FLIGHT PRICE PREDICTION USING MACHINE LEARNING

BY: MRS. SWATI AMIT MOTUGADE

PROBLEM STATEMENT

- Anyone who has booked a flight ticket knows how unexpectedly the prices vary. The cheapest available ticket on a given flight gets more & less expensive over time. This usually happens as an attempt to maximize revenue based on –
  1. Time of purchase patterns (making sure last-minute purchases are expensive)
  2. Keeping the flight as full as they want it (raising prices on a flight which is filling up in order to reduce sales and hold back inventory for those expensive last-minute expensive purchases)
- So, you have to work on a project where you collect data of flight fares with other features and work to make a model to predict fares of flights.

LITERATURE REVIEW

- India is the third-biggest avionics showcase in 2020 and the biggest by 2030.
- Customers are seeking to get the lowest price for their ticket, while airline companies are trying to keep their overall revenue as high as possible and maximize their profit.
- From the customer point of view, determining the minimum price or the best time to buy a ticket is the key issue. The conception of "tickets bought in advance are cheaper" is no longer working (William Groves and Maria Gini, 2013)

- According to Y. Chen et al. (2015) , predicting the actual ticket price is a more difficult task than predicting an optimal ticket purchase time due to various reasons.
- The higher the level of competition, the weaker of the market power of an airline, & then the less likely the chance of the airline fare increases.
- Tziridis et al. In his comparison, Bagging Regression Tree is identified as the best model, which is robust and not affected by using different input feature sets.
- The presence of LCC in a market has had a substantial impact on the total passenger volume and the air ticket price.

- It is a common practice for airlines to pass the cost of aviation fuel to the customer by adjusting the fare to compensate for the fluctuation of crude oil price.
- In Another finding , When the flight is at a difference of 2-3 days' time the ticket price starts increasing again.
- Short distance flights are more elastic (more price sensitive) than long distance flights
- Business class flights are more inelastic as compared to leisure class as business customers have less flexibility to change or cancel their travel date.

WEB SCRAPING THEORY USED

- Selenium will be used for web scraping data from www.yatra.com
- Flights on route of New Delhi to Mumbai in duration of 23 Jan 2022 to 4 Feb 2022.

- Data is scrap in three categories:
  - Economy class flight price extraction
  - Business class flight price extraction
  - Premium Economy class price extraction
- Selecting features to be scrap from website.
- In next part web scraping code executed for above mention details. Exporting final data in Excel file.

DATASET INFORMATION

- Dataset contain flight detail of around 2955 Flights on route New Delhi to Mumbai.
- Dataset has 12 features like Airlines, flight, Aero plane etc.

```
data= pd.read_excel('FlightDetails_Dataset.xlsx')
```

check for shape of dataset

```
print('No. of Rows :',data.shape[0])
print('No. of Columns :',data.shape[1])
pd.set_option('display.max_columns',None)
data.head()
```

```
No. of Rows : 2222
No. of Columns : 12
```

DATA PREPROCESSING

- Conversion of Duration column from hr & Minutes format into Minutes .

```
data['Duration'] = data['Duration'].map(lambda x : x.replace('05m','5m'))
```

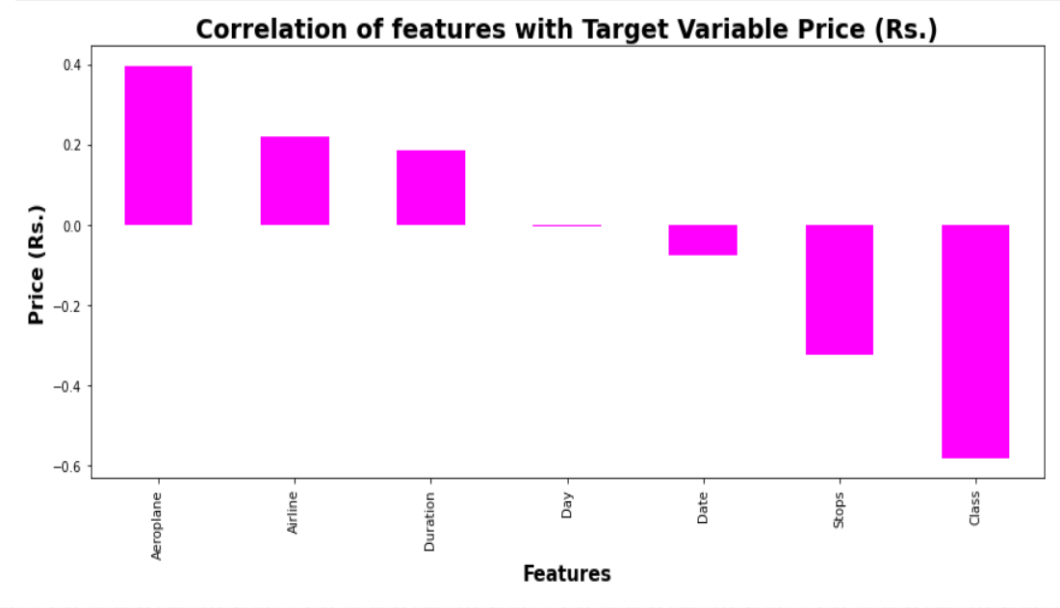
```
# Conversion of Duration column from hr & Minutes format to Minutes
data['Duration'] = data['Duration'].str.replace('h','*60').str.replace(' ','+').str.replace('m','*1').apply(eval)

# convert this column into a numeric datatypes
data['Duration']= pd.to_numeric(data['Duration'])
```

- New column for 'Day' & 'Date' is extracted from Date column.

```
#first first 3 characters are showing day so here we will take only those
data['Day']= data['Date'].map(lambda x :x[:3])
```

```
#the remaining characters are showing date of the flifgt so we will select only these characters now
data['Date']= data['Date'].map(lambda x :x[4:])
```

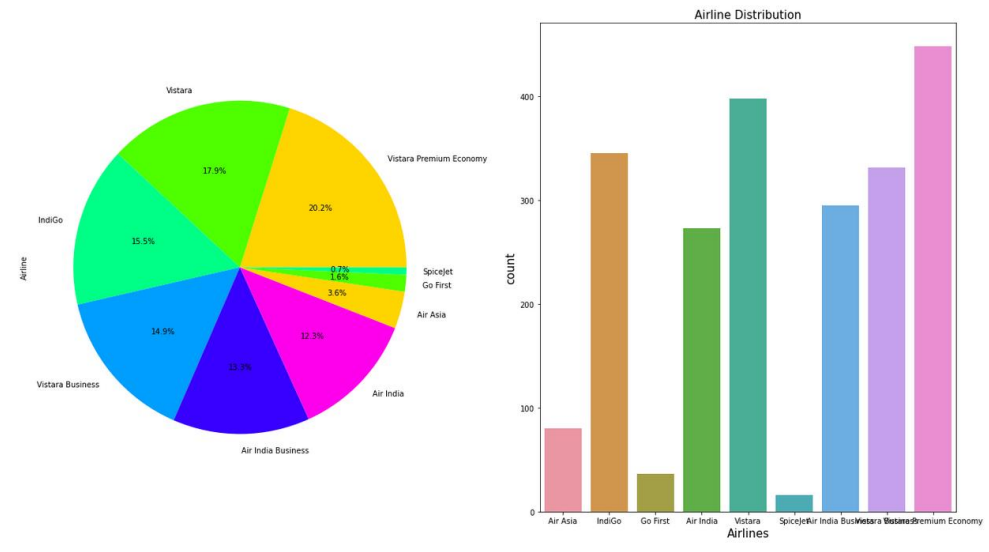
CORRELATION WITH TARGET VARIABLE



- We can see that class feature is correlated for more than -0.6 with target variable Price.
- Remaining feature are poorly correlated with target variable price.

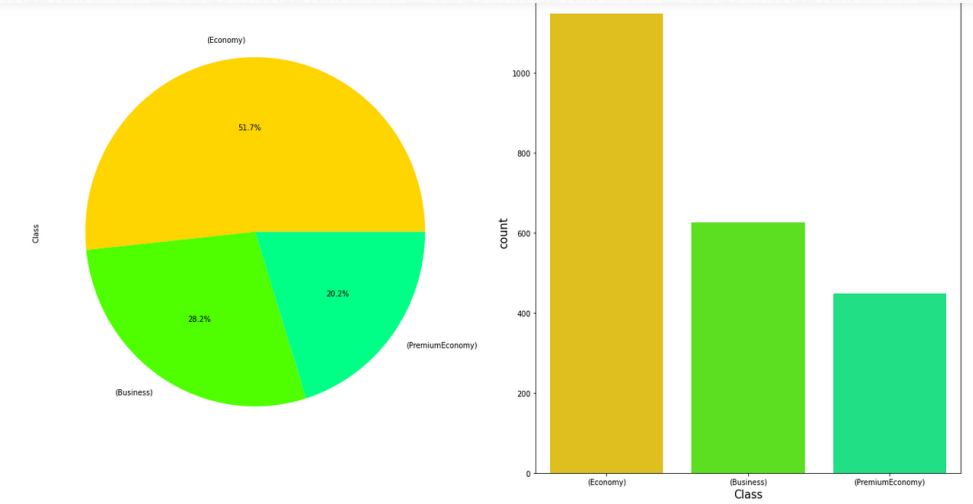
EXPLORATORY DATA ANALYSIS

AIRLINE-WISE DISTRIBUTION OF FLIGHTS



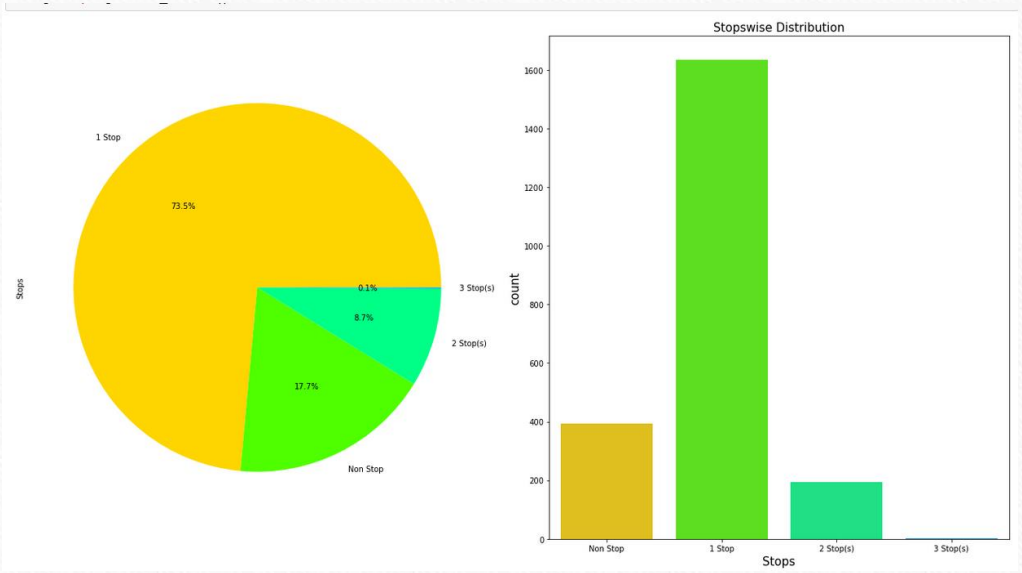
- 1.We can see maximum number of flights are of Business Class. Maxi mum number of flights run by Vistara Premium Economy while minimum Flights run by SpiceJet.
- 2. Around 28% of flights of Business Class.

CLASS-WISE DISTRIBUTION OF FLIGHTS



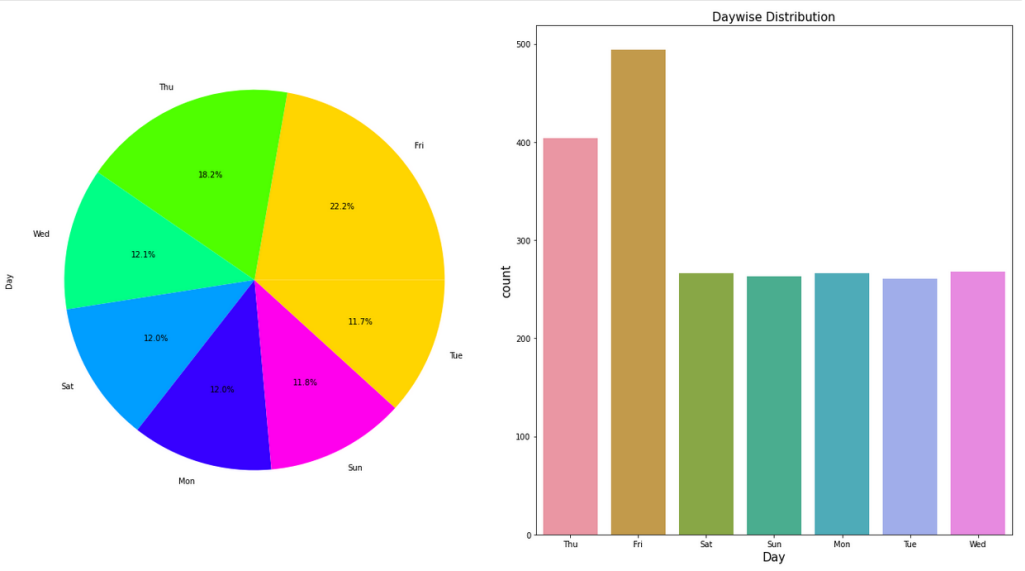
- 1.51.7% flights are of Economy class, as they are low cost of flight & most of people prefer it.
- 2.There are more business class flights than Premium Economy flights. It strange because Business class is costlier than Premium Economy class.

STOP-WISE DISTRIBUTION OF FLIGHTS



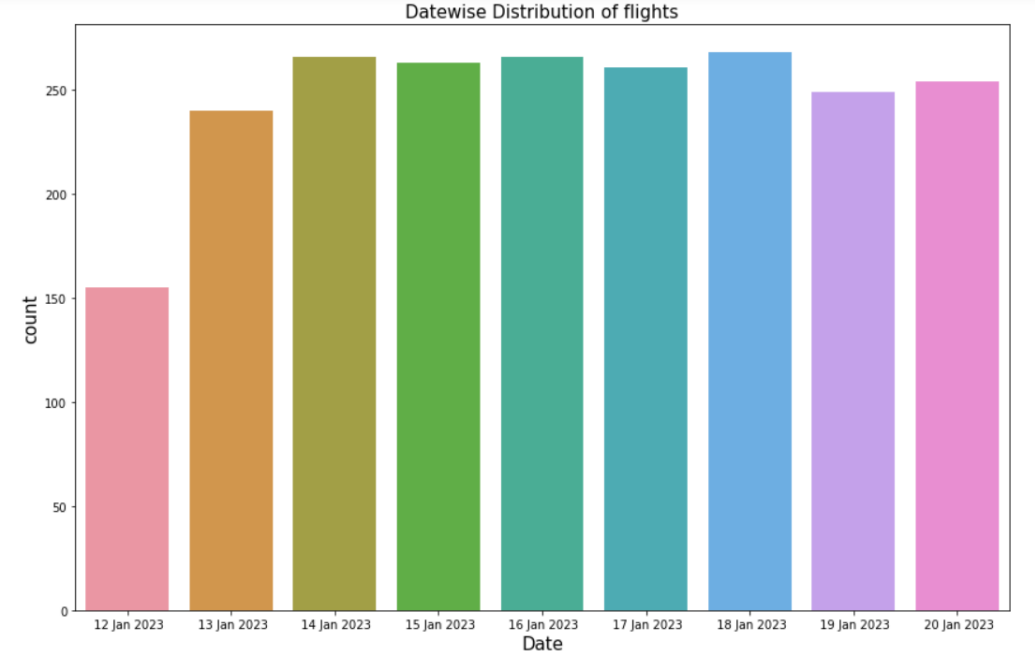
- 1. 73.5% flights take single stop in there way from Mumbai to Bangalore. It is also possible that these flights may have high flight duration compare to Non-stop Flight
- 2. 17.7% of flights do not have any stop in their route.

DAY WISE DISTRIBUTION



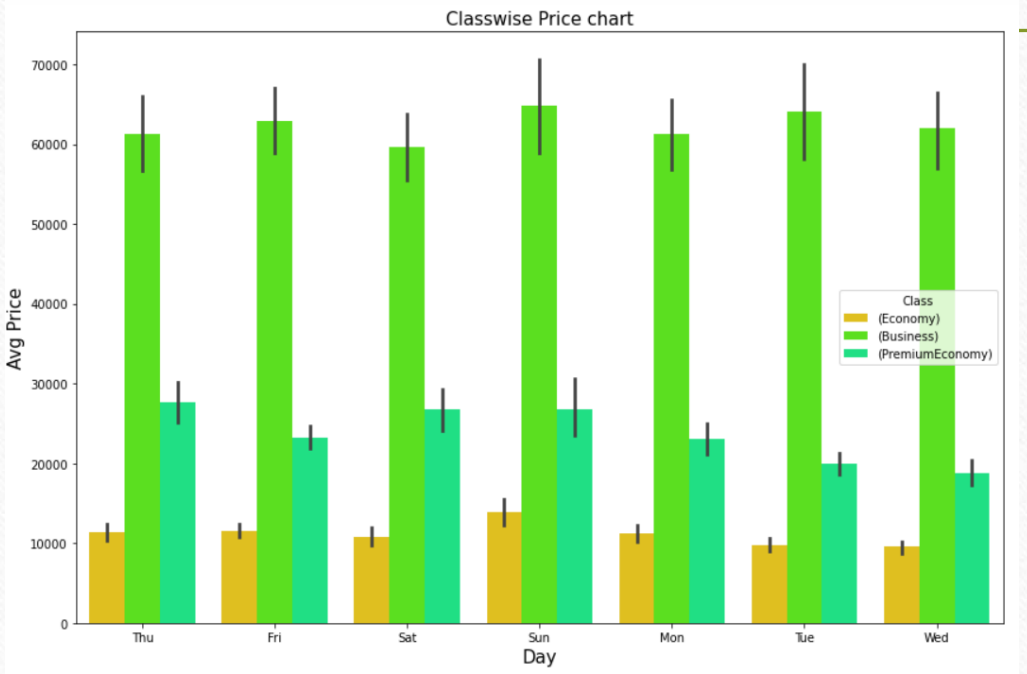
1.Maximum number flights runs on Friday where minimum number of flights runs on Tuesday.

DATE WISE DISTRIBUTION



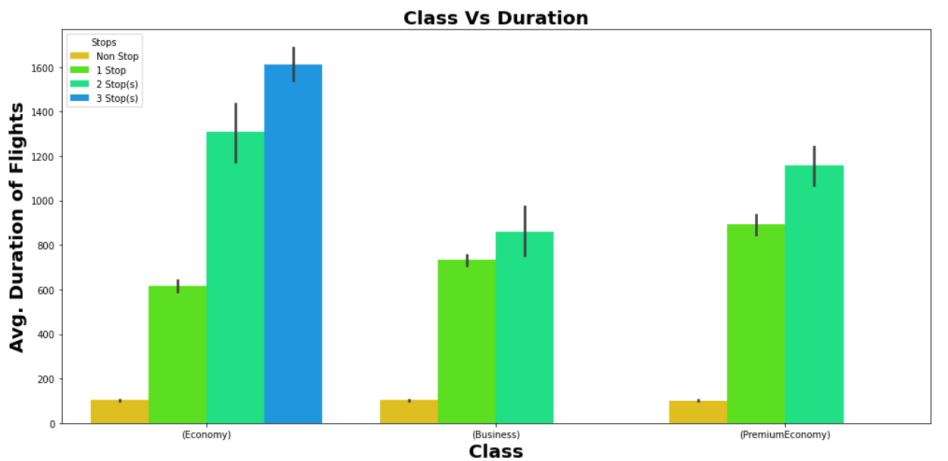
1. Maximum number of flights are available on 18 Jan 2023 and minimum flights are available on 12 Jan 2023.

DAY VS PRICE



- 1. Maximum average price for Business class is for Sunday and minimum for Saturday.
- 2. Maximum average price for Economy class is for Sunday and minimum for Tuesday.
- 3. Maximum average price for Premium Economy class is for Thursday and minimum for Wednesday.

CLASS VS DURATION



As the number of stops increases the duration also increases.

MACHINE LEARNING MODEL BUILDING

MACHINE LEARNING MODEL BUILDING

- This problem can be solved using regression-based machine learning algorithm.

- Methodology to Build Machine Learning Model:
  - Encoding Categorical data into Numerical data
  - Scaling data using Standard Scalar
  - Splitting data in training & test data using train_test_split from model_selection
  - Implementing various Regression Based Algorithm to build ML Model
  - Conducting 10 fold Cross validation
  - Hyper Parameter tuning of best Model
  - Saving Final Tuned Model using Joblib

REGRESSION ALGORITHMS IMPLEMENTATION

The different regression algorithm used in this project to build ML model are as below:

- Linear Regression
- Random Forest Regressor
- Decision Tree Regressor
- XGB Regressor
- Extra Tree Regressor

HYPERPARAMETER OPTIMIZATION

The 10-Fold cross validation performed over all models. We can see that XGB Regressor gives maximum R2 score of 94.501069 and maximum cross validation score. Among all model we will select XGB Regressor as final model and we will perform hyper parameter tuning over this model to enhance its R2 Score.

```
from sklearn.model_selection import GridSearchCV
X_train, X_test, Y_train, Y_test = train_test_split(X_scale, Y, random_state=366, test_size=0.3)
```

```
parameter = {'n_estimators':[400,500],'gamma':np.arange(0,0.2,0.1),
             'booster' : ['gbtree','dart','gblinear'], 'max_depth':[6,8],
             'eta' : [0.01, 0.1] }
```

```
GCV = GridSearchCV(XGBRegressor(),parameter,verbose =10)
```

```
GCV.fit(X_train,Y_train)
```

```
GCV.best_params_
```

```
{'booster': 'dart',
 'eta': 0.1,
 'gamma': 0.0,
 'max_depth': 6,
 'n_estimators': 400}
```

FINAL MODEL

Final model is built with best params got in hyper parameter tuning

```
final_model=XGBRegressor(booster='gbtree', max_depth=6, eta=0.1,
                         gamma=0.1, n_estimators=400)
final_model.fit(X_train,Y_train)
pred=final_model.predict(X_test)
print('R2_Score:',r2_score(Y_test,pred)*100)
print('mean_squared_error:',mean_squared_error(Y_test,pred))
print('mean_absolute_error:',mean_absolute_error(Y_test,pred))
print("RMSE value:",np.sqrt(mean_squared_error(Y_test, pred)))
```

```
R2_Score: 94.09917724102289
mean_squared_error: 40884283.70619134
mean_absolute_error: 3867.83636758281
RMSE value: 6394.081928329612
```

After hyper parameter tuning Final Model is giving us R2 Score of 94.09917% which is slightly decreased compare to earlier R2 score of 94.50106%.

FINAL EVALUATION MATRIX

ALGORITHM	R2 SCORE	CV SCORE
Linear Regression	70.79928	-4.45956
Extra Trees Regressor	93.06071	0.20531
Random Forest Regressor	94.04038	0.43168
Decision Tree Regressor	89.80232	0.10380
XGB Regressor	94.50106	0.48723
Final Model (XGB Hyperparameter Tuned)	94.09917	

LIMITATIONS & FUTURE SCOPE

- In this study we focus on flights on route of New Delhi to Mumbai, more route can incorporate in this project to extend it beyond present investigation.
- This investigation focus on short timeframe (8 days prior flights take off) which can be extended variation over larger period.
- Time series analysis can be performed over this model.

THANK YOU