FLIP ROBO

Flight Price Prediction Project

Submitted by
Akhil Vangeti

Acknowledgement

I sincerely thank Ms. Sapna Verma for providing me valuable information and guidance for completion of this project. His assistance has helped me a lot during the course of entire project.

I also thank datatrained institute for helping me gain hands on experience about machine learning.

Thanks to yatra.com for allowing their data to be used in this study.

Introduction

Air travel is one of the fastest way to travel but it comes with a tradeoff with cost. The price of the airline ticket changes over time very frequently to maximize revenue.

Therefore, there comes a need of analyzing the factors which affect the change in prices and find a right time to buy the tickets while saving few bucks. In this project we are building a model which can predict the prices of airlines.

Analytics and Modelling

The steps followed in this study are as follows:

1) Data collection
2) Data Study and Analysis
3) Data preprocessing
4) Modelling

Data Collection:

The data for this study is collected from the online used car trader CARS24.

The data which is collected from this website are as follows:

a) Source
b) Destination
c) Departure date
d) Departure day
e) Airline
f) Duration
g) Stops
h) Departure time
i) Arrival time
j) Price

The Selenium library is used for the scraping process. After the scraping process is complete, we had around 11000 data points.

The data then is saved into a csv file. This data is used in the next sections.

The ipynb file for the above is named as flights-price-scraping and can be found using this link: https://github.com/vangetiakhil/FRT3/tree/main/flight_price_prediction

Data Study and Analysis:

The data used in this study consists of continuous, categorical data.

The format of data is integer and object type. There are a total of 10 variables in this dataset and are also mentioned in Data collection section. These variables are:

a) Airport_from (source)
b) Airport_to (Destination)
c) Departure_date
d) Departure_day
e) Airline
f) Duration
g) Stops
h) Departure_time
i) Arrival_time
j) Price

The dataset contains 10905 rows.

The first step taken in the analysis part is to check for the missing values and it is found that there are no missing values.

From the exploratory data analysis, the following information has been deducted:

a) No of stops and duration are positively correlated.
b) No of stops, duration shows a positive correlation with price.
c) Indigo has the most flights working followed by Air india.
d) The average prices of Air India are higher than the other airline average prices.
e) Air india is the only airlines where a passenger might be taking more than 2 stops for a journey.

Data preprocessing:

The Price column is our target variable and the data in it is in integer format but with a value far greater than the predictors values. This column values if are bigger the model may not work as it is intended to be. The errors in predictions will be more.

Few columns are categorical variables and needs to be encoded to be fed into the model. This is done by using dummies method.

The continuous variable columns are duration, stops, days for departure and Price needs to be normalized before modelling.

To handle the bias occurring due to large values in price column, a log transformation has been applied on the price column.

Modelling:

The target variable is continuous. Therefore, this is a regression problem.

The different models which have been used for modelling are:

1) Linear regression
2) Decision tree regression
3) Kneighbors regression
4) SVR

The regularization models which have been used are:

1) ElasticNet
2) Ridge
3) Lasso
4) Bayesian Ridge
5) Huber Regressor

The Boosting/Ensemble models which have been used are:

1) Random forest regressor
2) AdaBoost regressor
3) Gradient Boosting regressor
4) XGB regressor

The above mentioned modeling techniques are applied and their r2 scores ranges from 50 - 80%. To check which model performs better, cross validation has been used.

After the cross validation scores have been checked, the following observations are made:

a) Decision Tree Regressor has an r2 score of 0.71.
b) Random forest and XGB boosting techniques are doing good.

To choose the best fit model, 4 metrics have been used. They are:

a) Mean absolute error
b) Mean squared error
c) R2 score
d) Difference between r2 score and cross val score

The best model is selected as XGBRegressor with r2_score 0f of 0.77.

GridsearchCV is used to find the best parameters for the model. A new model has been built using the best parameters found using GridSearchCV. The final model is then saved using Pickle library.

Conclusion

The data has been collected by scraping the website of Yatra. The data has been cleaned, analyzed, encoded and models based on different techniques have been built.

The final model is predicting with an r2 score of more than 75%. This can be further improved by expanding the data collection from many other sources and cities.

Note:- The full details of the project (ipynb file) can be found in this link: https://github.com/vangetiakhil/FRT3/tree/main/flight_price_prediction