Analysis of Bike Rental Demand in Philadelphia

Team: Hacket
Authors: Dennis Wiesbrock (7349539), Aleksandar Garkov (7339054), Lars Hohmann (7355359), Karim Dilling (7371323), Marlon Engels (7364513)
Supervisor: Univ.-Prof. Dr. Wolfgang Ketter
Co-Supervisor: Karsten Schroer

Link to Repository: `https://github.com/lars-hmn/DSML_Hacket/blob/main/DSML_Team_Code_Hacket.ipynb`

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

September 12, 2022

Contents

List of Figures

1 Workload

1. Data Collection: Lars Hohmann
2. Data Preparation: Dennis Wiesbrock / Lars Hohmann
3. Visualize data for temporal demand patterns and seasonality
   - Average usage per weekday: Dennis Wiesbrock
   - Seasonal usage: Karim Dilling
   - Trip duration per weekday: Lars Hohmann
   - Seasonal patterns of usage on weekdays: Aleksander Garkov
   - Temperature and usage correlation: Marlon Engels
4. Geographical demand patterns: Dennis Wiesbrock / Lars Hohmann
5. Key performance indicators: Dennis Wiesbrock
6. Predictive Analytics
   - Feature Engineering: Lars Hohmann
   - Polynomial Ridge Regression: Dennis Wiesbrock
   - Gradient Boost: Aleksander Garkov
   - Random Forests: Lars Hohmann
   - Evaluation and outlook: Lars Hohmann
7. Project report
   - One page executive summary for non technical people: Marlon Engels
   - Problem description (business goal and data science goal): Karim Dilling
   - Data description: Karim Dilling
   - Brief data preparation details: Karim Dilling
   - Data analytics: Analytical methods applied and performance evaluation: Karim Dilling
   - Conclusions: Karim Dilling
   - Latex Formation: Marlon Engels

2 Executive Summary

The way people use transportation is changing. Especially in big cities like Philadelphia the use of sharing systems for Cars, Scooters and Bikes instead of owning those vehicles is increasing in popularity. Even people who own some kind of vehicle use sharing systems because of the given flexibility. This paper focuses on bike sharing systems. These are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able to rent a bike from one location and return it to a different place on an as-needed basis. The data generated by these systems makes them attractive for researchers because the duration of travel, departure location, arrival location, and time elapsed is explicitly recorded. Bike sharing systems therefore function as a sensor network, which can be used for studying mobility in a city. The market depends on a variety of factors hence why companies e.g. often struggle to accurately meet the demand. It is important to understand user behaviour itself and the effects of external influences like weather on this behaviour in order to optimize business strategies. This is why we developed a handful of models visualizing the data given. We worked with a dataset containing information about bike rental and weather in Philadelphia in the year 2019. As a first measure we had to get a deeper understanding of the topic and prepared the data adequate to our requirements. Then we merged weather data into the dataset to compare it to bike rental data. Now we just had to clean the data and fill missing values and could work with it. Our main goal was to find correlations and causes of user behaviour by visualizing the data. For example by plotting the use of bike rental per weekday we could verify our hypothesis on whether or not the usage increases. This information can then help to provide accurate demand for bikes. We then proceeded to discover what the Key Performance Indicators are. These should always be monitored by the fleet operator and consist of System Utilisation, Docks Availability and Bike Availability. Using the information we got until now, we then looked at Predictive Analytics using machine learning. First off, we build three different models consisting of Polynomial Ridge Regression, Gradient Boost and Random Forests. Evaluating those models we came to the conclusion to choose the Random Forests model for deployment. Speaking of Business Recommendations earlier there are multiple things we found to be important. Bikes are rented at predictable specific times and places and demand should be supplied accordingly. Maintenance of bikes should therefore be done during times of less rental demand. In summary, analysis of bike rental demand can be extremely important in forms of decision-supporting systems and can potentially lower expenses for the company.

3 Problem Description

The greenhouse gas emissions due to transport make up for the second largest chunk of total EU emissions. Hence it will be a requirement to change our approach to mobility. Additionally to the pollution aspect there are safety concerns and large amounts of space required for roads, parking and traffic congestion. Thus the mobility landscape is changing fast. Recently the trend of mobility as-a-service (MaaS) and on-demand (MoD) has been growing. Bike sharing platforms are one major example for this. Therefore we investigate the ubiquitous amounts of real bike sharing data to monitor and optimize operations, boost profitability and increase service level. Our assumption is that data science could enable fleet operators to improve their operations quality to target a real demand-focused bike sharing system, which possibly increases the use of environmentally-friendly bikes. In this paper we will focus on two core aspects:

1. System monitoring: Get a deep understanding of the operational performance of the fleet to make better business and operational decisions.
2. Demand prediction: Predicting the future demand as accurately as possible is important to provide great availability.

4 Data Description

The provided data set gives us detailed information about bike rentals from indego in the city Philadelphia 2019. The data set consists of 744260 different rentals and 8 different attributes. Additionally we added weather statistics for Philadelphia from 2015 to 2020. It consists of 43848 entries and 4 attributes.

5 Data Preperation

As a first step we made sure that the time stamp data of the columns “start_time” and “end_time” of the bike sharing dataset is in the right datetime format. Then we added a duration attribute to the data that measures the duration of each ride. Additionally we added an attribute that shows what day of the week each data was for further analysis of differences between weekday and weekend. After that we merged the bike dataset with the weather data. Because we had weather data from 2015 to 2020, we only merged the relevant weather data from 2019. The next step was data cleaning. Therefore we collected the number of missing values and detected that 35 rows had a missing value. Compared to the number of 744260 totals rows this is an insignificant number and thus we decided to drop these rows. Next we dealt with outliers. Possible outliers are rides with a very high or low duration. We dropped rows with a duration $<=0$ or a duration with $>$ 3 standard deviations because these are not representative. For the geographical demand patterns and KPIs we collected data from the indigo API to get real-time bike station data. Firstly we transformed the data into a data frame to get an overview. Then we added coordinates of the start and end station to our already prepared data frame. Null values were dropped.

6 Descriptive Analysis

6.1 Temporal Demand Patterns and Seasonality

6.1.1 Average Usage per Weekday

To get an overview about the weekdays with the most demand we plotted the average number of bike trips for every weekday in a barplot (Figure 1). While during the week (Monday-Friday) the trip count varies marginal, you can see that the demand for bike trips on weekends drops. This could be explained by people not needing to be at their workplace. The boxplot graph (Figure 9) with the hourly number of trips shows that there are outbreaks on around 8am and around 5pm. This is probably around the time that most people have to get to their workplace or leaving their workplace to go home. In between these spikes the number of bike trips is also elevated compared to earlier and later than these rush hours. This could be explained by business traffic (e.g. meeting clients) and also by people being more likely to ride somewhere during the daytime.

6.1.2 Trip Duration per Weekday

When plotting the average trip duration per weekday (Figure 2) you can see that while the average usage per weekday dropped as shown in 6.1.1, the duration of the trips is longer at weekend. This can be explained by people not making short trips to their workplace, instead they are making longer bike-tours with friends.

6.1.3 Seasonal Usage

The number of bike trips per season (Figure 7) shows that the favourite time for bike trips is clearly the summer and the least favourite time is the winter while spring and fall are pretty equal in the middle between those two and on a similar level. One can conclude that the weather plays a significant role in whether people are renting a bike or not.

6.1.4 Usage per Month

The bar plot with the sales on a monthly basis (Figure 10) shows that the sales are significantly higher in the warmer part of the year. The most sales are made between June and October. Before and afterwards sales drop significantly. In the winter months the sales are even less than half as much as in the summer. The sales per weekday barplot graph shows that on weekends the sales are lower. So in conclusion with the graph of the trip duration in 6.1.2 one can say that the trips longer trips do not have a positive impact on sales.

6.1.5 Weather and Usage Correlation

The regression plot (Figure 4) shows that there is a positive correlation between the temperature and the number of bike trips and up to 20°C even a fairly linear one. For precipitation (Figure 5) there is a negative correlation between the amount of precipitation and the number of bike trips. So both regression plots confirm that the better the weather is, the more bike trips are made. The bar plot for the temperature and usage correlation (Figure 3) shows that the most popular temperature for a renting bike is between 20 and 25°C. From there on, the colder or warmer it is the fewer bike trips are made. This further confirms what we found for the seasonal usage.

6.2 Geographical Demand Patterns

6.2.1 Heatmaps

When comparing the heatmap with the start locations of the bike trips with the one that displays the end locations one can at first see that the end locations are a little more spread out than the start locations. One can also see that on the heatmap for the start locations there are more bikes in the north of the city while on the heatmap for the end locations there are more bikes towards the western side of the city. This is interesting data that helps on how to distribute the bikes in the city across the stations. At the start of the day it is probably better to have more bikes in the north of the city.

6.2.2 Map with station names

The map with markers for the station names shows that the most popular ones are at 15th & Spruce and Rodin Museum. Other popular stations can be found at 23rd & South, University City Station, 17th & Locust, Philadelphia Museum of Art and the Pennsylvania & Fairmount Perelman Building. They are mostly placed at the city center or close to it. Cultural facilities or universities are points of interest and thus they have higher bike traffic.

7 Key Performance Indicators

The following KPIs should be displayed on a dashboard for the fleet operator. Because we made API requests about the station information, the data will be in real-time. With the KPIs the fleet operator will be able to get an overview of the total system utilisation and also about the total docks and bike availability. Additionally we implemented a function to highlight stations with low bike availability (<15%). This will help the fleet operator to manage e.g. at which stations bikes are highly demanded. High dock availability should correlate with low bike availability at the same station and vice versa. Therefore the fleet operator can either monitor the dock availability or the bike availability by stations. For the observed time frame 595 out of 1677 bikes were not available for a ride (on ride or not available for ride, e.g. repair). This means a total utilization of approximately 35%. This means a total utilization of approximately 35%. If the total utilization will increase extremely, the fleet operator could send new bikes out of the headquarter to highly requested stations. The total docks availability at all stations is around 61%. If the bike availability goes below 15% the station should by watched by the fleet operator and new bikes should be sent if the bike availability decreases further. The total bike availability at all stations is around 34% and the number of stations with bike availability below 15% is at 37 from 153 stations.

8 Predictive Analytics

8.1 Feature Engineering

In the descriptive analytics part inputs like the time of day or the weather have an impact on the bike sharing demand. We create a new dataframe with hourly values to predict the total system-level demand in the next hour. This will be our dependent variable y. Further we defined our independent variables, which we concluded out of the descriptive analytics: “is_weekend”, “is_rushhour” (usual working hours), “is_daytime” and “temp” (average temperature) and “is_rain”.

8.2 Model Building

We decided to go with polynomial regressions, because linear regressions would not be appropriate to the fluctuating bike demand. We made Polynomial Ridge Regression, Gradient Boost and Random Forests and later compared the models. Further we decided to split our dataset into a 70% training set and a 30% validation set to prevent an overfitting of our Regression models.

8.2.1 Polynomial Ridge Regression

The Polynomial Ridge Regression belongs to the regularized regression methods. This means that we add a term to the canonical machine learning problem and this term will penetalize the magnitude of coefficients, which leads to a more generalized model. With this approach we counteract the high impact of high polynomial features and prevent overfitting, which is one of the great advantages of this regression. Overfitting occurs when the trained model performs well on the training dataset and performs poorly on the testing dataset. An disadvantage could be that the Ridge Regression trades variance for bias. You can see the high bias compared to the other regressions in the model evaluation part. There you can see that the MAE and MSE are significantly higher than the others.

8.2.2 Gradient Boost

Gradient boosted machines (GBMs) are extremely popular machine learning algorithms. GBMs build an ensemble of shallow and weak successive trees one at a time, where each new tree helps to correct previously made errors and therefore improve previously trained trees . With each tree added, the model becomes even more expressive. There are usually three parameters- number of trees, depth of trees and the learning rate of each tree. Gradient Boost often provides predictive accuracy that cannot be beaten by other algorithms. Due to the fact that it can handle missing data, imputation of data is not required. It also offers a lot of flexibility, because it can optimize on different loss functions. One of the biggest disadvantages of GBMs is that it continues improving to minimize all errors in the trees. This can overemphasize outliers and cause overfitting. A way to neutralize this, is to use cross-validation.Secondly, using this algorithm might be computationally expensive, because it often requires many trees, which can be time and memory exhaustive.

8.2.3 Random Forests

The random forest algorithm builds multiple decision trees of different architecture and merges them together to get a more accurate prediction. In contrast to the gradient boosting algorithm each tree gets trained separately. For regression problems the predicted outcome is the mean computation from all trees. An advantage is that it tackles the bias problem with single decision trees. Furthermore the tendency to overfit gets smaller the more trees you add. On the other hand random forests are hard to interpret due to their randomness and require a relatively large dataset to work properly.

8.3 Model Evaluation

When evaluating the different models one can state that regarding mean absolute error (MAE), mean squared error (MSE) and the accuracy (R2) the gradient boosting regression performs the best as it has the lowest MAE and MSE as well as the highest precision. The random forest regression is not far off though. It has very similar metrics. Only the ridge regression performs poorly compared to the other two. We also found out that the ridge regression and the gradient boost regression predict a negative demand for some inputs which is in contrast to a real life scenario where the minimum demand would simply be zero. So when comparing all three regressions overall we come to the conclusion to choose random forests as our model for deployment as it is very close to the gradient boost regression in terms of MAE, MSE and R2 while not having the issues with negative demand and as such it creates a more realistic prediction.

8.4 Outlook

Machine learning models can be refined by either tweaking the input features or the hyperparameters of the algorithm. So one way to improve the model can be collecting more data in order to have a bigger feature set to extract features from. This could enable us to create new features such as "wind_speed" which could also have an impact on the bike sharing demand. It would also allow for examining features in terms of their influence and correlation. This could lead to combining or even dropping some of the features. When regarding the hyperparameters we only adjusted the max_depth (= maximum number of levels in each tree) and used the default settings for the rest. Our model could be improved by adding and tweaking some of the other hyperparameters such as min_samples_leaf (= minimum number of data points allowed in a leaf node).

8.5 Business Recommendations

Overall our observation is that the time, weather and location play a significant role in the bike demand. The peak hours at around 8am and 5pm have the highest bike demand by far. The fleet operator has to ensure that bikes are available at that time and not e.g. brought to repair. Additionally before and after that time would be a good time for maintenance work. On weekends the duration of the trips increases while the amount of trips decreases a bit. This also means that there is some free space for bike maintenance. In general, since the bike trips on weekends tend to be longer, it is important that the bikes are checked and maintained properly beforehand. In general the warmer the weather and the less rain there is, the more bike demand there is (unless it gets to really warm temperatures around 30°C or higher). This also correlates with summer being the time with the highest demand. So one has to make sure that the bikes are properly maintained for the summer, so that most bikes will be available during that time. Lastly the heatmaps showed us that during the start of the day it is probably better to have more bikes in the north of the city. In general popular places with high demand are close to cultural facilities and close to the city center so the stations there need to be populate with more bikes compared to stations in the outer area of the city.

A Appendix

A.1 Figures

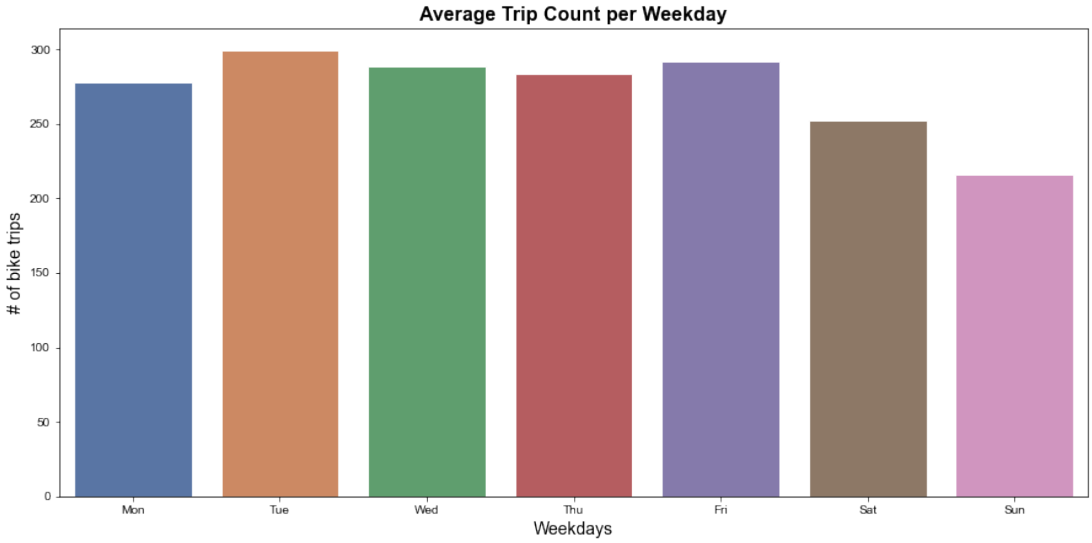

Figure 1: Average Trip Count per Weekday

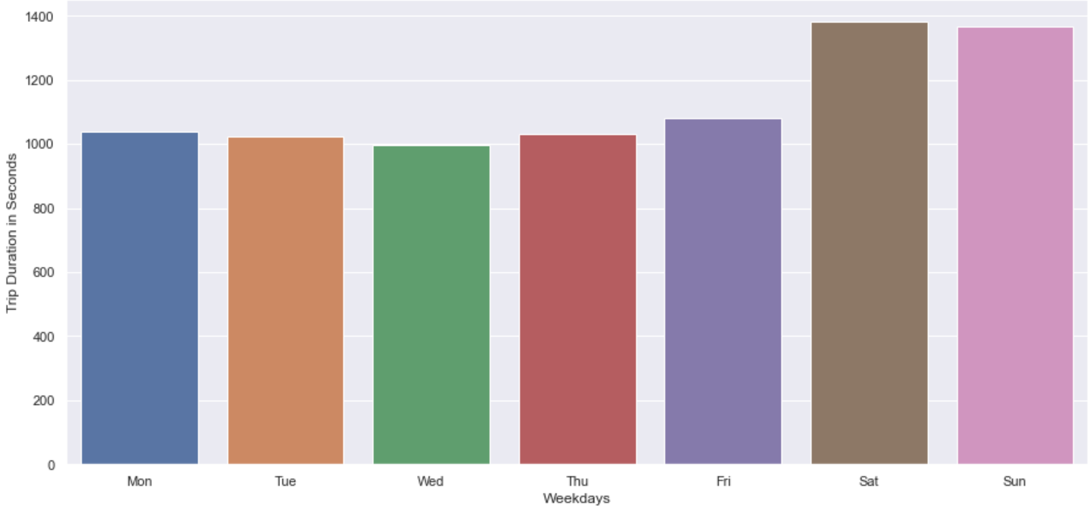

Figure 2: Average Trip Duration per Weekday

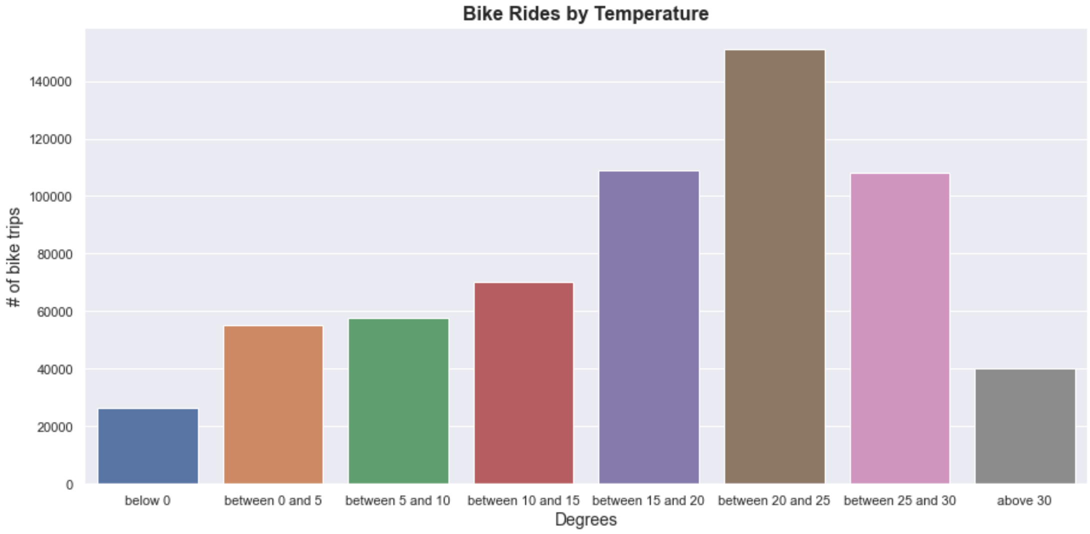

Figure 3: Bike Rides by Temperature

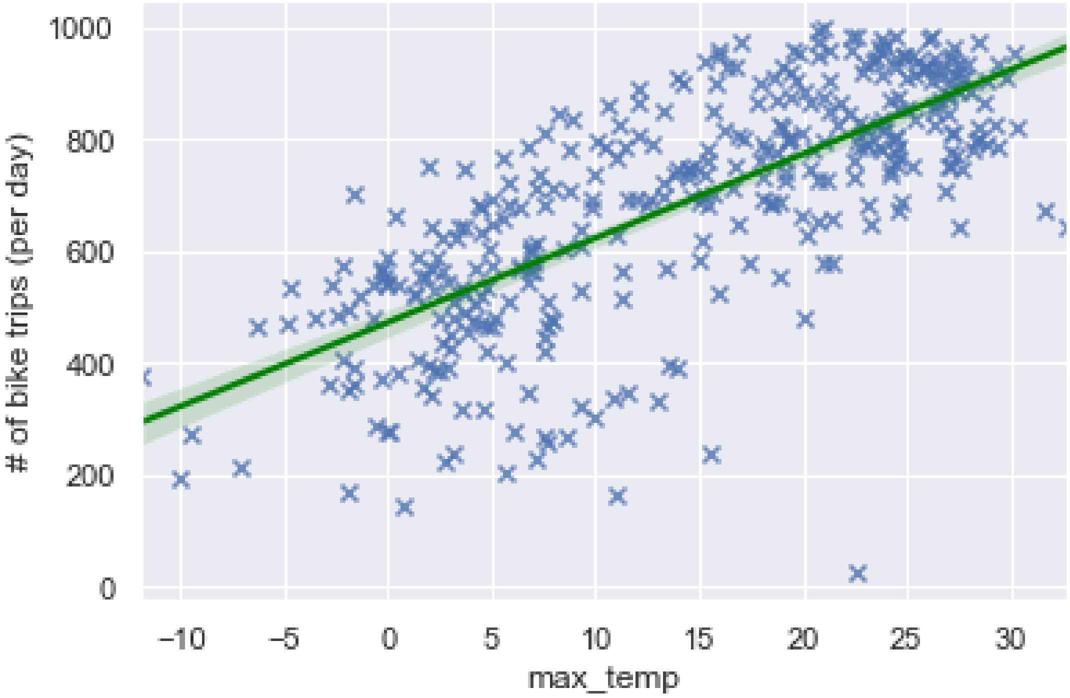

Figure 4: Correlation between bike trips and temperature

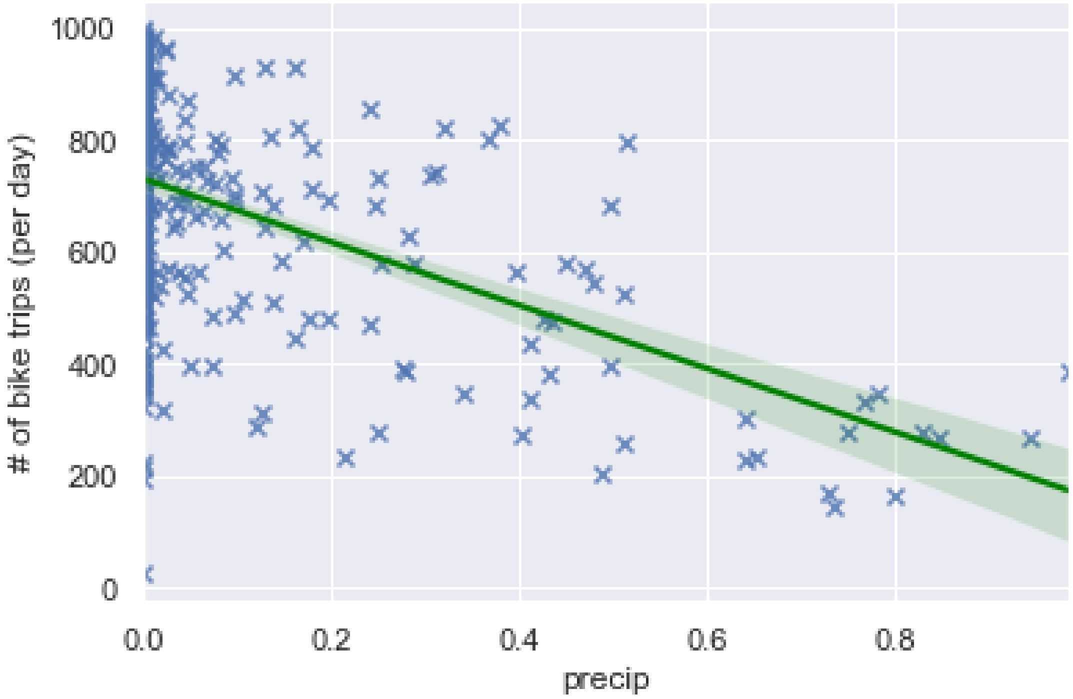

Figure 5: Correlation between bike trips and precipitation

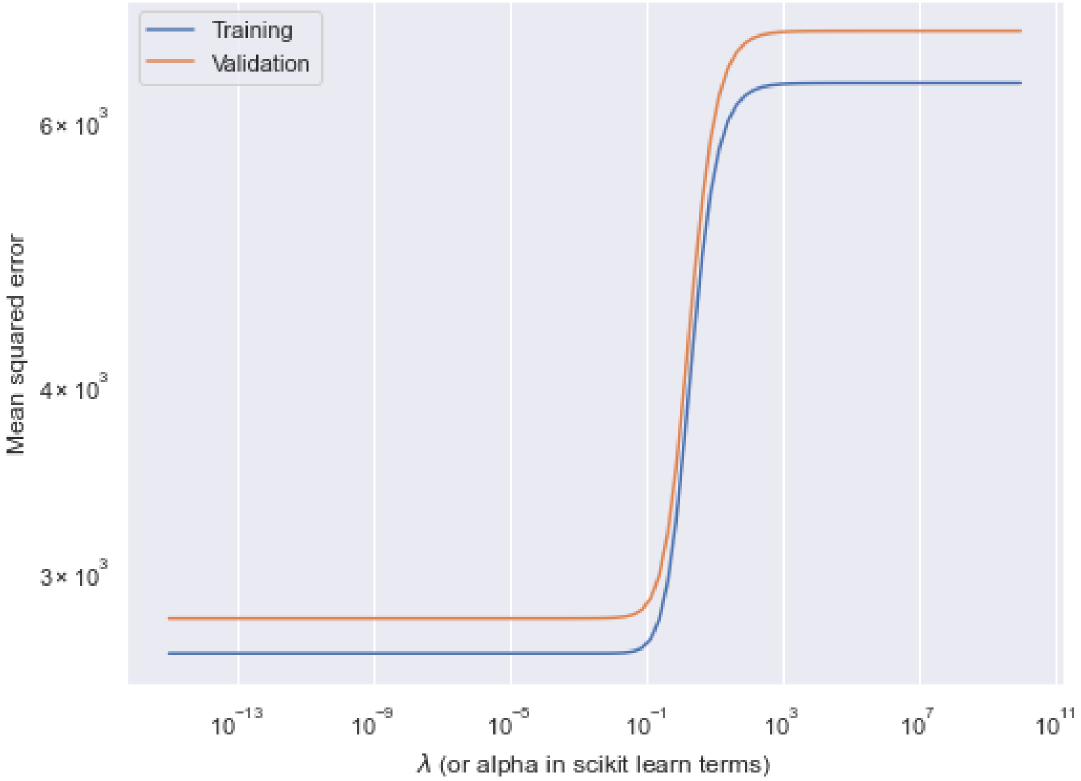

Figure 6: Polynomial Ridge Regression

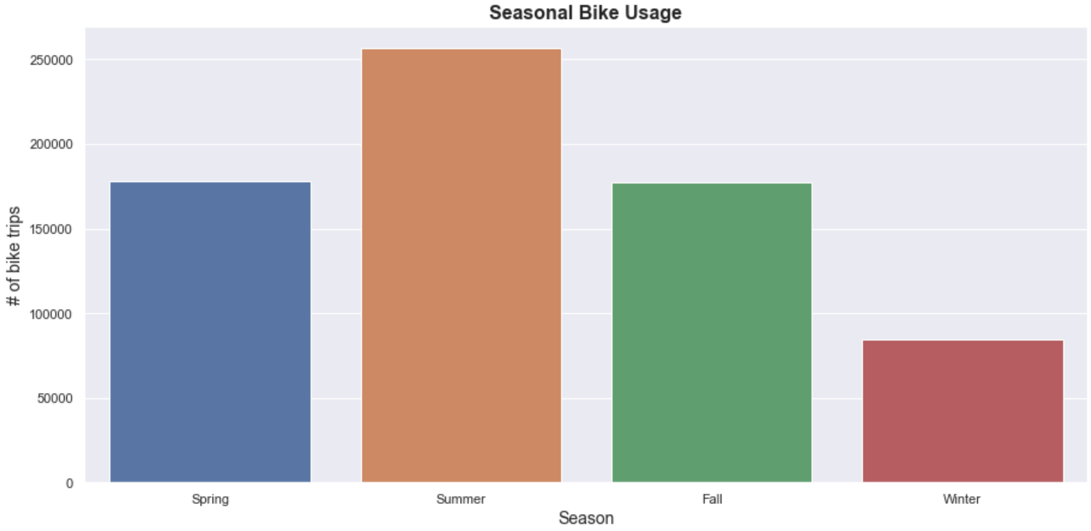

Figure 7: Seasonal Bike Usage

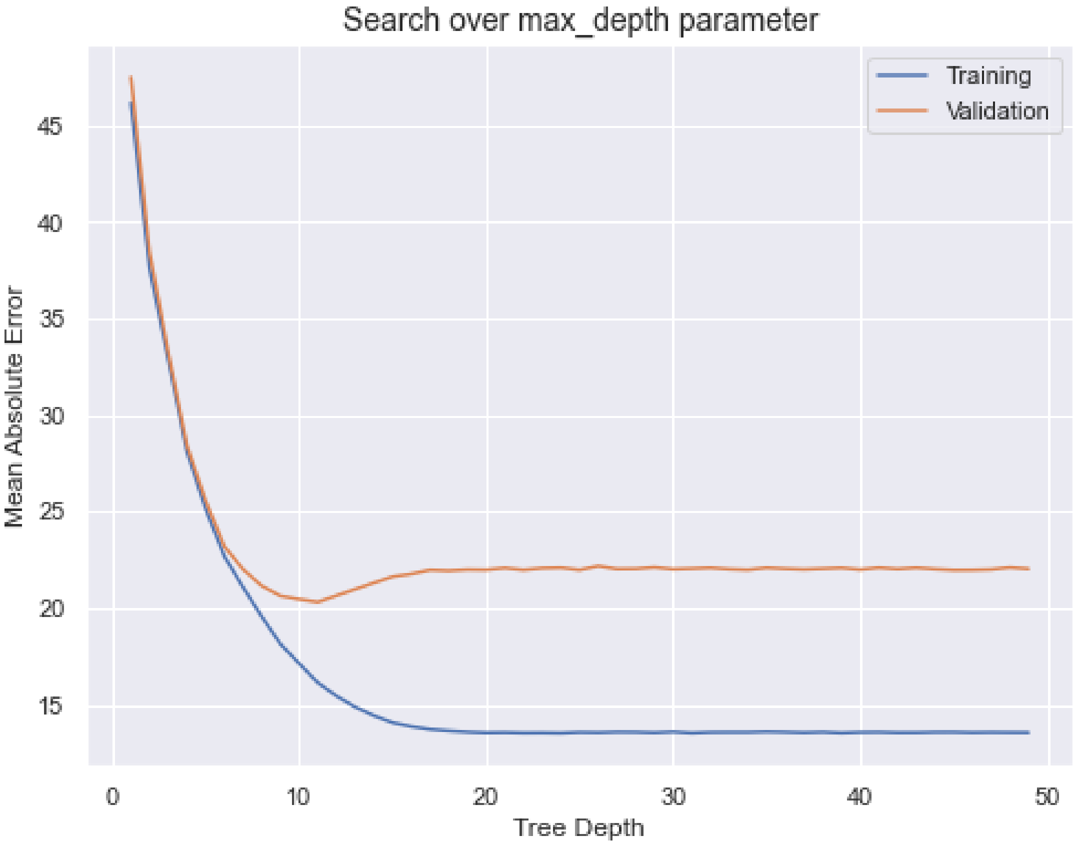

Figure 8: Tree Depth

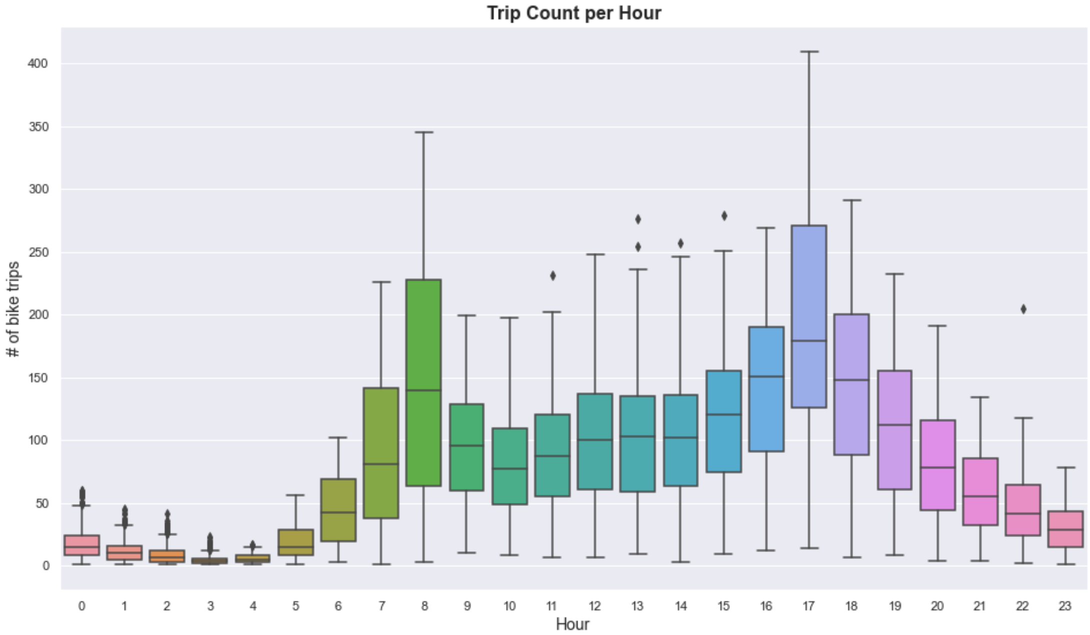

Figure 9: Trip Count per Hour

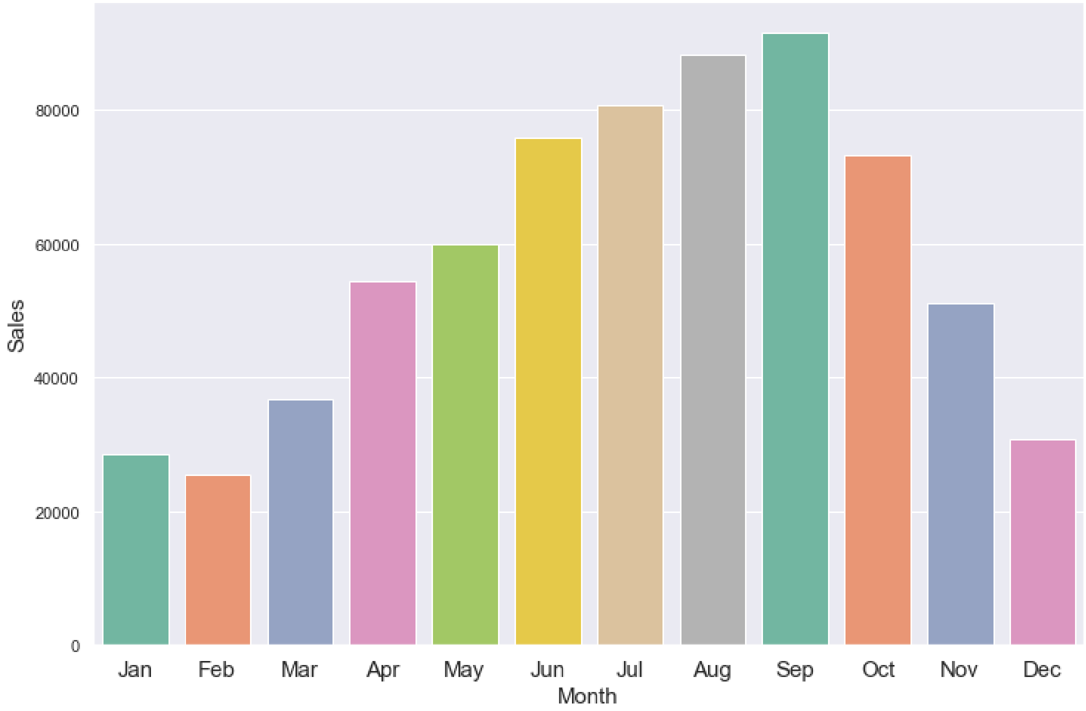

Figure 10: Total Sales per Month (Barplot)

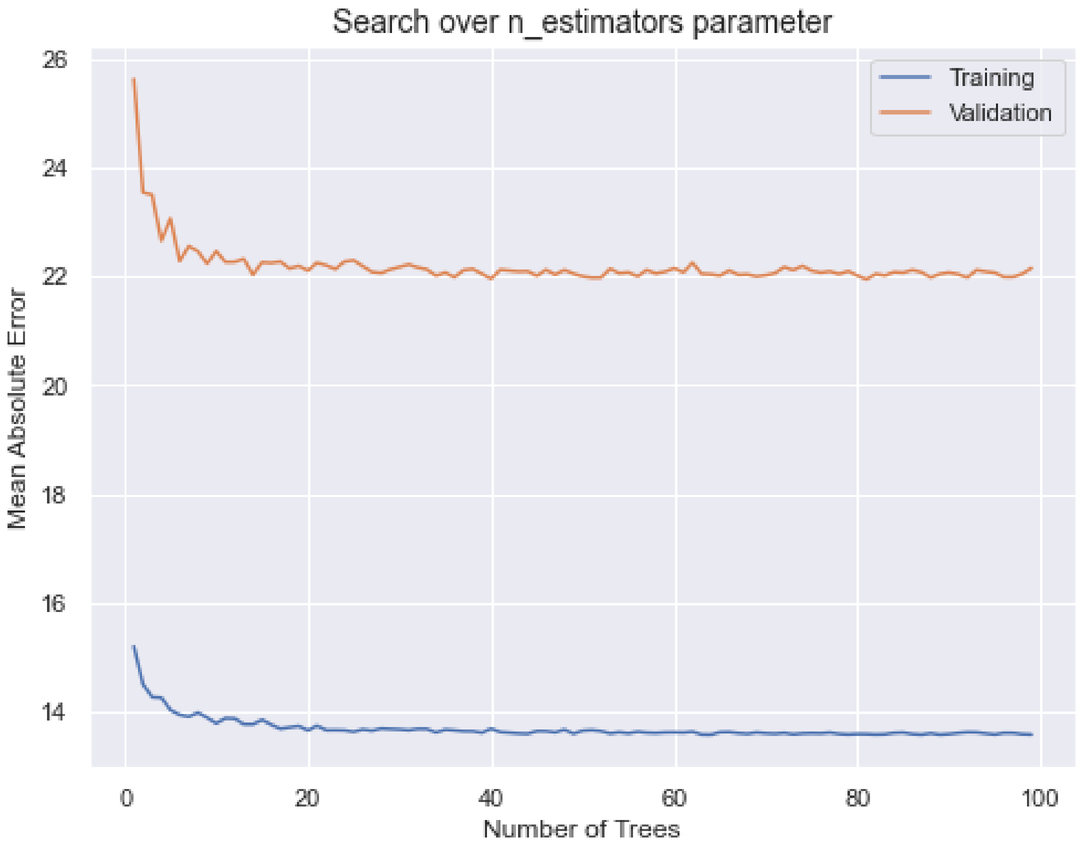

Figure 11: Random Forests