SUPERVISED MACHINE LEARNING
(REGRESSION)

PROJECT ON

TRANSPORT DEMAND PREDICTION

Name of the participants:

CHARAN

Data science trainees,
Alma Better, Bangalore

Table of Contents

Abstract:

Nairobi is one of the most heavily congested cities in Africa. Each day thousands of Kenyans make the trip into Nairobi from towns such as Kisii, Keroka, and beyond for work, business, or to visit friends and family. The journey can be long, and the final approach into the city can impact the length of the trip significantly depending on traffic. We need to analyze how do traffic patterns influence people's decisions to come into the city by bus and which bus to take. And knowing the traffic patterns in Nairobi help anticipate the demand for particular routes at particular times.

It is the dataset of tickets purchased from Mobi ticket for the 14 routes from "up country" into Nairobi between 17 October 2017 and 20 April 2018. This dataset includes the variables: ride id, seat number, payment method, payment receipt, travel date, travel time, travel from, travel to, car type, max capacity. This resulting model can be used by Mobi ticket and bus operators to anticipate customer demand for certain rides, to manage resources and vehicles more efficiently, to offer promotions and sell other services more effectively or even improve customer service by being able to send alerts and other useful information to customers.

1. Problem Statement

This challenge asks you to build a model that predicts the number of seats that Mobi ticket can expect to sell for each ride, i.e., for a specific route on a specific date and time. There are 14 routes in this dataset. All of the route's end in Nairobi and originate in towns to the North-West of Nairobi towards Lake Victoria.

The towns from which these routes originate are:

- Awendo
- Homa Bay
- Kehancha
- Kendu Bay
- Keroka
- Keumbu
- Kijauri
- Kisii
- Mbita
- Migori
- Ndhiwa

- Nyachenge
- Oyugis
- Rodi
- Rongo
- Sirare
- Sori

The routes from these 14 origins to the first stop in the outskirts of Nairobi takes approximately 8 to 9 hours from time of departure. From the first stop in the outskirts of Nairobi into the main bus terminal, where most passengers get off, in Central Business District, takes another 2 to 3 hours depending on traffic.

The three stops that all these routes make in Nairobi (in order) are:

1. **Kawangware: the first stop in the outskirts of Nairobi**
2. **Westlands**
3. **Afya Centre: the main bus terminal where most passengers disembark**

All of these points are mapped here.

Passengers of these bus (or shuttle) rides are affected by Nairobi traffic not only during their ride into the city, but from there they must continue their journey to their final destination in Nairobi wherever that may be. Traffic can act as a deterrent for those who have the option to avoid buses that arrive in Nairobi during peak traffic hours. On the other hand, traffic may be an indication for people's movement patterns, reflecting business hours, cultural events, political events, and holidays.

To analysis and predict the above given database, the current project is developed.

2. Introduction

Nairobi is one of the most heavily congested cities in Africa. Each day thousands of Kenyans make the trip into Nairobi from towns such as Kisii, Keroka, and beyond for work, business, or to visit friends and family. The journey can be long, and the final approach into the city can impact the length of the trip significantly depending on traffic. How do traffic patterns influence people's decisions to come into the city by bus and which bus to take? Does knowing the traffic patterns in Nairobi help anticipate the demand for particular routes at particular times?
The aim is to create a predictive model using traffic data provided from **Uber Movement** and historic bus ticket sales data from **Mobi ticket** to predict the number of tickets that will be sold for buses into Nairobi from cities in "up country" Kenya.

Nairobi Transport Data.csv (zipped) is the dataset of tickets purchased from Mobi ticket for the 14 routes from "up country" into Nairobi between 17 October 2017 and 20 April 2018. This dataset includes the variables: ride id, seat number, payment method, payment receipt, travel date, travel time, travel from, travel to, car type, max capacity.

Uber Movement traffic data can be accessed here. Data is available for Nairobi through June 2018. Uber Movement provided historic hourly travel time between any two points in Nairobi. Any tables that are extracted from the Uber Movement platform can be used in your model.

3. Understanding the Problem Statement

The data used to train the model will be historic hourly traffic patterns in Nairobi and historic ticket purchasing data for 14 bus routes into Nairobi from October 2017 to April 2018, and includes the place or origin, the scheduled time of departure, the channel used for the purchase, the type of vehicle, the capacity of the vehicle, and the assigned seat number.
This resulting model can be used by Mobi ticket and bus operators to anticipate customer demand for certain rides, to manage resources and vehicles more efficiently, to offer promotions and sell other services more effectively, such as micro-insurance, or even improve customer service by being able to send alerts and other useful information to customers.

Some of the key attributes consisting those variables which are used for this project are listed below:

Attributes and their Description

- Ride id: unique ID of a vehicle on a specific route on a specific day and time.
- Seat number: seat assigned to ticket
- Payment method: method used by customer to purchase ticket from Mobi ticket (cash or Mpesa)
- Payment receipt: unique id number for ticket purchased from Mobi ticket
- Travel date: date of ride departure. (MM/DD/YYYY)
- Travel time: scheduled departure time of ride. Rides generally depart on time.
- Travel from: town from which ride originated
- Travel to: destination of ride. All rides are to Nairobi.
- Car type: vehicle type (shuttle or bus)
- Max capacity: number of seats on the vehicle

4. Data Pre-processing

Before jumping to the exploration stage, we need to perform basic data pre-processing steps like null value imputation and removal of unwanted data.

4.1. Libraries Used

Pandas, NumPy, Matplotlib, Seaborn, Sklearn

```
import NumPy as np
import pandas as pd
```

```
# For visualizations
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.feature_selection import VarianceThreshold
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import RepeatedKFold
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.metrics import mean_absolute_error, mean_squared_error,r2_score
```

The dataset contains 51645 rows and 10 columns. But we only need information about important columns. So we are dropping other unwanted columns and reducing the dataset to 7 columns.

4.2. Handling Null Values

There are a few null values in the dataset. The major part of the dataset is null values which triggers many stumbling blockages in processing the necessary columns, therefore it is required to identify the columns containing lump sum null values and get dropped. So, let's drop these null values columns which contain more than 80 percent of null values.

5. PREPARING DATA FOR EXPLORATORY DATA ANALYSIS(EDA)

5.1. data loading

We have loaded the data into our drive and then finally loaded it to google colab notebook using pandas library.

5.2. Renaming the columns

This method is useful because it lets you modify a column heading without having to create a new column.

5.3. Removing Constant Features

We need to first remove the constant features. In this dataset the destination i.e., "travel to" is constant so removing that column.

6. EXPLORATORY DATA ANALYSIS:

6.1 Shape:

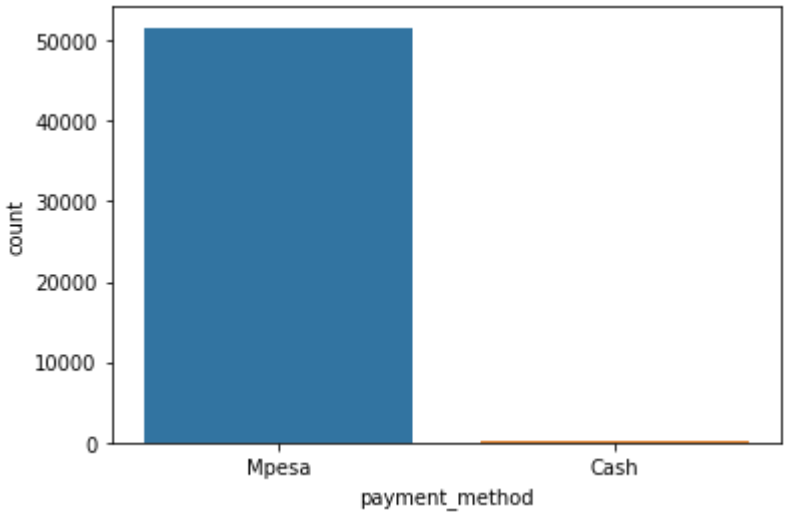
The shape attribute tells us a number of observations and variables we have in the data set. It is used to check the dimension of data. The given data set has more than 50000 observations and 10 variables in the data set.

6.2. Describe:

It computes a summary of statistics pertaining to the Data Frame columns. This function gives the mean, std and IQR values. And, function excludes the character columns and gives a summary about numeric columns.

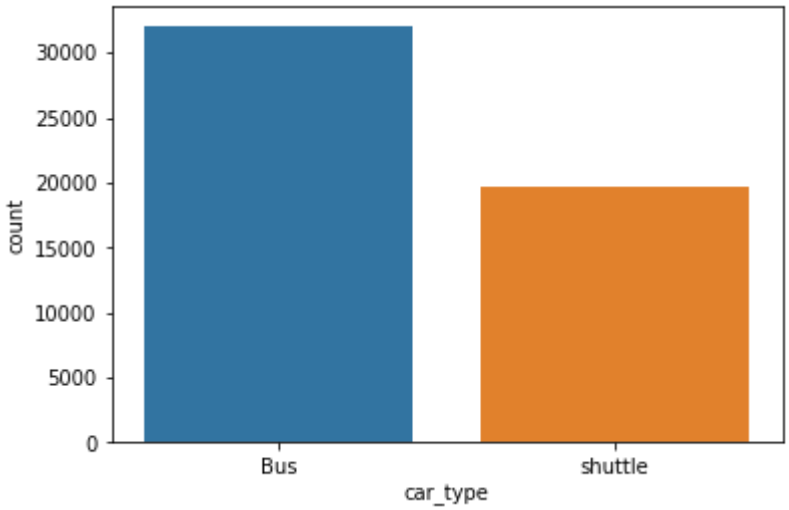
6.3. Analysis of different variables

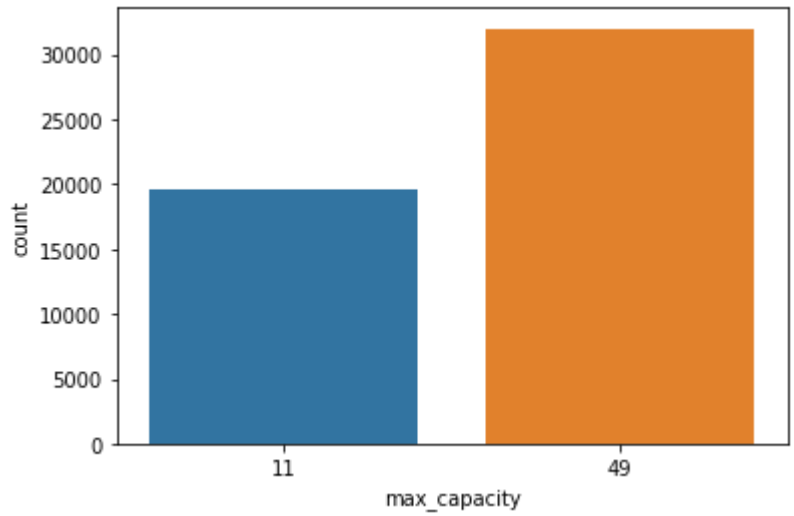
6.3.1. Analyzing the payment method

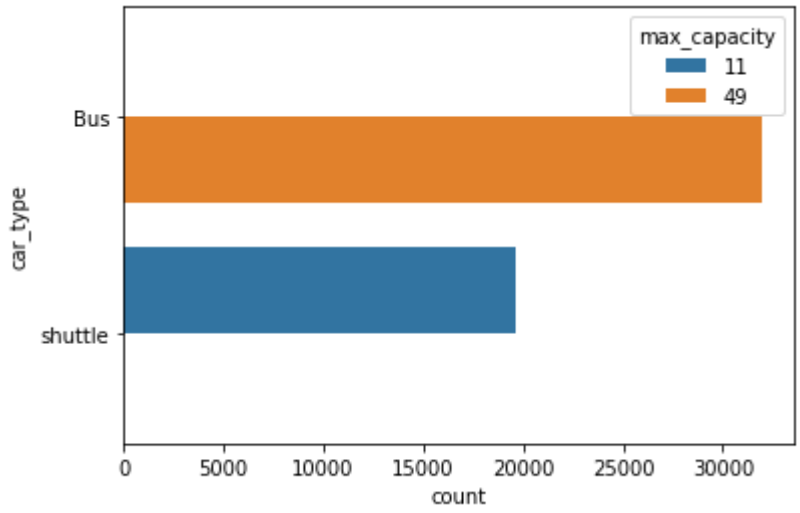


Travelers have used 2 types of payment method. They are Mpesa and cash. And the most of the people have used Mpesa to pay for their ticket.

6.3.2. Analyzing the vehicle type used by the customer and number of seats on the vehicle sell







There are totally 61 unique seats in this dataset.

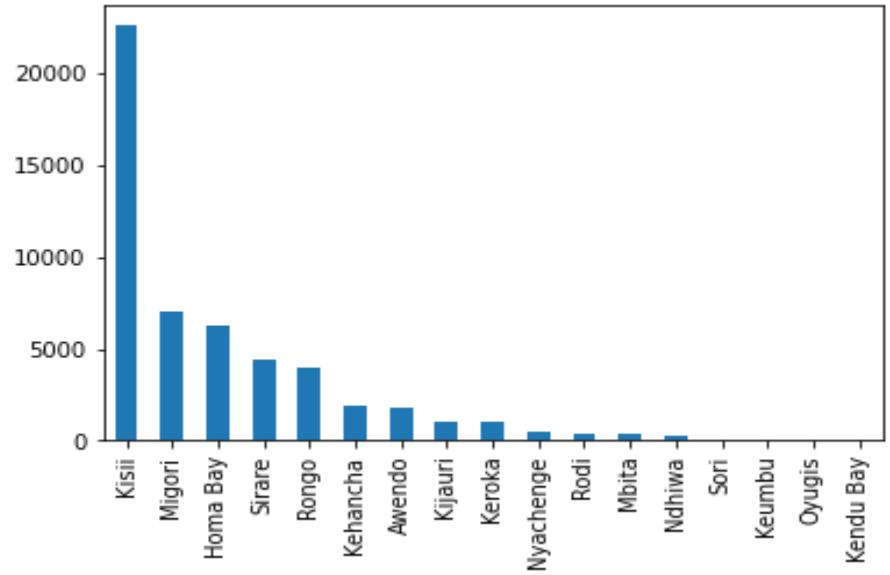
The record of 149 unique days is present in this dataset out of 2 years.

There are 2 different types of cars used (shuttle and bus) and most of them are bus.

There are two type of payment methods people have used to buy the tickets.

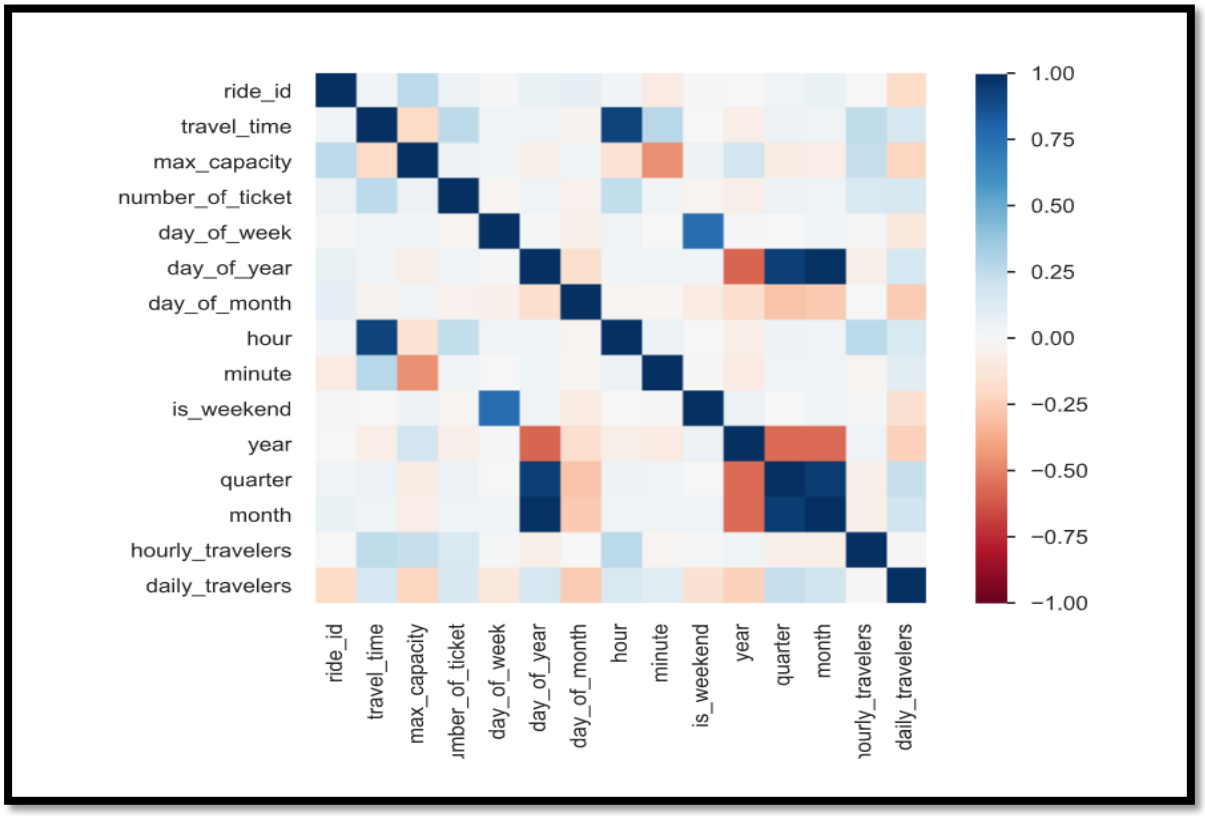
There are two type of cars Bus and shuttle and the maximum capacity of the bus is 49 while shuttle can contain 11 travelers.

6.3.3. Analyzing the travel from town from which ride originated

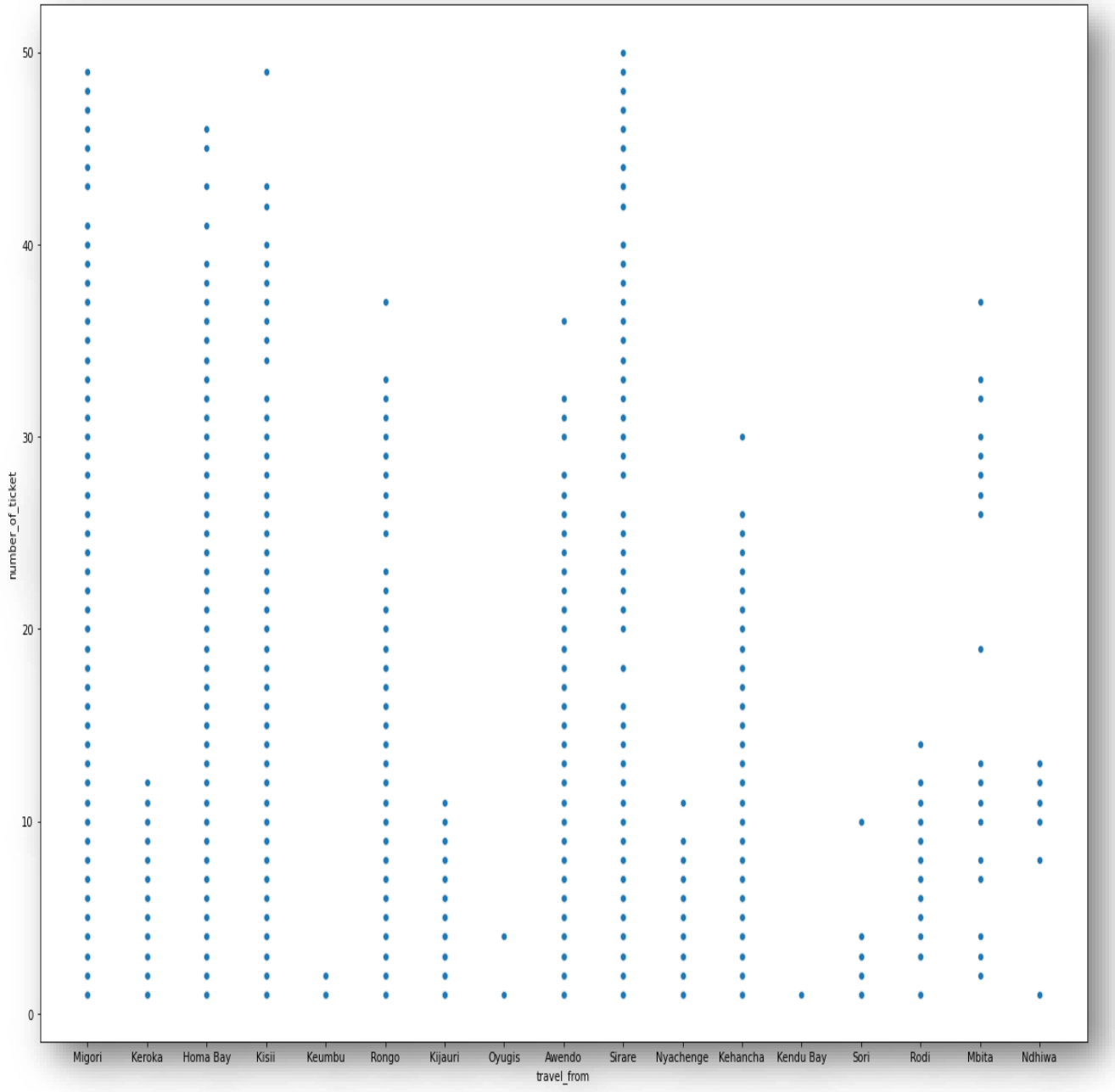


Most customers travel from Kisii town to Nairobi.

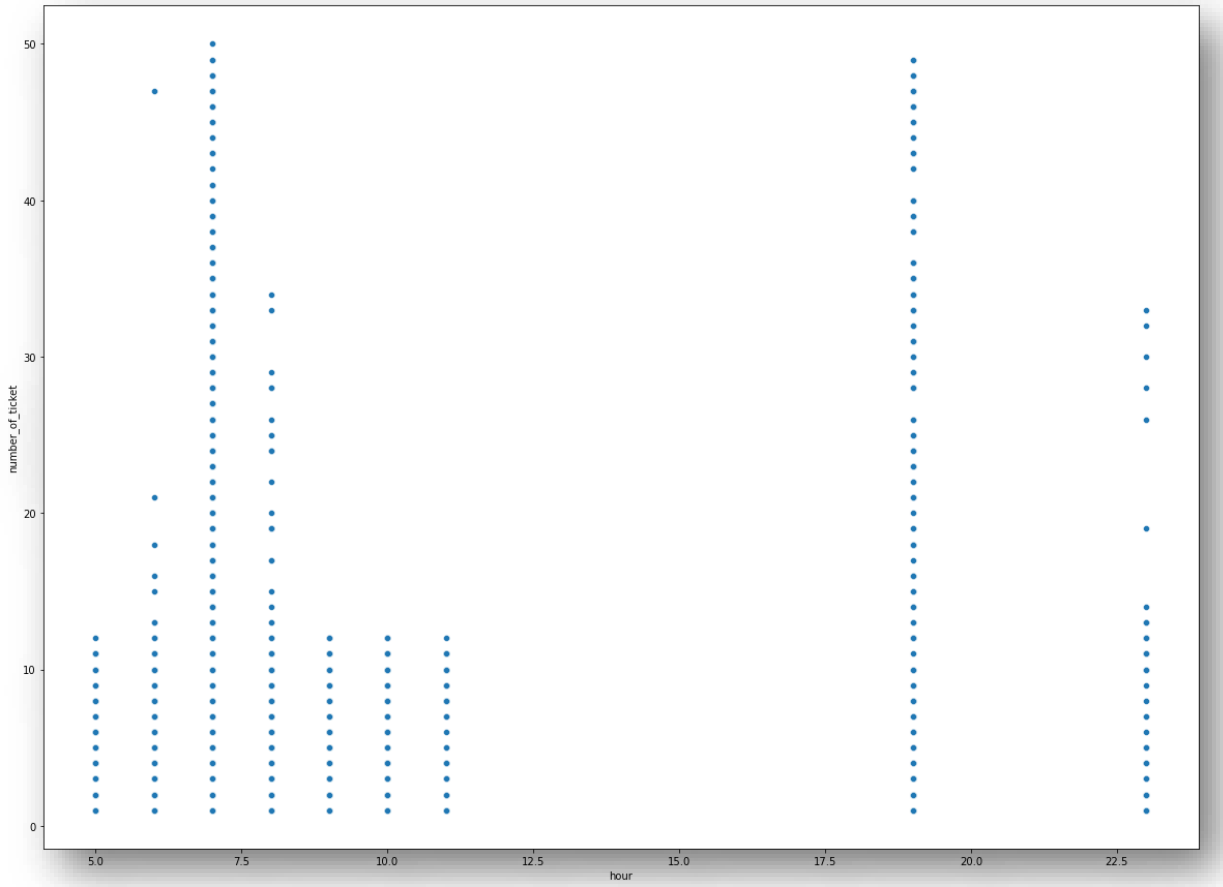
Correlation



Scatter plot of number of tickets sold to customers from different cities



7. Feature Engineering



We can see that most of the tickets were sold at 7 AM and 8 PM. And that seems true because in the morning most of the people go to the work and office.

From the above we can say that there is not ride between 12pm to 5.30Pm

8. Training the models

8.1. Fitting different models

For modeling we tried various classification algorithms like:

➤ Linear Regression

➤Regularized linear regression (Ridge and Lasso)

➤GBM

➤Random Forest Regressor

➤XG boost regressor

8.1. Algorithms:

Linear Regression:

It is one of the most-used regression algorithms in Machine Learning. A significant variable from the data set is chosen to predict the output variables (future values). Linear regression algorithm is used if the labels are continuous, like the number of flights daily from an airport, etc. The representation of linear regression is y = b*x + c.

Logistic Regression:

Logistic Regression is actually a classification algorithm that was given the name regression due to the fact that the mathematical formulation is very similar to linear regression.

The function used in Logistic Regression is sigmoid function or the logistic function given by:

f(x)= 1/1+e ^(-x)

Regularized linear regression (Ridge and Lasso):

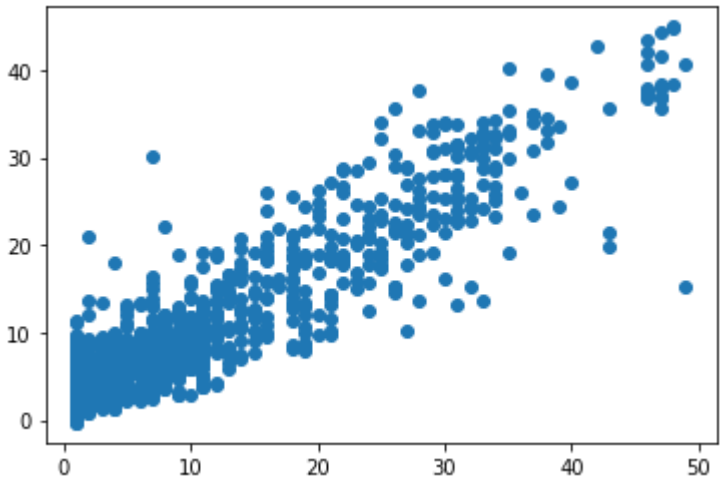
GBM:

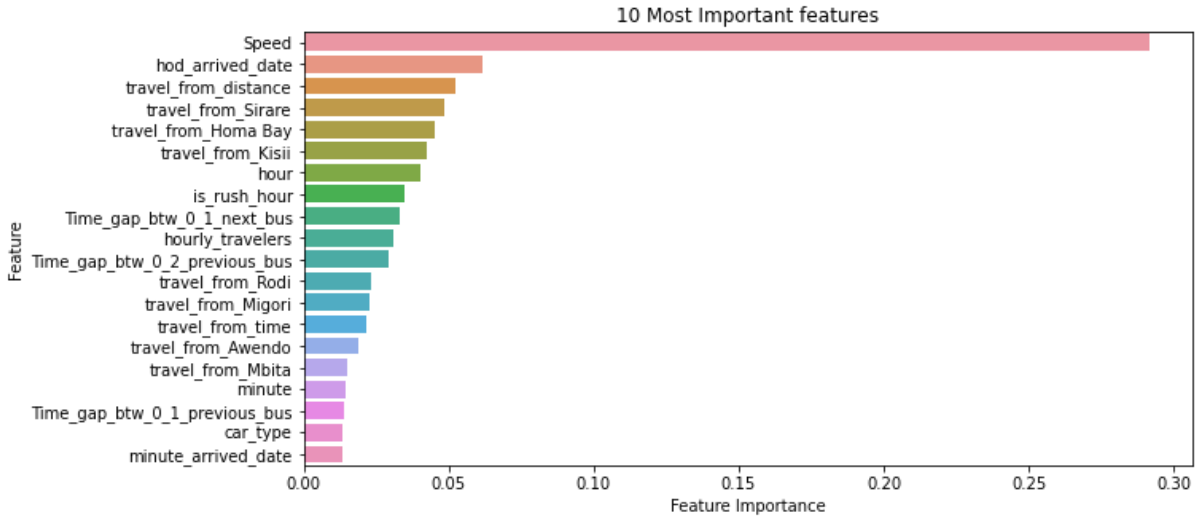
Random Forest Regressor:

Random Forests are an ensemble(combination) of decision trees. It is a Supervised Learning algorithm used for classification and regression. The input data is passed through multiple decision trees. It executes by constructing a different number of decision trees at training time and outputting the class that is the mode of the classes (for classification) or mean prediction (for regression) of the individual trees.

XG boost Regressor:

Scatter plot of test and predicted values





9. Evaluating the model

After the model is built, if we see that the difference in the values of the predicted and actual data is not much, it is considered to be a good model and can be used to make future predictions.

Few metric tools we can use to calculate error in the model
1.MSE (Mean Squared Error)
2.RMSE (Root Mean Squared Error)
3.MAE (Mean Absolute Error)
4.MAPE (Mean Absolute Percentage Error)
5.R_2(R – Squared)
6.Adjusted R_2

1.Mean Squared Error (MSE)
MSE or Mean Squared Error is one of the most preferred metrics for regression tasks. It is simply the average of the squared difference between the target value and the value predicted by the regression model.

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

2. Root Mean Squared Error (RMSE)
RMSE is the most widely used metric for regression tasks and is the square root of the averaged squared difference between the target value and the value predicted by the model. It is preferred more in some cases because the errors are first squared before averaging which poses a High penalty on large errors.

$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{N}}$$

3. MAE (Mean Absolute Error)

MAE is the absolute difference between the target value and the value predicted by the model. The MAE is more robust to outliers and does not penalize the errors as extremely as MSE. MAE is a linear score which means all the individual differences are weighted equally. It is not suitable for applications where you want to pay more attention to the outliers.

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}.$$

were

Yi is the prediction and xi the true value

4. MAPE (Mean Absolute Percentage Error)

The mean absolute percentage error (MAPE), also known as mean absolute percentage deviation (MAPD), is a measure of prediction accuracy of a forecasting method in statistics, for example in trend estimation, also used as a loss function for regression problems in machine learning.

$$M = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

where n=number of observations, At=actual value and Ft=forecast value

4. R2(R – Squared)

Coefficient of Determination or R^2 is another metric used for evaluating the performance of a regression model. The metric helps us to compare our current model with a constant baseline and tells us how much our model is better. The constant baseline is chosen by taking the mean of the data and drawing a line at the mean. R^2 is a scale-free score that implies it doesn't matter whether the values are too large or too small, the R^2 will always be less than or equal to 1.

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

5. Adjusted R-square

Adjusted R^2 depicts the same meaning as R^2 but is an improvement of it. R^2 suffers from the problem that the scores improve on increasing terms even though the model is not improving which may misguide the researcher.

$$R^2_{adj} = 1 - \left[\frac{(1 - R^2)(n-1)}{n-k-1} \right]$$

10. Conclusion

Starting with loading the data so far, we have done EDA, null values treatment, encoding of categorical columns, feature selection and then model building.

We used different types of regression algorithms to train our model like Linear Regression, Regularized linear regression (Ridge and Lasso), GBM, Random Forest Regressor, XG boost regressor. And also, we tuned the parameters of Random Forest regressor and XG boost

regressor and also found the important features for training the model. Out of them XG Boost with tuned hyperparameters gave the best result.
So, the accuracy of our best model is 79% which can be said to be good for this large dataset. This performance could be due to various reasons like no proper pattern of data, too much data, and not enough relevant features.