FLIGHT PRICE PREDICTION PROJECT

Submitted by:

Snehal Kumawat

INTRODUCTION

- ## Business Problem Framing

Flight ticket prices can be something hard to guess, today we might see a price, check out the price of the same flight tomorrow, and it will be a different story.To solve this problem, we have been provided with prices of flight tickets for various airlines using which we aim to build a model which predicts the prices of the flights using various input features.

- ## Conceptual Background of the Domain Problem

Anyone who has booked a flight ticket knows how unexpectedly the prices vary. Airlines use using sophisticated quasi-academic tactics known as "revenue management" or "yield management". The cheapest available ticket for a given date gets more or less expensive over time.

- ## Review of Literature

It is very difficult for the customer to purchase a flight ticket at the minimum price. For this several techniques are used to obtain the day at which the price of air ticket will be minimum. Most of these techniques are using sophisticated artificial intelligence(AI) research is known as Machine Learning.

- ## Motivation for the Problem Undertaken

So, if we could inform the travellers with the optimal time to buy their flight tickets based on the historic data and also show them various trends in the airline industry we could help them save money on their travels. This would be a practical implementation of a data analysis, statistics and machine learning techniques to solve a daily problem faced by travellers.

- ## Data Sources and their formats

Data Collection is done by Indian companies like yatra.com and stores it as a CSV file. We decided to scrape data using Selenium webdriver and we had to decide the parameters that might be needed for the flight prediction algorithm.

1. Source City
2. Destination City
3. Departure Date
4. Departure Time
5. Arrival Time
6. Total stops
7. Duration
8. Airline name

In Data Cleaning the data was further processed

Data was collected from websites such as yatra.com. And the data was then store in excel and csv file. The dataset was in the form of a csv file, so I used the read csv file function from the pandas module. The picture of the dataset I just have given below, you can observe that it consists of 1880 rows (records) and 10 columns (features).

	Unnamed: 0	airline_name	date_of_journey	source	destination	departure_time	arrival_time	duration	total_stops	price
0	0	SpiceJet	1/2/2022	New Delhi	Mumbai	06:20	08:40	2h 20m	Non Stop	5,953
1	1	Go First	1/2/2022	New Delhi	Mumbai	13:30	15:35	2h 05m	Non Stop	5,954
2	2	Go First	1/2/2022	New Delhi	Mumbai	15:30	17:35	2h 05m	Non Stop	5,954
3	3	Go First	1/2/2022	New Delhi	Mumbai	17:00	19:05	2h 05m	Non Stop	5,954
4	4	Go First	1/2/2022	New Delhi	Mumbai	07:00	09:10	2h 10m	Non Stop	5,954
...	...	...	...	...	...	...	...	...	...	...
1875	1875	Go First	13/2/2022	New Delhi	Mumbai	11:50	18:10	6h 20m	1 Stop	5,954
1876	1876	Go First	13/2/2022	New Delhi	Mumbai	10:50	18:20	7h 30m	1 Stop	5,954
1877	1877	Go First	13/2/2022	New Delhi	Mumbai	13:45	22:20	8h 35m	2 Stop(s)	5,954
1878	1878	Go First	13/2/2022	New Delhi	Mumbai	10:50	19:40	8h 50m	1 Stop	5,954
1879	1879	Go First	13/2/2022	New Delhi	Mumbai	09:15	18:10	8h 55m	1 Stop	5,954

1880 rows × 10 columns

fig 1.Sample data before data pre-processing

1. Data Preparation and Cleaning

Fig. 1 captures the research framework for the flight price prediction problem. It includes five major blocks, namely data collection, data preparation, feature processing, mode training, and model evaluation. These blocks of the diagram are explained in detail in the next subsections.

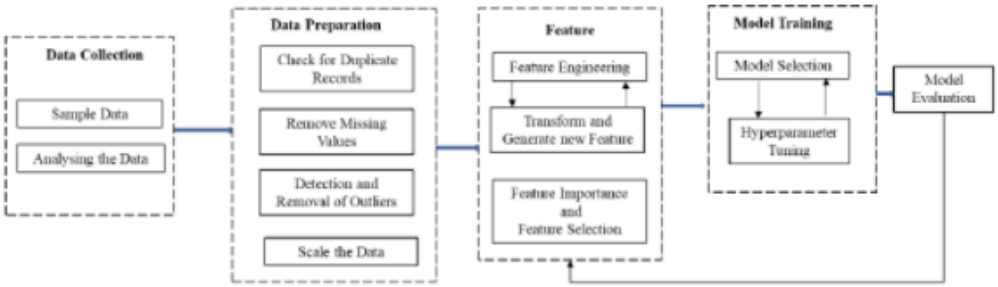

fig 2. Research framework

In this model, we need to feed the available independent variables to the model will predict the possible fare of flight. For designing the model the machine learning method I used is multiple regression model and the tool I used for coding is jupyter notebook.

About Project

Machine Learning Regression model is developed to predict the fare of flight. Data contains 1880 entries each having 10 variables.

Packages used: Pandas, Numpy, Seaborn, Matplotlib, Scikit and Stats models.

In this dataset, it consist of this columns such as Unnamed: 0. So I drop this column.

Data Preprocessing

This step is one of the important steps in supervised machine learning. It includes the following.

In the exploratory data analysis step, we cleaned the dataset by removing the duplicate values and null values. If these values are not removed it would affect the accuracy of the model. We gained further information such as distribution of data.

Next step is data pre-processing where we observed that most of the data was present in string format. Data from each feature is extracted such as day and month is extracted from date of journey in integer format, hours and minutes is extracted from departure time.

Features such as source, destination and airline name needed to be converted into values as they were of categorical type. For this One hot-encoding and label encoding techniques are used to convert categorical values to model identifiable values.

	total_stops	price	Journey_day	Journey_month	Dep_hour	Dep_min	Duration_hours	Duration_mins	airline_name_Air India	airline_name_Go First	airline_name_IndiGo
0	0	5953	2	1	6	20	2	20	0	0	0
1	0	5954	2	1	13	30	2	5	0	1	0
2	0	5954	2	1	15	30	2	5	0	1	0
3	0	5954	2	1	17	0	2	5	0	1	0
4	0	5954	2	1	7	0	2	10	0	1	0

fig3. Sample data after data preprocessing

2. Data Analysis

Data Visualization

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

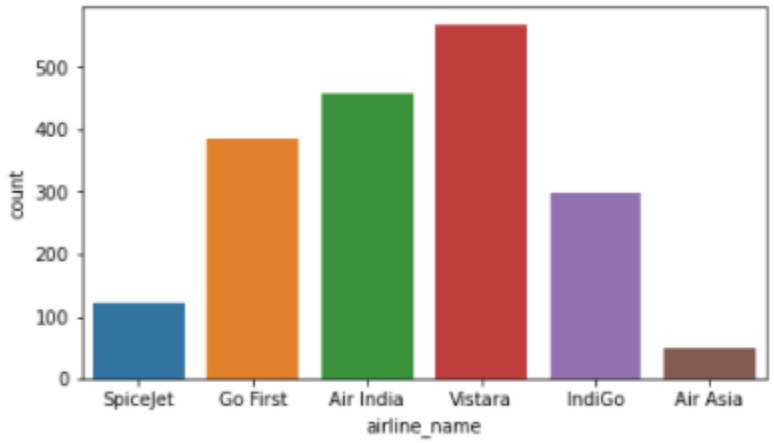

fig.4. Sample plot

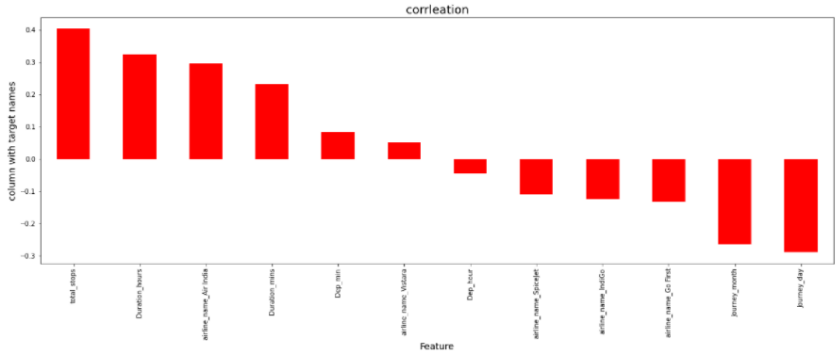

fig.5. Correlation of feature with target column

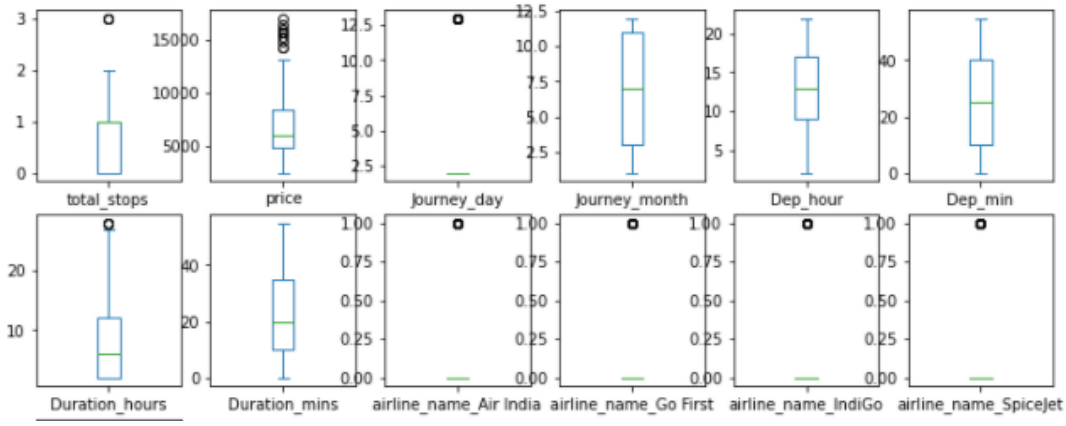

fig.6 figure showing outliers

From the above plot it is observed that data is not normalized and having some outliers.

3. Methodology

We utilized several classic and state-of-the-art methods, including ensemble learning techniques, with a 90% - 10% split for the training and test data. To reduce the time required for training, we used five thousand examples from our dataset. Linear Regression, Random Forest and Gradient Boost were our baseline methods. For most of the model implementations, the open-source Scikit-Learn package was used.

Linear Regression

Linear Regression was chosen as the first model due to its simplicity and comparatively small training time. The features, without any feature mapping, were used directly as the feature vec tors. No regularization was used since the results clearly showed low variance. Using this we get

```
Train Score 47.402954704926096
Test Score 47.39528916238773
```

Ridge Regression

A Ridge regressor is basically a regularized version of Linear Regressor. The regularized ter m has the parameter ‘alpha’ which controls the regularization of the model i.e helps in reducin g the variance of the estimates. Using this we get

```
R2 Score: 47.43312009405909
Cross Val Score: -172.71173369814014
```

Lasso Regression

The “LASSO” stands for Least Absolute Shrinkage and Selection Operator. Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean.

```
Cross validation score is:- -218.69765004402475
R2_score is :- 47.3952808278595
```

Random Forest

Random Forest is an ensemble learning based regression model. It uses a model called decision tree, specifically as the name suggests, multiple decision trees to generate the ensemble model which collectively produces a prediction. The benefit of this model is that the trees are produced in parallel and are relatively uncorrelated, thus producing good results as each tree is not prone to individual errors of other trees. This model was hence chosen to account for the large number of features in the dataset and compare a bagging technique with the following gradient boosting methods.

Using this we get

```
R2 Score 63.85157160852184
```

Gradient Boost

Gradient Boosting is another decision tree based method that is generally described as “a met hod of transforming weak learners into strong learners”. This means that like a typical boostin g method, observations are assigned different weights and based on certain metrics, the weigh ts of difficult to predict observations are increased and then fed into another tree to be trained. In this case the metric is the gradient of the loss function. This model was chosen to account f or non-linear relationships between the features and predicted price, by splitting the data into 100 regions.

Using this we get

```
R2 Score: 73.2769779792599
Cross Val Score: -87.57528592789285
```

XGBoost

Extreme Gradient Boosting or XGBoost is one of the most popular machine learning models i n current times. XGBoost is quite similar at the core to the original gradient boosting algorith m but features many additive features that significantly improve its performance such as built in support for regularization, parallel processing as well as giving additional hyperparameters to tune such as tree pruning, sub sampling and number of decision trees. A maximum depth of 16 was used and the algorithm was run on all cores in parallel.

Using this we get

```
R2 Score: 65.69078897272438
Cross Val Score: -65.21514441037999
```

From the above observation it is cleared that the best model is Random Forest Regressor.

Since the difference between the percentage score of cross validation and r2_score is optimum.

4. Conclusion

A proper implementation of this project can result in saving money of inexperienced people by providing them the information related to trends that flight prices follow and also give the m a predicted value of the price which they use to decide whether to book ticket now or later.

And here is actual and predicted values.

	Actual	Predict
1804	3569	3709.250
1411	5584	7238.570
1824	3855	3882.620
1703	5013	4264.470
19	5954	6435.840
...	...	...
1279	2410	4738.575
852	8728	7935.840
731	10470	10564.500
1874	5954	5691.060
581	9000	9065.100

5. Limitations of this work and Scope for Future Work

Currently, there are many fields where prediction-based services are used such as stock price predictor tools used by stock brokers and service like Zestimate which gives the estimated value of house prices. Therefore, there is requirement for service like this in the aviation industry which can help the customers in booking tickets.