**Mini Project**

On

**BIKE SHARING PREDICTION**

**Submitted in partial fulfillment of the requirements for the award of degree of**

**BACHELOR OF TECHNOLOGY**

**In**

**COMPUTER SCIENCE & ENGINEERING**

**BY**

**18WH5A0506 SAMA NISCHALA**

**17WH1A0521 DANNAPANENI TRIPURA**

**18WH5A0510 KOBBA JAGADISHWARI**

**Under the esteemed guidance of**

**Mr. U Chandrasekhar**

**Associate Professor**



**Department of Computer Science & Engineering**

BVRIT HYDERABAD

COLLEGE OF ENGINEERING FOR WOMEN

**(NBA Accredited EEE.ECE.CSE.IT B.Tech Courses)**

**(Approved by AICTE, New Delhi and Affiliated to JNTUH, Hyderabad)**

**Bachupally, Hyderabad – 500090**

**Nov, 2019**

BVRIT HYDERABAD

COLLEGE OF ENGINEERING FOR WOMEN

**(NBA Accredited EEE.ECE.CSE.IT B.Tech Courses)**

**(Approved by AICTE, New Delhi and Affiliated to JNTUH, Hyderabad)**

**Bachupally, Hyderabad – 500090**

Department of Computer Science & Engineering



**CERTIFICATE**

This is to certify that the mini project entitled “**Bike Sharing Prediction**” is a bonafide work carried out by **Ms. SAMA NISCHALA (18WH5A0506), Ms. DANNAPANENI TRIPURA (17WH1A0521), Ms. KOBBA JAGADISHWARI (18WH5A0510)** in partial fulfillment for the award of B.Tech degree in **Computer Science & Engineering**, **BVRIT HYDERABAD College of Engineering for Women, Bachupally, Hyderabad**, affiliated to Jawaharlal Nehru Technological University Hyderabad, Hyderabad under my guidance and supervision. The results embodied in the project work have not been submitted to any other University or Institute for the award of any degree or diploma.

**Internal Guide Head of the Department**

**U. Chandrasekhar Dr. Ch. Srinivasulu**

**Associate Professor, CSE Professor, CSE**

**DECLARATION**

We hereby declare that the work presented in this project entitled **“Bike Sharing Prediction”** submitted towards completion of Project work in IV Year of B.Tech of CSE at **BVRIT HYDERABAD College of Engineering for Women,** Hyderabad is an authentic record of our original work carried out under the guidance of **U Chandrasekhar, Associate Professor, Department of CSE.**

Sign with Date:

SAMA NISCHALA

(18wh5a0506)

Sign with Date:

DANNAPANENI TRIPURA

(17wh1a0521)

Sign with Date:

KOBBA JAGADISHWARI

(18wh5a0510)

**ACKNOWLEDGEMENT**

We would like to express our sincere thanks to **Dr.K.V.N.Sunitha**, **Principal, BVRIT HYDERABAD College of Engineering for Women,** for her support by providing the working facilities in the college.

Our sincere thanks and gratitude to **Dr.Ch.Srinivasulu, Head, Department of CSE, BVRIT HYDERABAD College of Engineering for Women,** for all timely support and valuable suggestions during the period of our project.

We are extremely thankful to our Internal Guide, **U Chandrasekhar, Associate Professor, CSE, BVRIT HYDERABAD College of Engineering for Women,** for his constant guidance and encouragement throughout the project.

Finally, we would like to thank our Mini Project Coordinator, all Faculty and Staff of CSE department who helped us directly or indirectly. Last but not least, we wish to acknowledge our **Parents** and **Friends** for giving moral strength and constant encouragement.

**SAMA NISCHALA (18wh5a0506)**

**DANNAPANENI TRIPURA (17wh1a0521)**

**KOBBA JAGADISHWARI (18wh5a0510)**

**ABSTRACT**

In this project we tried to apply machine learning algorithm into a real world problem – “Bike Sharing Prediction”. This dataset contains the hourly and daily count of rental bikes between years 2011 and 2012 in capital bike share system with the corresponding weather and seasonal information.

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **S.NO** | **Figure Description** | **Page no** |
| 1. | Design | 9 |
| 2. | Blot Plots | 13-17 |
| 3. | Correlation Analysis | 19 |
| 4. | Train Model | 21 |
| 5. | Feature Importance | 23 |

**LIST OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **S.no** | **Topic** | **Page no.** |
| 1. | Introduction | 8 |
| 1.1) | Problem Statement | 8 |
| 1.2) | Objective | 8 |
| 2. | Requirements | 8 |
| 2.1) | Software | 8 |
| 2.2) | Hardware | 8 |
| 3. | Design | 9 |
| 4. | Implementation | 10 |
| 4.1) | Imports | 10 |
| 4.2) | Data Analysis | 11 |
| 4.3) | Blot Plots | 12 |
| 4.4) | Correlation Analysis | 17 |
| 4.5) | Data Split | 20 |
| 5. | Train Model | 21 |
| 5.1) | Model with RandomForestRegressor | 22 |
| 6 | Feature Importance | 23 |
| 7 | Conclusion | 24 |
| 8 | Reference | 24 |
| 9 | Project Link | 24 |

**1) INTRODUCTION**

* 1. **PROBLEM STATEMENT**

Make an explorative data analysis and build a model for the hourly utilization “cnt” of the given dataset. Report the mean absolute deviations

* 1. **OBJECTIVE**

The main objective is to predict or classify the bike rental count hourly or daily based on the environmental and seasonal settings by using regression and classification.

For this purpose, we decided to treat the task as both regression and classification problem with each group member experimenting with a different algorithm for regression and classification.

**2) REQUIREMENTS**

**2.1) SOFTWARE**

**-** Google Collaborator

- Dataset (17379×17)

- Python Language

**2.2) HARDWARE**

**-** Intel core Pentium processor

- RAM 4GB

- Hard Disk Drive 1 TB

**3.) Design**:

Bike Sharing Prediction

Fetch & Analyze the data

Fetch & Analysis the data

­­

Visual Representation

Result Analysis

Gradient Boosting Regression

Correlation Analysis

Random Forest Regression

Train the Model Data

Linear Regression

Feed the data to the model

Building a Prediction Model

Create new Factors

Partition Data

Remove Missing data

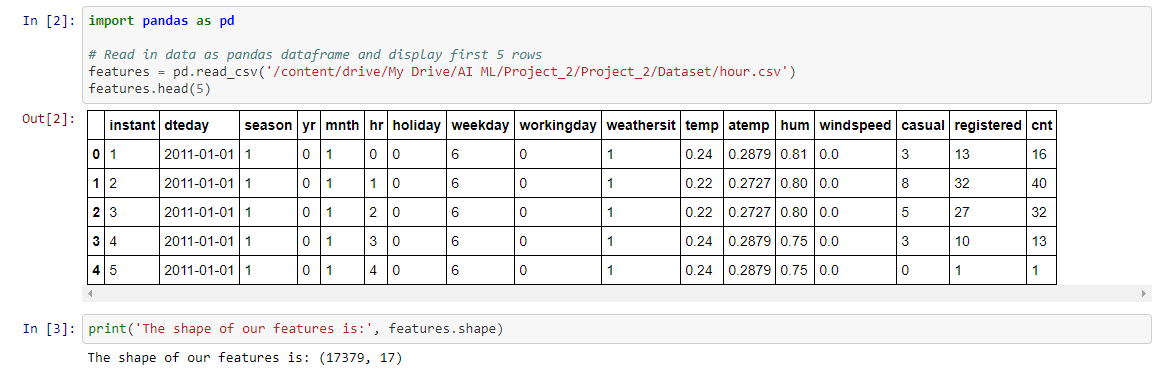
Datsa

Clean Data

Preprocessing

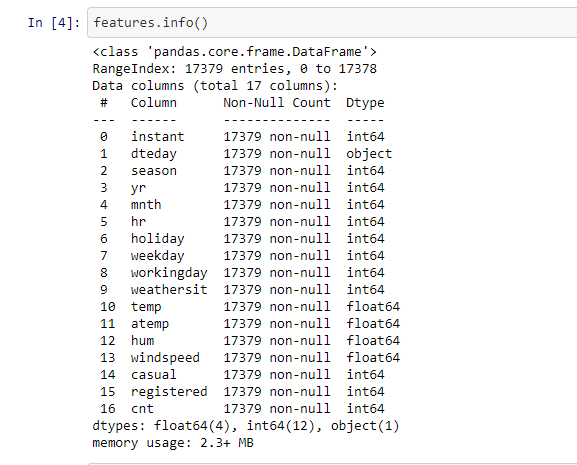
**4) IMPLEMENTATION:**

**4.1) IMPORTS**

****

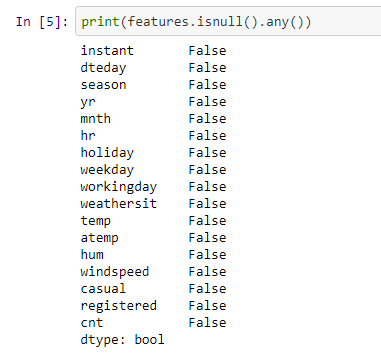
**4.2) DATA ANALYSIS:**

Data Analysis is a process of understanding the data. In this we find patterns and try to obtain inferences due to which the underlying patterns are observed. It is a process of inspecting, cleansing, transforming, and modeling data with the**goal of discovering useful information, suggesting conclusions, and supporting decision-making.**

****

Now, Here we check for any null values that are present or not.

**Null Value**: A **NULL value** is a special marker used in SQL to indicate that a data **value** does not exist in the database.

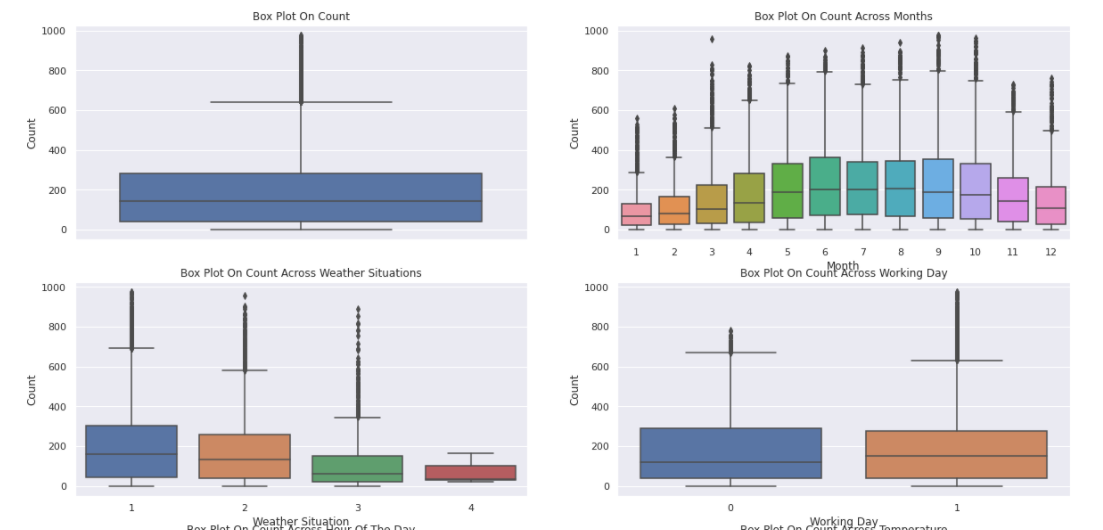
****

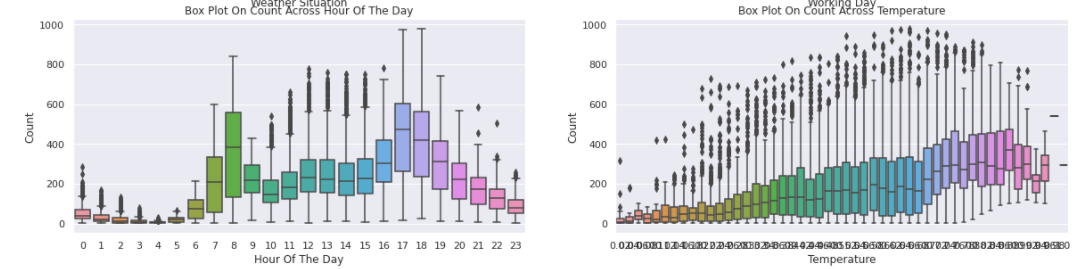
Here we can see that they is no null values are present in this dataset.

**BOX PLOTS:**

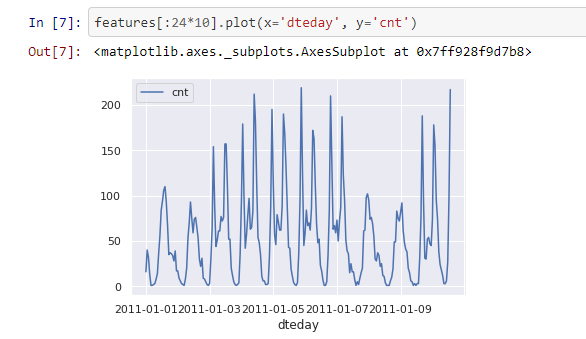
A boxplot is used below to analyze the relationship between a categorical feature and a continuous feature. Here we use various plots to analyze the data.

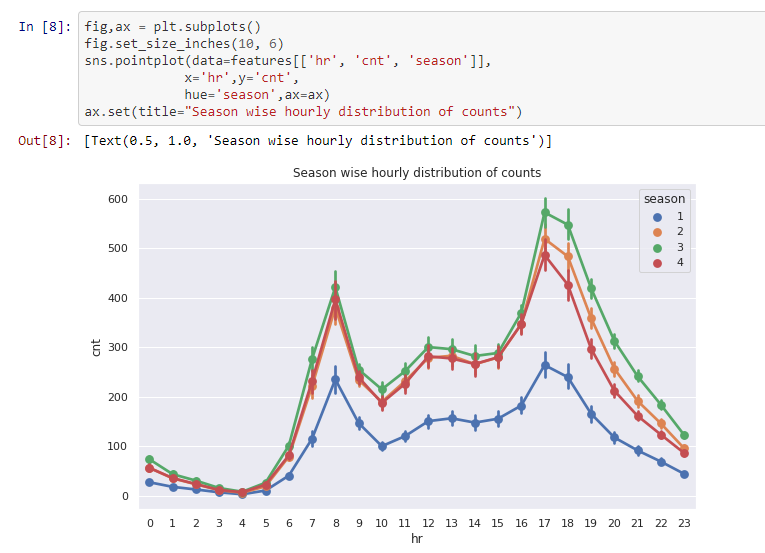
****





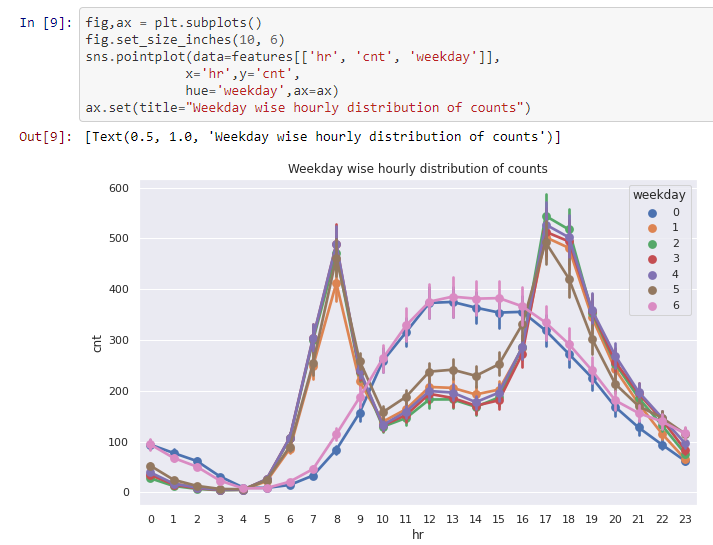
 The working day and holiday box plot indicates that more bicycles are rent during normal working days than on weekends or holidays. The hourly box plots show a local maximum at 8 am in the morning and 5pm and 6 pm in the evening. Another important factor seems to be the temperature: higher temperatures lead to an increasing number of bike rents and lower temperatures not only decrease the average number of rents but also shows more outliers in the data.





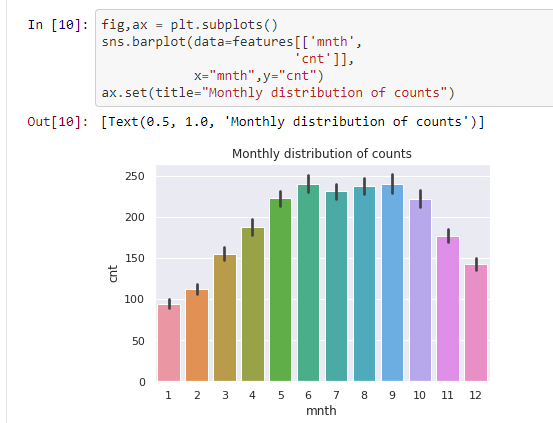
* Here, blue colour indicates the “spring” season.
* Orange colour indicates the “summer” season.
* Green colour indicates the “Rainy” season.
* Red colour indicates the “winter” season.

Hence, we observe that spring has the lowest count throughout the year.

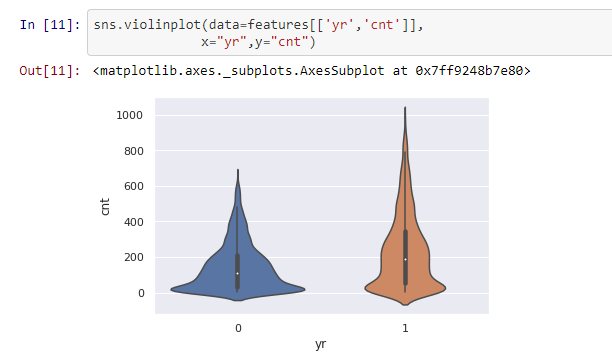


* Here blue colour indicates “Sunday”.
* Orange colour indicates “Monday”.
* Green colour indicates “Tuesday”.
* Red colour indicates “Wednesday”.
* Blue colour indicates “Thursday”.
* Violet colour indicates “Friday”.
* Pink colour indicates “Saturday”.

Days 0 and 6 have similar trend.

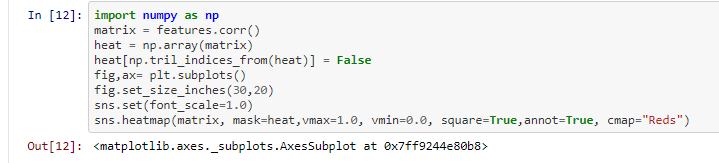


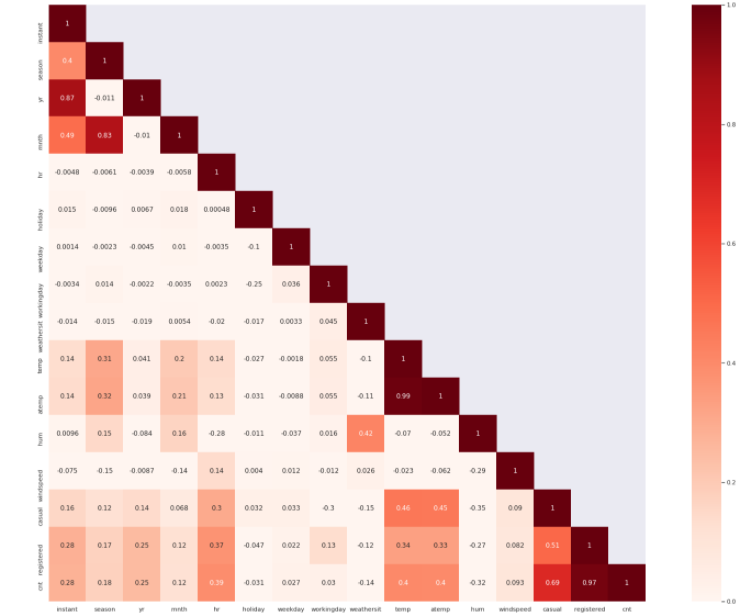
June and September see the highest bike users.



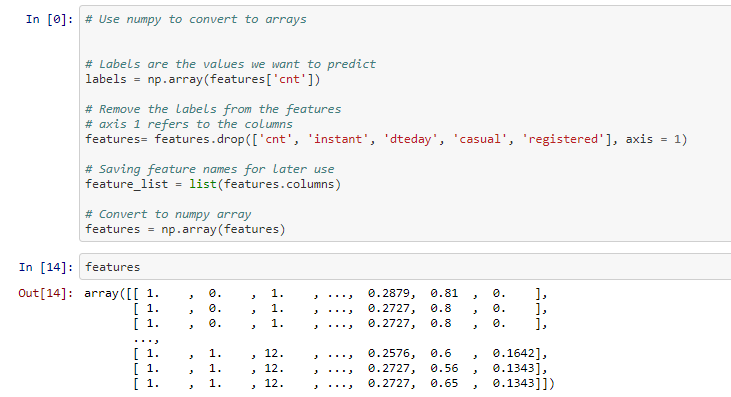
**4.4) CORRELATION ANALYSIS**

Correlation Analysis is a statistical method that is used to discover if there is a relationship between two variables and how strong that relationship may be.

****

****

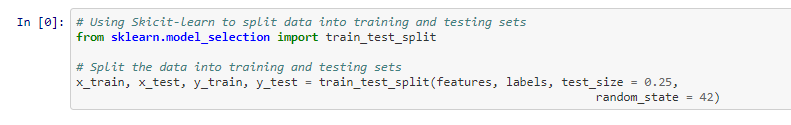
**4.5) DATA SPLIT:**

****

**train\_test\_split ():** It method from the sklearn package is used to split training and testing data.

**test\_size** — this parameter decides the size of the data that has to be split as the test dataset. This is given as a fraction. For example, if you pass 0.5 as the value, the dataset will be split 50% as the test dataset. If you’re specifying this parameter, you can ignore the next parameter.

**Random\_state** — here you pass an integer, which will act as the seed for the random number generator during the split.

****

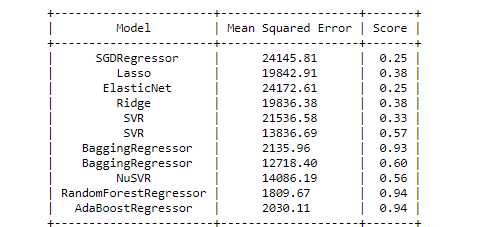
Hence, we take 25% data for test data with a random state 42.

**4.6) TRAIN MODEL**

1. You configure a model, by choosing a particular type of algorithm, and defining its parameters or hyper parameters. Choose any of the following model types:
   * [Classification models](https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/machine-learning-initialize-model-classification), based on neural networks, decision trees, and decision forests, and other algorithms.
   * [Regression models](https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/machine-learning-initialize-model-regression), which can include standard linear regression, or which use other algorithms, including neural networks and Bayesian regression.
2. Provide a dataset that is labelled, and has data compatible with the algorithm. Connect both the data and the model to Train Model.
3. After training is completed, use the trained model with one of the [scoring modules](https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/machine-learning-score), to make predictions on new data.

* Here, we use regression model as ‘cnt’ is a continuous output variables.

****

****

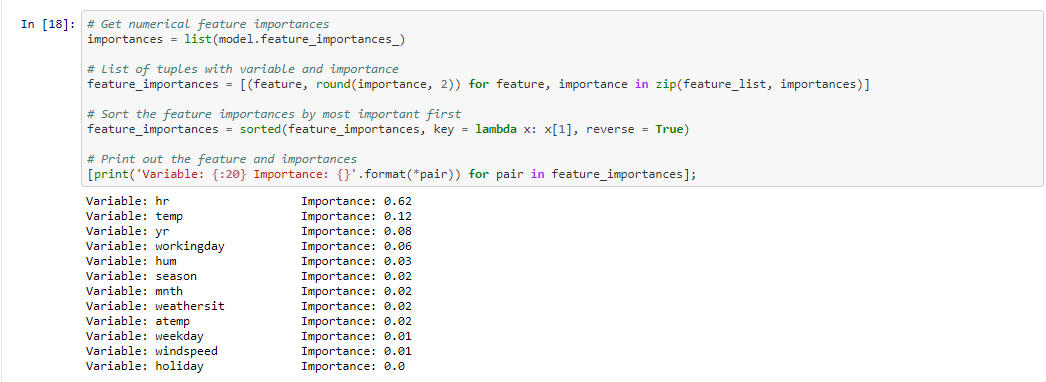
In the above table we notice that ensemble Regression models (RandomForestRegressor, AdaBoostRegressor and BaggingRegressor) have highest accuracy score. So, we choose RandomForestRegressor as it has least mean square error when compared to other regression models.

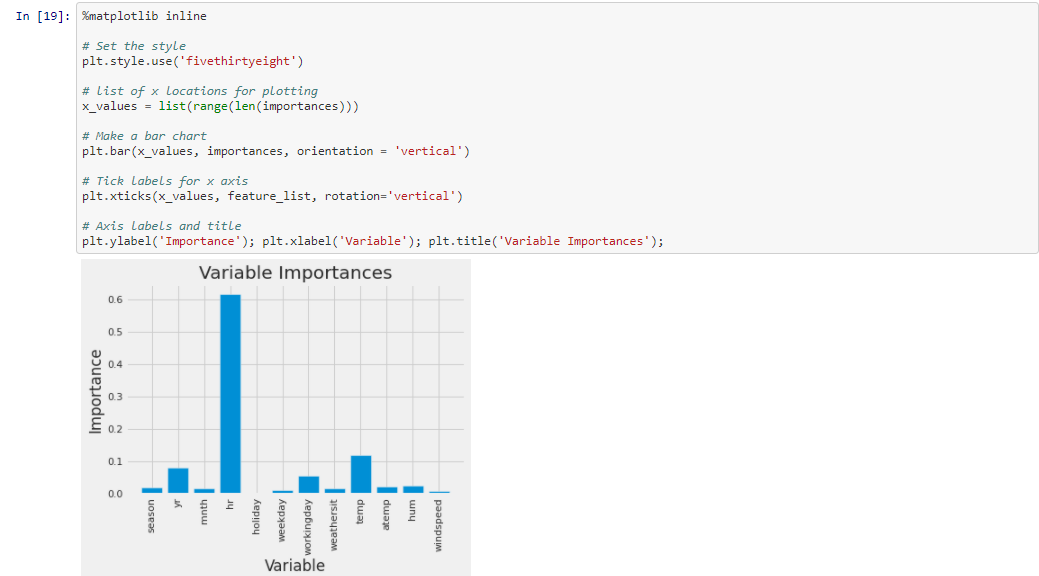
**MODEL WITH RANDOM FOREST REGRESSOR**

Random Forest is a learning method that operates by constructing multiple decision trees. The final decision is made based on the majority of the trees and is chosen by the random forest.

****

**Feature Importance:**

****

****

**CONCLUSION:**

 The result corresponds to the high correlation of the hour and temperature variable with the bicycle sharing count in the feature correlation matrix.

**REFERENCES:**

* Github
* Kaggle

**PROJECT LINK:**

https://github.com/d-b-tripura/Bike\_Sharing/blob/master/BikeShare.ipynb