Project:-

Flight Price Prediction

Submitted by:- Mohit Meel

Acknowledgement

The project allocated by the company contains the steps or process how to complete the project as it contains the file of sample documentation and problem statement which shows the process to do the project.

The data sources used is the websites of flight ticket booking and the tool used to scrape the data is web scraping selenium.

The references and professionals that are helping with the project is our SME who is always ready to help us to complete the project whenever we need.

Introduction

- The problem statement is about to know about the fluctuations in the flight prices. As we all know the flight prices are usually high when we book them at the time of travel or before 2 days to travel and at the time when there are festivals coming and also at the time when the holidays come. The prices of flights are less when we book before 1 month to travel. All such scneario which fluctuates prices of flight is the main part of our study of the project.
- The data collected for the project is purely primary as it is scraped from the most famous websites which Paytm/flights. The data is showing the departure time, departure station, arrival time, arrival station, any stoppages, travel time, airline names and prices of different airlines, the class preference is the economy and also the date to travel is of this month such as 23,24,25,26 and then return flights of date is 23,24,25 of this month.
- The aim to work with the project is to ascertain the best time to book the tickets so the prices is cheapest and if it is emergency then which websites should be used to book the tickets in cheaper rates

Analytical Problem Framing

- The data source used to collect the data is primary source as it scraped from the website Paytm/flights. The technique used is web scraping selenium. The format of data is csv. There are 1502 rows and 9 columns. The rest data description is shown in the introduction part.
- For the data cleaning, first the data is analzsed by the running code:- .head, .tail, .shape, . dtypes, .columns, .is null, .info, .describe, . corr . After that the assumptions where that the datatype of the project is object therefore it is converted into numerical. The outliers and skewness present in the dataset is not removed as the datatype is object. Standard scaler is used for scaling the dataset.
- The relationship between the input and output variable is somewhere less relation. As the input variable affect the output variable in terms of ascertaining the output variable which is prices, the distance between two stations, time travel to complete the journey, the airline name and the class too.

Model Development and Evaluation

The approaches I used to solve the problem is the regression models as the prediction variable(price) which is continuous.

The models used for training and testing is the linear regression, lasso, ridge, support vector regressor, random forest regressor.

The models listed above is run and the results come up are surprising as the only model which is random forest regressor scored 97% and rest not even scored more than 15%.

The visualization part done on the project is the heatmap for knowing null values and correlation between the input and output variable, the boxplot for knowing the outliers present, the kdeplot, histplot for checking the skewness in the dataset, the barplot for the lasso and ridge and the scatterplot showing the best fit line of random forest regressor.

Conclusion

The key findings and observations from the whole problem is that the prices of ticket are high when we book at the time of travel such within the 2-3 days before travel thus the prices of different airline flights are somewhere same but the tickets which are done of flights running at early morning and at late night are less as compared to the flights running at morning, afternoon, and evening. There are so many websites which book tickets but according to me the best one that stands on the flights ticket booking is the skyscanner. The tickets also come high when these are book at the time of festivals, at the summer and winter vacations. The best time to purchase is before one month of travel.

There were no such problems while cleaning the data and building and evaluation of model.
The best model that come is the random forest regressor with the score of 97%. The prices of flights after doing the prediction by the machine is showing somewhere slight difference within the prices of predicted and actiual prices.

The limitations are that the class of ticket is not there in the study of data as all the data is of economy class and the most important is the services offered by different airlines to customers after purchasing tickets at the boarding and in the flights. These points should be taken while predicting the prices for better results.