_Outgoing	0.02
t_min150_Outgoing	0.01	t_min270_Outgoing	0.01

Table 8.2

Feature importance for the prediction of rented shared bicycles in and outside peak hours for the Jaarbeursplein storage derived from the RFR

Peak		**Outside peak**	
feature	importance	feature	importance
t_min30_Rentals	0.41	t_min30_Rentals	0.38
t_min420_Returns	0.10	t_min60_Rentals	0.17
t_min60_Rentals	0.10	t_min90_Rentals	0.07
Halfuurnum	0.07	Halfuurnum	0.05
HH	0.07	t_min120_Rentals	0.04
t_min450_Returns	0.04	HH	0.03
t_min90_Rentals	0.03	day_of_week_nr	0.02
t_min390_Returns	0.03	t_min150_Rentals	0.02
t_min150_Rentals	0.02	t_min30_Returns	0.01
t_min30_Returns	0.01	temperature	0.01

5.4 User type correlation

Sub research question three aims in gaining insight in the influence of different user types on the bicycle facilities around Utrecht central station. Different user types are defined by their time of arrival and departure in the bicycle storages or time of rental and return in the BSS, obtained from the raw dataset. Definitions are presented in section 4.3. Consequently, a correlation matrix is constructed in order to see correlations between the incoming number of bicycles and the incoming number per user type. The results are presented in table 9.

Table 9

Correlation matrix of user types

Stationsplein				**Jaarbeursplein**			
Rentals		Incoming		Rentals		Incoming	
Short use	0.65	Short stay	0.45	Short use	0.64	Short stay	0.37
Activity side	0.86	Home side	0.97	Activity side	0.91	Home side	0.96
Home side	0.12	Activity side	-0.05	Home side	0.07	Activity side	-0.01
Long term	0.45	Long term	0.92	Long term	0.42	Long term	0.83
Evening users	0.25	Evening users	0.02	Evening users	0.18	Evening users	0.05
Other	0.12	Other	0.04	Other	0.05	Other	0.05

For the BSS part of the storages, rentals of activity-side users yield the highest correlation with the total number of rentals. For the Stationsplein storage this number is 0.86 and for the Jaarbeursplein storage 0.91. These numbers reflect the high share of activity side users making use of the BSS parts of the bicycle facilities near Utrecht central station. Moreover, a high correlation is found for short use, indicating that shared bicycles are often rented for short trips that do not last longer than 5 hours. For the common bicycle storage part, home side users represent the largest part among the user types. For the Stationsplein storage this number is 0.97 and for the Jaarbeursplein storage 0.96. Furthermore, long term users of the common bicycle storage part of both facilities yield a high correlation with incoming bicycles, indicating that a reasonable number of users is storing their bicycles for a longer period than 15 hours. Gaining insight in these user types per storage helps understanding the peaks in usage and can be useful for making better predictions.

6. Discussion

This section will evaluate the results regarding the research questions.

6.1 Research questions

- *What is the average incoming and outgoing number of bicycles and shared bicycles for certain periods of time such as day of the week and part of the day?*

To answer the first sub research question, insight in bicycle usage patterns is acquired. Exploratory data analysis resulted in visualizations of these patterns in figure 1 and 2. Moreover, a linear regression model that functioned as baseline was created to understand relationships between bicycle facility usage and different time features, such as peak hour and weekday. The results of this linear model are discussed in section 5.1 and indicated that a relationship is present between incoming and outgoing bicycles and weekday. Moreover, a relationship was found between incoming and outgoing bicycles and peak hour. This also applies to the rentals and returns of the BSS part of the storages. These results demonstrate that peak hours can be considered important for the influence of bicycle facility usage. During weekdays, the majority of users is storing their bike in morning peak hours and is collecting their bike in evening peak hours. This is in compliance with the study of Vogel et al. (2011), who observed this pattern for BSS usage and indicates that the BSS part and the common bicycle storage part are predominantly used by commuters. However, usage peaks slightly differ between the BSS part and common bicycle part since dominance of user types vary. This will be addressed in the answer on sub research question 3.

- *Which features yield the highest influence for the prediction of bicycle facility usage during peak hours and outside peak hours?*

In order to give answer to the second research question, subsets of the timeseries data are made for peak hours and non-peak hours. Consequently, feature importances were generated by the Random Forest model and showed that inventory features prove to be important when predicting bicycle facility usage. Moreover, time features such as hour of the day yielded high predictive power. Surprisingly, when looking at external features employed in this study, weather features do not prove high importance for the prediction of bicycle facility usage inside and outside peak hours. This applies to both the common bicycle storage part and the BSS part of the facilities and is not in agreement with existing literature (Giot and Cherrier, 2014; El-Assi, Mahmoud and Habib, 2017; Saneinejad, Roorda and Kennedy, 2012). One of the reasons for this, is that the facilities studied in this work are found near a railway station that is mainly used by commuters, who tend not to be affected by the weather. However, this would then only apply for the subsets made for periods inside peak hours. Another reason could be that a large share of the people in the Netherlands use bicycles as primary form of transportation. Dutch cycling culture is unique, people make use of them for everyday purposes and do not let adverse weather conditions influence themselves.

- *To what extent does user type influence the usage of bicycle facilities around Utrecht central station?*

The third research question is answered in section 5.4, where correlations between incoming bicycles and incoming bicycles per user type are presented. Results show that for the common bicycle storages, home-side users are highly important considering usage of these facilities. In addition, rentals of activity-side users yield the highest correlation with the total number of rentals for the BSS part of the storages. These results confirm existing literature (Keijer and Rietveld, 2000; Martens, 2004; Jonkeren et al., 2018) and demonstrate the high influence of activity-side users for the BSS part and the high influence of home-side users for the common bicycle storage part. Home-side users tend to utilize the bicycle facilities at the beginning of their trip during morning peak hours, whereas they collect their bicycle at the end of their trip during evening peak hours. Since the BSS part is dominated by activity side users, its pattern is opposite. These differences result in slightly shifted peaks which can be useful for management of the facilities (e.g. redirection of storage space when shared bicycles are almost not present during daytime). Moreover, understanding behavior of these user types allows for making better predictions in future studies. Departing trains could be an interesting factor for predicting bicycle facility usage where home-side users dominate, whereas arriving trains possibly are interesting for the prediction of BSS usage.

- *Which model yields the best results for the prediction of bicycle facility usage?*

In order to answer this question, three models were trained and evaluated on a test set. A Random Forest Regressor, Multilayer Perceptron and Support Vector Regressor were employed for the prediction of bicycle facility usage. Results of the model performance can be found in Appendix B. For all fixed points in the future for the set time horizon, the Random Forest Regressor yielded best performance results. This applies to both facilities for the common bicycle storage part as well as the BSS part. Additionally, the SVR and the MLP outperformed the baseline model as well, but could not compete with the RFR. Hence, results of the current study validate existing literature that found similar results with regard to the best performing algorithm (Dias, Bellalta and Oechsner, 2015; Lozano, De Paz, Villarrubia González, Iglesia, and Bajo, 2018; Yang et al., 2016; Yin, Lee, and Wong, 2012). Predictions for both storages yielded similar results for all of the employed models. A moderate difference in prediction results however is present between the common bicycle storage part and the BSS part when considering the R^2. This difference can be explained due to differences in the size of the raw datasets and a more stable pattern for the common bicycle storage part.

- *To what degree can future bicycle storage usage be predicted for the secured bicycle storages of Utrecht central station?*

Finally, the main question of this study aims to give clarity on how well bicycle facility usage can be predicted. Looking to the best performing model, the current number of bicycles for the Stationsplein and the Jaarbeursplein storage can be predicted with a RMSLE of 0.35. The prediction of the current number of outgoing bicycles yields a RMSLE of 0.41 and 0.39 for the Stationsplein and Jaarbeursplein storage respectively. The MAE yielded by the RFR for the prediction of incoming and outgoing bicycles did not exceed 15 for the Jaarbeursplein storage and 21 for the Stationsplein storage on the entire time horizon. For the prediction of BSS usage on the entire horizon, the MAE did not exceed 6 for the Jaarbeursplein storage and 5 for the Stationsplein storage. Whether these prediction results are precise enough, depends on the implementation by different instances such as the NS, ProRail and the municipality of Utrecht. Suggestions on improvement of these results are given in the conclusion.

6.2 Limitations

Several limitations were present in this study. The data used for the current study only contained information about users arriving and departing from two facilities around Utrecht central station. Therefore, these numbers cannot be used to give an estimation about when a storage capacity has reached its maximum. Different systems are integrated in the bicycle facilities that aim to count the number of stored bicycles in a facility. However, data collection of these systems do not (yet) allow for data analysis and may be inaccurate.

Secondly, a separate membership area exists in the facilities investigated in this study. Since members are not required to check-in or check-out, no data is present of these users. In order to give better advices to the management of these facilities, it would be desirable to acquire data of their usage as well.

Lastly, datasets used in this work did not contain information about the origin of the users. When using the bicycle facilities studied in this paper, a public transportation card is needed (OV-chipkaart). These cards are also required for the usage of trains and, thus, contain a lot of information about the origin and destination of an individual, that would be valuable for predicting bicycle facility usage.

7. Conclusion

This thesis explored ways to improve the management of bicycle facilities using data of its usage. The main goal was to investigate the degree to which bicycle facility usage can be predicted. Predictions of the incoming and outgoing number of bicycles and shared bicycles for the storages around Utrecht central station provide valuable information for the management of these facilities.

In order to predict bicycle facility usage, multiple state-of-the-art algorithms were trained and evaluated on data from two bicycle storages. Results showed that a Random Forest Regressor performed best and confirmed its applicability for predicting bicycle facility usage. Moreover, inventory features proved their importance for time-series forecasting whereas weather features did not yield high predictive power for this specific task, contradictory to existing literature. A user type study gave insight into the different groups that make use of the facilities and showed a clear pattern. Activity-side users dominate the bicycle sharing systems whereas home-side users dominate usage of common bicycle storages. In future work, it would be interesting to include departing and arriving trains, as they could possibly be interesting factors for predicting bicycle facility usage. Also, including occupation-level data of facilities could improve predictions as it enhances understanding the patterns of incoming and outgoing bicycles. Lastly, it would be interesting to add more trip information to the facility user check-ins in order to improve performance of the predictions.

References

Awad, M., & Khanna, R. (2015). Support vector regression. In *Efficient Learning Machines* (pp. 67-80). Apress, Berkeley, CA.

Bachand-Marleau, J., Lee, B. H., & El-Geneidy, A. M. (2012). Better understanding of factors influencing likelihood of using shared bicycle systems and frequency of use. *Transportation Research Record*, *2314*(1), 66-71.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Dias, G. M., Bellalta, B., & Oechsner, S. (2015). Predicting occupancy trends in Barcelona's bicycle service stations using open data. *arXiv preprint arXiv:1505.03662.*

El-Assi, W., Mahmoud, M. S., & Habib, K. N. (2017). Effects of built environment and weather on bike sharing demand: a station level analysis of commercial bike sharing in Toronto. *Transportation*, *44*(3), 589-613.

Faghih-Imani, A., Eluru, N., El-Geneidy, A. M., Rabbat, M., & Haq, U. (2014). How land-use and urban form impact bicycle flows: evidence from the bicycle-sharing system (BIXI) in Montreal. *Journal of Transport Geography*, *41*, 306-314.

Faghih-Imani, A., & Eluru, N. (2015). Analysing bicycle-sharing system user destination choice preferences: Chicago's Divvy system. *Journal of transport geography*, *44*, 53-64.

Fishman, E., Schepers, P., & Kamphuis, C. B. M. (2015). Dutch cycling: quantifying the health and related economic benefits. *American journal of public health*, *105*(8), e13-e15.

Gebhart, K., & Noland, R. B. (2014). The impact of weather conditions on bikeshare trips in Washington, DC. *Transportation*, *41*(6), 1205-1225.

Giot, R., & Cherrier, R. (2014). Predicting bikeshare system usage up to one day ahead. In *Computational intelligence in vehicles and transportation systems (CIVTS), 2014 IEEE symposium on* (pp. 22-29). IEEE.

Hulot, P., Aloise, D., & Jena, S. D. (2018). Towards Station-Level Demand Prediction for Effective Rebalancing in Bike-Sharing Systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 378-386). ACM.

Jonkeren, O., Harms, L., Jorritsma, P., Huibregtse, O., Bakker, P., & Kager, R. (2018). *Waar zouden we zijn zonder de fiets* [where would we be without bicycles]. retrieved from Kennisinstituut voor Mobiliteitsbeleid: https://www.kimnet.nl/publicaties/rapporten/2018/07/12/waar-zouden-we-zijn-zonder-de-fiets-en-de-trein

Keijer, M. J. N., & Rietveld, P. (2000). How do people get to the railway station? The Dutch experience. *Transportation Planning and Technology, 23*(3), 215-235.

KNMI (n.d.). *About the KNMI*. Retrieved on October 16, 2018 from https://www.knmi.nl/over-het%20knmi/over

Koskela, T., Lehtokangas, M., Saarinen, J., & Kaski, K. (1996). Time series prediction with multilayer perceptron, FIR and Elman neural networks. In *Proceedings of the World Congress on Neural Networks* (pp. 491-496). INNS Press San Diego, USA.

Lozano, Á., De Paz, J. F., Villarrubia González, G., Iglesia, D. H., & Bajo, J. (2018). Multi-Agent System for Demand Prediction and Trip Visualization in Bike Sharing Systems. *Applied Sciences*, *8*(1), 67.

Martens, K. (2004). The bicycle as a feedering mode: experiences from three European countries. *Transportation Research Part D: Transport and Environment, 9*(4), 281-294.

Martens, K. (2007). Promoting bike-and-ride: The Dutch experience. *Transportation Research Part A: Policy and Practice, 41*(4), 326-338.

NS (n.d.). *Biggest, smallest and fastest growing stations*. Retrieved on November 9, 2018 from https://nieuws.ns.nl/grootste-kleinste-en-snelst-groeiende-stations/

NS (2015) Annual report 2014, Dutch Railway company, Utrecht

NS (2018) Annual report 2017, Dutch Railway company, Utrecht

O'brien, O., Cheshire, J., & Batty, M. (2014). Mining bicycle sharing data for generating insights into sustainable transport systems. *Journal of Transport Geography, 34*, 262-273.

Raviv, T., Tzur, M., & Forma, I. A. (2013). Static repositioning in a bike-sharing system: models and solution approaches. *EURO Journal on Transportation and Logistics, 2*(3), 187-229.

Saneinejad, S., Roorda, M. J., & Kennedy, C. (2012). Modelling the impact of weather conditions on active transportation travel behaviour. *Transportation research part D: transport and environment, 17*(2), 129-137.

Sarkar, A., Lathia, N., & Mascolo, C. (2015). Comparing cities' cycling patterns using online shared bicycle maps. *Transportation, 42*(4), 541-559.

Singla, A., Santoni, M., Bartók, G., Mukerji, P., Meenen, M., & Krause, A. (2015). Incentivizing Users for Balancing Bike Sharing Systems. In *AAAI* (pp. 723-729).

Vogel, P., Greiser, T., & Mattfeld, D. C. (2011). Understanding bike-sharing systems using data mining: Exploring activity patterns. *Procedia-Social and Behavioral Sciences, 20*, 514-523.

Yang, Z., Hu, J., Shu, Y., Cheng, P., Chen, J., & Moscibroda, T. (2016). Mobility modeling and prediction in bike-sharing systems. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services* (pp. 165-178). ACM.

Yin, Y. C., Lee, C. S., & Wong, Y. P. (2012). Demand Prediction of Bicycle Sharing Systems.

Zhou, Y., Wang, L., Zhong, R., & Tan, Y. (2018). A Markov Chain Based Demand Prediction Model for Stations in Bike Sharing Systems. *Mathematical Problems in Engineering, 2018*.

Appendix A

Table A1

Regression results for Incoming bicycles Stationsplein storage

	coef	std err	t	P>|t|	[0.025	0.975]
Const	27.447	2.307	11.898	0.000	22.925	31.969
Half hour	-0.0034	0.001	-2.703	0.007	-0.006	-0.001
Day_of_week_number	0.4682	0.190	2.459	0.014	0.095	0.841
Week	13.113	1.270	10.324	0.000	10.623	15.603
Weekend	14.334	1.286	11.148	0.000	11.814	16.855
Peak	58.905	1.548	38.045	0.000	55.870	61.939
Non peak	-31.457	1.338	-23.514	0.000	-34.079	-28.835
Hour	0.1577	0.121	1.298	0.194	-0.080	0.396
Wind speed avg	-0.0061	0.024	-0.256	0.798	-0.053	0.041
Temp	-0.2736	0.070	-3.922	0.000	-0.410	-0.137
Sunshine	0.5379	0.168	3.194	0.001	0.208	0.868
Rain duration	-0.7368	0.246	-2.996	0.003	-1.219	-0.255
Rain sum	-0.0943	0.108	-0.872	0.383	-0.306	0.118
Cloud cover	0.1652	0.148	1.118	0.264	-0.125	0.455
Humidity	0.0809	0.043	1.889	0.059	-0.003	0.165
Rain	-0.0124	1.099	-0.011	0.991	-2.167	2.142
Warm day	-5.474	2.388	-2.292	0.022	-10.154	-0.793

Table A2

Regression results for Outgoing bicycles Stationsplein storage

	coef	std err	t	P>|t|	[0.025	0.975]
Const	17.398	2.290	7.596	0.000	12.909	21.887
Half hour	0.0077	0.001	6.052	0.000	0.005	0.010
Day_of_week_number	-0.2603	0.189	-1.377	0.168	-0.631	0.110
Week	7.428	1.261	5.890	0.000	4.956	9.900
Weekend	9.970	1.277	7.810	0.000	7.468	12.473
Peak	35.446	1.537	23.060	0.000	32.433	38.459
Non peak	-18.048	1.328	-13.589	0.000	-20.651	-15.444
Hour	0.5984	0.121	4.964	0.000	0.362	0.835
Wind speed avg	-0.0555	0.024	-2.337	0.019	-0.102	-0.009
Temp	-0.0485	0.069	-0.701	0.483	-0.184	0.087
Sunshine	-0.0804	0.167	-0.481	0.631	-0.408	0.247
Rain duration	-0.0310	0.244	-0.127	0.899	-0.510	0.448
Rain sum	0.0140	0.107	0.131	0.896	-0.197	0.225
Cloud cover	-0.1146	0.147	-0.781	0.435	-0.402	0.173
Humidity	-0.1685	0.043	-3.962	0.000	-0.252	-0.085
Rain	-0.2301	1.091	-0.211	0.833	-2.369	1.909
Warm day	-4.109	2.371	-1.733	0.083	-8.756	0.538

Table A3

Regression results for rentals in BSS Stationsplein storage

	coef	std err	t	P>|t|	[0.025	0.975]
Const	2.9762	0.336	8.845	0.000	2.317	3.636
Half hour	-0.0005	0.000	-3.008	0.003	-0.001	-0.000
Day_of_week_number	0.2066	0.029	7.186	0.000	0.150	0.263
Week	0.9930	0.181	5.484	0.000	0.638	1.348
Weekend	1.9832	0.183	10.811	0.000	1.624	2.343
Peak	5.5783	0.213	26.159	0.000	5.160	5.996
Non peak	-2.6020	0.189	-13.756	0.000	-2.973	-2.231
Hour	-0.0098	0.016	-0.616	0.538	-0.041	0.021
Wind speed avg	-0.0046	0.004	-1.271	0.204	-0.012	0.003
Temp	-0.0041	0.011	-0.389	0.698	-0.025	0.017
Sunshine	0.1519	0.025	6.194	0.000	0.104	0.200
Rain duration	-0.0662	0.038	-1.758	0.079	-0.140	0.008
Rain sum	0.0068	0.016	0.424	0.671	-0.025	0.038
Cloud cover	0.0321	0.021	1.493	0.136	-0.010	0.074
Humidity	-0.0044	0.006	-0.710	0.478	-0.017	0.008
Rain	-0.1496	0.170	-0.881	0.378	-0.482	0.183
Warm day	0.5588	0.306	1.828	0.067	-0.040	1.158

Table A4

Regression results for returns in BSS Stationsplein storage

	coef	std err	t	P>|t|	[0.025	0.975]
Const	1.2085	0.250	4.832	0.000	0.718	1.699
Half hour	0.0014	0.000	11.326	0.000	0.001	0.002
Day_of_week_number	-0.0556	0.021	-2.604	0.009	-0.098	-0.014
Week	0.5594	0.135	4.156	0.000	0.296	0.823
Weekend	0.6492	0.136	4.761	0.000	0.382	0.916
Peak	2.0590	0.159	12.989	0.000	1.748	2.370
Non peak	-0.8505	0.141	-6.048	0.000	-1.126	-0.575
Hour	0.1193	0.012	10.128	0.000	0.096	0.142
Wind speed avg	-0.0068	0.003	-2.522	0.012	-0.012	-0.002
Temp	0.0058	0.008	0.743	0.458	-0.010	0.021
Sunshine	0.1138	0.018	6.242	0.000	0.078	0.150
Rain duration	0.0059	0.028	0.212	0.832	-0.049	0.061
Rain sum	-0.0189	0.012	-1.588	0.112	-0.042	0.004
Cloud cover	0.0346	0.016	2.167	0.030	0.003	0.066
Humidity	-0.0279	0.005	-5.992	0.000	-0.037	-0.019
Rain	0.0028	0.126	0.023	0.982	-0.245	0.250
Warm day	-0.0003	0.227	-0.001	0.999	-0.446	0.445

Appendix B

Table B1

Baseline results for Stationsplein storage

Type of usage	Forecast horizon	R^2	MAE	RMSE	RMSLE
Rental	t = 0	0.74	4.90	9.20	0.76
	t + 240	0.24	8.34	15.83	1.05
	t + 480	0.53	6.95	12.39	1.00
Return	t = 0	0.82	4.11	6.19	0.68
	t + 240	0.74	5.40	7.41	0.89
	t + 480	0.73	5.42	7.50	0.88
Incoming	t = 0	0.92	27.17	47.48	1.25
	t + 240	0.35	79.54	138.31	1.91
	t + 480	0.69	44.49	95.09	1.50
Outgoing	t = 0	0.88	25.67	48.77	0.93
	t + 240	0.94	22.75	32.88	1.00
	t + 480	0.94	23.30	33.24	0.93

Table B2

Baseline results for Jaarbeursplein storage

Type of usage	Forecast horizon	R^2	MAE	RMSE	RMSLE
Rental	t = 0	0.70	6.98	13.92	0.86
	t + 240	0.21	12.13	22.54	1.31
	t + 480	0.51	9.82	17.69	1.13
Return	t = 0	0.83	5.03	7.78	0.67
	t + 240	0.74	6.72	9.64	0.85
	t + 480	0.73	6.78	9.79	0.90
Incoming	t = 0	0.90	20.75	31.44	1.09
	t + 240	0.39	53.57	78.21	1.63
	t + 480	0.64	35.48	59.73	1.47
Outgoing	t = 0	0.86	18.39	29.67	0.90
	t + 240	0.86	21.06	29.78	1.03
	t + 480	0.87	20.53	29.09	1.02

Table B3

Random Forest results for Stationsplein storage

Type of usage	Forecast horizon	R^2	MAE	RMSE	RMSLE
Rental	t = 0	0.87	3.53	6.57	0.45
	t + 240	0.74	4.37	9.18	0.49
	t + 480	0.77	4.22	8.68	0.49
Return	t = 0	0.87	3.28	5.20	0.43
	t + 240	0.88	3.36	5.07	0.46
	t + 480	0.88	3.33	5.06	0.44
Incoming	t = 0	0.98	11.36	24.46	0.35
	t + 240	0.91	20.10	50.14	0.49
	t + 480	0.93	17.88	45.04	0.46
Outgoing	t = 0	0.97	12.67	24.06	0.41
	t + 240	0.97	12.90	22.26	0.47
	t + 480	0.97	13.32	22.67	0.49

Table B4

Random Forest results for Jaarbeursplein storage

Type of usage	Forecast horizon	R^2	MAE	RMSE	RMSLE
Rental	t = 0	0.87	4.51	9.23	0.48
	t + 240	0.75	5.86	12.58	0.55
	t + 480	0.75	5.88	12.57	0.55
Return	t = 0	0.88	4.00	6.46	0.44
	t + 240	0.86	4.33	7.01	0.50
	t + 480	0.86	4.36	7.09	0.47
Incoming	t = 0	0.96	9.29	18.77	0.35
	t + 240	0.86	14.31	36.96	0.40
	t + 480	0.92	12.57	28.77	0.41
Outgoing	t = 0	0.94	10.70	20.03	0.39
	t + 240	0.93	11.55	20.74	0.47
	t + 480	0.93	11.91	20.79	0.47

Table B5

Multilayer Perceptron results for Stationsplein storage

Type of usage	Forecast horizon	R^2	MAE	RMSE	RMSLE
Rental	t = 0	0.85	3.90	7.04	0.56
	t + 240	0.67	5.25	10.42	0.71
	t + 480	0.74	4.78	9.24	0.69
Return	t = 0	0.86	3.48	5.39	0.51
	t + 240	0.87	3.61	5.22	0.61
	t + 480	0.85	3.83	5.62	0.63
Incoming	t = 0	0.97	14.42	29.41	0.64
	t + 240	0.78	37.41	80.11	1.10
	t + 480	0.92	22.00	49.73	0.78
Outgoing	t = 0	0.94	16.14	34.83	0.64
	t + 240	0.97	15.63	24.31	0.65
	t + 480	0.97	15.63	23.62	0.73

Table B6

Multilayer Perceptron results for Jaarbeursplein storage

Type of usage	Forecast horizon	R^2	MAE	RMSE	RMSLE
Rental	t = 0	0.85	4.90	9.73	0.65
	t + 240	0.62	7.31	15.55	0.73
	t + 480	0.73	6.66	13.18	0.82
Return	t = 0	0.86	4.27	7.04	0.54
	t + 240	0.86	4.60	7.16	0.63
	t + 480	0.84	4.78	7.54	0.66
Incoming	t = 0	0.97	10.55	18.28	0.61
	t + 240	0.75	21.89	49.61	0.73
	t + 480	0.86	16.44	37.49	0.69
Outgoing	t = 0	0.90	13.31	25.61	0.53
	t + 240	0.92	14.07	22.78	0.71
	t + 480	0.92	14.56	23.06	0.73

Table B7

Support Vector Regressor results for Stationsplein storage

Type of usage	Forecast horizon	R^2	MAE	RMSE	RMSLE
Rental	t = 0	0.77	4.85	8.62	0.72
	t + 240	0.50	6.26	12.82	0.85
	t + 480	0.70	5.38	9.88	0.79
Return	t = 0	0.79	4.19	6.61	0.61
	t + 240	0.84	4.05	5.84	0.66
	t + 480	0.83	4.24	6.08	0.69
Incoming	t = 0	0.95	17.72	39.57	0.67
	t + 240	0.62	40.19	105.97	1.06
	t + 480	0.83	24.53	70.75	0.84
Outgoing	t = 0	0.93	18.40	36.63	0.60
	t + 240	0.96	16.42	26.94	0.69
	t + 480	0.96	17.05	27.43	0.75

Table B8

Support Vector Regressor results for Jaarbeursplein storage

Type of usage	Forecast horizon	R^2	MAE	RMSE	RMSLE
Rental	t = 0	0.81	5.40	11.13	0.69
	t + 240	0.49	8.06	18.08	0.91
	t + 480	0.72	6.65	13.37	0.79
Return	t = 0	0.84	4.65	7.54	0.54
	t + 240	0.84	4.83	7.56	0.64
	t + 480	0.83	5.04	7.71	0.69
Incoming	t = 0	0.95	11.70	22.89	0.58
	t + 240	0.73	22.66	51.63	0.81
	t + 480	0.85	16.24	39.10	0.72
Outgoing	t = 0	0.89	12.91	25.80	0.48
	t + 240	0.91	13.53	23.29	0.62
	t + 480	0.92	13.41	23.07	0.64