FLIGHT PRICE PREDICTION

Submitted by:

POOJASHREE D S

ACKNOWLEDGMENT

• Ultimate guide to deal with text data by Shubham Jain featured in Analytics vidya

• Visualizing text data using wordcloud by Deepika Singh featured in Pluralsight.

INTRODUCTION

- ## Business Problem Framing

  The main aim of the project is to price of the flight ticket

- ## 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. 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, 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.

- ## Motivation for the Problem Undertaken

  The objectives of the project can broadly be laid down by the following

- Flight Trends Do airfares change frequently? Do they move in small increments or in large jumps? Do they tend to go up or down over time?.
- Best Time To Buy What is the best time to buy so that the consumer can save the most by taking the least risk? So should a passenger wait to buy his ticket, or should he buy as early as possible?
- 3. Verifying Myths Does price increase as we get near to departure date? Is Indigo cheaper than Jet Airways? Are morning flights expensive?

Analytical Problem Framing

- Data Sources and their formats

  Data source: yatra.com.

The data is mainly collected from the yatra.com using selenium and I scrapped the data for the ticket price from New Delhi to Mumbai.

- Data Preprocessing Done

  The data set consists of 1644 rows and 10 columns but we have to remove unnamo:0 column and I converted the price value which is in string format to float . The important step in the data peprocessing is to check is there an null value present in the data .

- Hardware and Software Requirements and Tools Used

  pandas and NumPy are the basic libraries imported. Matplotlib.pyplot and seaborn are used for visualization.

  Sklearn is used for pre-processing and model building steps. All algorithms for processing are imported. Sklearn.metrics will provide the needed evaluation metrics such as accuracy score, classification report and confusion matrix.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  The size of the dataset is about one thousand six hundred and the dataset does contain ten features

- Testing of Identified Approaches (Algorithms)

  Decision Tree Regressor

  Random Forest Regressor

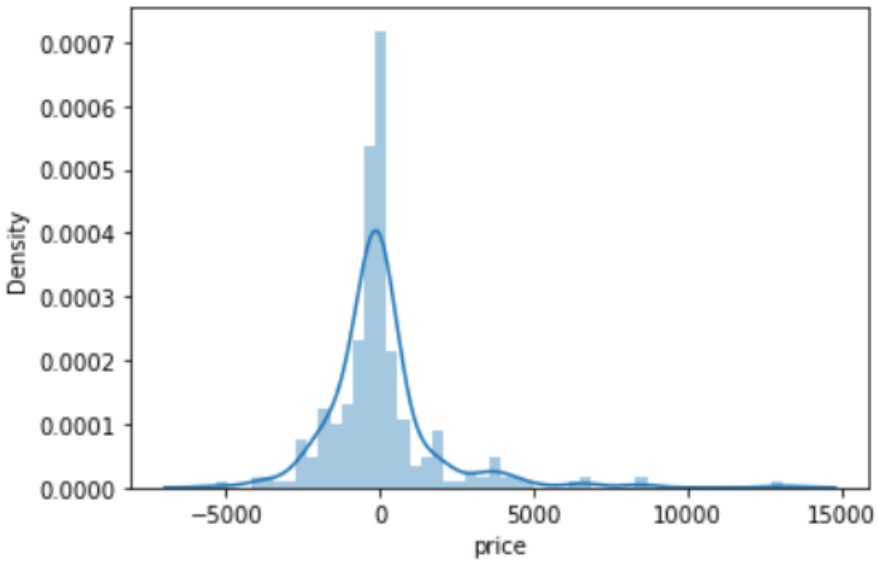
  Voting Regressor

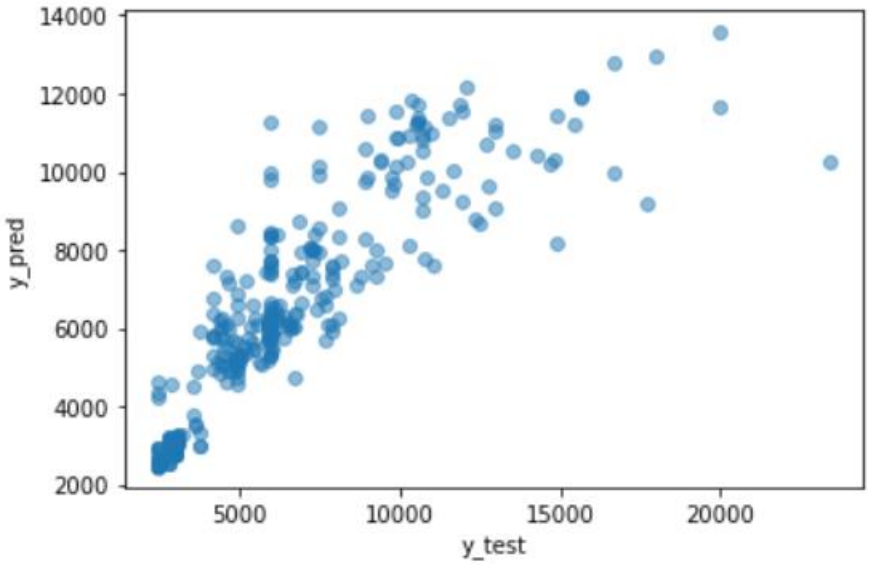
- Run and Evaluate selected models
- The above algorithms are fitted to x_train and y_train which is generated by train test split method. The performance of the model is validated based on the predictions made for x_test.

```
from sklearn.ensemble import RandomForestRegressor
rf_reg = RandomForestRegressor(n_estimators=400,min_samples_split=15,min_samples_leaf=2,
max_features='auto', max_depth=30)
rf_reg.fit(X_train, y_train)
y_pred=rf_reg.predict(X_test)

print("Random Forest Score on Training set is",rf_reg.score(X_train, y_train))#Training Accuracy
print("Random Forest Score on Test Set is",rf_reg.score(X_test, y_test))#Testing Accuracy

accuracies = cross_val_score(rf_reg, X_train, y_train, cv = 5)
print(accuracies)
print("Accuracy: {:.2f} %".format(accuracies.mean()*100))
print("Standard Deviation: {:.2f} %".format(accuracies.std()*100))

mae=mean_absolute_error(y_pred, y_test)
print("Mean Absolute Error:" , mae)

mse=mean_squared_error(y_test, y_pred)
print("Mean Squared Error:" , mse)
```

```python
accuracies = cross_val_score(rf_reg, X_train, y_train, cv = 5)
print(accuracies)
print("Accuracy: {:.2f} %".format(accuracies.mean()*100))
print("Standard Deviation: {:.2f} %".format(accuracies.std()*100))

mae=mean_absolute_error(y_pred, y_test)
print("Mean Absolute Error:" , mae)

mse=mean_squared_error(y_test, y_pred)
print("Mean Squared Error:" , mse)

print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))

print('The r2_score is', metrics.r2_score(y_test, y_pred))

sns.distplot(y_test-y_pred)
plt.show()

plt.scatter(y_test, y_pred, alpha = 0.5)
plt.xlabel("y_test")
plt.ylabel("y_pred")
plt.show()
```

```
Random Forest Score on Training set is 0.8352993875401366
Random Forest Score on Test Set is 0.7354831941246596
[0.69425732 0.6613142  0.74558153 0.68705689 0.69180519]
Accuracy: 69.60 %
Standard Deviation: 2.74 %
Mean Absolute Error: 1027.133932995749
Mean Squared Error: 3232763.5486412505
RMSE: 1797.9887509773944
The r2_score is 0.7354831941246596
```



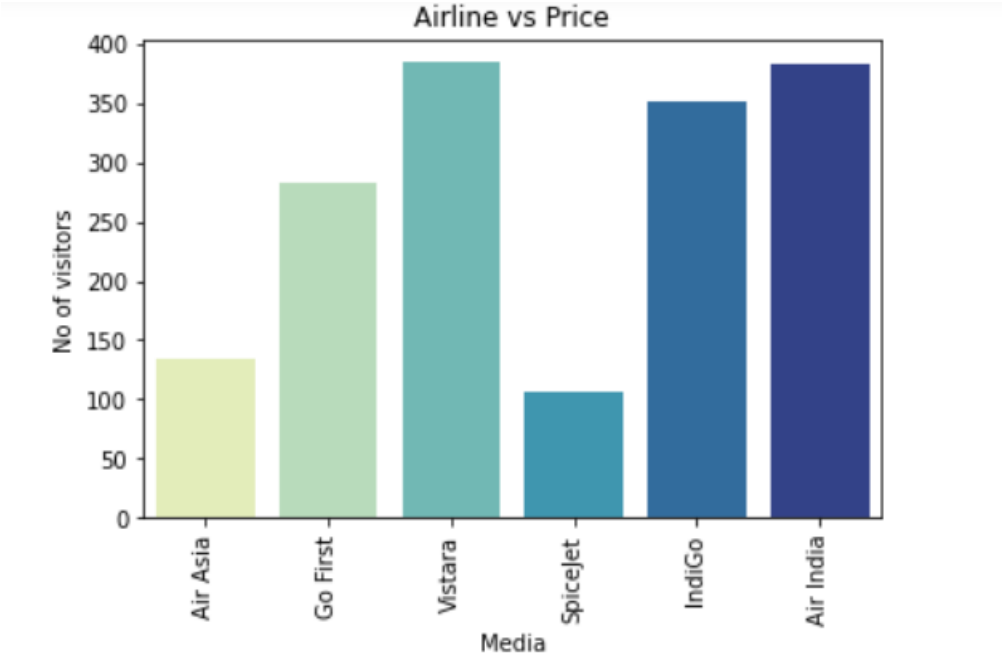


- Key Metrics for success in solving problem under consideration

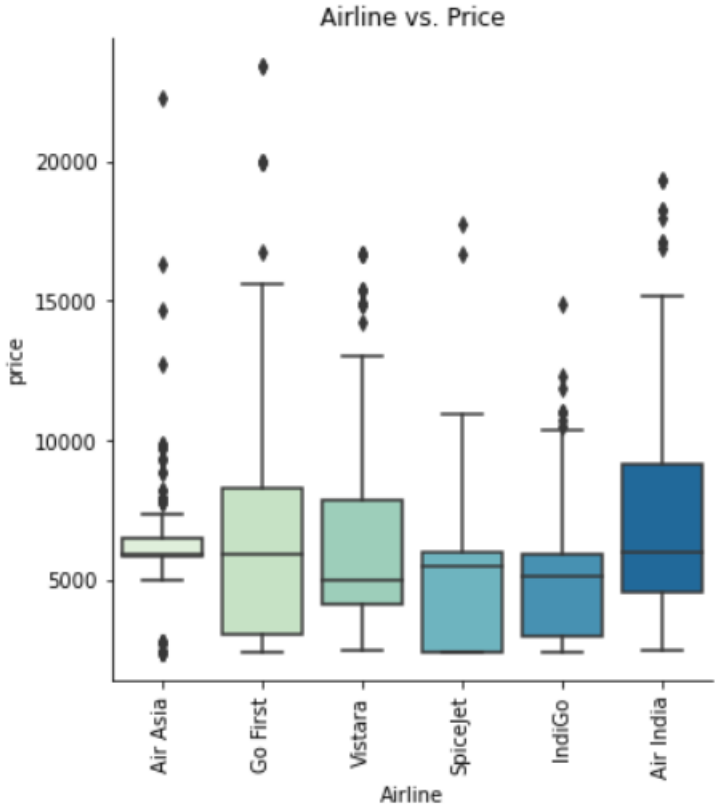
  Random Forest Regressor gives better accuracy than oher two regressor. The metrics used are accuracy score, classification report and confusion matrix. Accuracy score gives the performance of both classes as a whole. Classification report comprising of precision, recall and f1 score help us in understanding the efficiency of model with respect to each class. Confusion matrix places a distinction between false positives and false negatives.

- Visualizations

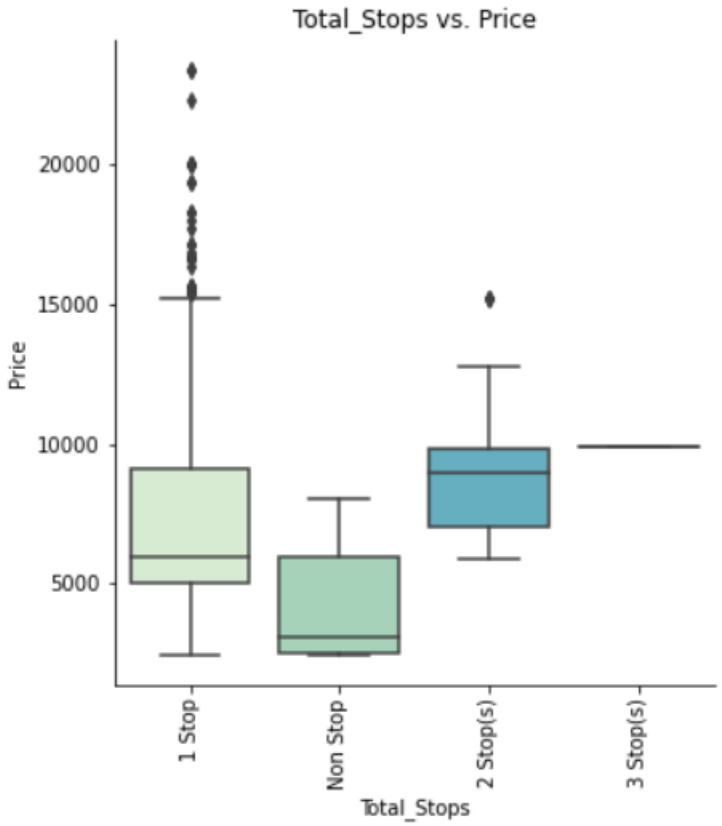
  Bar graph are generated to understand the price of in different flights



From the above graph we can able to conclude Air india and Vistara have more number of passengers then others.



The above graph tells that the Air india have highest price range than the other.



Price of the ticket decreases as the booking is earlier in the present situation December is cheapest because it is after three months

CONCLUSION

The aim is to predict the price of the ticket. Air india and Vistara have more number of passengers then others. Air india have highest price range than the other. Price of the ticket decreases as the booking is earlier in the present situation December is cheapest because it is after three months