Analytics and Application WS2019-20
Team Assignment

The AA Team

Authors:
Tim Schäfer - 7369518 (Descriptive Analysis)
Malte Hain - 7364092 (Descriptive Analysis)
Lukas Ripperger - 7364457 (Descriptive Analysis)
Philipp Page - 7367085 (Data Preparation and Descriptive Analysis)
Lukas Humpe - 7363827 (Data Preparation and Predictive Analysis)

Department of Information Systems for Sustainable Society
Faculty of Management, Economics and Social Sciences
University of Cologne

January 29, 2020

Executive Summary

The avoidance of greenhouse gas emissions is currently more topical than ever thanks to initiatives such as Fridays-for-Future. Politicians, companies and scientists are dealing with the topic and working on possible solutions. One of the major areas where greenhouse gas emissions play a role is the mobility sector. At a time when more and more people are moving to cities, new concepts are needed to transport the masses of people and to react to challenges such as air quality. As a possible partial solution, bicycle rental offers in the sense of the mobility as a service concept have emerged in many cities in Germany.

In order to convince more people of such offers, the availability of bicycles must be ensured by offering potential customers who want to get from one place to another at a certain time a bicycle at a rental station. Since there are stations in cities that have more returns than rentals at certain times. A manual distribution of bicycles must be carried out by the provider. In order to better organize such re-distributions with regard to customer satisfaction, the present project work investigated the usage behaviour of the provider Call a Bike in Kassel in the years 2015 and 2016.

One task in the project was to investigate patterns in bookings with regard to customer groups and to develop a forecast model to predict the demand for bicycles. Since the booking data are time series, a special approach is required using an algorithm from Facebook. In addition, other algorithms were used which were evaluated and compared based on their performance.

One finding that was made during the course of the project is that the period with the highest usage is in summer. When looking at the individual day, the afternoon represents the most frequented part of the day. In the city of Kassel, the ICE station is the station where more bicycles are returned than rented. The other way round, the station at the university is the station where more bicycles are borrowed than returned. This is an important finding for the operative business of Call a Bike for the redistribution of bikes in the city area. The model developed to forecast bookings was able to predict on a daily basis the demand for bicycles with a Mean Absolute Error (MAE) of 55 bikes. In an hourly forecast the MAE is 5 bikes.

The knowledge gained in the course of this project about the usage behaviour as well as the developed prediction model gives interesting insights, which have to be further refined for the operational use. Among other things, the fact that it is assumed that a bicycle is available at a station is one of the most important factors. If none is available but a potential customer wants to rent one, this is currently not taken into account in the demand.

Contents

1 Introduction

Subject of this project work are the data and the business model of the company Call a Bike which operates a bicycle rental system in Kassel, among other cities in Germany. Call a Bike is one of many providers currently active in the mobility market and offers new services for passenger transport. The bicycle rental system is characterized by fixed rental stations, which are distributed throughout the city. Bicycles can be rented and returned by customers independently at these stations. The customer is billed by rental duration. These bookings can be obtained anonymously for several years via the Open Data Portal (*Call A Bike - Open-Data-Portal – Deutsche Bahn Datenportal*, 2016), which is the basis for this project. The following sections describe the visualization, clustering and prediction techniques applied to gain insights for tackling the business problem of predicting the bike rental demand and distinguishing customer segments.

2 Problem Description

A supplier of a bicycle rental system faces several challenges in its management. First and foremost, he must optimally serve the demand for bicycles in order to generate turnover and not leave any sales opportunity unattended. Furthermore, a high degree of attractiveness to the customer must be ensured in order to differentiate from competitors. Since the purchase of many bikes is capital-intensive and further purchases also increase maintenance costs, a supplier is interested in an optimal and demand-oriented distribution of his bikes. The suppliers in the market therefore have service teams which redistribute the bikes between the stations in order to maintain the offer in a customer-friendly way (Ricker, Meisel, & Mattfeld, 2012). Several issues have to be considered in this task: At which stations is there a supply surplus? At which stations do potential customers go away empty-handed? When do these extremes occur? In order to support the operator in meeting these challenges, it would help him to know how demand is developing and what patterns of behaviour exist among the customers. This could help to better plan the routes of the service staff and ultimately provide the customer with a better experience when using the bike rental system. Precisely formulated: A prediction of the exact demand for bicycles, both in terms of stations and time, is necessary for the successful and customer-oriented operation of such a system. In addition, the knowledge and understanding of the customers' usage behavior is valuable for the strategic development and expansion of the system in the city.

3 Data Description

The data obtained by Deutsche Bahn contains information about transactional bike bookings of Call a Bike. Each row in the data set represents one booking and can be associated to a customer and a vehicle. An additional data set contains a set of all rental stations containing information like the geographical coordinates. For the purpose of an in depth descriptive and predictive analysis of that data, two distinct data sets are required.

Table 1 shows the final data set which has been used as a basis for the descriptive analysis. The rental zone coordinates have been obtained by an additional data set, provided on the Open-Data-Portal by the Deutsche Bahn (*Call A Bike - Open-Data-Portal – Deutsche Bahn Datenportal*, 2016).

The predictive analysis requires an aggregation of the booking data to different resolutions in order to make a prediction for different time frames. Therefore, all columns, shown in Table 2, are aggregated values obtained by the raw bookings.

Details about the data preparation process leading to the final data sets are outlined in Section 4.

Table 1: Data set for descriptive analysis

Column Name	**dtype**	**Description**
BOOKING_HAL_ID	int64	The unique ID that identifies a booking
VEHICLE_HAL_ID	int64	The unique ID that identifies a bike
CUSTOMER_HAL_ID	object	The unique ID that identifies a customer
DATE_FROM	object	The date and time a booking has started
DATE_UNTIL	object	The date and time a booking has ended
START_RENTAL_ZONE	object	The name of the rental zone, where the booking has started
START_RENTAL_ZONE_HAL_ID	int64	The unique ID that identifies the rental zone, where the booking has started
END_RENTAL_ZONE	object	The name of the rental zone, where the booking has ended
END_RENTAL_ZONE_HAL_ID	int64	The unique ID that identifies the rental zone, where the booking has ended
START_LONGITUDE	float64	The longitude for the location of the start rental zone
START_LATITUDE	float64	The latitude for the location of the start rental zone
END_LONGITUDE	float64	The longitude for the location of the end rental zone
END_LATITUDE	float64	The latitude for the location of the end rental zone
DISTANCE	float64	The air distance in km between the start and end rental zone
BOOKING_DURATION	float64	The duration in minutes between the start and end time of booking

Table 2: Data set for predictive analysis

Column	**dtype**	**Description**
COUNT	int64	The number of bikes rented in the next period
SUNSHINE_DURATION	float64	The amount of sun shining in the next period in minutes
AIR_TEMP	float64	The mean temperature at two meters above ground in the next period in degress
HUMIDITY	float64	The mean humidity at two meters above ground in the next period in percent
PRECIPITATION_HEIGHT	float64	The amount precipitation in the next period in mm
RAIN_FALLEN	float64	Binary variable indicating whether rain has fallen in the next period
MEAN_WIND_SPEED	float64	The mean wind speed in the next period in kmph
BANK_HOLIDAY	int64	Binary variable indicating whether there is a bank holiday in the next period
SCHOOL_HOLIDAY	float64	Binary variable indicating whether there is a school holiday in the next period

4 Data Preparation

4.1 Cleansing

Since the raw data set contains bookings of many different cities and years, the data has been filtered to obtain the relevant entries for the given task (Kassel, 2015 and 2016). Moreover, columns which contain duplicate information or are not relevant for further analysis have been dropped. For instance, the DATE_BOOKING column is equivalent to the DATE_FROM column and has therefore been dropped. Additionally, a very small subset of rows (0.0046%) has missing values and was dropped as well. Finally, the data set is ready for further task specific feature engineering, which is why the process is splitted.

4.2 Clustering

In a first step, the cleaned data is enriched with location information for the rental stations. With this additional data it was possible to append the data with two columns containing the distance between the start and end rental zone of each booking and the time of the booking duration for faster and easier usage in later steps. Surprisingly, latitude and longitude in the raw station data set are interchanged. The mistake has been identified after a first station visualization on a map based on the coordinates. This mistake has been addressed by renaming the particular columns in the final data frame. The distance between the start and end rental zones has then been calculated utilizing the geodesic function of the geopy[1] python package. Finally, the booking duration has been calculated by subtracting the DATE_FROM column from the DATE_UNTIL column and converting the result to an integer value in minutes.

4.3 Prediction

According to the task, the bike rental demand should be predicted in different temporal resolutions. Bike demand is defined as the number of bookings in a certain time period, e.g. one hour. Therefore, the bookings have been re-sampled to one hour by saving the booking count for each hour as separate row. This resolution has been chosen because it allowed for a convenient addition of climate data, as this data is also available in a one hour resolution. Since there is no direct weather data available for Kassel, the nearest station “Schauenburg Elgershausen”, which is eleven km away from Kassel, has been attached to the aggregated data frame. The weather data provider is the German Meteorologi-

[1] https://pypi.org/project/geopy/

cal Service Provider[2]. The handling of missing weather data has been addressed by using a time linear interpolation leading to a realistic data imputation. Important to note is that sunshine duration was a missing value during night and has been filled with zeros. A further potentially important variable are holidays which is why bank as well as school holiday data have been added as binary coded variable. At this point, the one hour data set is ready to use for prediction.

For the purpose of long term predictions, copies of the one hour data set have been created and aggregated to two hours, six hours and 24 hours. The aggregation function for each column is an important factor for a reliable up scaling. For example, to scale the column MEAN_WIND_SPEED up, the mean of the wind speed has to be used instead of the sum. For details about the utilized aggregation functions, consult the Jupyter Notebook (Data Preparation for Analysis).

5 Data Analysis

5.1 Descriptive

To gain a first understanding of the data the trajectory of the time series resampled at different resolutions has been plotted. One of the main indications of the plots is that the bike rental demand is highly volatile and shows seasonal patterns. The most obvious pattern is that rental demand and its variance increases until the middle of the year and decreases again. Figure 1 illustrates this finding.

Figure 2 indicates that there is a strong sub-daily seasonality (e.g. demand is low during night hours and high during the late afternoon). In addition, there is a strong seasonal pattern in regards to the four seasons of the year with summer depicting the highest mean demand over all.

Considering station wise demand patterns, one can identify popular departure and arrival stations. Figure 3 shows that the ICE train station and the main station are popular departure stations while the two universities in Kassel are the major destinations. Though, we cannot be sure if bikes are re-distributed from the university to train stations in order to make the service available to customers with more purchasing power. In general, one observers a bike demand shift from west to east which might indicate that Call a Bike re-distributes its bike manually to serve the customer demands.

Figure 4 depicts the result of a cluster analysis using booking duration and distance as discriminating features. One main finding is that customers can be differentiated by driving speed and booking duration. Customers who rent

[2] https://cdc.dwd.de/portal

approximately between 0 and 50 minutes can be divided by their speed. First, there are the slow customers (red cluster), who only travel up to 1 km. Then, there are customers, who travel with a medium speed (yellow cluster), who travel between one and three kilometers. Finally, there are fast customers (light blue cluster), who travel up to 8 km in the same amount of time as the slow customers. Customers, who took more time than those three types, are further divided into two clusters. These are hard to interpret due to their size. However, the they could be read as customers who do trips for a leisure activity or are tourist, visiting the city.

5.2 Predictive

Predictive analysis has been carried out on four different resolutions. Steps taken for each individual resolution are similar even though there are minor differences in the approach taken for sub-daily resolution and daily resolution.

Figure 5 illustrates the correlation between variables in the data set. It can be seen that all variables have some kind of reasonable influence on the target variable (e.g. increasing bike rental demand with increasing temperatures). In addition, Figure 6 shows that there is a strong autocorrelation between the trajectory and its own lagged values. This has been the case for all resolutions. Therefore all features as well as lagged values were considered in prediction.

In the case of daily resolution the use of time series algorithms was suitable, therefore SARIMAX and Facebook's Prophet algorithm have been chosen since they allow for the use of external regressors as well as time series decomposition. In the case of sub-daily resolution, however, the data set was too volatile to use traditional time series algorithms. Therefore, linear and random forest regression have been conducted.

The performance of all algorithms at all resolutions was evaluated carrying out rolling one-step-ahead cross-validation yielding the following results. The results clearly suggest that more complex models lead to better prediction results. This is especially true for daily predictions since not only the prediction accuracy increase

Table 3: Comparison of mean absolute error (MAE) of different algorithms at different resolutions

Temporal resolution / Algorithm	**Daily**	**Hourly**	**Two-hourly**	**Six-hourly**
SARIMAX	63.69	-	-	-
Prophet	55.37	-	-	-
Linear regression	-	5.45	8.60	29.58
Random forest regression	-	6.43	11.76	32.47
Random forest regression (optimized)	-	5.17	8.23	27.64

by about eight bikes a day on average but also prediction time is lower by using the Prophet algorithm (for detailed information, consult the Jupyter Notebook - Daily Data). Therefore, Prophet beat SARIMAX not only in effectiveness but also in efficiency.

Also, at sub-daily resolutions the random forest regression outperforms the simple linear regression model. In this case, however, better performance was only achieved in terms of effectiveness since fitting the random forest regression model requires more computational resources and time but increases prediction performance in terms of accuracy. Looking at Figure 8 it becomes clear that random forest is a better model representation of the underlying process since not only predictions but also the fit increases using this method.

Table 4: Comparison of normalized mean absolute error (nMAE) of different algorithms at different resolutions

Temporal resolution Algorithm	**Daily**	**Hourly**	**Two-hourly**	**Six-hourly**
SARIMAX	0.13	-	-	-
Prophet	0.12	-	-	-
Linear regression	-	0.27	0.21	0.25
Random forest regression	-	0.32	0.29	0.27
Random forest regression (optimized)	-	0.26	0.21	0.23

In terms of prediction accuracy at different resolutions the normalized mean absolute error has been used, because there were instances that containing zero values as their actual value, resulting in a division by zero error. Table 4 shows that daily prediction has the lowest nMAE compared to the metrics of sub-daily predictions. This finding combined with insights from Figure 7 indicates a high prediction accuracy at a daily resolution.

6 Summary and Limitations

This work has been the first attempt to provide reliable insights and predictive capabilities to enable Call a Bike to re-balance their bike fleet in accordance to demand patterns. With the help of descriptive analysis first valuable insights for the operational business could be gained unveiling hidden repeating demand patterns that could guide the re-balancing. The developed forecast provides good results for overall demand as a limitation however, direct empirical recommendations of how many bikes should be re-distributed to and from specific stations cannot be made yet. For a value-added use for Call a Bike, the demand should be predicted station wise as a next step. Further research should address the complex relationships that occur due to the fact that bikes can only be rented if they are available. Making use of station wise prediction is a non-trivial challenge, since the re-balancing and demand of bikes depends heavily on the underlying dynamics of this process, which should be examined as a refinement of this analysis.

A Appendix

A.1 Figures

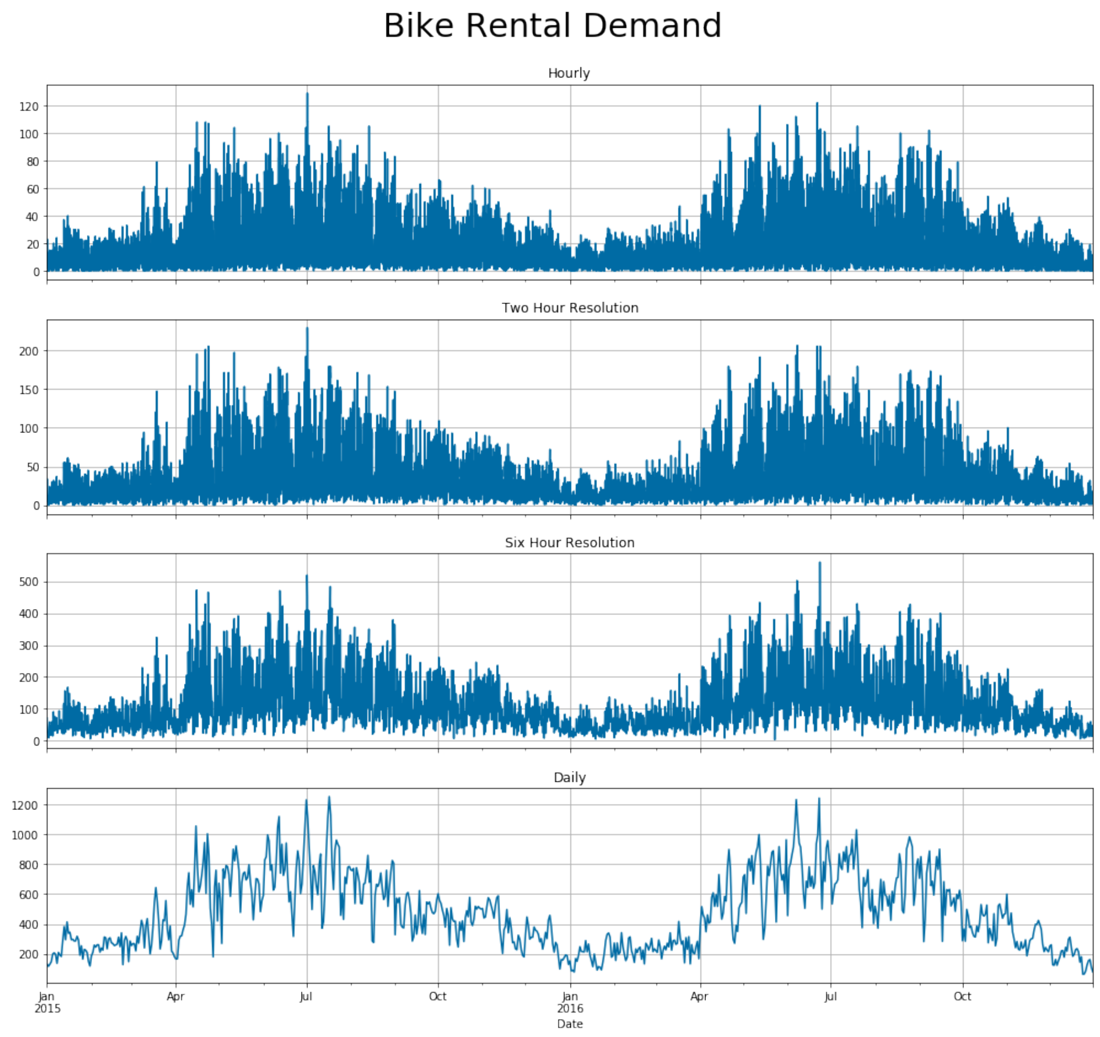

Figure 1: Bike rental demand in a one, two, six and 24 hour resolution for a two years period

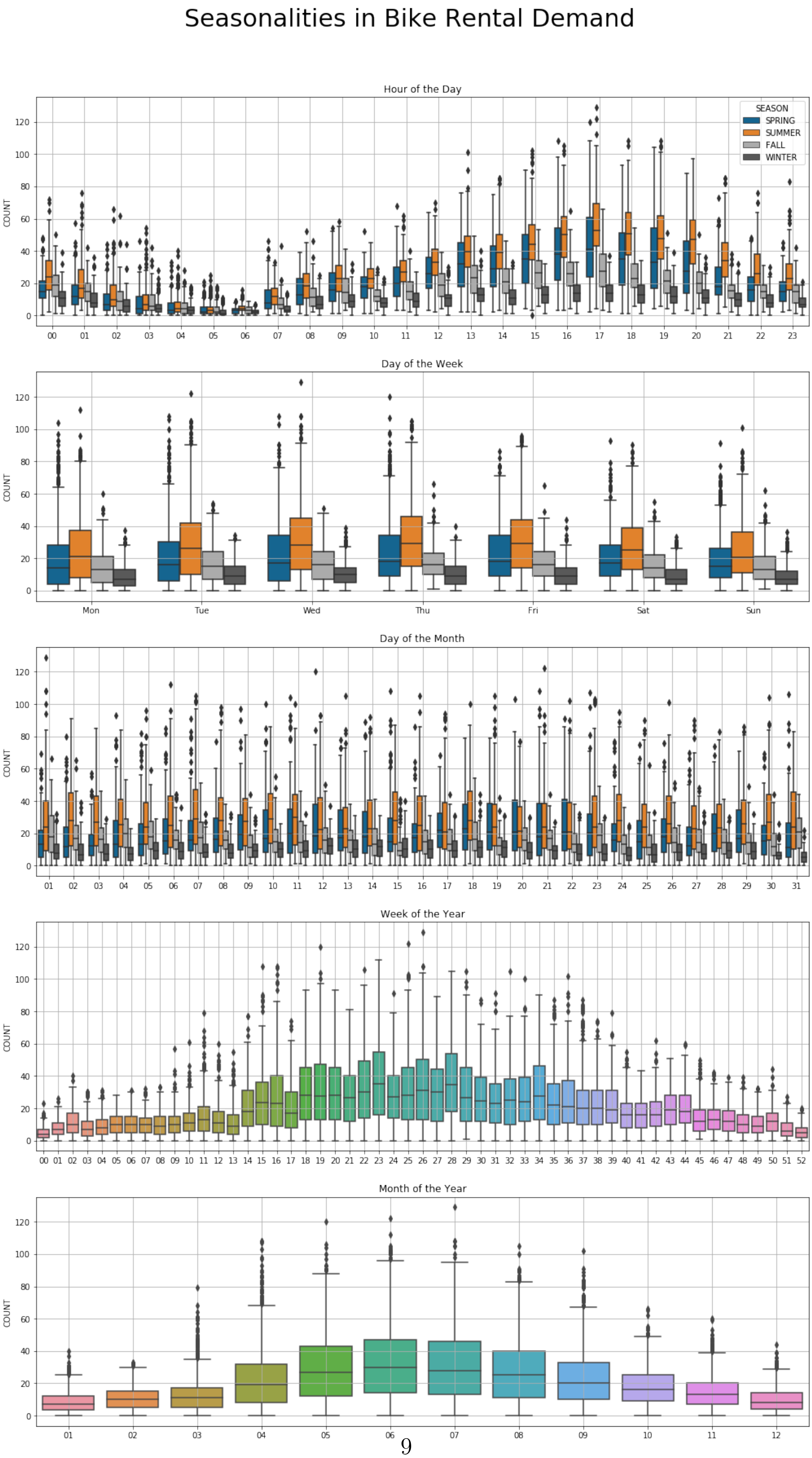

Figure 2: Mean shifts of bike rental demand over time visualized in a box plot

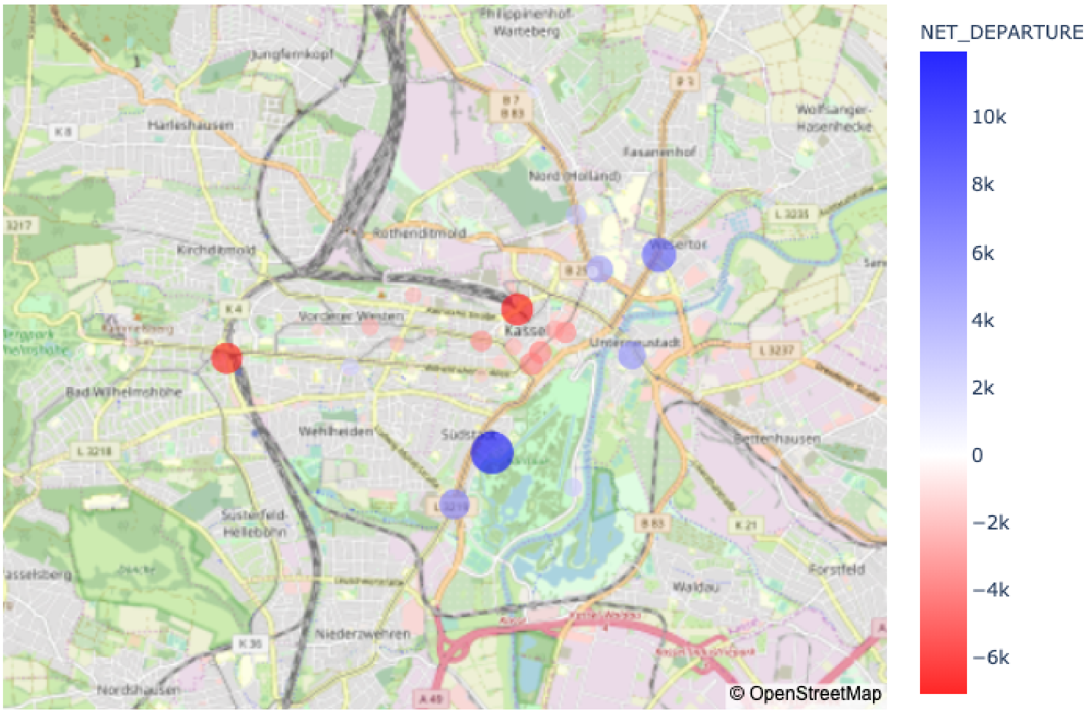

Figure 3: Station wise departures over the whole time

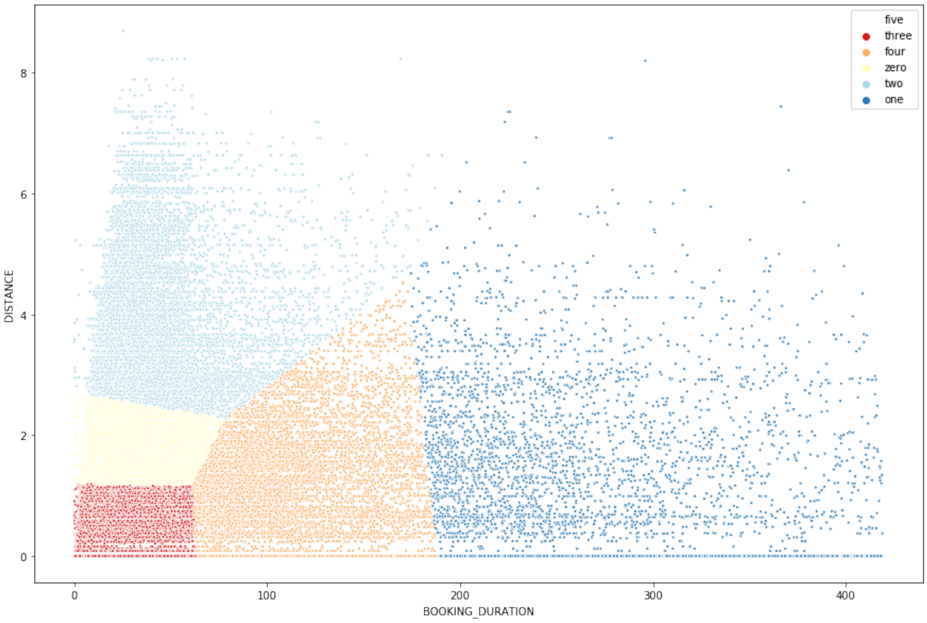

Figure 4: Customer clusters by booking duration and air distance

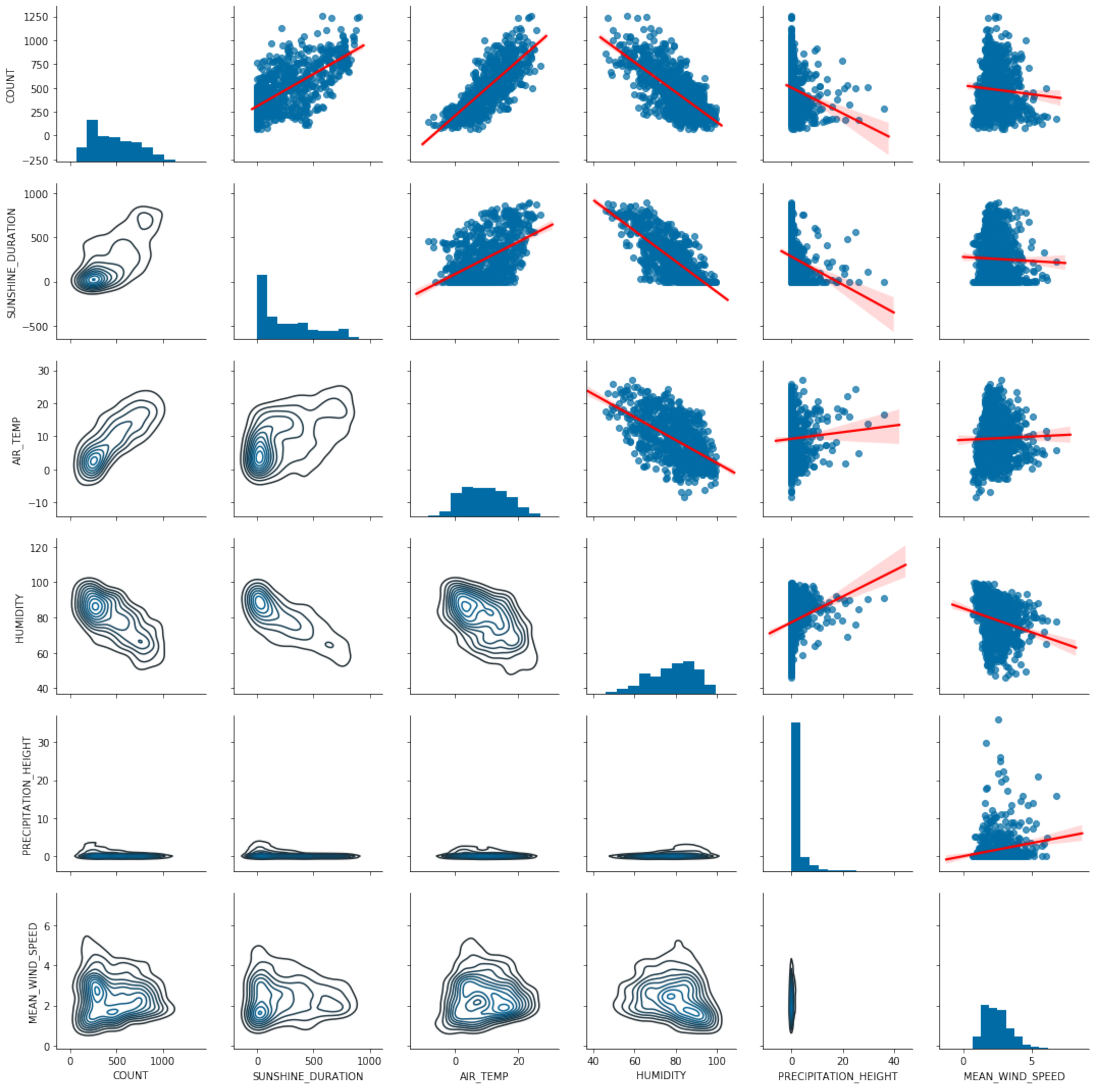

Figure 5: Pair plot of feature correlation

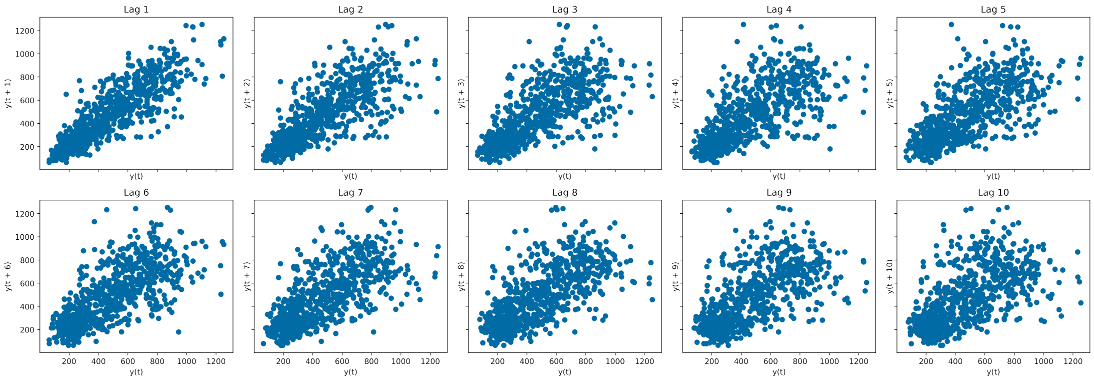

Figure 6: Autocorrelation of daily time series

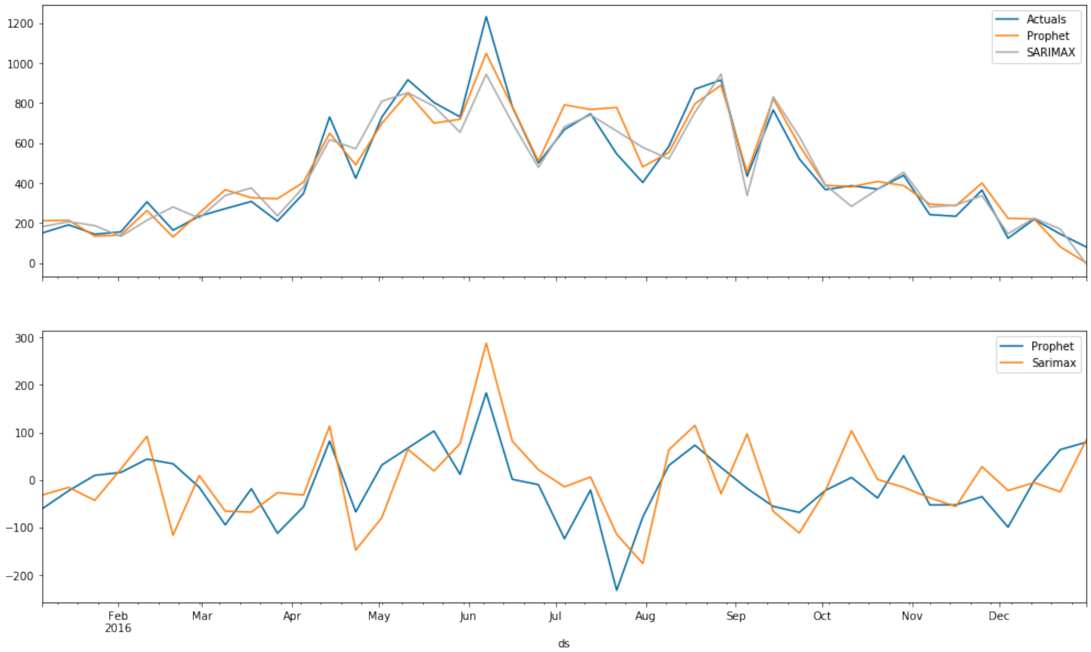

Figure 7: Daily prediction accuracy and residuals

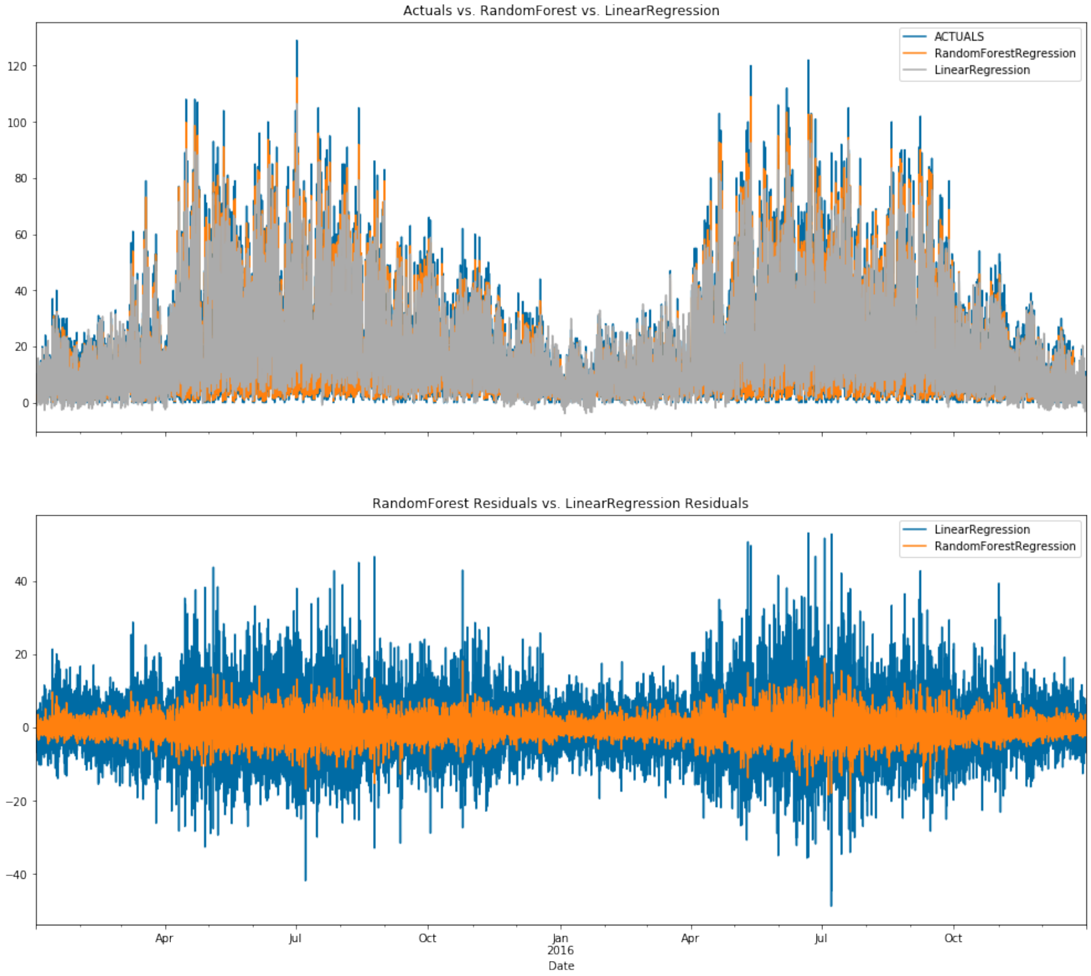

Figure 8: Performance comparison of linear regression and random forest regression

References

Call a bike - open-data-portal – deutsche bahn datenportal. (2016). Retrieved 28.12.2019, from `https://data.deutschebahn.com/dataset/data-call-a-bike`

Ricker, V., Meisel, S., & Mattfeld, D. (2012).

Optimierung von stationsbasierten Bike-Sharing Systemen, p. 217. Retrieved 28.12.2019, from `https://web.winforms.phil.tu-bs.de/paper/ricker/2012_ricker_opt_bss.pdf`