Seoul Bike Sharing Demand Prediction

Rohan A.G
Data science trainee,
AlmaBetter, Bangalore

Abstract:

As more number of rented bikes are being used in the cities nowadays, it becomes important for the company to predict the number of required rental bikes required across a day so that no demand supply gap would be generated for rental bikes. This project aims at providing necessary solution to predict the rental bikes demand using machine learning algorithms so that all the stakeholders of the business can be satisfied.

1.Problem Statement:

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.

Variables

- Date: year-month-day
- Rented Bike count - Count of bikes rented at each hour • Hour - Hour of the day
- Temperature-Temperature in Celsius
- Humidity - %
- Wind speed - 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 - NoFunc(Non Functional Hours), Fun(Functional hours)

2. Introduction

The present scenario is about how good is the customer service in any industry as the number of options at the customer's disposal are unlimited. So, it becomes extremely important to make sure that the customers will not be made to wait for the rental bikes. It would also not be practical to keep lot of bikes even when the demand is low. Hence, with the help of machine learning, this project aims at predicting the rental bike demand so that no problems arise.

3. Steps involved:

- **Exploratory Data Analysis:**

  The first step of our project is performing the EDA process on the dataset so that we can get the idea about the dataset i.e., the number of variables, the data type of the variables, visualize the dataset for better understanding and decide the suitable methods and algorithms that might produce desired Outcomes.

- **Data Preprocessing:**

  The dataset was imported and read the csv file, null values were taken care of at first, Distribution was created using histogram for numerical features, later outliers were found out with the help of box plot and treated by log transformation.
  Label encoding of categorical values was done. Correlation heat map was generated to understand the correlation among the variables and removed the features which has high correlation.

- **Building Machine Learning Model:**

After the data preprocessing is done then the data will be ready to be fit into machine learning models. We have used algorithms such as Linear Regression, Polynomial Regression, Random Forest Regression for the prediction and used Regression Evaluation Metrics to find out the best fit models such as mean squared error, mean absolute error, root mean squared error and r2 score.

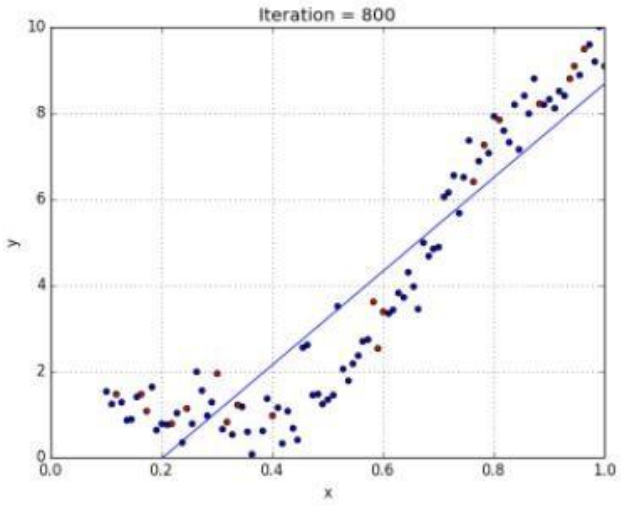
- **Summary:**

At last the summary of the project is described to have brief look over the project.

4.Algorithms:

1.Linear Regression Model:

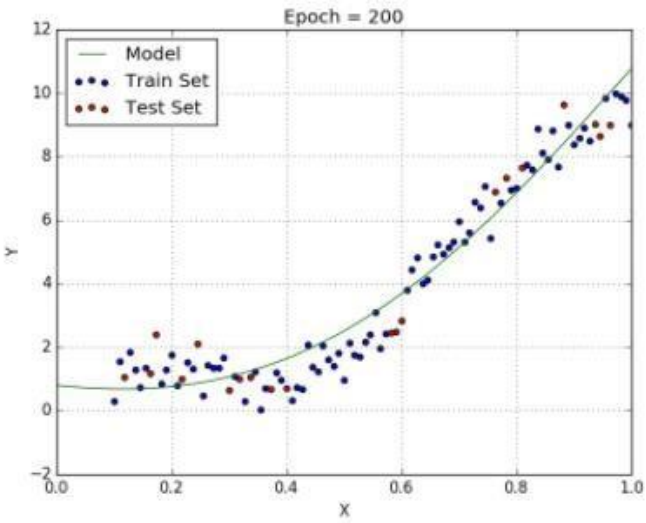
- **Linear Regression** is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables.
- It is mostly used for finding out the relationship between variables and forecasting.
- Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.
- Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.



Fig(a): Example of Linear Regression

2.Polynomial Regression Model:

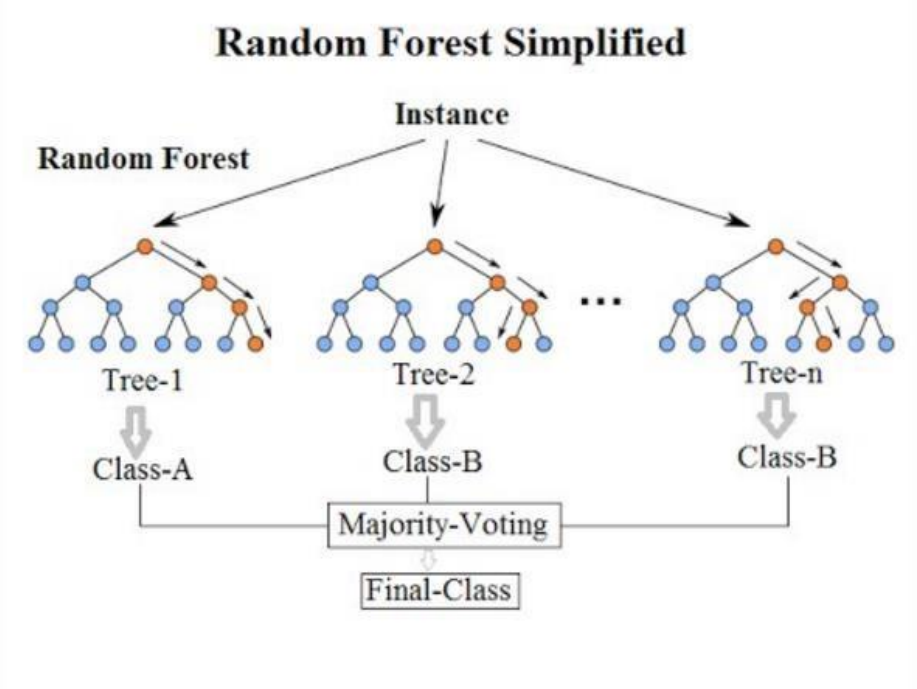
- Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial.
- It is also called the special case of Multiple Linear Regression in ML. Because we add some polynomial terms to the Multiple Linear regression equation to convert it into Polynomial Regression,
- It is a linear model with some modification in order to increase the accuracy.
- The dataset used in Polynomial regression for training is of non-linear nature.
- It makes use of a linear regression model to fit the complicated and nonlinear functions and datasets.
- Hence, "In Polynomial regression, the original features are converted into Polynomial features of required degree (2,3,..,n) and then modelled using a linear model."



Fig(b): Example of polynomial regression.

3. Random Forest Regressor:

- Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML.
- It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.
- Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.
- Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.
- The greater number of trees in the random forest leads to higher accuracy and prevents the problem of overfitting.



5.Model performance:

Mean Squared Error:

Mean Squared Error, or MSE for short, is a popular error metric for regression problems. It is also an important loss function for algorithms fit or optimized using the least squares framing of a regression problem. Here "least squares" refers to minimizing the mean squared error between predictions and expected values.

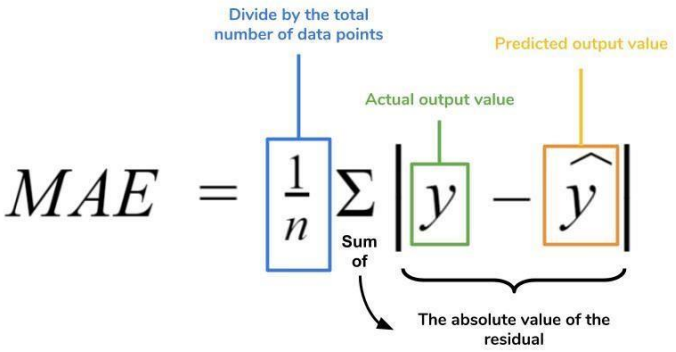
The MSE is calculated as the mean or average of the squared differences between predicted and expected target values in a dataset.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

MSE = mean squared error
n = number of data points
Y_i = observed values
$\hat{Y}_i$ = predicted values

Mean Absolute Error:

Mean Absolute Error is a model evaluation metric used with regression models. The mean absolute error of a model with respect to a test set is the mean of the absolute values of the individual prediction errors on over all instances in the test set. Each prediction error is the difference between the true value and the predicted value for the instance.



$$MAE = \frac{1}{n}\sum \left| y - \hat{y} \right|$$

R2 Score:

R2 score is a metric that tells the performance of your model, not the loss in an absolute sense that how many wells did your model perform.
In contrast, MAE and MSE depend on the context as we have seen whereas the R2 score is independent of context. So, with help of R squared we have a baseline model to compare a model which none of the other metrics provides.

The same we have in classification problems which we call a threshold which is fixed at 0.5. So basically, R2 squared calculates how must regression line is better than a mean line. But how to interpret R2 score.
The normal case when the R2 score is between zero and one like 0.8 which means your model is capable to explain 80 per cent of the variance of data.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

6.Hyper parameter tuning:

Grid Search CV:

Grid Search combines a selection of hyper parameters established by the scientist and runs through all of them to evaluate the model's performance. Its advantage is that it is a simple technique that will go through all the programmed combinations. The biggest disadvantage is that it traverses a specific region of the parameter space and cannot understand which movement or which region of the space is important to optimize the model.

7.Conclusion:

1. We have used three regression models in our analysis to predict the rented bike count that are Linear Regression, Polynomial Regression and Random Forest Regressor.
2. We have used Hyperparameter tuning on Random Forest regressor using GridSearchCV to find the best parameters.
3. Random Forest Regressor is the best model among the three with least errors and highest R2 score of 98.21% for training set and 86.98% for test dataset.
4. The peak time is 6pm where highest number of bookings are rented.
5. Approximately between 10 degree Celsius to 20 degree Celsius the rent booking is peak.
6. Number of bookings are very much higher in summer season.
7. 'Rainfall' and 'Snowfall' has a huge impact on number of bikes rented.
8. Bike renting is high on Functional days (ie No holiday).