Seoul bike rental

Seoul bike rentals

Jacob Jacobsen, Jaime Bendala Oliveira, Mathieu Boers, Pietro Degasperi

Exam numbers: 4112340, 4112595, 4112204, 451097

22-12-2022

**Data Science and Machine Learning**

Julia Pahl, Jie Cai, Marina Meireles Pereira

30 pages

37932 characters

Abstract

By analyzing the Bike rental count in Saul this paper found that bike rental is heavily used during commuting hours. The weekly pattern is divided into a Monday-to-Friday and Saturday-Sunday pattern. Further, a correlation between the weather temperature and the bike rental count is found. To support the usage of bikes, a prediction model is suggested using the random forest tree-based algorithm. The model has a mean average error of 144 and based on the range of bike rentals the model could be used to predict the bike rental count.

Contents

[1. Introduction 2](#_Toc122512349)

[2. Data presentation 3](#_Toc122512350)

[2.1 Data cleaning 4](#_Toc122512351)

[3. Methodology 6](#_Toc122512352)

[3.1 Framework 6](#_Toc122512353)

[3.2 Research Methodology 7](#_Toc122512354)

[4. Exploratory Data Analysis 8](#_Toc122512355)

[4.1 Boxplot 8](#_Toc122512356)

[4.2 Variation of bike rentals 9](#_Toc122512357)

[4.3 Seasons 10](#_Toc122512358)

[4.4 Days and hours 11](#_Toc122512359)

[4.5 Correlation of the parameters 12](#_Toc122512360)

[4.6 Temperature 13](#_Toc122512361)

[5. Prediction method 15](#_Toc122512362)

[5.1 Random Forest 16](#_Toc122512363)

[5.1.1 Feature importance 20](#_Toc122512364)

[5.1.2 Prediction with Random Forest 21](#_Toc122512365)

[5. Discussion 23](#_Toc122512366)

[7. Conclusion 24](#_Toc122512367)

[Bibliography 25](#_Toc122512368)

[Appendix 1. Data cleaning steps 27](#_Toc122512369)

[Appendix 2. Data Science lifecycle 29](#_Toc122512370)

# 1. Introduction

Data science is the combination of multidisciplinary techniques that aims to extract and analyse meaningful insights from data, which then could be applied and used to guide decision making and strategic planning (Education, 2022).

For the following report, a given dataset “*Seoul Bike Sharing Demand Data Set*”, describing the number of bikes shared in Seoul, will be analysed. Furthermore, the objective of the report is to further expand the knowledge learnt during the *Data Science* course, therefore, the theory used during the report will be linked to that one given during the course and self-research. Consequently, the dataset will be analysed using Python.

As mentioned before, the main topic of the report is *bike sharing in Seoul.* Nonetheless, bike sharing is not just a relevant topic for the city of Seoul, but worldwide. In fact, the bicycle sharing system has grown significantly during the last few decades, almost everywhere where the infrastructure allowed, becoming a common method of transportation for individuals. The bicycle sharing system allow the public to commute utilizing a sustainable and convenient transport, increasing the mobility comfort (Cho, 2020).

The bike-sharing scheme has been implemented in cities across different continents, creating a positive impact on urban traffic, greenhouse emissions, public health and increasing the cycling community (Sathishkumar V E, 2020). In the case of Seoul, a gradual plan to cover the entire population (9,975,000 people in 2022 (Review, 2022)) was made, increasing the number of stations from 150 to 1500, and 20,000 bicycles during 2016 (Cho, 2020).

The following report is divided into 7 different sections, containing a presentation of the data, a description of the methodology followed, an exploratory analysis, a model fit, classification methods, a discussion, and a conclusion, furthermore, the report also includes the bibliography and appendices.

The objective of the report is to predict the number of bikes rented during a given period by implementing a model in Python. This prediction could be used to guide decision-making and strategic planning in terms of maintenance, rental prediction and multiple other topics.

Furthermore, a set of research questions linked to the aim of the report have been settled, thus providing some hypothetical aim to the analysis. The following questions relate to the maintenance, prediction and what affects bike rental.

* *“When should the maintenance take place on the bikes?”*
* *“What conditions lead to an enhanced use of bikes in Seoul?”*
* *“Could the prediction be accurate enough to use in a real case scenario?”*

By having in mind, the previously presented questions, the analysis can be aimed to provide answers to these, and therefore solve the assignment presented during the *Data Science and Machine Learning* course.

# 2. Data presentation

The “Seoul Bike Sharing Demand Data Set” Dataset is retrieved from the UCI archive (Repository, 2020) and includes statistics for an entire year.

The most interesting value in the dataset whose trend wants to be understood is the number of bike rentals, the dataset provided gives this information in form of the number of bicycles rented every single hour.

Several other variables are available, which should be investigated and analysed by comparing them with trends in the number of rentals, all the variables included in the dataset are shown in **Error! Reference source not found.Error! Reference source not found.**.

|  |  |  |  |
| --- | --- | --- | --- |
| Name of the variable | Data type | Units | Description |
| Date |  | YYYY-MM-DD | Input variable |
| Hour | 0 - 24 | H | Input variable |
| Rented Bikes | quantitative | Bikes/hour | Output variable |
| Temperature | quantitative | °C | Input variable |
| Humidity | quantitative | % | Input variable |
| Wind Speed | quantitative | m/s | Input variable |
| Visibility | quantitative | m | Input variable |
| Dew point temperature | quantitative | °C | Input variable |
| Solar radiation | quantitative | MJ/mq | Input variable |
| Rainfall | quantitative | mm | Input variable |
| Snowfall | quantitative | cm | Input variable |
| Season | categorical | Winter/Spring/Summer/Fall | Input variable |
| Holyday | binary | Y/N | Input variable |
| Functional | binary | Y/N | Input variable |

Table 1. Type of variables in the studied dataset.

The data set consists of 8760 observations (hours in a year) with 14 variables where most of which are part of weather conditions and time/date information, there is only one output variable which is the number of rented bikes.

As previously mentioned, the Python programming language is going to be used for the study of this data frame, moreover, the libraries mentioned below will be downloaded and installed, to allow the work to be done more efficiently and quickly.

In particular, it is important to mention the Pandas library, which once downloaded can invoke using the following command: "import pandas as pd" Pandas is a free downloadable library that allows to manipulate and analyze of large amounts of data, offering pre-programmed functions to manipulate tables of different types of values, for more information, see the official website (Pandas, 2022).

Another very relevant and powerful plugin was used to create the graphs: matplotlib, which can be imported with the command "import matplotlib.pyplot as plt" This library allows to create a wide variety of visualizations that can be static, animated and even interactive, with a style similar to the Matlab program (Matplotlib, 2022).

Last, the NumPy library can be useful for working with large tables of numbers and matrices, offering functions to simplify work on arrays; the callback method is “import numpy as np” (Numpy, 2022).

## 2.1 Data cleaning

Thanks to the Python programming language conjugated with the above plugins, it is possible to create an exhaustive data visualization. Nonetheless, before starting to analyse and graphing, it is necessary to clean the data to ensure that it is accurate, consistent, and ready for use. This includes identifying and correcting errors and inconsistencies in the data, as well as formatting the data in a way that would be suitable for the analysis. The following steps were taken to clean the data:

* Correctly importing the data
* Checking for missing values
* Creating dummies variables
* Normalizing the data

A more detailed description of the above-mentioned steps and how they have been performed can be found in Appendix 1.

By cleaning and preparing the data, the quality and reliability of the data are improved, which will help to ensure that the analysis is accurate and informative.

Inaccurate data can lead to incorrect or misleading conclusions: If the data is not accurate, it can lead to incorrect or misleading conclusions about the relationships between different variables. Furthermore, data that is not consistent can be difficult to analyze and compare, as it may not be clear how different data points are related to each other.

After validating the data frame, the values of the output variable ("Rented bikes") have been plotted in a scattered plot in relation to its date of use ("Date"), a first result can then be obtained visually and is presented in Figure 1.

Chart, scatter chart

Description automatically generated

Figure 1 - Rented bikes during a year.

It is still early to draw conclusions from this scattered plot, further investigation and analysis has to be made in order to discover more meaningful information.

# 3. Methodology

During the selection of the methodology and structuring for the report, the *Data Science Life Cycle* (see Appendix 2) was used as referenced to create a structured and compelling analysis. Furthermore, an exploratory approach has been selected for the completion of the study. Therefore, to complete the mentioned analysis and subsequent report, a framework was created.

## 3.1 Framework

The following framework, seen on figure 2, establishes and describes the different stages and phases the project is expected to go through to be completed.

Stage 1, Stage 2, and Stage 3 consist of the preliminary work needed to be completed before the analysis could be conducted. During these 3 stages, there was no data handling, besides an introduction to the data set and an introduction to the given case study, this was later used to properly identify the problem statement and therefore the objective of the analysis. Furthermore, a preliminary literature review on different models was completed to understand which one could be a better fit, nonetheless, this part has not been included in the report.

Stage 4 consists of a larger stage, which is also divided into 4 different phases. During this stage, the data is presented and cleaned, preparing this one to be analysed. Moreover, different classification methods have been used to select the best-fitting model(s). These methods are KNN, Linear Regression, AdaBoost, and Random Forest. The data has been split in three identical sections, where 2/3 have been used for the training of the model(s), and the remaining 1/3 has been used for testing. The results of the study will be discussed and concluded at the end of the report.

Table

Description automatically generated with medium confidence

Figure 2 - Framework for the report (Kannan, 2018).

## 3.2 Research Methodology

As mentioned before, most of the knowledge needed for the study was acquired during the lectures and preparation for *Machine Learning and Data Science* course. Nonetheless, self-research was also conducted to expand this knowledge and give more depth to the learning of the report. Part of this research was done through the reading of online forums, where similar models were completed. Consequently, the research methodology consisted in gathering the necessary information from different sources, lectures, and self-research, to successfully complete the assignment, and hypothetically respond to the research questions.

# 4. Exploratory Data Analysis

In this section a better understanding of the dataset is given with Exploratory Data Analysis (EDA). Within EDA graphical representations and statistical techniques are used to obtain insights from the dataset. EDA helps first with the formulation of the hypotheses and can help with the validation of them (Ghosh, 2018). EDA will in this project be used to better understand the dataset, this will be done in a graphical way. Next to this EDA helps in this project to formulate hypotheses and the research questions.

## 4.1 Boxplot

The next step is to look at the graphical representations to better understand the data. The first step is to create a plot where the boxplots of the bike rentals are shown per month. This will give an inside in the minimum observation, 25th percentile (lower quartile), median, mean, 75th percentile (upper quartile) the Maximum observation. As shown in Figure 3. Explanation Boxplot (Institute, 1999). (Institute, 1999).

Chart, box and whisker chart

Description automatically generated

Figure 3. Explanation Boxplot (Institute, 1999).

As shown in Figure 4. Boxplots of the bike rentals over the months. the minimum number of bikes rented are zero. When looking at the maximum observations, there are a few potential outliers. These are mostly occurring when there is a big growth, like in March and when the summer season is.

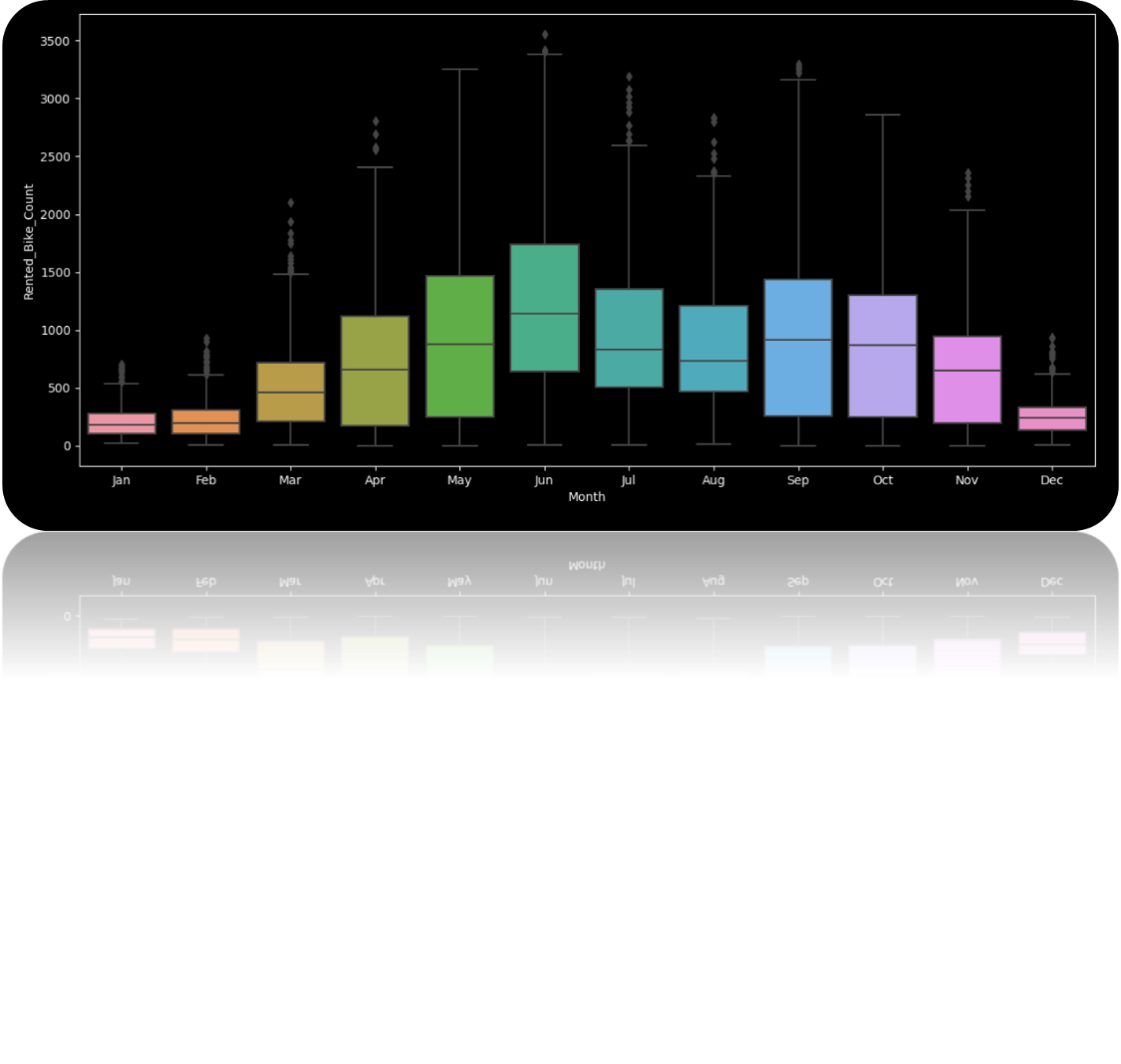


Figure 4. Boxplots of the bike rentals over the months.

## 4.2 Variation of bike rentals

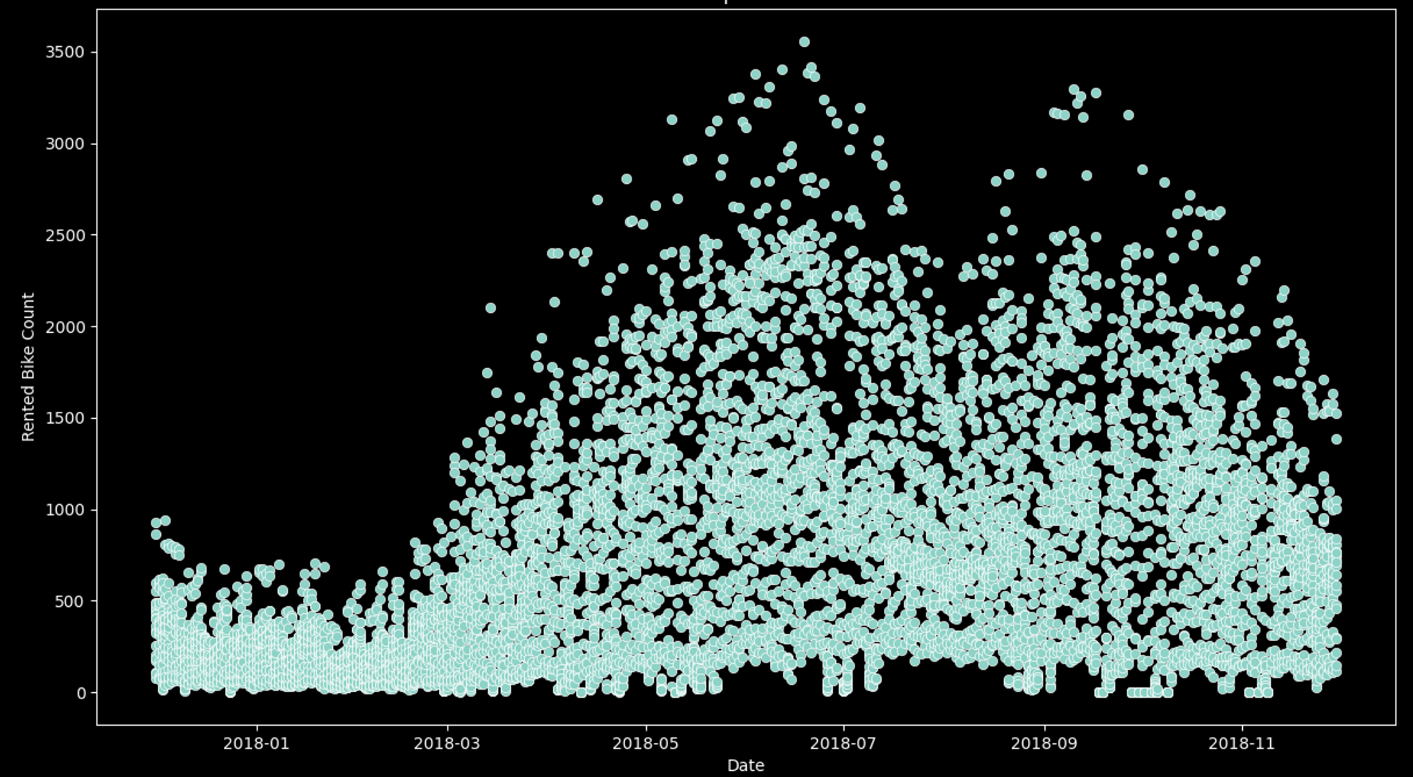
The boxplot gave a first perspective on the variation within the days. To better understand the variation in Figure 5. Scatterplot of the number of bike rentals on each date. a scatterplot is given the variation of the bike rentals per day. A conclusion can be made that the variation in the beginning of the year are smaller then in the middle and around September. Also, can be concluded that in these months there are only a few outliers that cause this higher range of variation (red circles). The last conclusion that can be made is that from May until November there are more days with a higher bike rental than zero in the other months (white circles).

Figure 5. Scatterplot of the number of bike rentals on each date.

## 4.3 Seasons

The knowledge and experience of the authors with the seasons in Europe gave the interest to see if in South Korea the seasons the same way is divided. There are four different seasons:

1. Spring, from April to June.
2. Summer, from July to August.
3. Autumn, from September to November.
4. Winter, from December to March (When is the best time to visit South Korea?, n.d.).

In order to be able to make a season pie the seasons are divided like the months are spread in Europe, because the South Korean seasons are almost the same way distributed as the European one. This will help to better understand which season is more popular for bike rentals.

Figure 6. Percentage of rented bikes according to the season. Shows the season pie and what can be concluded from it, is that in the summer the most bikes are rented, the spring and autumn almost the same number of bikes and in the winter the least. This knowledge will be used the next section, where the correlation of the characteristics is being analysed.

Figure 6. Percentage of rented bikes according to the season.

## 4.4 Days and hours

Chart, bar chart

Description automatically generatedBetween the months there is a shown difference, now between the days and hours are relevant to analyse. This will show if some days/hours are more popular for bike rental. In Figure 7. Bar chart of the average bike rentals per day. a bar chart is shown with the mean values of the days over the year., a conclusion can be made that there is almost no difference between the days. Only the Sunday looks slightly lower then the other days.

Figure 7. Bar chart of the average bike rentals per day.

Chart, histogram

Description automatically generatedTo see if there is a difference between the hours and days Figure 8. A plot of the bike rentals per hour per day. is made. As shown, there is a difference between the week-/workdays (Monday, Tuesday, Wednesday, Thursday and Friday) and the weekends (Saturday and Sunday) (Dictionary, n.d.). This gives an interesting perspective, that the bike is more popular on the week-/workdays and on these days, they are used the most on the commuting hours. This will mean that they are mostly used to go and back to work. While the number of bikes during the weekend is higher between 10 AM and 4 PM compared to workdays.

Figure 8. A plot of the bike rentals per hour per day.

## 4.5 Correlation of the parameters

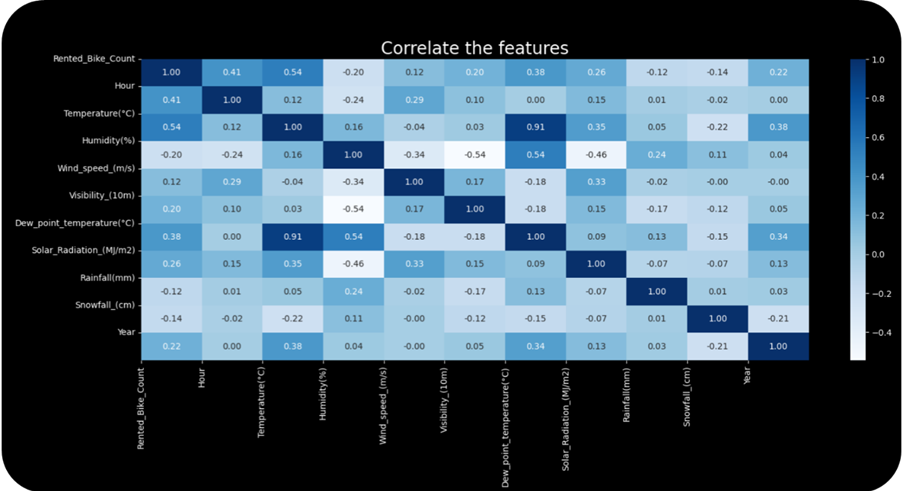
The first part of the paragraph is focused on the first graphical analysis. The next step is to see if the parameters have a correlation with other parameters. This will help showing where the project should focus on. In Figure 9. Correlation map of parameters.. The “heatmap” of the correlation is shown.

Figure 9. Correlation map of parameters.

The outcome of the heatmap is focussed on the parameters that have a correlation with the number of bike rentals. There is a correlation between the bike rentals and the hours, temperature, and the Dew\_point\_temperature. The Dew\_point\_temperature is based upon temperature and dew, so it was a correlation within the attribute. This is important to know when looking at the correlation with the amount of bike rentals and the attributes.

For the next analysis the knowledge of the “heatmap” is used to create a ranked correlation by highest to lowest correlation. The hour is taken out of this analysis, because this is not an attribute that is influenced by the weather. In Figure 10. Correlation of the parameters., the analysis is shown, and it can be concluded that the temperature has the biggest influence on the amount of bike rentals.

Graphical user interface, application

Description automatically generated

Figure 10. Correlation of the parameters.

## 4.6 Temperature

To get a greater understanding of the correlation between the temperature and the bike rentals new graphical analysis are created. The first one is a graph Figure 11. Temperature per month. that shows the amount of bike rentals and the temperature per month. The conclusion is that the temperature is rising in the summer months and is declining in the winter months. Next to this when there is a peak in the temperature, the amount of bike rentals is decreasing.

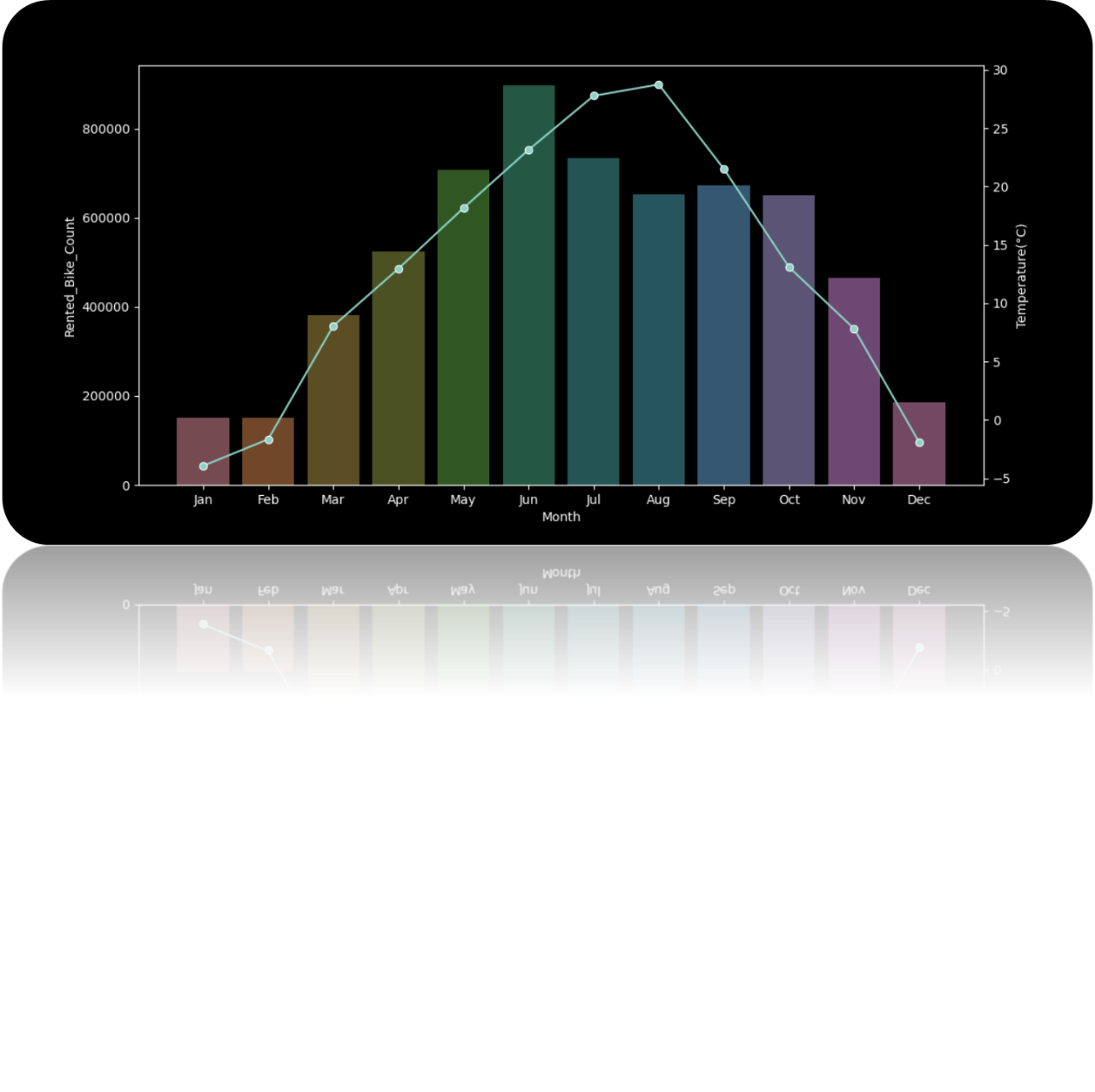


Figure 11. Temperature per month.

To better understand this, Figure 12. bike rental shown over the temperature. shows the amount of bike rentals at each given temperature. What can be concluded is that the correlation is strong when the temperature is lower than thirty-four degrees Celsius.

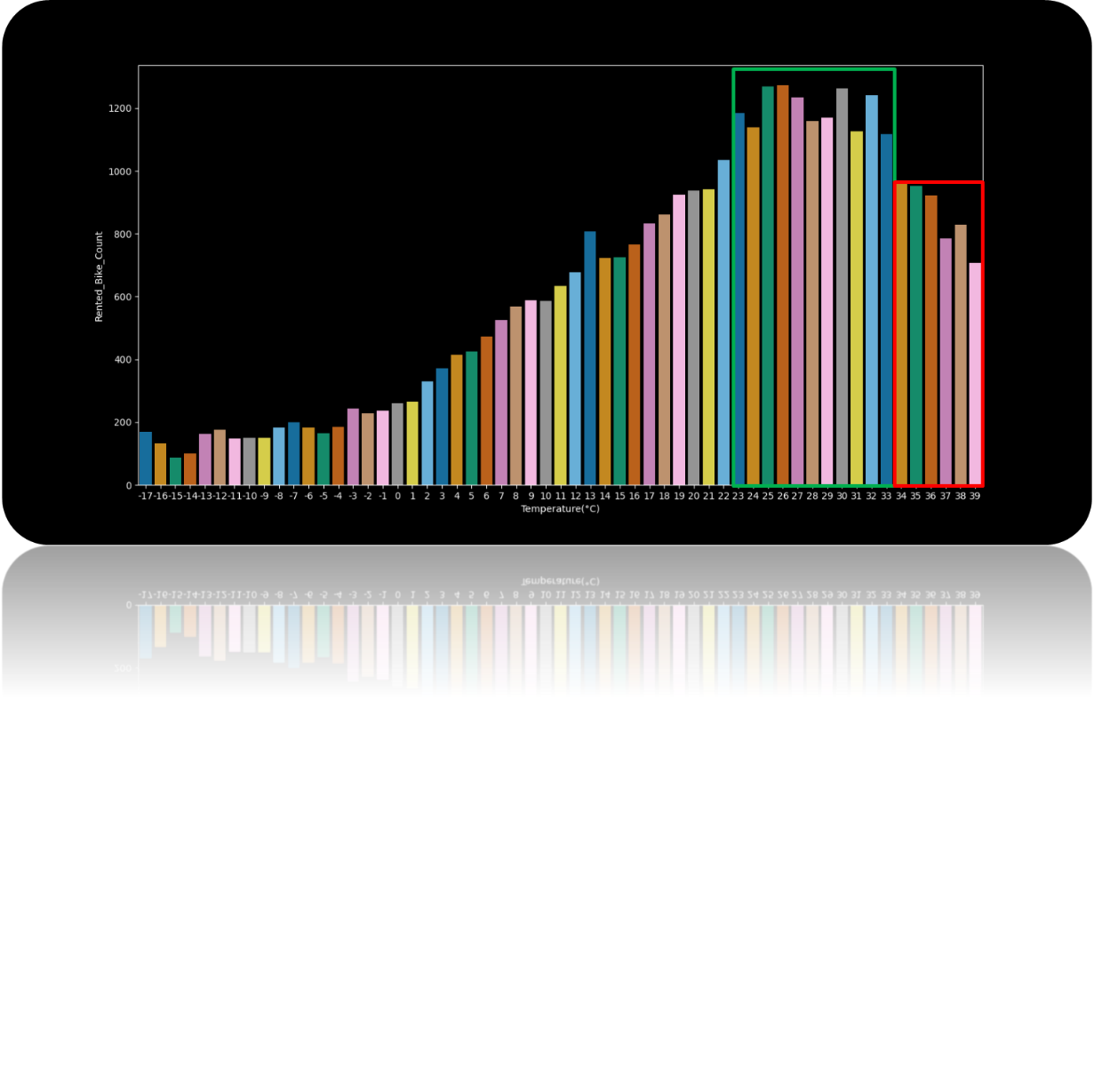


Figure 12. bike rental shown over the temperature.

# 5. Prediction method

During the following section an assessment of different prediction methods will be performed, concluding on the best fitting method. This will be done through a developed framework for the selected model

The selection is based upon testing different prediction methods such as Linear regression, KNN and adaboost. These methods are opted out based on their R2 value. If the R2 value was under 60% and therefor is not qualified to explain enough of the data, see Table 2. Testing different prediction methods opted out based on their R2 value.

|  |  |
| --- | --- |
| **Model Name** | **R2 Value [%]** |
| KNN | 52,26 |
| Linear Regression | 55,04 |
| AdaBoost | 59,87 |
| Random Forest | 76,76 |

Table 2. Testing different prediction methods opted out based on their R2 value.

To assess the linear model, the relationship between the input variables and the output is assessed by using exhaustive search and comparing the R2 adjusted value with the Akaike information criterion (AIC) value, see Figure 2 - Framework for the report (Kannan, 2018).

The exhaustive search method is used based on the number of variables. It is possible to test all combinations of variables because the data set only includes 14 variables, and the time spend on cross check all combinations is insignificant. If the number of variables is higher and the dataset includes more data, the forward or backword selection could be alternative options.



Table 3. An exhaustive search on the 14 variables of the dataset.

The exhaustive search shows that the combination of all variables, only excluding the visibility gives the best combination when comparing the R2 adjusted and AIC value to a linear regression.

## 5.1 Random Forest

Random forest one of the ensembles learning methods that can be used for classification and regression task. It operates by constructing a multitude of decision trees and then aggregate these decision trees together to a single result. For regression task the mean prediction of the individual trees is used for, and for the classification task the prediction is the class selected by most trees. For this prediction model, regression is used.

To begin with, a data split is made to train and test the model build. This is done by creating a data frame with variables that should be included in the model. This variable is extracted from the original data frame. To accommodate the variable the value to predict is chosen as the variable: Y\_Series and are the Rented Bike Count from the original data. Next to this, the variables shown in Figure 14. Grind search criteria. are used to create the training and test data.

Variables\_Frame = df[[

        "Temperature(°C)",

        "Dew\_point\_temperature(°C)",

        "Hour",

        "Humidity(%)",

        "Wind\_speed\_(m/s)",

        "Solar\_Radiation\_(MJ/m2)",

        "Rainfall(mm)",

        "Snowfall\_(cm)",

        "Seasons",

        "Holiday",

        "Functioning\_Day"

        ]]

Y\_Series = df["Rented\_Bike\_Count"]

Figure 13. Created variables to perform the model building.

The Training data is now used to build the Random Forest prediction model.

rfr\_model = RandomForestRegressor()

Figure 14. Grind search criteria.

x\_train, x\_test, y\_train, y\_test = train\_test\_split(variables\_Frame, Y\_Series, test\_size=0.33)

By calling the Random Forest Regressor model it is now possible to use Grid Search. Grid Search is used when values for two or more parameters must be evaluated. The grind search makes an exhaustive search over the specified parameter values included in the grid search (scikit-learn, 2020). The included parameters are the “param\_grid” variable, while the actual grid search has the variable “clf”.

The parameter “max\_depth” limits the number of nodes in the tree. The best value depends on the interaction of the input variables. (Sklearn, 2022)

The parameter “n\_estimators” is the number of trees in the forest. (boosting stages that will be performed.) (Sklearn, 2022)

The parameter “bootstrap” is set to True and False. This means that that the grid search will test if the training of the model is better with or without bootstrap enabled. (Sklearn, 2022)

The variable verbose is set to print warnings if they occur while performing the parameter search. To help reduce the search time, n\_jobs help to activate parallel CPU usage. (Sklearn, 2022)

param\_grid = {

'max\_depth': [2, 4, 6, 8, 10, 12, 14, 16],

'n\_estimators': [50, 100, 150, 200, 250, 300],

      'bootstrap': [True, False]}

clf = GridSearchCV(

rfr\_model, param\_grid, verbose=1,

n\_jobs = multiprocessing.cpu\_count() // 2

)

Figure 15. Included parameters within the grid search.

The grid search is now performed by using the training data with the command shown in Figure 15. Included parameters within the grid search.

Showing in Figure 16. The best parameters for building the model. that the best parameters for building the model, is to train the model with bootstrap enabled and a max forest depth on 10 with 150 estimators. The best explained variance score (similar to the R^2 score) is 86,2% on training data.

Figure 16. The best parameters for building the model.

clf.best\_params\_ = {'bootstrap': True, 'max\_depth': 10, 'n\_estimators': 150}

Theis hyperparameters are used to guide the parameter calibration in the right direction. Therefore, the best parameters from the grid search are used to see if there is an increase or reduction around the parameter area.

Figure 17: Trees in the forest shows what happens if we increase the number of trees in the forest and the depth of those trees, and thereby the complexity of the random forest model. The green line indicates that the model starts to overfit a little after 300 estimators and the performance is stalling after the tree depth hits 16.

* N\_values = Mean Absolute Error
* Depth\_values = Mean Absolute Error

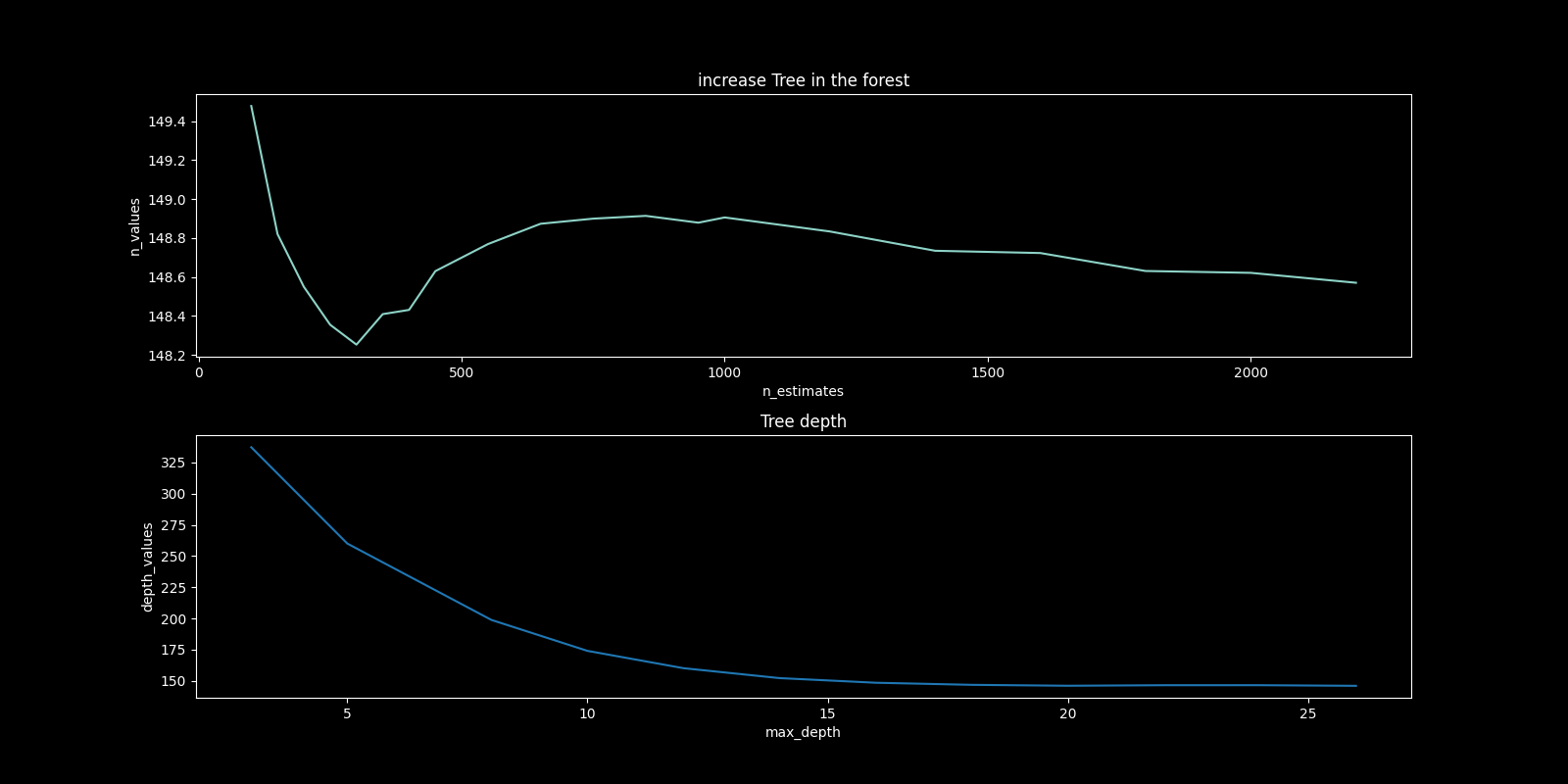


Figure 17: Trees in the forest

The grid search is then adjusted to include the newfound parameters

The best parameters are a max depth at 16 with 200 estimators and bootstrap set to False.

Theis hyperparameters are then used to cross validate the Random Forest parameters by using K-Fold, as shown in Figure 18. K-Fold.

Figure 18. K-Fold.

kfold = KFold(n\_splits=5, shuffle=True)

kf\_cv\_scores = cross\_val\_score(

clf,

x\_train,

y\_train,

cv = kfold,

scoring="explained\_variance")

K-Fold cross validation takes all datapoint and splits it into K folds (5 in this analysis) so there is an equal amount of datapoints in each fold. four of the five folds are included and used to train the model and 1-fold is excluded from the training and used to test the model. Then the train and test fold are shuffled and tested again. This is done five times and the average of the five folds explained variance score is given. By using kf\_cv\_scores.mean() the cross-validation explained variance score is 87% with the hyperparameters given from the grid search. An important note here is that the cross-validation has only included train data. This means that there is still 33% of the original data the model has not seen.

This data is used for testing the Random Forest model to see how the model performs.

Figure 19. Rain Forrest model.

RF\_Model = RandomForestRegressor(

bootstrap = True,

max\_depth = 16

n\_estimators = 200)

RF\_Model.score(x\_test, y\_test)

The model is now created, and the score is based on the test data. The explained variance score is 86.7%. By comparing the explained variance scores generated we can see if the model is under- or overfitted.

|  |  |
| --- | --- |
| **Method** | **Score** |
| GridSearchCV average score [Train Data] | 86,2 % |
| K-fold CV average score [Train Data] | 87,0 % |
| RandomForestRegressor [Test Data] | 86,7 % |

Table 4. Variance score of Grid search, K-Fold and Random Forest regressor.

The hyperparameter for the random forest model is now tuned and the best explained variance scores is found.

### 5.1.1 Feature importance

To prune the trees in the random forest model, recursive feature elimination is used. Recursive feature elimination trains the random forest with all features and considering smaller and smaller sets of features while an external estimator assigns wights to the individual features. Features that do not contribute to a performance increase are pruned. Table 5 shows the result of recursive feature elimination. All features have a positively performance increase to the random forest model and therefore are all features included into the model

|  |  |  |  |
| --- | --- | --- | --- |
| # | **feature** | **support** | **ranking** |
| **0** | Temperature(°C) | SAND | 1 |
| **1** | Dew\_point\_temperature(°C) | SAND | 1 |
| **2** | Hour | SAND | 1 |
| **3** | Humidity(%) | SAND | 1 |
| **4** | Wind\_speed\_(m/s) | SAND | 1 |
| **5** | Solar\_Radiation\_(MJ/m2) | SAND | 1 |
| **6** | Rainfall(mm) | SAND | 1 |
| **7** | Snowfall\_(cm) | SAND | 1 |
| **8** | Seasons\_Spring | SAND | 1 |
| **9** | Seasons\_Summer | SAND | 1 |
| **10** | Seasons\_Autumn | SAND | 1 |
| **11** | Holiday\_1 | SAND | 1 |
| **12** | Functioning\_Day\_1 | SAND | 1 |

Table 5: recursive feature elimination

The included features in the model have different importance. As mentioned in section 4, the Exploratory Data Analysis, the temperature has a significant importance on the count of bike rentals. This is further illustrated in Figure 20, which is a Pareto analysis of how important the features included in the Random Forest model are.

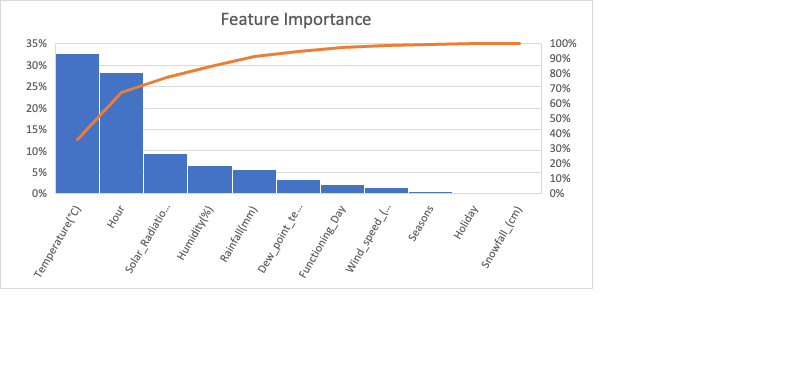


Figure 20. Pareto analysis of the important features in the Random Forest model.

Another importance feature included in the model is the time of day, included as Hour. The pareto diagram illustrate its almost equal important to include in the Random Forest model. The Solar Radiation is showed as almost 10%, but is secretly correlated with the temperature, because the more the sun is shining the hotter the temperature and the more radiation occurs.

### 5.1.2 Prediction with Random Forest

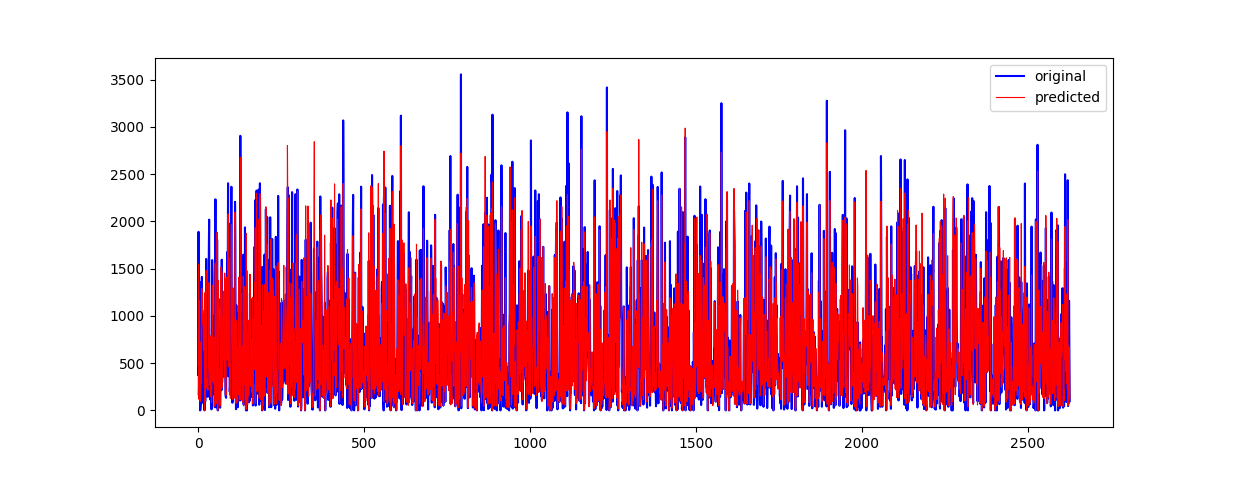
By using the predicting method on the RF\_Model variable, the model can make a prediction, see Figure 21. The figure shows that the model can explain most of the sporadically bike demand but have some trouble explaining the high values occurring.

Figure 21. Prediction model Random Forrest from hour 0.

A closer look at the model is shown in Figure 22. The model follows the overall rented bike count pattern, some of the higher count could be explained by variables not included in the dataset, and therefore is not predicted.

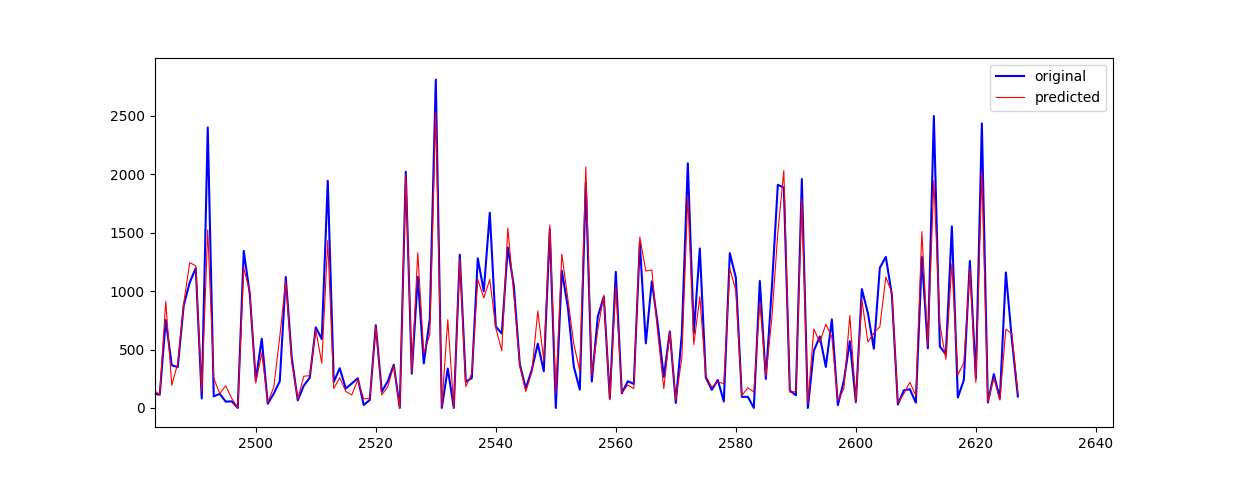
The models overall Rooted Mean Squared Error (RMSE) is 229 and the mean absolute Error is 142, which indicated that the model on average is 142 bikes from the true value.

Figure 22. Prediction model Random Forrest from hour 2500 until 2620.

# 5. Discussion

During the following section, a discussion on the study will be concluded. Throughout this discussion, the results and findings of the assignment will be reviewed, furthermore, a set of delimitations will be presented, and the future of the analysis will be argued.

After the analysis of different models to predict the correct bike count, Random Forest was selected as the best fitting one. The initial selection was made based on the R2 value, but further analysis was completed. Further criteria for the selection of the model were critical for the final selection of the model, as some other methods as SKboost predicted negative values at some points and had a negative impact when miss predicting a values, therefore, reducing the accuracy of the other models. After the selection of the method, the final code was written and verified that work using the training data created from the original data sheet. Where the model was optimized, nonetheless, in this model there has not been a direct prone applied on the data from the data sheet, as all features had a positive impact in the accuracy of the model. In this case the prone of the data was controlled through the depth of the model trees, thus limiting the complexity of the model. Later on, the model was then validated using the rest of the data that was not used for training. Nonetheless, the prediction of this model does not reach the minimum real data, therefore overestimating the rentals in some way.

Moreover, other factors not existent in the data sheet used for the prediction of the data could have an impact on the predictions. These factors could be the population trend in the city, as the larger the population, the higher the number of rentals should be, or the number of cars in the city over time, as the lower the number of cars, again, the higher the number of rentals. Furthermore, other less tangible features could be described, in the event these should one day be part of the model somehow. These features could be political and legal factors, as new laws, sanitary emergencies, government grants and so on.

In the case of this model, seasonality represents the most important factor for the bike rentals. This importance comes as it has the highest significance when looking at the number of bike rentals. This could be influenced by the higher tourism during the summer months, but also to the direct corelation of other factors to this one, as it is the temperature and weather, which could be also linked to the hour.

All things considered, the model present a solution to the aimed study, predicting the number of bikes rented during a given period of time. Furthermore, the model can explain most of the demand when using the actual data saved for the validation of the model. Consequently, it could be assumed the model could be trusted and therefore used to solve some of the hypothetical research questions presented during the report. Thus, by implementing a model like this one, the planning maintenance, the operational purchase planning and other tasks could be optimized, improving the overall output.

Some limitations to the project were the no exploration of further data, as presented just before when considering the population of Seoul, a relevant factor for the prediction of future demand. This further exploration of data could be incorporated in future research, which could use this one as the base for the upcoming one.

# 7. Conclusion

The following research questions were formulated and will be answered in this section: *When should the maintenance take place on the bikes? What conditions lead to an enhanced use of bikes in Seoul?* and *Could the prediction be accurate enough to use in a real case scenario?*

It can be concluded that there is a Seasonality with a peak in June and a high bike rental from May to October. The winter only counts for 7% of the yearly rental, therefor that’s why yearly maintenance should be done int the winter period. The everyday maintenance should be done between 4-5 because the bike rental count is at the lowest at this point.

Monday to Friday has a higher rental count than Saturday and Sunday. This is because the general population is using the bikes to and from work in the community hours. This is due to the heavy use in the community hours, which is backed by the correlation between bike rental count and hours. The correlation between the bike rental count and temperature indicates that the higher the temperature is on the current day, the more bikes will be rented. If the temperature raises over 33 degrees, there is a decline in the rental count which indicates that the temperature is too high.

The best model selection was done by measuring different models explained variance value. The Random Forest had a beginning explained variance value of 76,76. By fine tuning the hyperparameters an increase to 87,0 in the cross validation was made. The hyperparameter was tuned by measuring the Mean Squared Value to avoid over- or underfitting. To finalize the model a test was done on data that was not included in the training of the model and the explained variance was stable at 86,7. The hyperparameter with the highest explained variance is used to prune the features included in the model. The recursive feature elimination is used and categorized all included features as rank 1, which means that all features is included in the final model build.

The random forest has 200 trees included with a max depth of 16 and does not use bootstrap. The general performance of the model is around 230 Root Mean Squared Error (RMSE) and has a Mean Absolute Error (MAE) of around 144 bikes.

On this performance basis, the conclusion is that the random forest model can be used to predict the rental bike count and can thereby help the Seoul bike rental to have bikes ready for use at the right time.

# Bibliography

Cho, S. V. (2020). A rule-based model for Seoul Bike sharing demandprediction using weather data. *European Journal of Remote Sensing*, 166-183.

Dictionary. (n.d.). *Weekday*. Retrieved from Dictionary: https://www.dictionary.com/browse/weekday#:~:text=Since%20the%20weekend%20is%20considered,of%20weekdays%20are%20called%20weeknights.

Education, I. C. (2022, August 2). *Data Science*. Retrieved from IBM Cloud Education: https://www.ibm.com/cloud/learn/data-science-introduction

Ghosh, A. (2018). A comprehensive review of tools for exploratory analysis of tabular industrial datasets. *Elsevier*, 19.

Institute, S. (1999). SAS/STAT user's guide Version 8. In S. Institute, *SAS/STAT user's guide Version 8.* (p. 43). Cary, N.C. : SAS Institute.

Kannan, D. (2018). Role of multiple stakeholders and the critical success factor theory for the sustainable supplier selection process. *International Journal of Production Economics*, 391-418.

Matplotlib. (2022). *Matplotlib: Visualization with Python*. Retrieved from Matplotlib: https://matplotlib.org/

Munnangi, J. (2022). *Data Science project life cycle*. Retrieved from Medium: https://medium.com/co-learning-lounge/complete-data-science-project-life-cycle-9eae6e4ed4c9

Numpy. (2022). *Numpy*. Retrieved from Numpy: https://numpy.org/

Pandas. (2022). *Pandas*. Retrieved from Pandas: https://pandas.pydata.org/

Repository, U. M. (2020). *Seoul Bike Sharing Demand Data Set*. Retrieved from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand

Review, W. P. (2022). *Seoul Population 2022*. Retrieved from World Population Review: https://worldpopulationreview.com/world-cities/seoul-population

Sathishkumar V E, J. P. (2020). Using data mining techniques for bike sharing demand prediction inmetropolitan city. *Computer Communications*, 353-366.

scikit-learn. (2020, 11 16). *scikit-learn*. Retrieved from scikit-learn: https://scikit-learn.org/stable/modules/generated/sklearn.model\_selection.GridSearchCV.html

*Sklearn*. (2022, 11). Retrieved from Sklearn GridSearchCV: https://scikit-learn.org/stable/modules/generated/sklearn.model\_selection.GridSearchCV.html

*When is the best time to visit South Korea?* (n.d.). Retrieved from Selective asia: https://www.selectiveasia.com/south-korea-holidays/weather#:~:text=There%20are%20only%20minimal%20regional,winter%20(December%20to%20March).

# Appendix 1. Data cleaning steps

To accomplish all these steps basic python functions were used:

**Step 1. Correctly importing the data:**

The original file with The Bicycle Rental Data is in a .csv format ("SeoulBikeData.csv") therefore, it is important to check what type of delimiter is used to correctly make the import into the python environment correct. The first three lines of the file contain the following text:

1) Date,Rented Bike Count,Hour,Temperature(°C),Humidity(%),Wind speed (m/s),Visibility (10m),Dew point temperature(∞C),Solar Radiation (MJ/m2),Rainfall(mm),Snowfall (cm),Seasons,Holiday,Functioning Day

2) 01/12/2017,254,0,-5.2,37,2.2,2000,-17.6,0,0,0,Winter,No Holiday,Yes

3) 01/12/2017,204,1,-5.5,38,0.8,2000,-17.6,0,0,0,Winter,No Holiday,Yes

Notice how the first line of the data file defines the names of the variables, while from the second onward the actual data of the data frame begins (for a total of 8761 lines including the index).

bikeData = pd.read\_csv(filePath, delimiter=”,”)

by using the function .shape it is possible to verify that the matrix have been correctly created:

bikeData.shape

-> (8760, 14)

**#bikeData.info()**

**Step 2. Checking for missing values**

It is highly suggested to drop the values that are missing with the .dropna function

bikeData.dropna

-> no missing values have been found

**Step 3. Creating dummy variables**

Dummy variables need to be created, that which means converting non-numerical categories (usually strings) to numerical binary values, for example transforming a binary category from values Yes/No to 1/0, or converting a multiple category variable to its dummy version: A,B,C / 100,010,001.

How to plot the scatter plot of the cleaned data frame:

date\_series = pd.to\_datetime(df["Date"])

y\_series = df["Rented\_Bike\_Count"]

plt.style.use('dark\_background')

fix, ax1 = plt.subplots(figsize=(16,8), )

sns.scatterplot(x=date\_series, y=y\_series, ax = ax1)

ax1.set\_title("Time Series - Scatterplot of Rented Bike Count")

ax1.set\_xlabel("Date")

ax1.set\_ylabel("Rented Bike Count")

plt.show()

# Appendix 2. Data Science lifecycle

Diagram

Description automatically generated

Figure 23. Data Science Lifecycle (Munnangi, 2022).