San Jose State University

CMPE 274 – Business Intelligence Technologies

Bike Demand Prediction System

Supervisor

Prof. Chandrashekar Vuppalapati

Dhruva Gajera (016040934)
Sriram Chinta (016002506)
Subhash Reddy (016003403)
Teja Sree Goli (016040986)

Abstract

Bike rental systems are a new means of transportation gaining popularity worldwide. It has already been implemented in several affluent countries and is gaining popularity in developing nations. The world's various entrepreneurs have attempted to establish a bike rental system but failed due to a lack of good data analytics. In the bike rental industry, forecasting demand is crucial. Past usage trends and weather data could be combined efficiently to estimate bike rental demand. Those who do not own a car can take advantage of bike rentals. It can be presented so that the number of bikes that can be hired can be predicted. It is possible to solve the challenge of forecasting the number of cycles rented at any given hour in a city. The problem will be referred to as " Rental Bike count" from now on. The forecasts are generated for each hour of the day to address one of the biggest causes of customer loss: a lack of bikes.

Index

Sr No	Table Of Content	Page No
1	**Problem formulation**	4
2	**Introduction**	4
3	**Literature Review**	5
4	**Architecture Design and Approach**	5
5	**Data Overview**	6
	5.1 Data Processing	7
	5.2. Exploratory Data Analysis	8
	5.2.1 Different Season Plots	10
	5.2.2 Different Month Plots	11
	5.2.3 Distribution of Windspeed	12
	5.2.4 Distribution of Temperature Values	12
	5.2.5 Average Demand for Bikes for different days of the week	13
	5.2.6. Demand for Bikes during Different Hours:	14
6	**Machine Learning Models Comparison/Prediction**	15
	6.1 Deep Neural Network	15
	6.2 K-Nearest Neighbors	16
	6.3 PLS Regression	19
	6.4 Decision Tree Regressor	20
	6.5 Gradient Boosting Regressor	21
	6.6 Logistic Regression	23
7	**Outcomes**	24
8	**Conclusion**	24
9	**Future Scope**	24
10	**Application GUI**	25
	REFERENCES	26

1. Problem formulation

The issues these businesses confront is maximizing profit and providing rides for customers. People may miss out on these bikes because they are unavailable. There are times when there is less need for these motorcycles, which need to be utilized. Dealing with these situations and comprehending the demand for bikes on different days and conditions becomes crucial.

Due to the present market crisis, US bike sharing has recently seen significant drops in its revenue. As a result, many bicycle rental businesses need help remaining viable in the current market. As a result, it has made the conscious decision to create a business plan that would help it increase income as soon as the economy and market conditions are stable again. Additionally, they intend to grow their company outside of the US. As a result, they need its potential partners to predict how many bikes customers will rent under certain circumstances.

2. Introduction

This machine-learning project will use various approaches to forecast bike demand. There are several bike rental firms, such as Zagster and City Bikes. It would be very beneficial for businesses to predict the bike market at specific times of day and for other factors like temperature and wind speed. Therefore, it is crucial to comprehend the demand for bikes at various points in time so that businesses can use this knowledge to increase profits and deliver bikes to multiple people as and when necessary. With the help of machine learning and data science, numerous bike loan companies would see a good increase in revenue.

As part of the research, we will analyze the Kaggle bike demand dataset to understand some of its properties better. Later, we would add a few more attributes to get a reasonable estimation of the parameters that must be considered when applying various machine learning and deep learning algorithms to visualization. To determine how well our model did overall, not just on the training data but also on the test data, we would then plot the graphs of the predictions and the actual values.

This project aims to use independent factors to model the demand for shared bikes. The management and potential partners will use it to comprehend how different features affect needs differently. They can adjust the business plan accordingly to fulfil the customer's expectations and demand levels. Additionally, the model will help management comprehend demand patterns in a new market.

Rental Bike prediction is based on various features like

- Weather (cloudy, windy, rainy)
- Season (Summer, Winter)
- Temperature

- Humidity
- Wind Speed,
- Holiday (Yes/No)

3. Literature Review

Various prediction models were utilized in empirical research to forecast demand for bike sharing. The models were often given historical data and different external inputs, such as weather, temporal, and spatial data [1]. The data collection has twelve different types of variables, making analysis difficult. Although each element may play a role in the outcome, the amount of data points for wind direction that are Nil is improbable. It is difficult to tell whether the data set is empty or contains missing data. Therefore, disregard the impact of wind speed. All categories—aside from wind speed are considered in our analysis. Sort the data into two categories: quantitative (temperature, temp, humidity, casual, registered, and count) and qualitative (season and weather), which are actual measurements expressed through descriptions in plain language instead of numbers. (Quantitative data needs to be measured and expressed in numbers rather than a natural language description. [6]]).

4. Architecture Design and Approach:

We used the following approach to train the model and build the application.

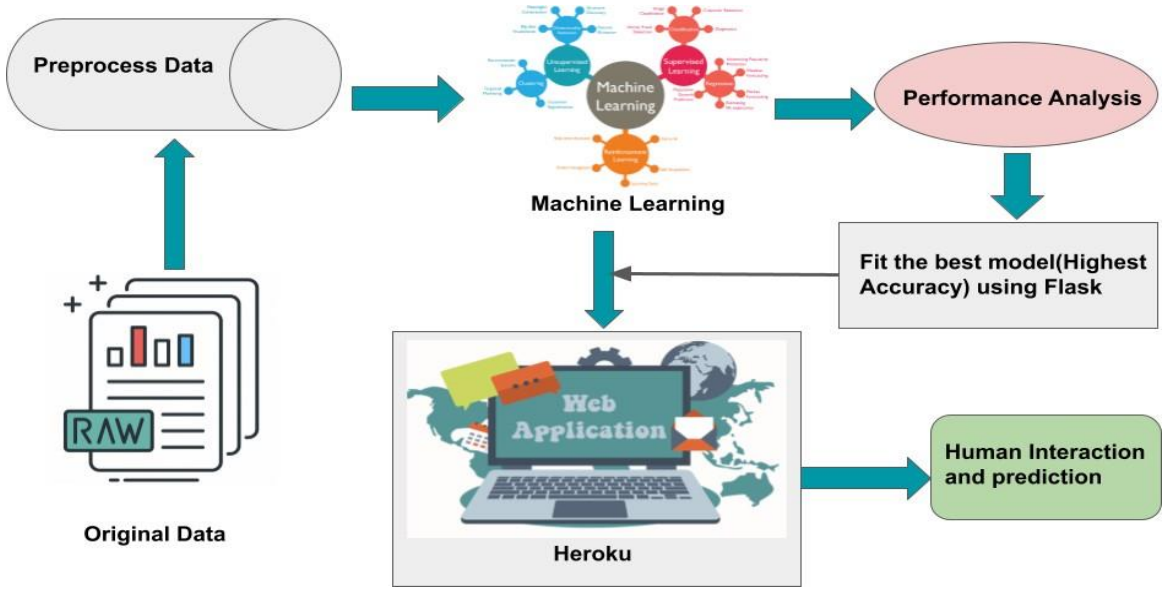


Figure 1: Bike Rental Architectural

Approach

- Understanding Business problems and getting and Visualizing data.
- Feature Engineering
- Training a machine learning model using scikit-learn.
- Create CSS and HTML layout
- Building and host a Flask web app on AWS/Heroku.
- A user has to put details like Humidity, Month, Temperature, and it is a holiday. Wind Speed, etc.
- Once it gets all the field information, the prediction is displayed as the rental information.

5. Data Overview

A new generation of traditional bike rentals, bike sharing systems automates the entire process from registration to rental to return. These systems make it simple for users to hire a bike from one location and return it to another. There are already around 500 bike-sharing schemes and over 500 thousand bicycles available worldwide [3]. These systems are of tremendous interest nowadays because of their crucial role in transportation, environmental, and health issues. Bike-sharing designs are appealing for research due to their data generation qualities and intriguing real-world applications.
In contrast to other modes of transportation like the bus or the subway, these systems openly record the distance traveled and the location of departure and arrival. This feature transforms the bike-sharing system into a fictitious sensor network that can track urban motion [1]. Thus, monitoring this data will allow the detection of the most significant occurrences in the city. This dataset includes the hourly and daily counts of rental bikes from the Capital Bikeshare program in Washington, DC, between 2011 and 2022, along with accompanying meteorological and seasonal data [6].

Data set obtained from Kaggle.

Provided a future set:

- Weather (cloudy, windy, rainy)
- Season (Summer, Winter)
- Temperature
- Humidity
- Wind Speed

- Holiday (Yes/No)

	instant	dteday	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
0	1	2011-01-01	1	0	1	0	0	6	0	1	0.24	0.2879	0.81	0.0	3	13	16
1	2	2011-01-01	1	0	1	1	0	6	0	1	0.22	0.2727	0.80	0.0	8	32	40
2	3	2011-01-01	1	0	1	2	0	6	0	1	0.22	0.2727	0.80	0.0	5	27	32
3	4	2011-01-01	1	0	1	3	0	6	0	1	0.24	0.2879	0.75	0.0	3	10	13
4	5	2011-01-01	1	0	1	4	0	6	0	1	0.24	0.2879	0.75	0.0	0	1	1

Table 1: Bike sharing Database

5.1 Data Processing:

Metrics

We analyze metrics that account for continuous output variables because this is a regression problem, and they base their estimates on the difference between the actual and expected output. The metrics utilized for this prediction are listed below.

- Mean Squared Error
- Mean Absolute Error

There are a large number of machine learning models used in the prediction of the demand for Bikes. Below are the models that were used for prediction.

- Deep Neural Networks
- K Nearest Neighbors
- Partial Least Squares (PLS) Regression
- Decision Tree Regressor
- Gradient Boosting Regressor
- Logistic Regression
- Long Short Term Memory (LSTM)

Machine Learning Predictions and Analysis

- To perform the machine learning analysis, it is crucial to be aware of some of the features in the data.
- To understand some of the underlying features, we will use a variety of data visualizations. Once we have a firm grasp of these features, we will use a variety of machine-learning models to estimate the demand for bikes based on these features.
- After receiving the machine learning predictions, we will employ several techniques that may help us produce the models that businesses can apply in various ways.
- Therefore, where the demand for bikes is predicted in advance using machine learning and deep learning, this would save bike rental companies a lot of time and money.

5.2. Exploratory Data Analysis

With the aid of "visual analytics" and "machine learning," bicycle rental companies will be able to understand the total number of bicycles that must be present at various periods in time and, as a result, be able to predict the need for bikes in the future.

Consequently, the companies would save vast amounts of money while giving various needy people the needed help.

5.2.1 Different Season Plots:

We'll illustrate the demand for bikes in this section for each season. We will additionally consider and count the total data points for the various seasons to gain a more comprehensive perspective. Understanding how much demand there is for motorcycles during different seasons might be crucial because of how closely this characteristic relates to the desire for specific bikes. This can also be seen in the actual world, where we witness more people riding bikes in one or more seasons than in others. As a result, taking this attribute into account would help us better comprehend its significance and the degree to which it would affect our machine-learning predictions.

Count plot of Different Seasons

We see below the count of different seasons and can understand that there are slightly more fall season values than the other seasons. That is because of the variation in the number of days present every month. Therefore, based on the seasonal data, we can get a good idea about bike demand.

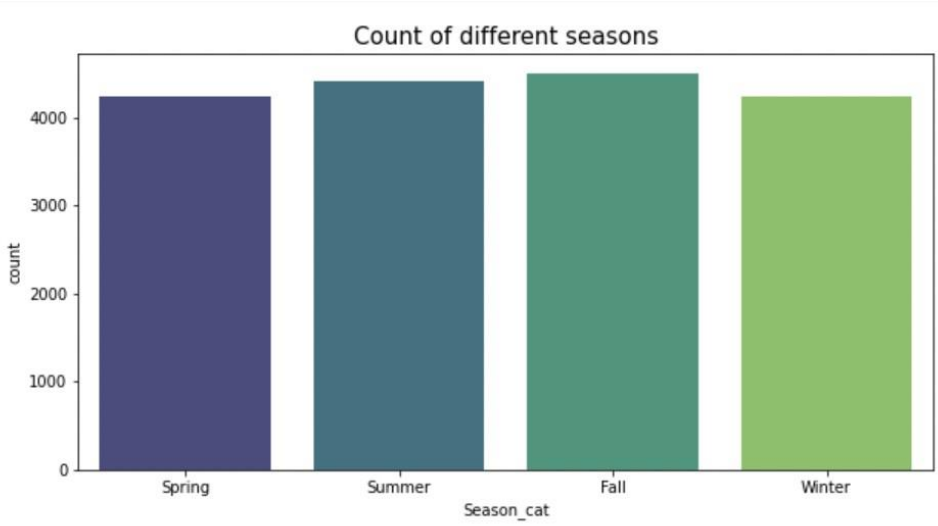


Figure 2. Different Seasons

Plot for Average Demand for Bikes during Different Seasons

According to the information below, the fall season has a higher demand for bikes than the other seasons. Additionally, it has been found that demand for motorcycles is the lowest.

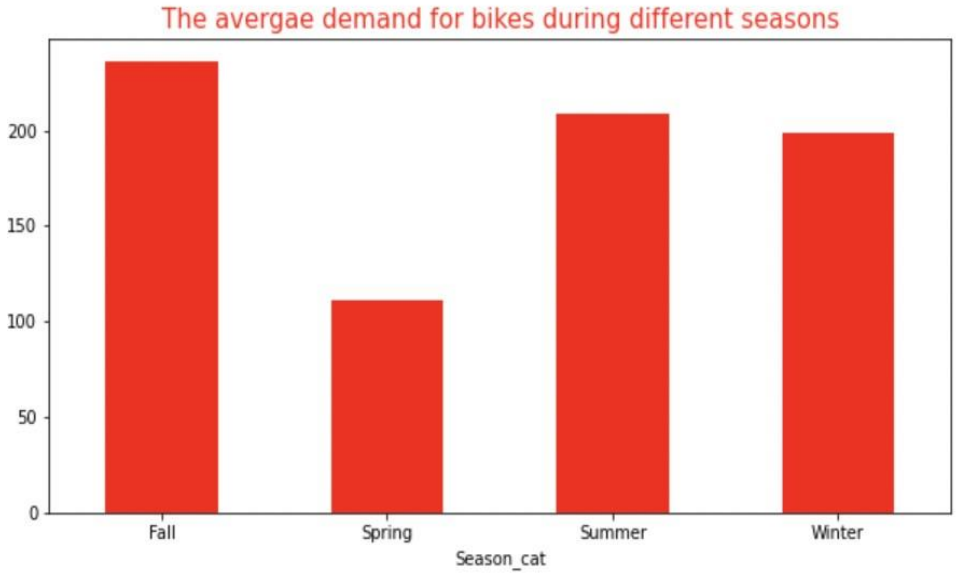


Figure 3. Average Demand for Bikes during different seasons

Total Demand for Bikes for Different Seasons: The data below shows that demand for motorcycles is most robust in the fall. On the other hand, the need for motorcycles is lowest in

spring. As a result, there will likely be a high demand for bikes in the fall and a relatively low demand for bikes in the spring.

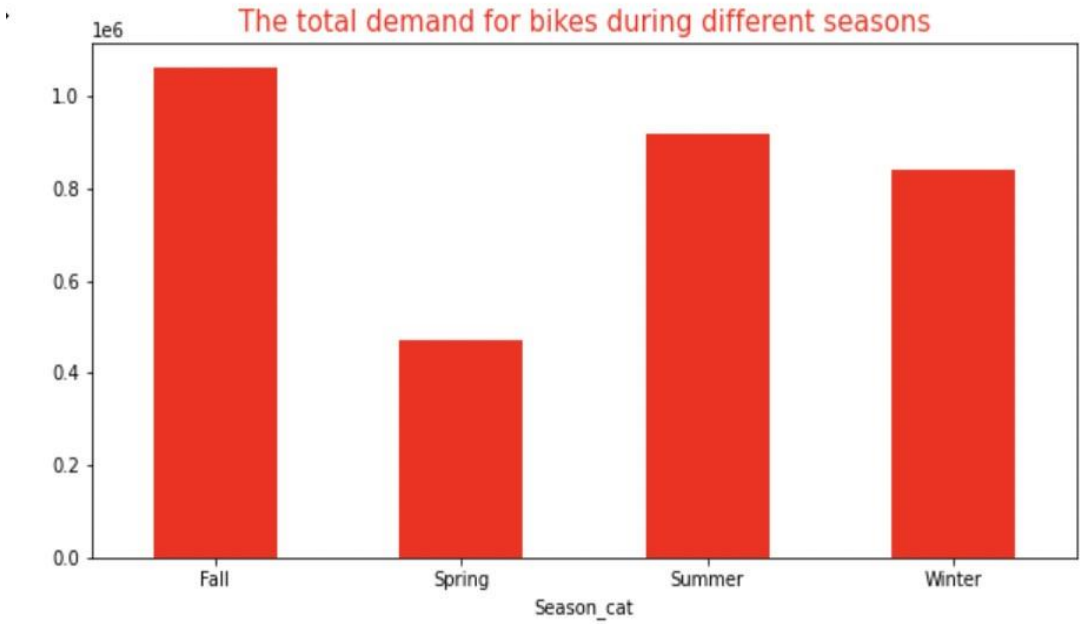


Figure 4: Total Demand for Bikes for Different Seasons

5.2.2 Different Month Plots

It's time to get more specific now and learn how the demand for motorcycles varies throughout the year. In general, we typically only anticipate a few bikes, much alone bikes from loan firms, when there is snow or rain in cities. Therefore, taking into account the various months during which there is a demand for bikes helps us comprehend the significance of this feature in our machine-learning analysis.

Countplot of Different Months

The graph below displays the overall values for several months. As a result, some months have fewer days than others, and as a result, their values are somewhat lower than those of the latter.

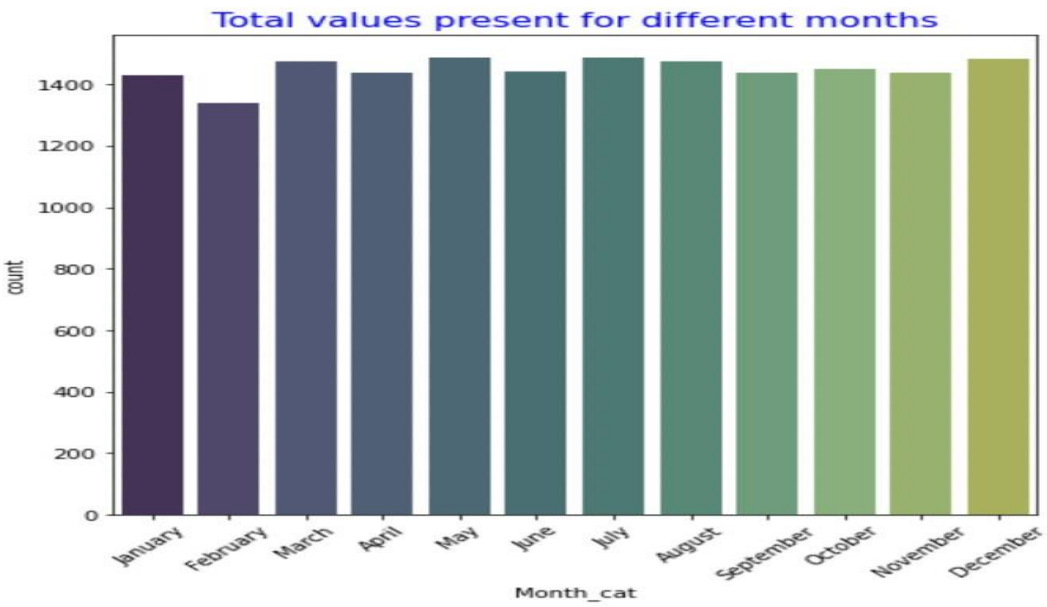


Figure 5: Total value present for other months

Average Demand for Bikes for Different Months

Bikes are most in demand in September, June, and August, respectively. Some have little need compared to the other months. We can also notice that demand for motorcycles is at its lowest on average in January. Therefore, since there will be quite a significant demand for bikes in September, action must be taken. However, we observe that there needs to be more demand for bicycles in January. Therefore, such cycles can be moved to a different area to guarantee that individuals can access them as needed.

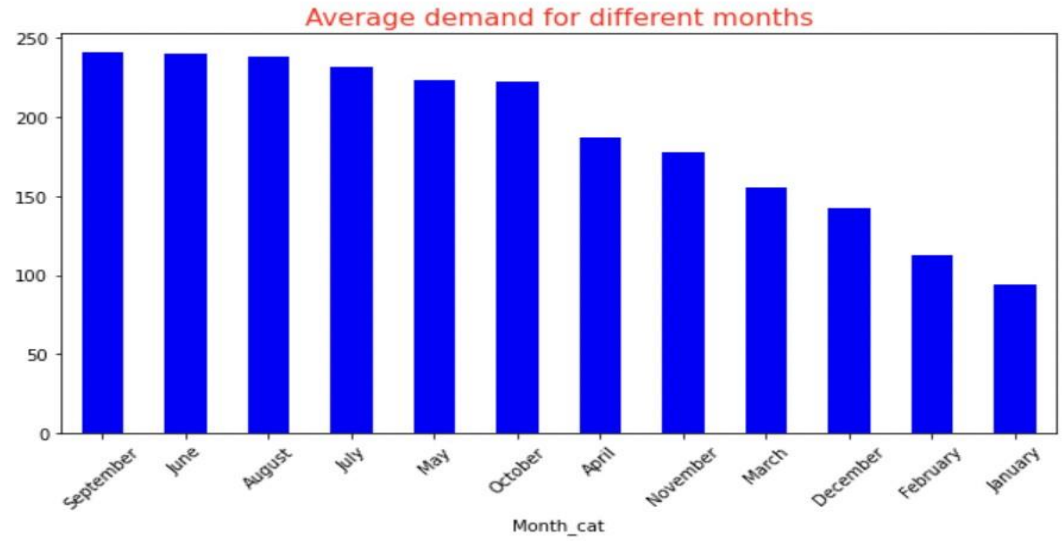


Figure 6: Average Demand for Bikes for Different Months

5.2.3 Distribution of Windspeed

When we examine the data distribution, we see that most of the windspeed readings we deal with are for lower speeds. There are several extreme values present. As a result, the majority of the entries in the dataset have windspeed values that could be higher.

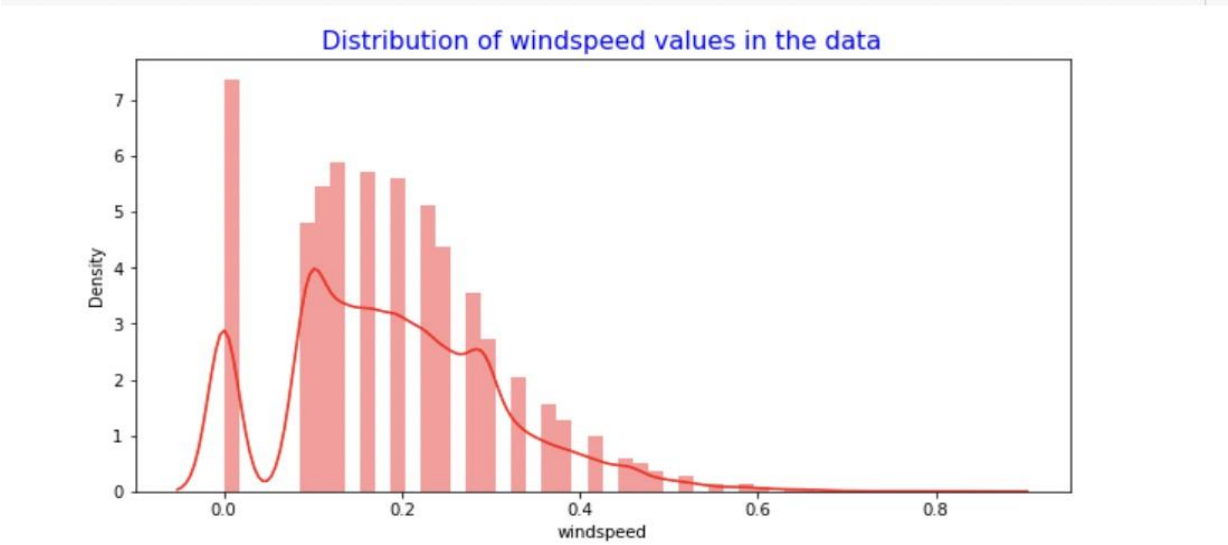


Figure 7: Distribution of Windspeed

5.2.4 Distribution of Temperature Values

The distribution of temperature values is pretty even, as can be seen. Because of this, we are using temperature numbers that are relatively equally spread. This would guarantee that we receive the demand for various temperature values.

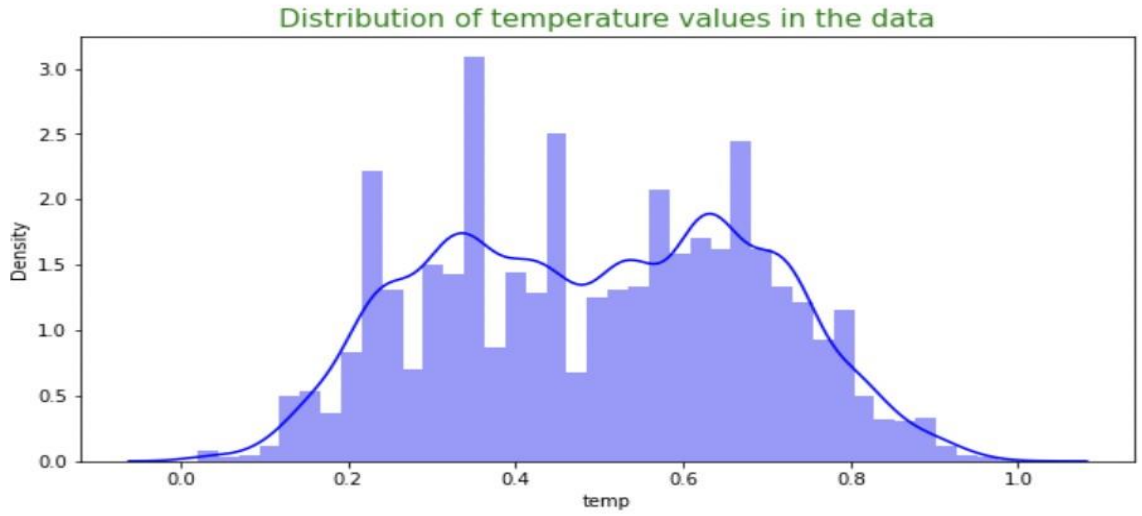


Figure 8: Distribution of Temperature Values

Temperature for Different Months

The side graph shows that the average monthly temperature is more significant in July and relatively low in February because it is springtime throughout those months. As a consequence, we obtain various monthly average temperatures. Using this knowledge, we will examine how the demand for bicycles might alter depending on the various numbers.

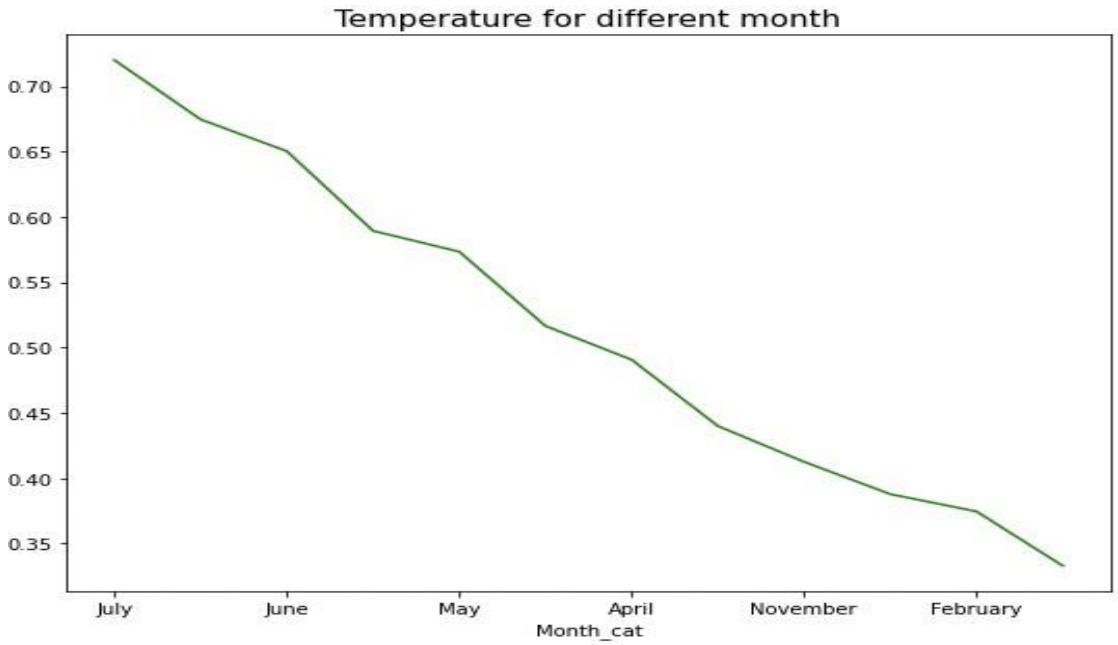


Figure 9: The temperature of different months

5.2.5 Average Demand for Bikes for different days of the week

The values of the average demand for motorcycles are nearly the same with only minor variations, as seen in the chart below, suggesting that there is little difference between different days of the week. Since it is difficult for us to tell the variations apart, this feature may not be very helpful in aiding our ability to forecast the overall demand for motorcycles.

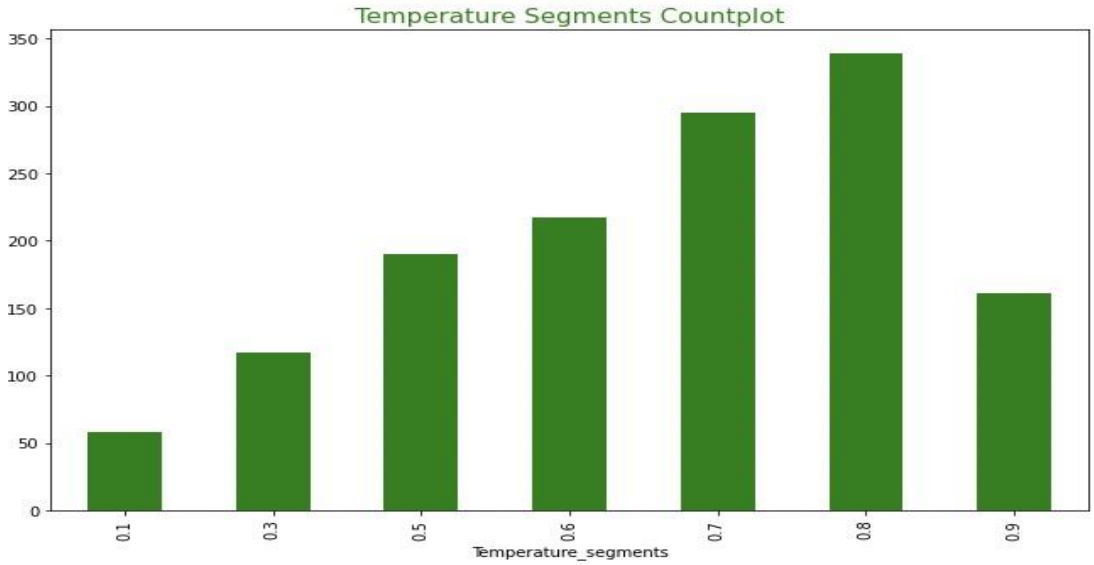


Figure 10: Demand for bikes for different days of the week

5.2.6. Demand for Bikes during Different Hours:

The demand for bicycles peaks at about 5 p.m. Washington time, as seen. As observed, the need for bikes in the early morning, from 1 am to 7 am, is relatively low. In addition, we can keep that there is a significant demand for bikes beginning at 8 am and that the market rises steadily. As shown in following figure 9.

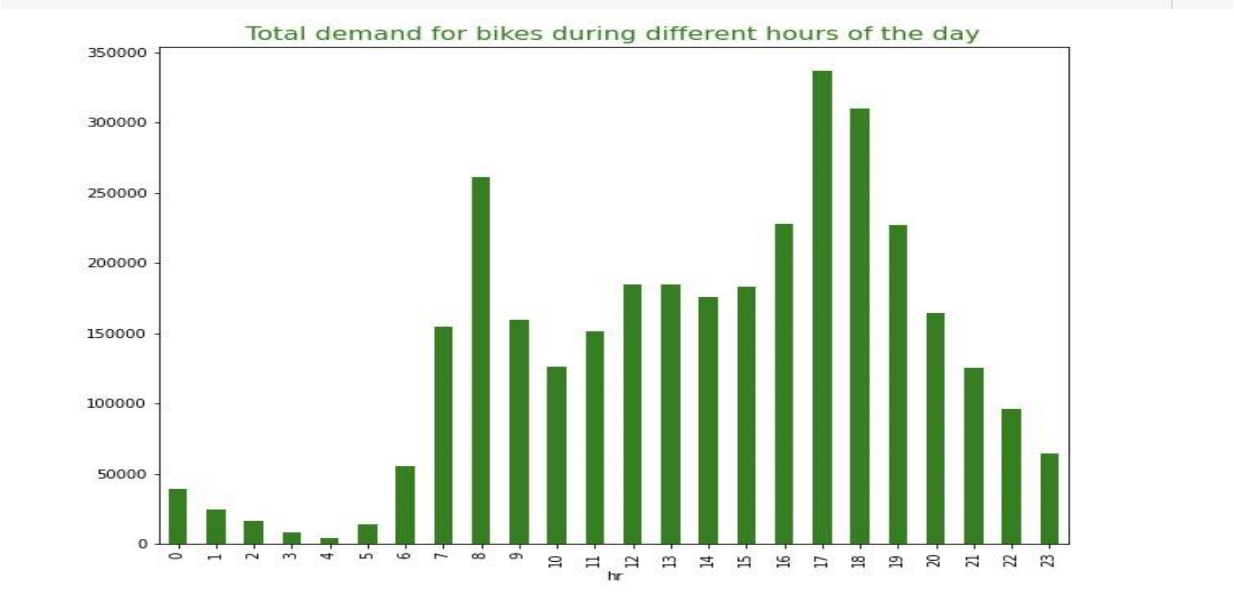


Figure 9. Bike demand at different hours of day

6. Machine Learning Model Comparison/Prediction

We used various machine learning analyses to ensure we received the most accurate forecasts for bike demand. The initial step would be to distinguish between the target variable and the input features by separating the X and y values. To generate the best predictions from the machine learning models, we, therefore, employ X and y as these variables.

To be able to adjust the hyperparameters for various machine learning models and guarantee that we obtain the best predictions for those models, it is crucial to divide the data into training and cross-validation groups. It's critical to get good hyperparameter settings that guarantee that various machine learning models perform favorably on the test set.

Also, to optimize the weights as limited as possible rather than having a varied weight range for distinct features, it is crucial to feature-trait the columns and ensure that the values of the input features fall between 0 and 1. We employ a " MinMaxScaler " scaler to take care of these responsibilities.

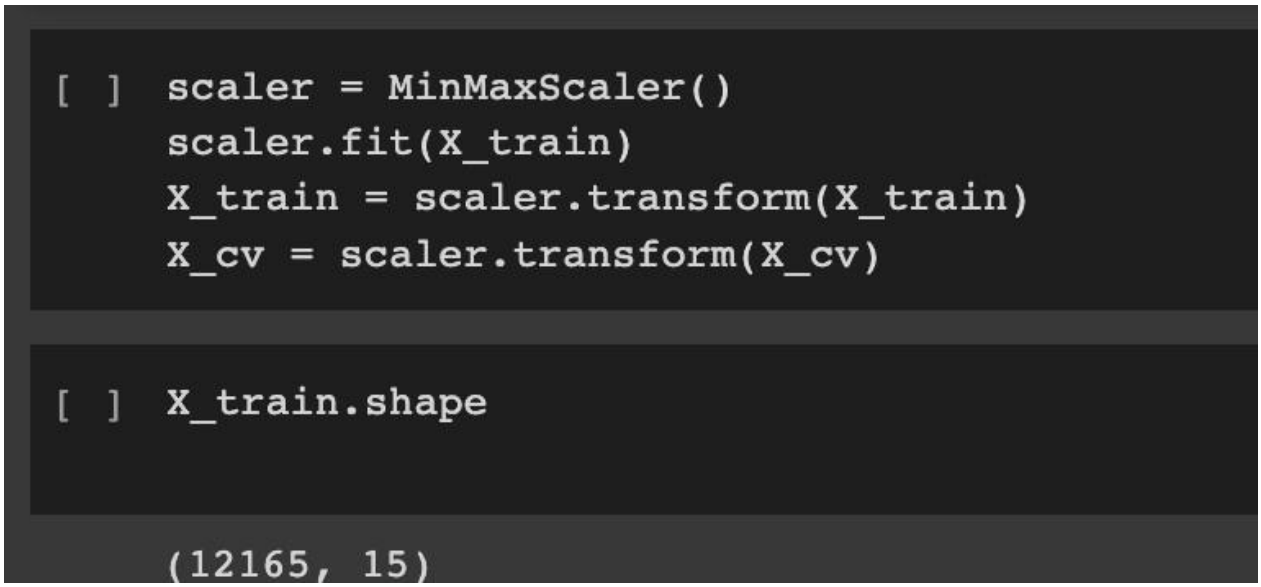


Figure 10: MinMaxScaler

6.1 Deep Neural Network

We have tried various models for different parts of our application. First, we will talk about Deep neural networks. The deep neural network is the machine learning model that we would employ initially. There are several hidden units, and we have decided to make the activation value "relu" in each

case. Before fitting the machine learning model for predictions in the cross-validation data, we chose a variety of optimizers and metrics.

To Perform this model, We trained the machine learning model for 200 iterations to optimize the weights and ensure that it performs well on the cross-validation data.

We train the machine learning model for 200 iterations to optimize the weights and ensure that it performs well on the cross-validation data.

Scatterplot between Y_test and Y_predictions:

Looking at the scatterplot, we can see that the predictions and actual bike demand are close. Therefore, we evaluate the machine learning model's performance on the test set using the cross-validation output.

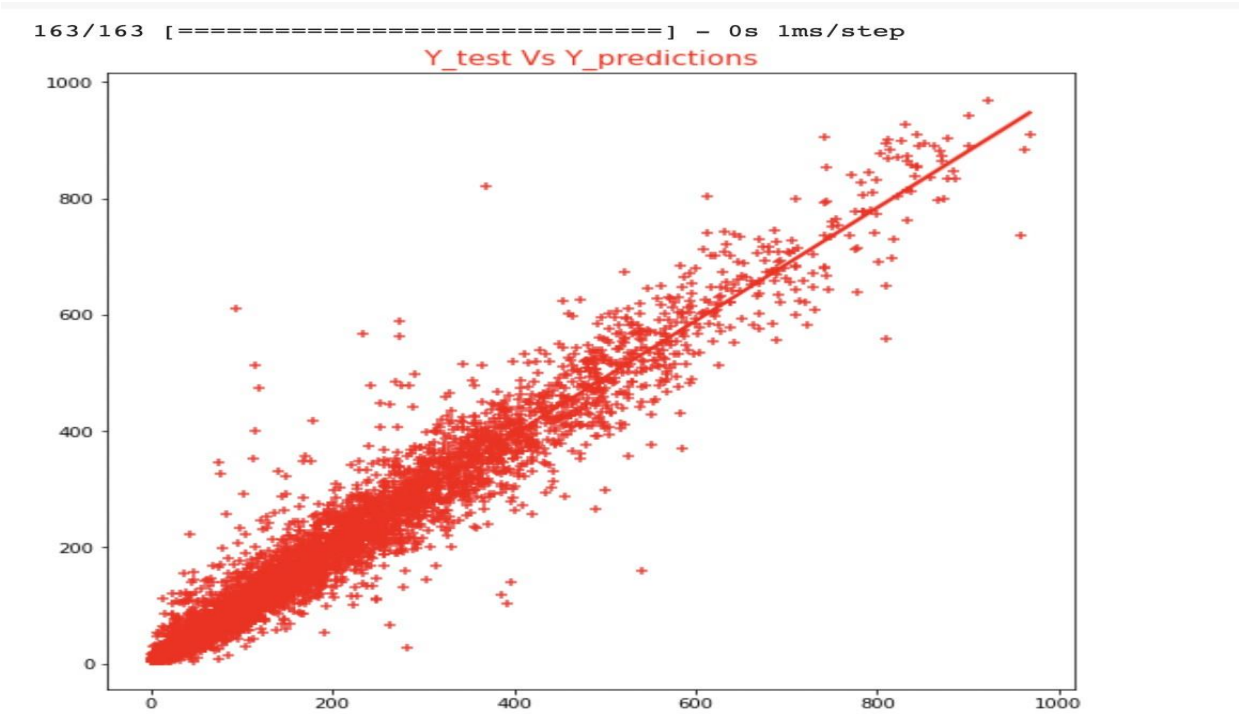


Figure 11. Scatterplot between Y_test and Y_predictions

6.2 K-Nearest Neighbors

We would use hyperparameter tuning to acquire the best neighbors for the k-nearest neighbor's algorithm. While using KNN, we must choose the number of nearest neighbors, and the algorithm will consider that number when making predictions. The algorithm would be more biased in favor of the majority class the higher the number of nearest neighbors because it would

base decisions on the number of nearest neighbors. We must choose the appropriate hyperparameter to guarantee that we acquire the best predictions from the cross-validation data. The iteration and various values for the hyperparameters are provided below so that the values may be fitted into the training set. We can obtain the values from the cross-validation data.

For different values of the nearest neighbors, we get another mean squared error and mean absolute error, respectively.

	K Nearest Neighbors	Mean Squared Error	Mean Absolute Error
0	2	12945.821634	72.429613
1	3	12988.911669	74.553062
2	5	12878.514538	77.040967
3	8	13143.848401	80.212936
4	10	13326.131830	81.903759
5	11	13415.819808	82.686892
6	15	13906.935362	85.300141
7	20	14263.887532	87.385021

Table 2. different values of the nearest neighbors

Line plot between Neighbors and the Mean Squared Error

The figure below shows an increase in the error with an increase in the value of K. As can be seen, K should be set at 5. Since that was the best K value out of the K values we evaluated during the cross-validation stage of our machine learning problem, we would use that value in our predictions as in the following figure 8.

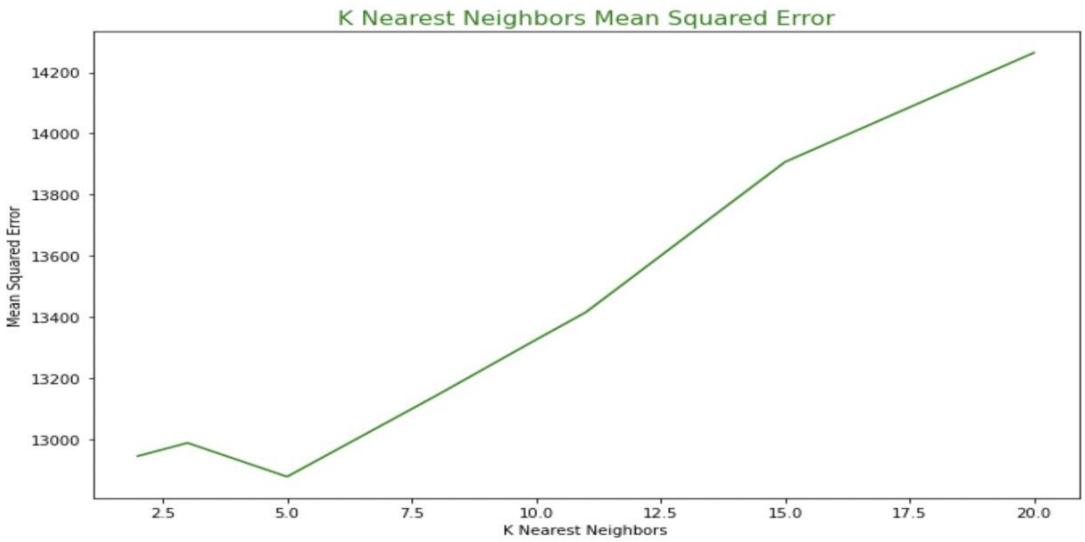


Figure 12. Line plot between Neighbors and the Mean Squared Error

Line plot between Neighbors and Mean Absolute Error.

This is in line with an upward trend where the mean absolute error increases proportionally to rising K values. As a result, the fundamental mean mistake increases significantly along with the value of K. K should be set to 2, as this is where the mistake is least. However, we'll continue to use K as its optimum value of 5 since Mean Absolute Error is a crucial indicator that helps us determine whether predictions are accurate, as shown in figure 9.

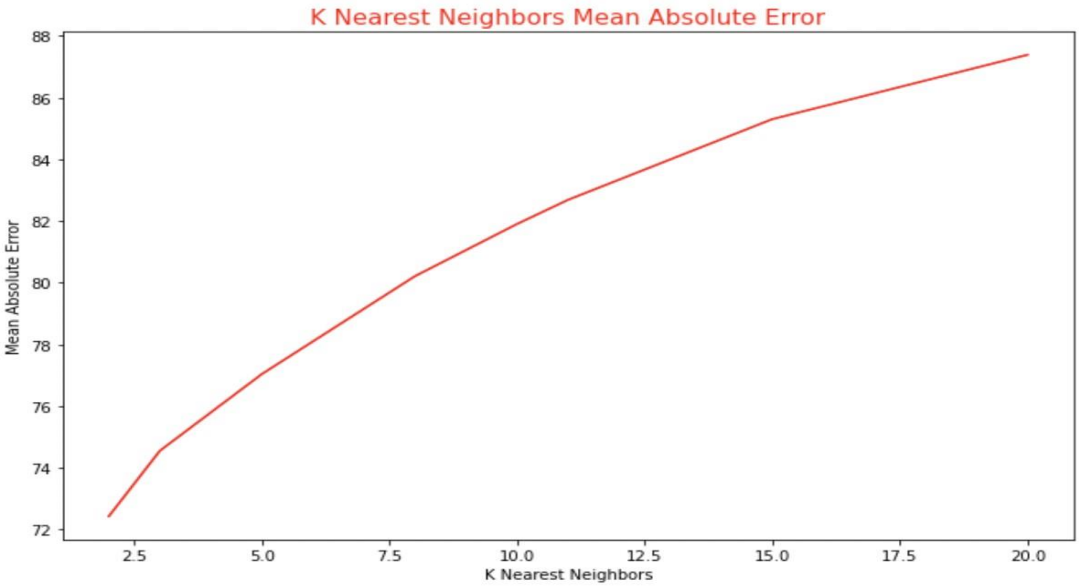


Figure 13: Line plot between Neighbors and Mean Absolute Error

Scatterplot between Y_test and Y_predictions

Since the values of the y test and y predictions do not fit in a straight line as displayed, we can observe that the K Nearest Regressor could perform better. Given that the values between the y test and y predictions are pretty dispersed, K Nearest Regressor is a good machine learning model for our task, as shown in Figure 10.

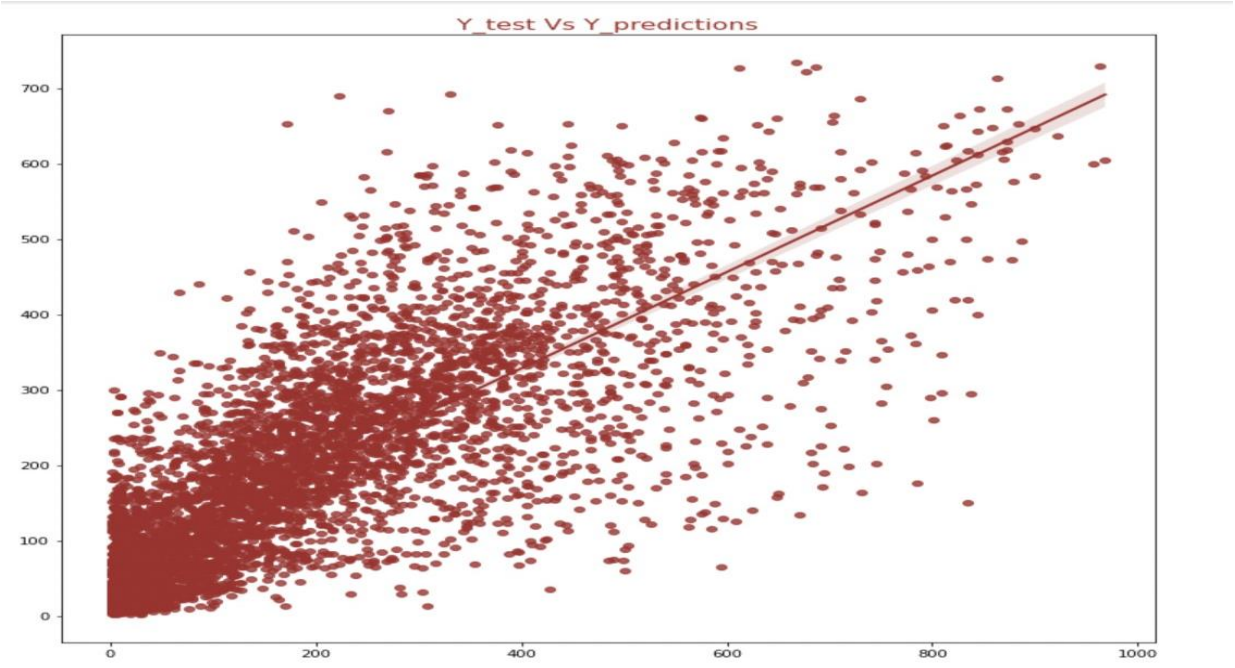


Figure 14: Scatterplot between Y_test and Y_predictions

6.3 PLS Regression

Another regression type that allows us to choose the number of components of a hyperparameter is this one. We would adjust the hyperparameter, test several values, and determine which would yield the most accurate forecasts.

Plots for PLS Regression

The mean absolute error plot and the mean squared error plot are what we will now examine, respectively. As can be seen, 5 in figure 5.3.1 is the ideal number of components because the mean absolute error is at its lowest and the mean squared error is at its lowest. We choose the optimal hyperparameter so that the error is standard for both the mean squared error and the mean absolute error, even though there are other values of components where the mean squared error is lower.

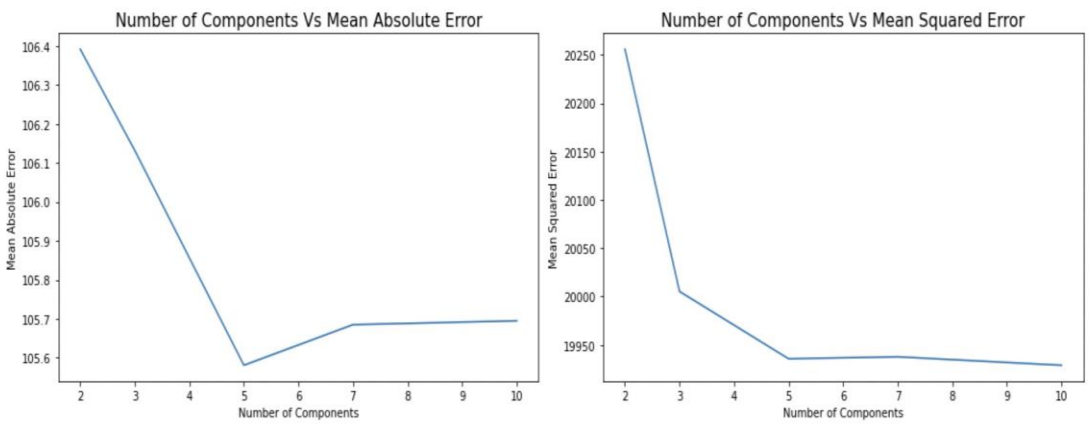


Figure 15: Plots for PLS Regression

6.4 Decision Tree Regressor

Using the Decision Tree Regressor, we have quite a few hyperparameters. Still, we would be working with just one hyperparameter, which is the depth of the tree. Let us choose different values for the max depth and see how the values influence different machine learning outcomes, respectively.

We would examine various decision tree regressor max depth settings and observe how the values would alter following multiple values. Additionally, we can watch that for the mean squared error and mean absolute error; the max depth value is optimum when it is 15 correspondingly. Therefore, 15, as per figure 12, is the best choice for the maximum depth because it has the lowest mean squared error and means absolute error.

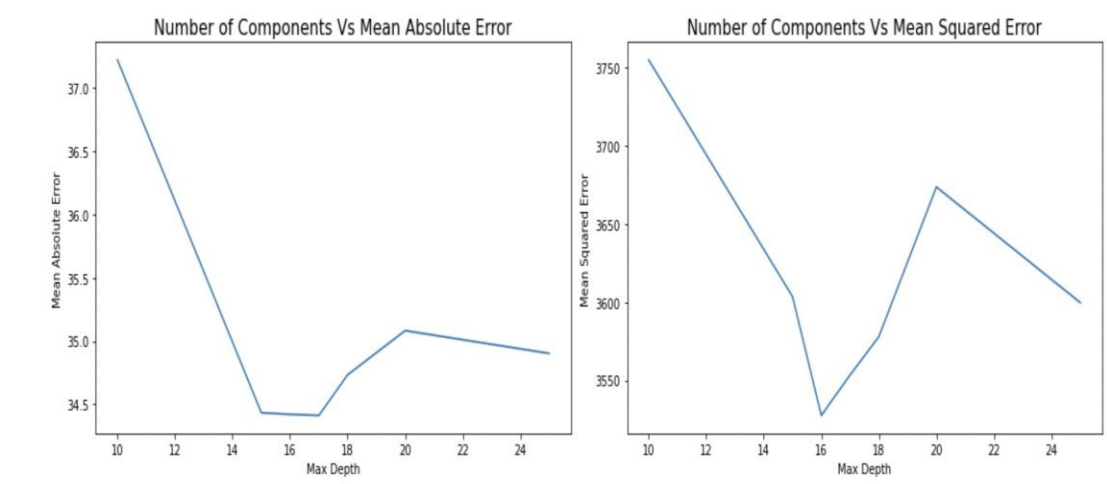


Figure 16: Decision Tree Regressor

Y_test and Y_predictions

The decision tree regressor produced very accurate predictions on the test set, as seen from the scatterplot below. The forecasts and the actual test output values are reasonably close. As a result, the outcome might be predicted using a decision tree regressor. However, let's also examine the performance of the other machine-learning models on the test set.

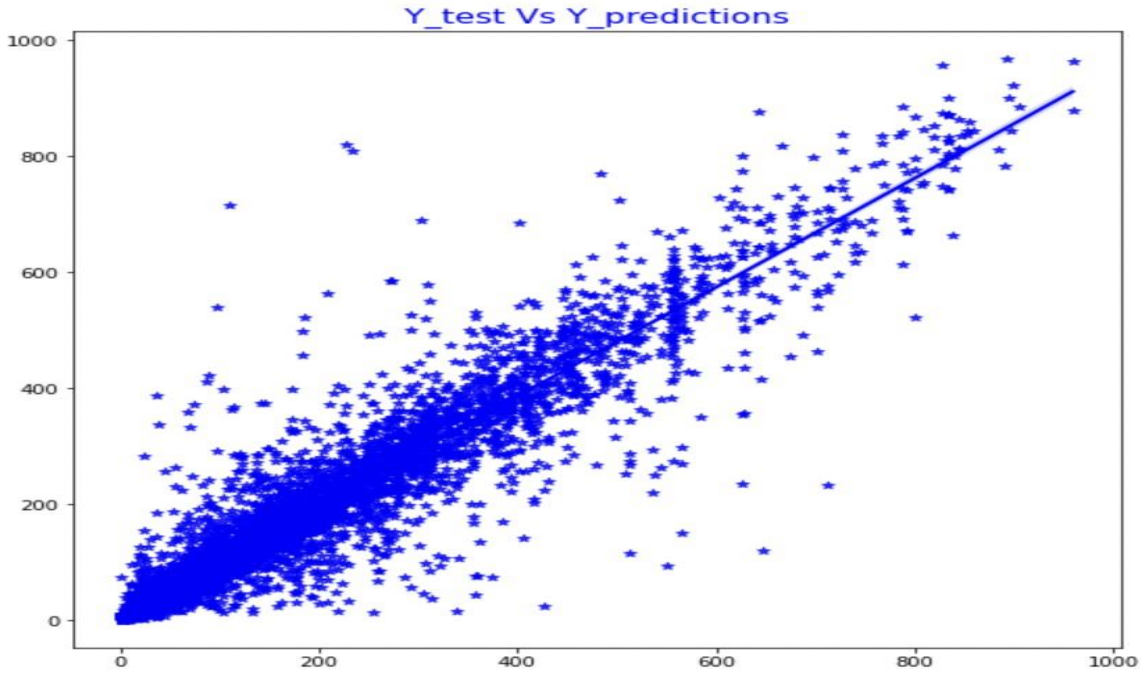


Figure 17: Y_test and Y_predictions

6.5 Gradient Boosting Regressor

Use the most popular machine learning model for predictions now. We would employ the gradient-boosting decision tree model to guarantee that we would obtain the most accurate forecasts. We can see that there are various estimators, each of which is a hyperparameter that needs to be tuned to get the mean squared error and mean absolute error with the lowest errors possible.

Plots of Mean Absolute Error and Mean Squared Error

The graph below clearly illustrates how the machine learning model advances as the number of estimators rises. As a result, we would train the model with various estimators to obtain the lowest error. To guarantee that we get the optimal machine learning models for prediction, we will learn and apply multiple components, as shown in Figure 16

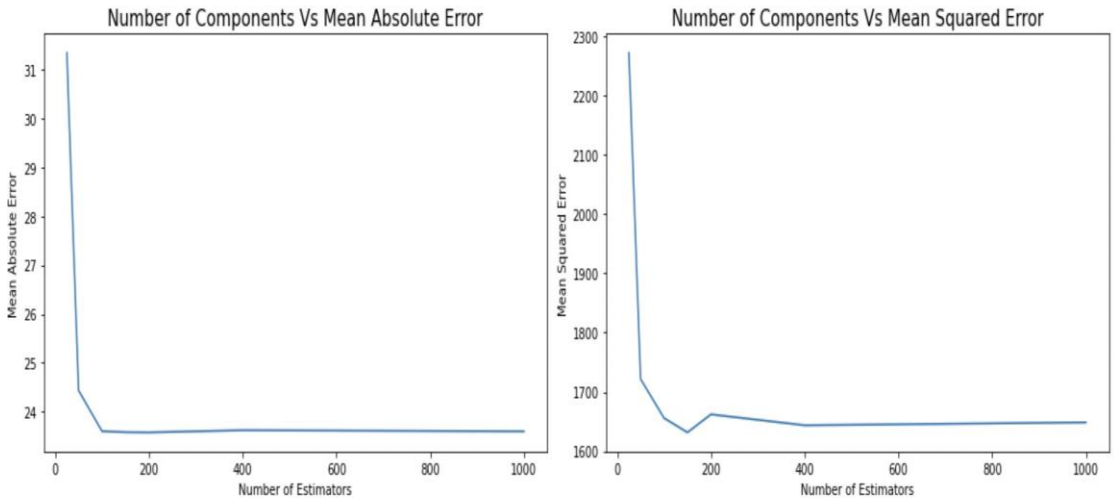


Figure 18: Mean Absolute Error and Mean Squared Error

Scatterplot between Y_test and Y_predictions

As you can see, a straight line can be made between the forecasts and the actual test results, respectively. In addition, we observe that the model outperforms other machine learning models reasonably well. Due to its positive effects on the cross-validation set, we would use this machine-learning model for deployment.

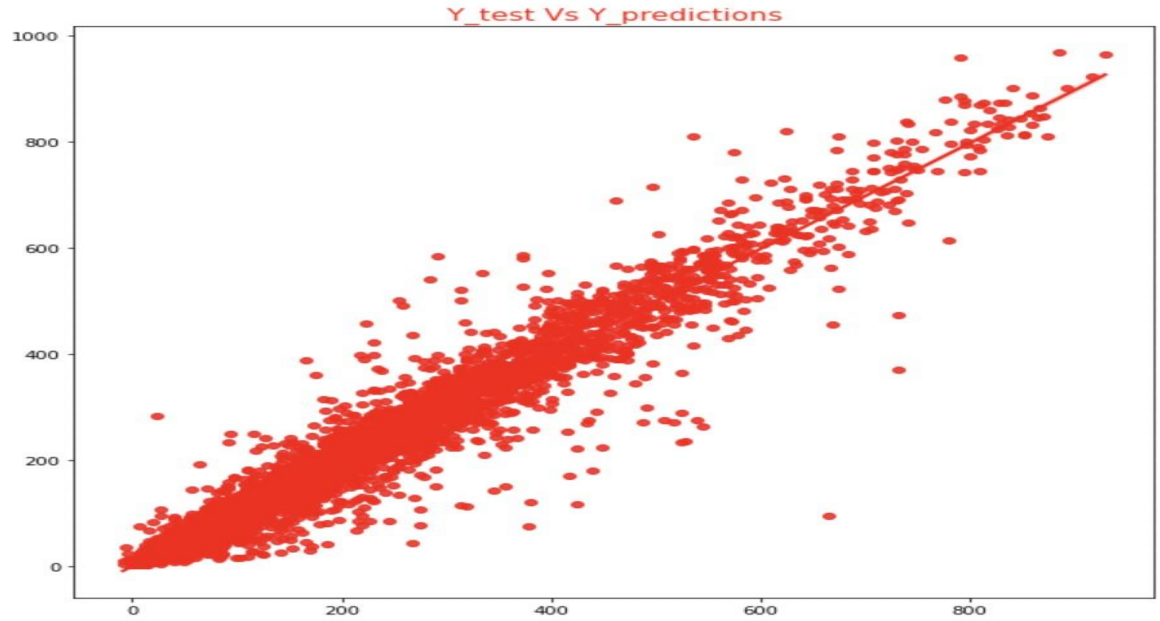


Figure 19: .Scatterplot between Y_test and Y_predictions

6.6 Logistic Regression

Ultimately, we use the Logistic Regression model and see how well the model does. The plot below shows that it performs less closely than the other machine learning models. Therefore, using different machine learning models for prediction is better.

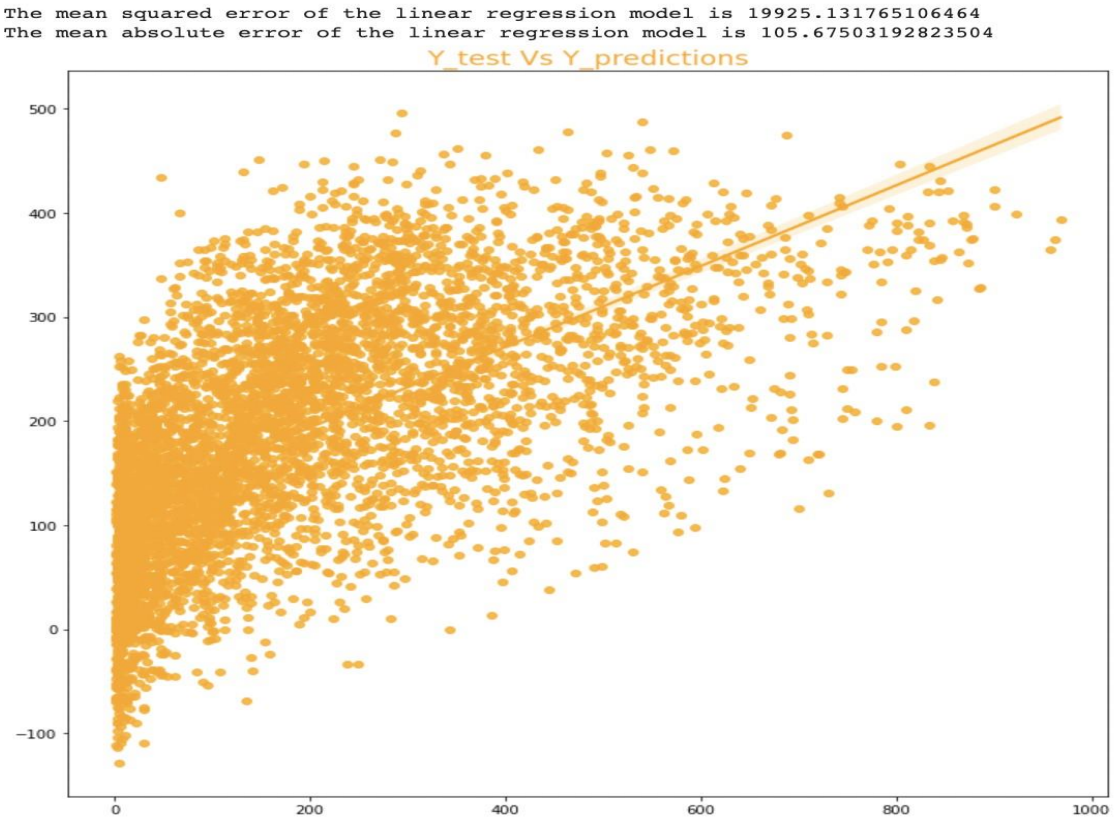


Figure 20: Scatterplot logistic Regration between Y_test and Y_predictions

7. Outcomes

After comparing with other models below is our outcome.

- Gradient Boosting Decision Regressor and Deep Neural Networks are the models that were able to do the best in anticipating the demand for bikes.
- To ensure a thorough grasp of all aspects and their respective contributions to the outcome variable, exploratory data analysis (EDA) was carried out.
- The mean absolute error (MAE) produced by the best machine learning models was roughly 23, which is excellent given the size of the issue at hand.

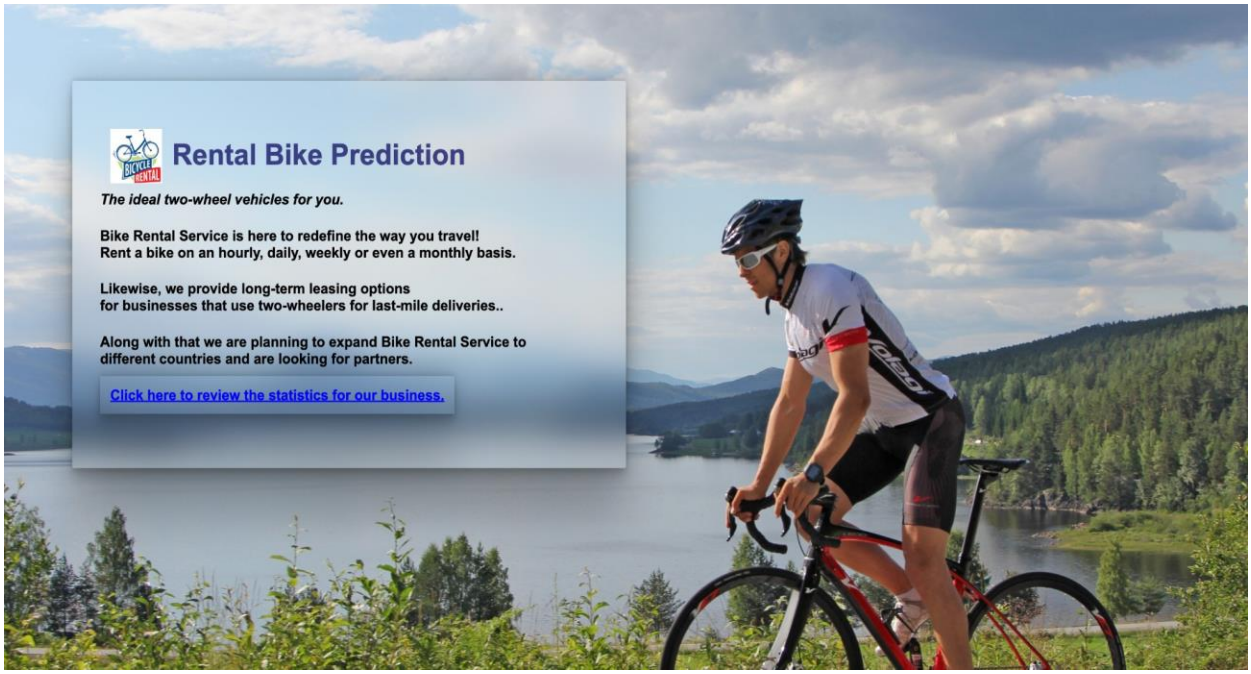
8. Conclusion:

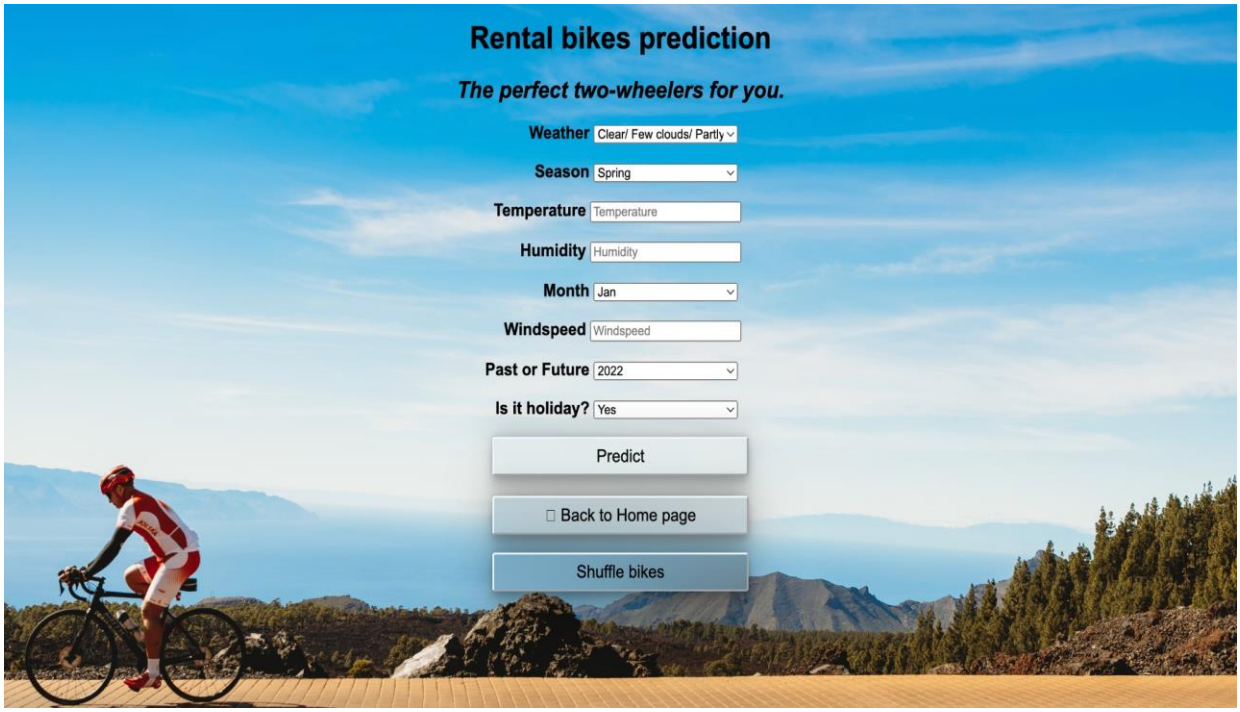
- The cross-validation data showed that the gradient-boosting approach outperformed the other machine-learning models. Although the gradient boosting tree was the best machine learning model for predicting bike demand, deep neural networks also scored well.
- Scatterplots were employed to obtain the forecasts' output and the actual test results.
- The values were changed to range between 0 and 1, accordingly.
- An excellent grasp of various features and how they affect the results for multiple datasets and distributions can be gained through exploratory data analysis.

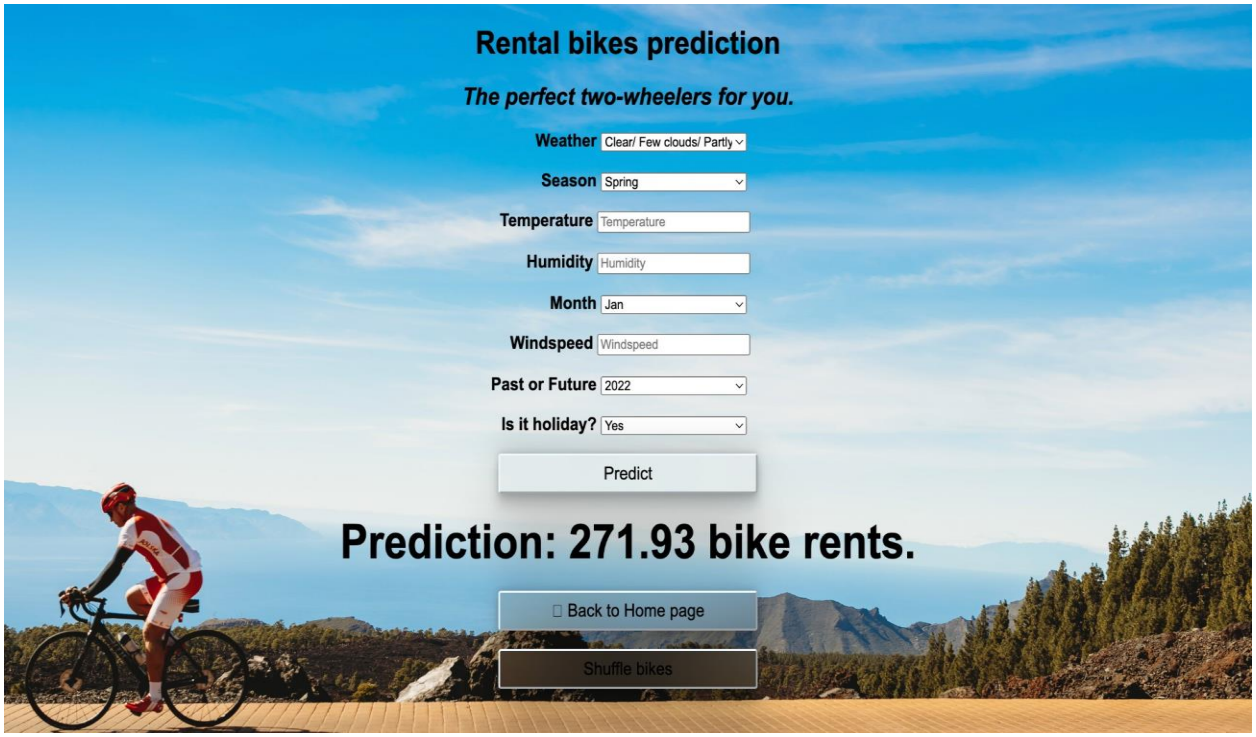
9. Future Scope

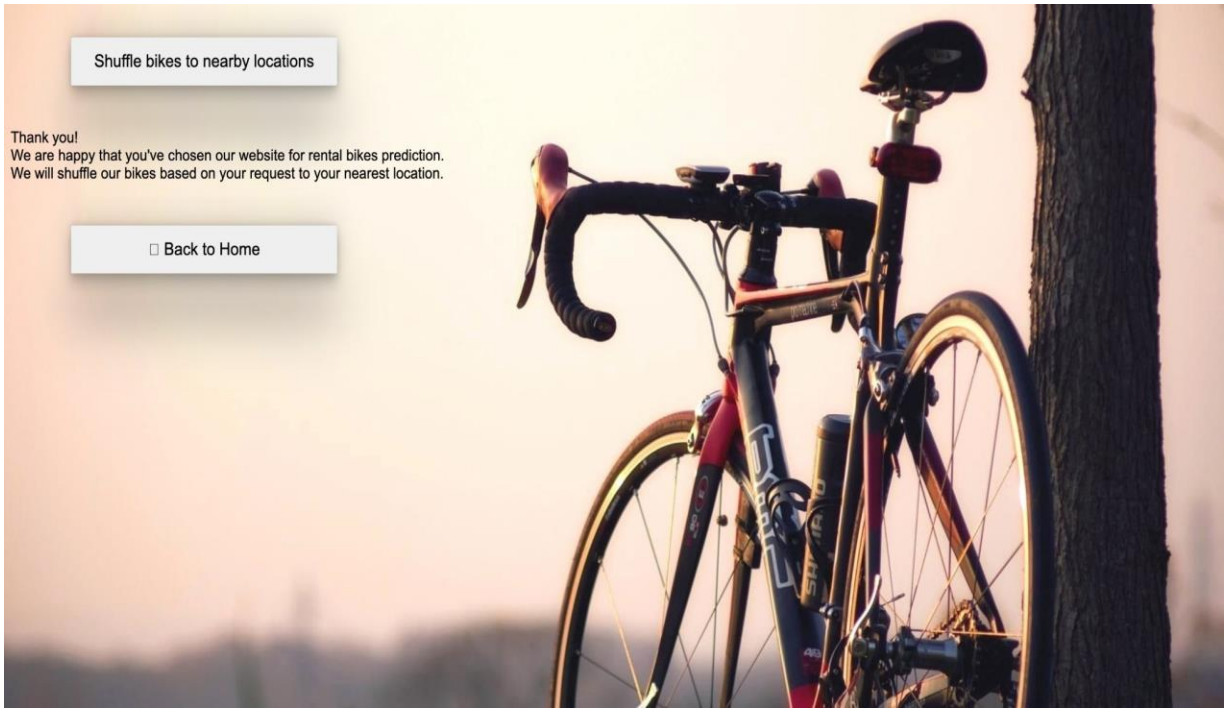
- Additional data like the street connection score and people's impressions of the biking environment might be included to get even better predictions from the models.
- The best machine learning models could be implemented in real-time, highlighting the demand for bikes so that administrators could act in response to it.
- Since the current dataset only includes data for Washington, D.C., our future goals would be to forecast the number of bikes that will be rented and to shuffle those rentals based on any place and region across the entire United States.

10. Application GUI:






Rental bikes prediction
The perfect two-wheelers for you.
Weather
Clear/ Few clouds/ Partly
Season
Spring
Temperature
Temperature
Humidity
Humidity
Month
Jan
Windspeed
Windspeed
Past or Future
2022
Is it holiday?
Yes
Predict
Prediction: 271.93 bike rents.
Back to Home page
Shuffle bikes


Shuffle bikes to nearby locations
Thank you!
We are happy that you've chosen our website for rental bikes prediction.
We will shuffle our bikes based on your request to your nearest location.
Back to Home

REFERENCES

1. IJISET - International Journal of Innovative Science, Engineering & Technology- Bike share demand prediction.
2. J. Larsen. (April 25, 2013). Plan B Updates - 112: Bike-Sharing Programs Hit the Streets in Over 500 Cities Worldwide.
3. P. DeMaio, "Bike-sharing: History, Impacts, Models of Provision, and Future," Journal of Public Transportation, vol. 12, no. 4, p. 3, 2009.
4. Abdelhalim, A., & Traore, I. (2009). A new method for learning decision trees from rules. In Machine Learning and Applications, 2009. ICMLA 09. International Conference on (pp. 693–698).
5. Alippi, C., & Roveri, M. (2010). Virtual k-fold cross validation: An effective method for accuracy assessment. In The 2010 International Joint Conference on Neural Networks (IJCNN)
6. Jing, C., & Zhao, Z. (2015). Research on Antecedents and Consequences of Factors Affecting the Bike Sharing System—Lessons From Capital Bike Share Program in Washington, DC. In the International Conference on Logistics Engineering, Management and Computer Science (LEMCS 2015). Atlantis Press.
7. Joelsson, S. R., Benediktsson, J. A., & Sveinsson, J. R. (2005). Random forest classifiers for hyperspectral data. In Geoscience and Remote Sensing Symposium, 2005. IGARSS 05. Proceedings. 2005 IEEE International (Vol. 1, p. 4–pp). IEEE.
8. Larsen, J. (2013). Bike-sharing programs hit the streets in over 500 cities worldwide. Earth Policy Institute, 25.
9. Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection, 4, 40–79.
10. YouLi Feng , ShanShan Wang “A Forecast for Bicycle Rental Demand Based on Random Forests and Multiple Linear Regression”
11. R. Giot, R.Cherrier“Predicting Bike Share Demand upto One hour ahead” 2013 IEEE 9th International Confrence on Data Management, France.