**ML Regression Capstone Project**

**Seoul Bike Sharing Demand Prediction**

Kritisha Panda

**Abstract:**

Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually, providing the city with a stable supply of rental bikes becomes a major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes. This report demonstrates the use of various ML algorithms and a comparison between them based on certain evaluation metrics to decide the best model that will give the most accurate predictions for rented bike count each hour.

Introduction:

The use of ML in business analytics has occupied various fields and many positions. Major reasons are higher volume, data availability and fast processing. Businesses are now making profit by use of ML and implementing it in their systems to compete. The major application in business is to assist in extracting information and knowledge from massive data sets. According to recent studies, it is expected that more than 60% of the population in the world tends to dwell in cities, which is higher than 50% of the present scenario. Some countries around the world are practising righteous scenarios, renderings mobility at a fair cost and reduced carbon discharge. On the contrary other cities are far behind in the track. Urban mobility usually fills 64% of the entire kilometres travelled in the world. It ought to be modelled and taken over by inter-modality and networked self-driving vehicles which also provides a sustainable means of mobility. Systems called Mobility on Demand has a vital part in raising the vehicles’ supply, increasing its idle time and numbers.

**Problem Description**

Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually, providing the city with a stable supply of rental bikes becomes a major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.

Today, bike-sharing systems are blooming across more cities around the world. To complete a short trip renting a bike is a faster way when compared to walking. Moreover, it is eco-friendly and comfortable too compared to driving.

**1. Data Description**

The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information

**Attribute Information**

* Date - year-month-day
* Rented Bike count - Count of bikes rented at each hour
* Hour - Hour of the day
* Temperature- Temperature in Celsius
* Humidity - %
* Windspeed - m/s
* Visibility - 10m
* Dew point temperature - Celsius
* Solar radiation - MJ/m2
* Rainfall - mm
* Snowfall - cm
* Seasons - Winter, Spring, Summer, Autumn
* Holiday - Holiday/No holiday
* Functional Day – No Func(Non Functional Hours), Fun(Functional hours)

**Feature Breakdown:**

**Date**: The date of the day, during 365 days from 01/12/2017 to 30/11/2018, formatting in DD/MM/YYYY, we need to convert into date-time format.

**Rented Bike Count**: Number of rented bikes per hour which our dependent variable and we need to predict that

**Hour:** The hour of the day, starting from 0-23 it's in a digital time format

**Temperature (°C):**  Temperature of the weather in Celsius and it varies from -17***°****C to 39.4****°****C*.

**Humidity (%)**: Availability of Humidity in the air during the booking and ranges from 0 to 98%.

**Wind speed (m/s):** Speed of the wind while booking and ranges from 0 to 7.4m/s.

**Visibility (10m):** Visibility to the eyes during driving in “m” and ranges from 27m to 2000m.

**Dew point temperature (°C)**: Temperature At the beginning of the dayand it ranges from -30.6**°**C to 27.2**°**C.

**Solar Radiation (MJ/m2):**  Sun contribution or solar radiation during ride booking which varies from 0 to 3.5 MJ/m2.

**Rainfall (mm):** The amount of rainfall during bike booking which ranges from 0 to 35mm.

**Snowfall (cm):** Amount of snowing in cm during the booking in cm and ranges from 0 to 8.8 cm.

**Seasons:** Seasons of the year and total there are 4 distinct seasons i.e summer, autumn, spring and winter.

**Holiday:** If the day is holiday period or not and there are 2 types of data that is holiday and no holiday

**Functioning Day:** If the day is a Functioning Day or not and it contains object data type yes and no.

**2.EDA on given Data set**

If we want to explain EDA in simple terms, it means trying to understand the given data much better, so that we can make some sense out of it. we using univariate frequency analysis was conducted to describe key characteristics of each feature including, minimum and maximum value, average, standard deviation and others. It was also used to produce a value distribution and identify missing values, and outliers.

EDA is a process of examining the available dataset to discover patterns, spot anomalies, test hypotheses, and check assumptions using statistical measures. In this chapter, we are going to discuss the steps involved in performing top notch exploratory data analysis

**2.1 Data Analysis:**

This is one of the most crucial steps that deals with descriptive statistics and analysis of the data. The main tasks involve summarizing the data, finding the hidden correlation and relationships among the data, developing predictive models, evaluating the models, and calculating the accuracies. Some of the techniques used for data summarization are summary tables, graphs, descriptive statistics, inferential statistics, correlation statistics, searching, grouping, and mathematical models

**2.2 Data Sourcing**

Data Sourcing is the process of finding and loading the data into our system. Broadly there are two ways in which we can find data.

1. Private Data
2. Public Data

Data collected from several sources must be stored in the correct format and transferred to the right information technology personnel within a company. As mentioned previously, data can be collected from several objects on several events using different types of sensors and storage tools.

**2.3 Data Pre-processing**

A dataset may contain noise, missing values, and inconsistent data; thus, pre-processing of data is essential to improve the quality of data and time required in the data mining.

**2.4 Data Cleaning**

After completing the Data Sourcing, the next step in the process of EDA is Data Cleaning. It is very important to get rid of the irregularities and clean the data after sourcing it into our system.

Irregularities are of different types of data.

Missing Values

1. Incorrect Format
2. Incorrect Headers
3. Anomalies/Outliers

**2.5 Data Deduplication**

It is very likely that your dataset contains duplicate rows. Removing them is essential to enhance the quality of the dataset.

**2.6 Missing Values**

There is a representation of each service and product for each customer. Missing values may occur because not all customers have the same subscription. Some of them may have a number of service and others may have something different. In addition, there are some columns related to system configurations and these columns may have null values but in our orange telecom data set there are no null values present

If there are missing values in the Dataset before doing any statistical analysis, we need to handle those missing values.

**2.7.Handling Outliers**

Outliers are data points that diverge from other observations for several reasons. During the EDA phase, one of our common tasks is to detect and filter these outliers. The main reason for this detection and filtering of outliers is that the presence of such outliers can cause serious issues in statistical analysis.

**2.8 Measures of Central Tendency**

The measure of central tendency tends to describe the average or mean value of datasets that is supposed to provide an optimal summarization of the entire set of measurements. This value is a number that is in some way central to the set. The most common measures for analysing the distribution frequency of data are the mean, median, and mode.

**2.9 Standardizing Values:**

To perform data analysis on a set of values, we have to make sure the values in the same column should be on the same scale. For example, if the data contains the values of the top speed of different companies’ cars, then the whole column should be either in meters/sec scale or miles/sec scale.

**2.10 Univariate analysis**

Univariate analysis is the simplest form of analysing data. It means that our data has only one type of variable and that we perform analysis over it. The main purpose of univariate analysis is to take data, summarize that data, and find patterns among the values. It doesn't deal with causes or relationships between the values. Several techniques that describe the patterns found in univariate data include central tendency (that is the mean, mode, and median) and dispersion (that is, the range, variance, maximum and minimum quartiles (including the interquartile range), and standard deviation).

**2.11 Bivariate Analysis**

If we analyse data by taking two variables/columns into consideration from a dataset, it is known as Bivariate Analysis.

**a) Numeric-Numeric Analysis:**

Analysing the two numeric variables from a dataset is known as numeric-numeric analysis. We can analyse it in three different ways.

* Scatter Plot
* Pair Plot
* Correlation Matrix

**b) Numeric - Categorical Analysis:**

Analysing the one numeric variable and one categorical variable from a dataset is known as numeric-categorical analysis. We analyse those mainly using mean, median, and box plots.

**3.Algorithms**

**3.1. Linear Regression**

Linear regression is a supervised machine learning model majorly used in forecasting. Supervised machine learning models are those where we use the training data to build the model and then test the accuracy of the model using the loss function.

Linear regression is one of the most widely known time series forecasting techniques which is used for predictive modelling. As the name suggests, it assumes a linear relationship between a set of independent variables to that of the dependent variable (the variable of interest).

We’re going to fit a line

**y = β0 + β1x**

to our data. Here, x is called the independent variable or predictor variable, and y is called the dependent variable or response variable. Before we talk about how to do the fit, let’s take a closer look at the important quantities from the fit:

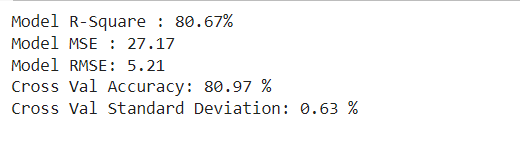
• β1 is the slope of the line: this is one of the most important quantities in any linear regression analysis

• β0 is the intercept of the line.

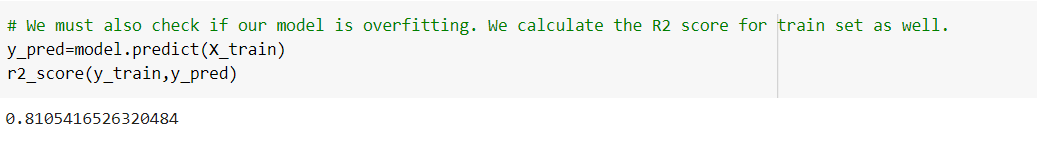


Fig: Linear Regression

Evaluation metrics of LR model



For train set:

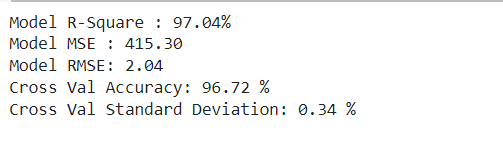


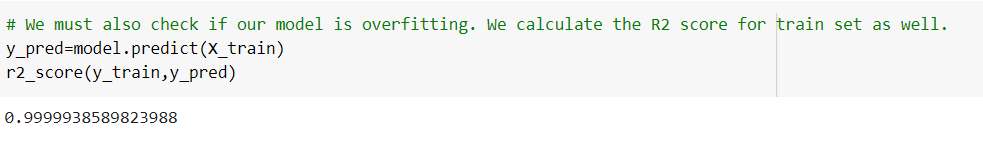
**3.2.Decision Tree**

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. A tree can be “learned” by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called*recursive* partitioning. Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, and then moving down the tree branch corresponding to the value of the attribute as shown in the above figure. This process is then repeated for the subtree rooted at the new node.



Evaluation metrics for Decision tree:

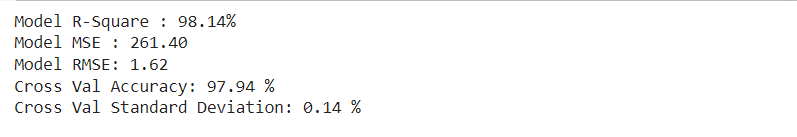


Accuracy of train set

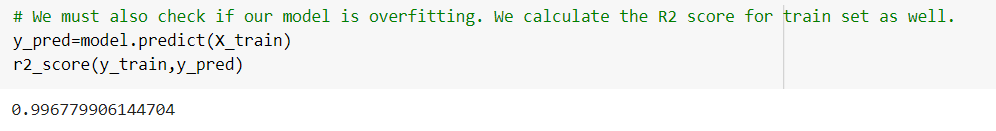
3.3.Random Forest:

Random Forest is a bagging type of Decision Tree Algorithm that creates a number of decision trees from a randomly selected subset of the training set, collects the labels from these subsets and then averages the final prediction depending on the greatest number of times a label has been predicted out of all

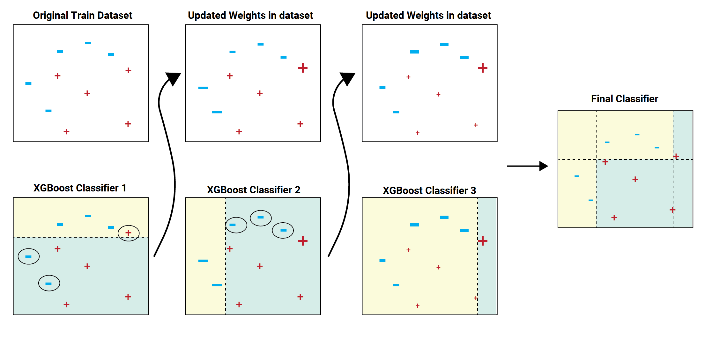
Evaluation metrics of Random Forest:



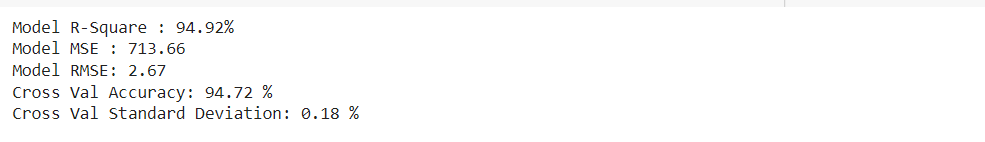
Accuracy of train set:



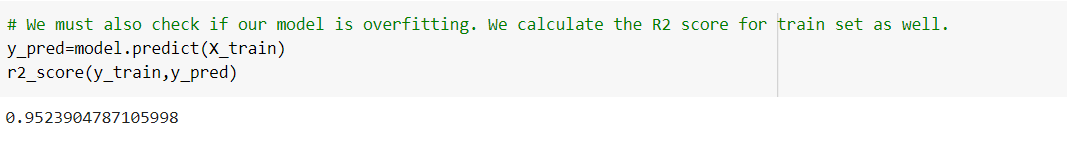
**3.4.Xg Boost**

**XGBoost**is an optimized distributed gradient boosting library designed to be highly **efficient, flexible**and**portable**. It implements machine learning algorithms under the [Gradient Boosting](https://en.wikipedia.org/wiki/Gradient_boosting) framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way. The same code runs on major distributed environment (Hadoop, SGE, MPI) and can solve problems beyond billions of examples.

Evaluation metrics for XG Boost:



Accuracy of train set:

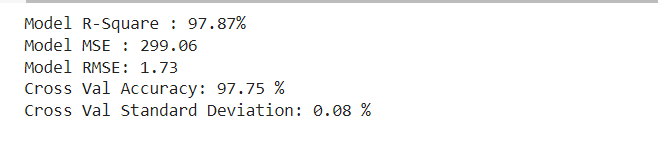


**3.5.Bagging**

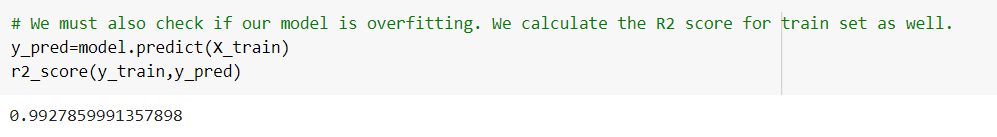
Bagging, also known as bootstrap aggregation, is the ensemble learning method that is commonly used to reduce variance within a noisy dataset. In bagging, a random sample of data in a training set is selected with replacement—meaning that the individual data points can be chosen more than once. After several data samples are generated, these weak models are then trained independently, and depending on the type of task—regression or classification, for example—the average or majority of those predictions yield a more accurate estimate.

As a note, the random forest algorithm is considered an extension of the bagging method, using both bagging and feature randomness to create an uncorrelated forest of decision trees.

**Evaluation metrics for Bagging**

****

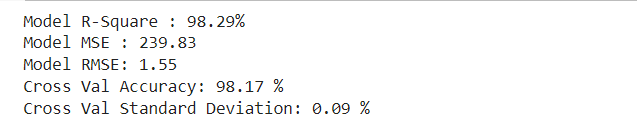
**Accuracy of train set**

****

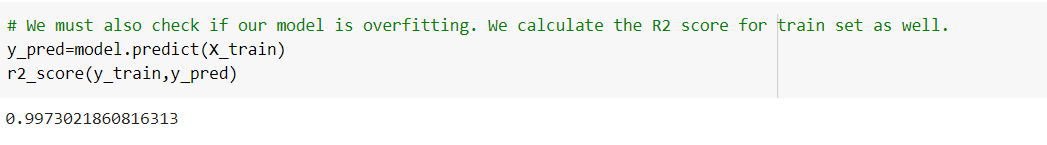
**3.6.Stacking**

Stacking (sometimes called Stacked Generalization) is a different paradigm. The point of stacking is to explore a space of different models for the same problem. The idea is that you can attack a learning problem with different types of models which are capable to learn some part of the problem, but not the whole space of the problem. So, you can build multiple different learners and you use them to build an intermediate prediction, one prediction for each learned model. Then you add a new model which learns from the intermediate predictions the same target.  
This final model is said to be stacked on the top of the others, hence the name. Thus, you might improve your overall performance, and often you end up with a model which is better than any individual intermediate model.

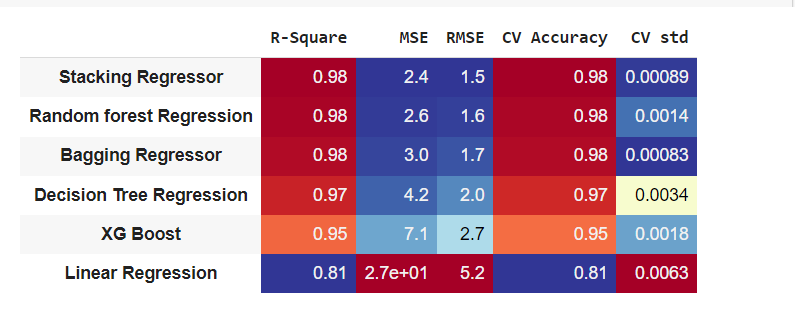
Evaluation metrics



Accuracy of train set



4.Model comparison chart



**5.Conclusion:**

Hence, from the model comparison chart we see that R2 score of Stacking Regressor, Random Forest and Bagging regressor models are the highest and most closer to 1 (=0.98) .R-Squared or the coefficient of determination is the most important evaluation metric. It is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model. Of all the models implemented, we see the R2 score of three models are the same, so we see the next evaluation parameter, MSE and RMSE.

MSE and RMSE don't provide insights individually, i.e just by looking at the number we can't conclude anything about the model. Rather its a comparison metric used to compare the performance of different models for the same data set. From the model comparison chart, we see that MSE and RMSE is minimum for Stacking Regressor model. Therefore, its the best model for our business problem. Also, RMPSE is 12% for the model which is quite acceptable.

For the given problem, I've also calculated the cross validation accuracy and standard deviation. This involves simply repeating the cross-validation procedure multiple times and reporting the mean result across all folds from all runs. This mean result is expected to be a more accurate estimate of the true unknown underlying mean performance of the model on the dataset, as calculated using the standard error. We find the CV accuracy of Stacking Regressor to be 0.98 or 98% which is quite a good value.

Therefore, Stacking Regressor is the best model for the given business problem