ALY6140 – ANALYTICS SYSTEM TECHNOLOGY

CAPSTONE PROJECT – FLIGHT PRICE PREDICTION

Team Members	
Name	NUID
SHIVANGKUMAR PANCHAL	002752793
SAI SACHIN CHANDRASEKAR	002934358
CHRISTINA NIKITHA STANLEY	002745073

DECEMBER 15, 2022

Northeastern University

INDEX

Introduction

In order, to extract useful information from the flight booking dataset from the "Ease My Trip" website, the study's goal is to analyse it and run various statistical hypothesis tests.

The dataset would be trained using the statistical algorithm "Linear Regression" to forecast a continuous target variable. Potential travellers can purchase tickets via the "Easemytrip"

website, which is a platform for booking flights. Passengers will greatly benefit from the unique insights that are discovered through a thorough analysis of the data.

Exploratory Data Analysis:

Data Extraction:

- Extracting the data using the pd.read_csv() function from the file location.

```
In [14]: import numpy as np
         import pandas as pd
         import matplotlib.pyplot as plt
         import seaborn as sns
         import warnings

         df = pd.read_csv("C:\\Users\\Rahul\\Desktop\\Clean_Dataset.csv\\Clean_Dataset.csv")
         df.head()
```

Out[14]:

	Unnamed: 0	airline	flight	source_city	departure_time	stops	arrival_time	destination_city	class	duration	days_left	price
0	0	SpiceJet	SG-8709	Delhi	Evening	zero	Night	Mumbai	Economy	2.17	1	5953
1	1	SpiceJet	SG-8157	Delhi	Early_Morning	zero	Morning	Mumbai	Economy	2.33	1	5953
2	2	AirAsia	I5-764	Delhi	Early_Morning	zero	Early_Morning	Mumbai	Economy	2.17	1	5956
3	3	Vistara	UK-995	Delhi	Morning	zero	Afternoon	Mumbai	Economy	2.25	1	5955
4	4	Vistara	UK-963	Delhi	Morning	zero	Morning	Mumbai	Economy	2.33	1	5955

Figure 1 Loading the dataset

The dataset contains 300153 entries with 12 columns which contains the data types like integer, object and float. The dataset has 6 unique airlines, source city, departure time, arrival time, destination city, and also contains unique stops and class.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 300153 entries, 0 to 300152
Data columns (total 12 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   Unnamed: 0        300153 non-null  int64
 1   airline           300153 non-null  object
 2   flight            300153 non-null  object
 3   source_city       300153 non-null  object
 4   departure_time    300153 non-null  object
 5   stops             300153 non-null  object
 6   arrival_time      300153 non-null  object
 7   destination_city  300153 non-null  object
 8   class             300153 non-null  object
 9   duration          300153 non-null  float64
 10  days_left         300153 non-null  int64
 11  price             300153 non-null  int64
dtypes: float64(1), int64(3), object(8)
memory usage: 27.5+ MB
```

Figure 2. Data info

Data clean-up:

Checking for missing values

```
data.isnull().sum()
```

```
Unnamed: 0          0
airline             0
flight              0
source_city         0
departure_time      0
stops               0
arrival_time        0
destination_city    0
class               0
duration            0
days_left           0
price               0
dtype: int64
```

Checking for duplicate values

```
[ ] data.duplicated().sum()
```

```
0
```

Figure 3. checking for missing and duplicate values

```
data.drop('Unnamed: 0', axis=1, inplace=True)
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 297720 entries, 0 to 300152
Data columns (total 11 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   airline           297720 non-null  object
 1   flight            297720 non-null  object
 2   source_city       297720 non-null  object
 3   departure_time    297720 non-null  object
 4   stops             297720 non-null  object
 5   arrival_time      297720 non-null  object
 6   destination_city  297720 non-null  object
 7   class             297720 non-null  object
 8   duration          297720 non-null  float64
 9   days_left         297720 non-null  int64
 10  price             297720 non-null  int64
dtypes: float64(1), int64(2), object(8)
memory usage: 27.3+ MB
```

Figure 4. Removing the unnecessary column

In the above figure 3 and 4. The dataset was checked for missing values, duplicate values and removing the unnecessary column. In which we have removed the Unnamed column from the dataset which makes the dataset clean.

```
[41] le=LabelEncoder
airline_le=LabelEncoder()
data['airline']=airline_le.fit_transform(data['airline'])
flight_le=LabelEncoder()
data['flight']=flight_le.fit_transform(data['flight'])
source_city_le=LabelEncoder()
data['source_city']=source_city_le.fit_transform(data['source_city'])
departure_time_le=LabelEncoder()
data['departure_time']=departure_time_le.fit_transform(data['departure_time'])
stops_le=LabelEncoder()
data['stops']=stops_le.fit_transform(data['stops'])
arrival_time_le=LabelEncoder()
data['arrival_time']=arrival_time_le.fit_transform(data['arrival_time'])
destination_city_le=LabelEncoder()
data['destination_city']=destination_city_le.fit_transform(data['destination_city'])
class_le=LabelEncoder()
data['class']=class_le.fit_transform(data['class'])
```

```
[47] data.head()
```

	airline	flight	source_city	departure_time	stops	arrival_time	destination_city	class	duration	days_left	price
0	4	1408	2	2	2	5	5	1	2.17	1	5953
1	4	1387	2	1	2	4	5	1	2.33	1	5953
2	0	1213	2	1	2	1	5	1	2.17	1	5956
3	5	1559	2	4	2	0	5	1	2.25	1	5955
4	5	1549	2	4	2	4	5	1	2.33	1	5955

Figure 5. Encoding the character column into numeric.

From the above figure 5. The character column is encoded in the numeric format to test with the price test column to predict the error. They are encoded in the following:

Airlines		source city		Departure time		Stops		Arrival time		Destination city		Class	
Air Asia	0	Bangalore	0	Afternoon	0	One	0	Afternoon	0	Air Asia	0	Business	0
Air India	1	Chennai	1	Early Morning	1	two or more	1	Early Morning	1	Air India	1	Economy	1
Go First	2	Delhi	2	Evening	2	Zero	2	Evening	2	Go First	2		
Indigo	3	Hyderabad	3	Late Night	3			Late Night	3	Indigo	3		
SpiceJet	4	Kolkata	4	Morning	4			Morning	4	SpiceJet	4		
Vistara	5	Mumbai	5	Night	5			Night	5	Vistara	5		

Exploratory data visualization

Price of seats in each class:

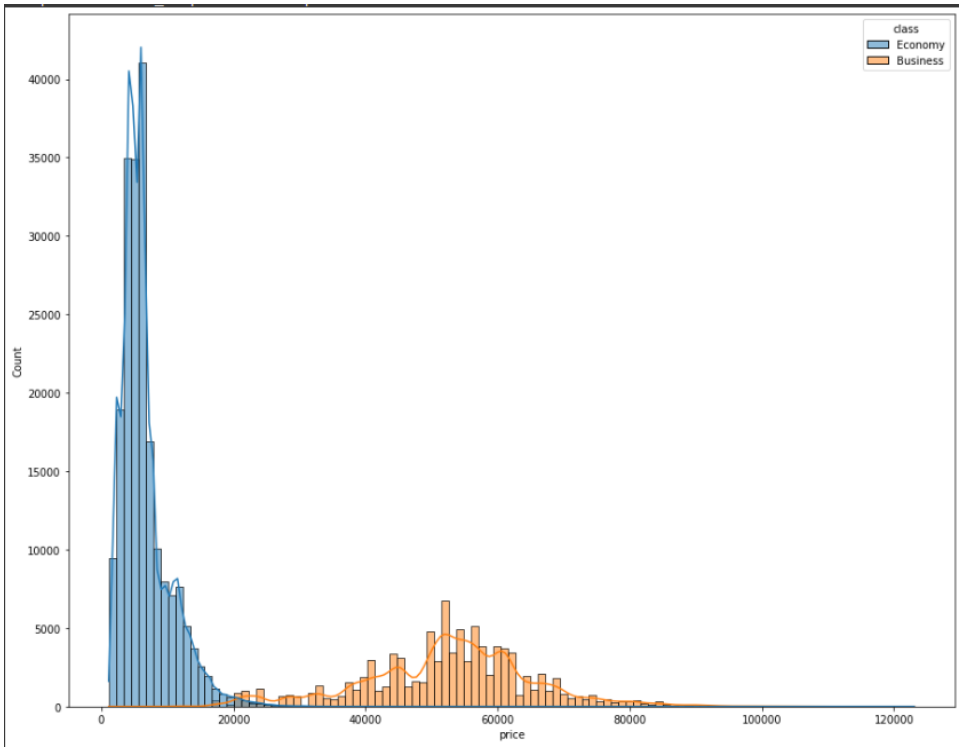

Figure 6. Price of seats in each class

The figure 6. Represents the price of the 2 different classes (Economy, Business). In this graph the number of counts in the Economy is higher than the Business class.

Class vs Price:

Figure 7. Price variation between the class

The figure 7. Represents the price variation between the classes (Economy and Business). This shows the price in business class is higher than the Economy class.

Classes of different airlines:

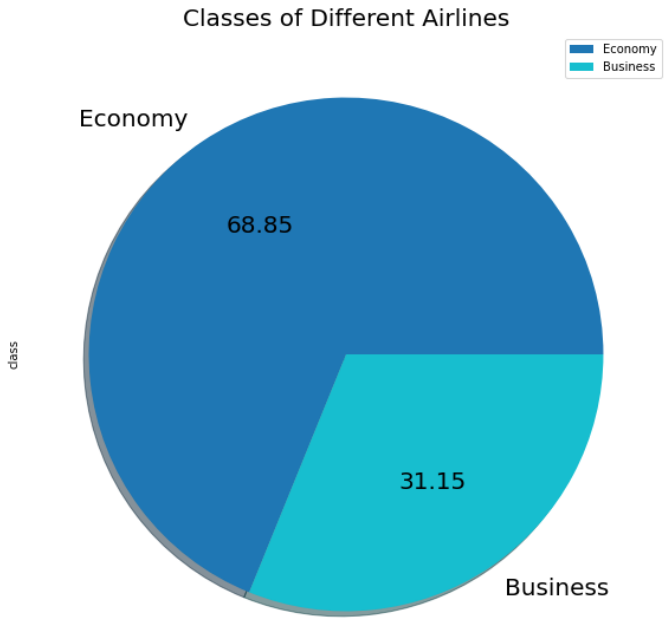

Figure 8. Percentage between different classes from the total number of seats in airlines.

The figure 8. Represents the total number of seats in airlines and how they are categorized. In the chart given the Economy class has the highest number of seats in airlines than the Business class because the cost of the Economy class ticket is cheaper than the Business class tickets.

Departure and arrival time vs ticket price:

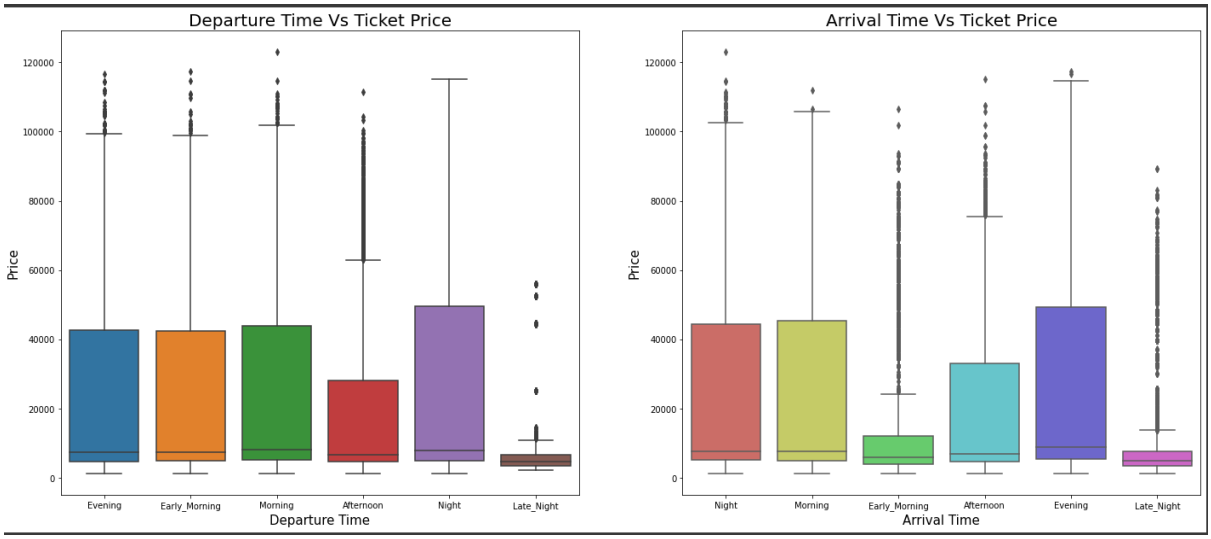

Figure 9. Price variation based on departure and arrival time of the flights.

From the above figure 9. The price varies based on the departure and arrival time of the flights. In which, Ticket Price is More for the Flights when the Departure Time is at Night Ticket Price is almost equal for flights Having Departure time at Early morning , Morning and Evening Ticket Price is Low for the Flights Having Departure Time at Late-night Ticket Price is More for the Flights when the Arrival Time is at Evening Ticket Price is almost equal for flights Having Arrival time is at Morning and Night Ticket Price is Low for the Flights Having Arrival Time at Late-night as same as Departure Time

Days left for departure vs ticket price:

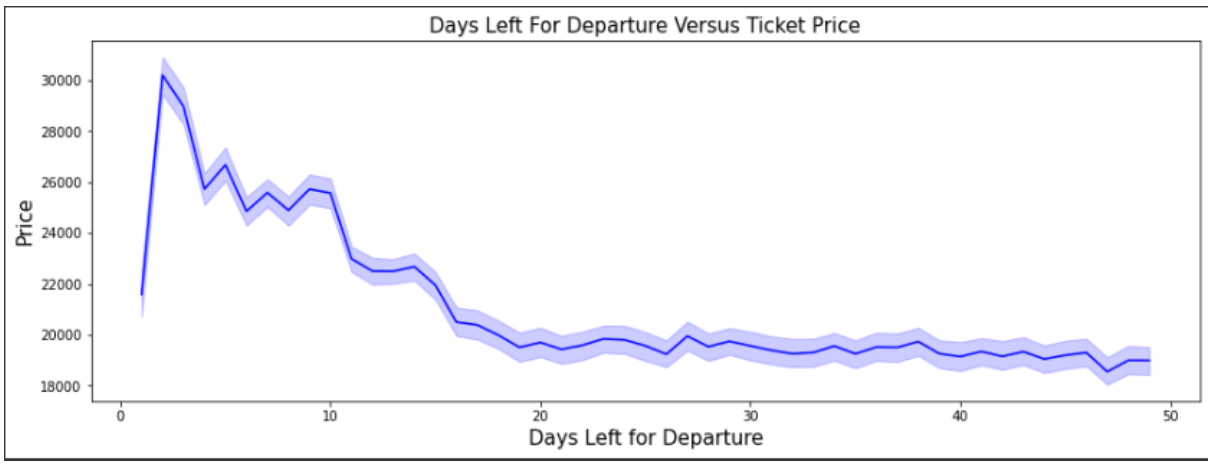

Figure 10. Fare rate variation between days left for departure.

From the above figure 10. The ticket price increases as the days left before departure decreases.

Number of stops vs the ticket price:

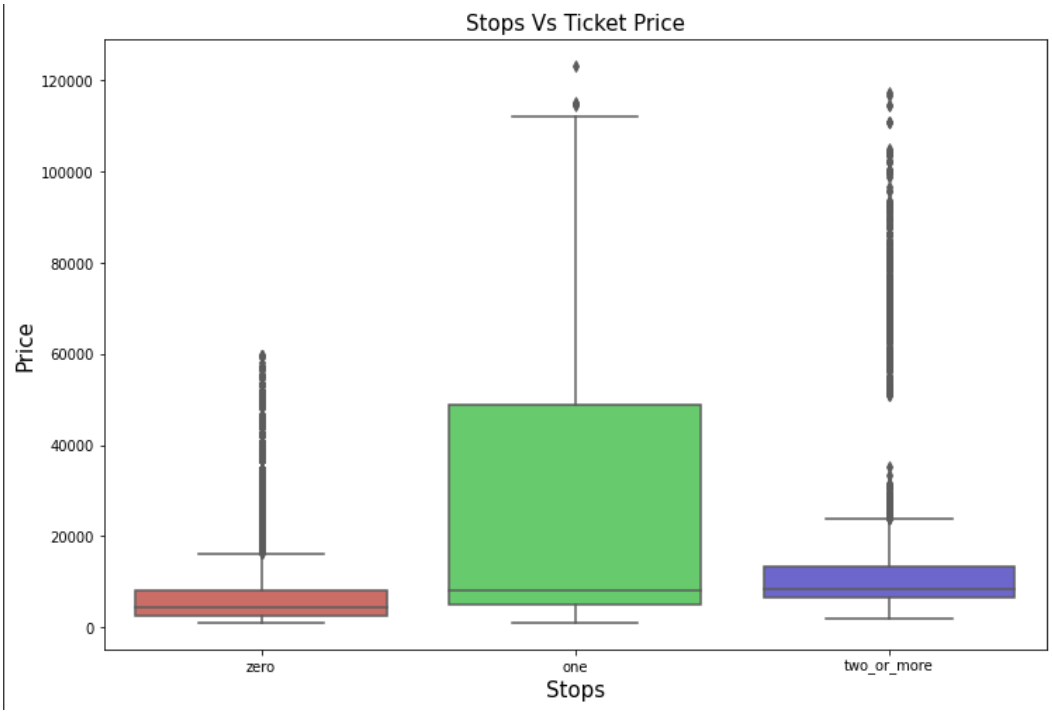

Figure 11. Number of stops vs ticket price.

From the figure 11. Flight which has zero stops has the lowest fare for the airlines when compared to one or more stops. The flight which has one stop has the highest price for the journey of the flight. The flight which has two or more stops also have the highest and lowest fare of the ticket depending on the classes of the airlines. The price of one or more stops varies because the flight will can be a international connecting flight or depending on the classes of the airlines. Even for two or more stop we find that it has outliers because it can be International connecting flight or depending on the class.

DATA MODELLING

Linear Regression

A variable's value can be predicted using linear regression analysis based on the value of another variable. The dependent variable is the one you want to be able to forecast. The independent variable is the variable you are using to forecast the value of the other variable.

Prediction values for Linear regression:

	Y_test	Y_pred
0	2278	-375.0299773
1	5402	3535.206826
2	49725	50943.6473502
3	24056	45288.111446
4	55983	55029.364172
5	3000	-2096.355029
6	4697	7070.864995

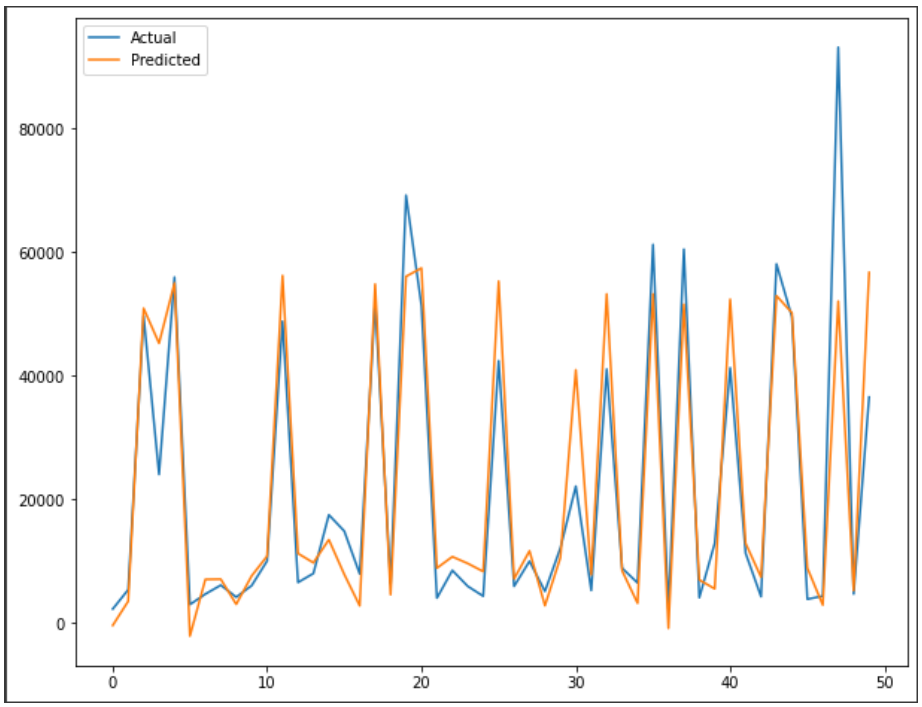

Figure 12. Linear regression using the predicted and actual values.

From the above figure 12. The graph represents the predicted and actual values of the price of the tickets. In which this prediction is used to find the future price of the airlines ticket. The actual value represents the independent variable whereas the predicted value is dependent on the actual value.

Lasso Regression

Shrinkage is used in the linear regression method known as lasso regression. When data values shrink toward a middle value, such as the mean, this is called shrinkage. Simple, sparse models are encouraged by the lasso approach (i.e., models with fewer parameters). When models exhibit significant levels of multicollinearity or when you wish to automate specific steps in the model selection process, such as variable selection and parameter elimination, this sort of regression is ideally suited.

Prediction values for Lasso Regression:

	Y_test	Y_pred
0	2278	-374.848302
1	5402	3541.304370
2	49725	50941.874212
3	24056	4528.0463888
4	55983	55023.227200
5	3000	-2094.643063
6	4697	7074.594895

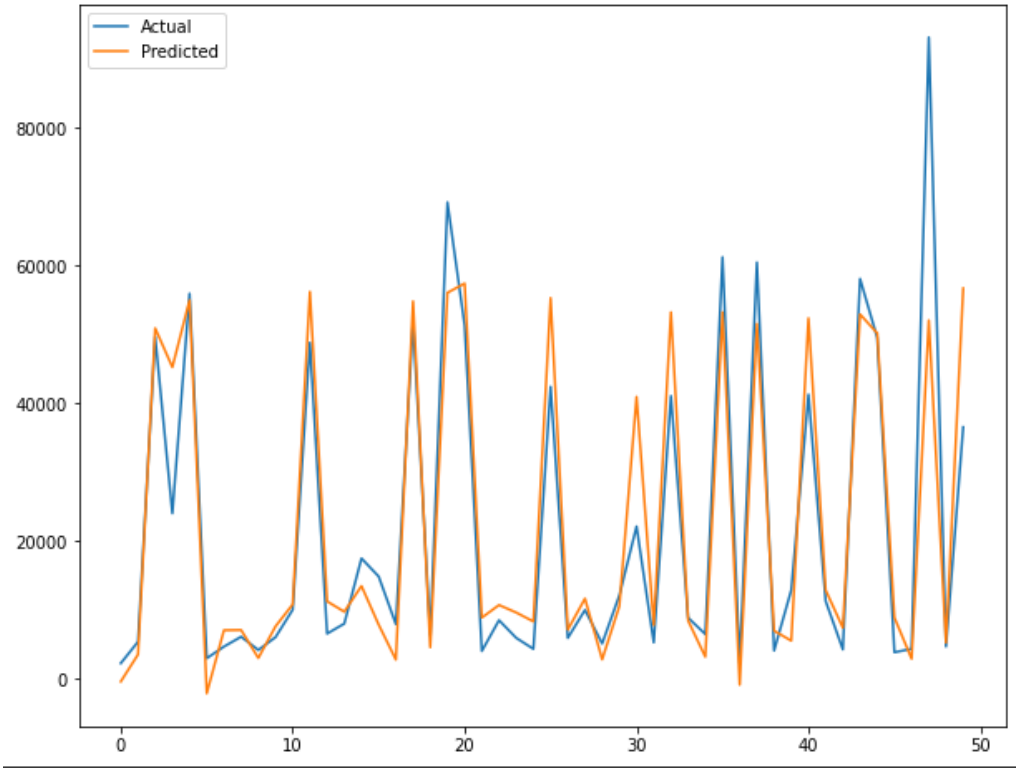

Figure 13. Actual and predicted value graph using Lasso Regression.

From the above figure 13. The values of the data shrink towards the middle value (mean value). This sort of regression is used to automate specific steps in the model selection process.

Decision Tree Regression

A non-parametric supervised learning technique for classification and regression is called a decision tree (DT). The objective is to learn straightforward decision rules derived from the data features to build a model that predicts the value of a target variable. A piecewise constant approximation of a tree can be thought of.

Prediction values for Decision Tree:

	Y_test	Y_pred
0	2278	3008.250332
1	5402	6050.993118
2	49725	48425.144941
3	24056	26776.889143
4	55983	60777.733367
5	3000	2550.634951
6	4697	5485.276010

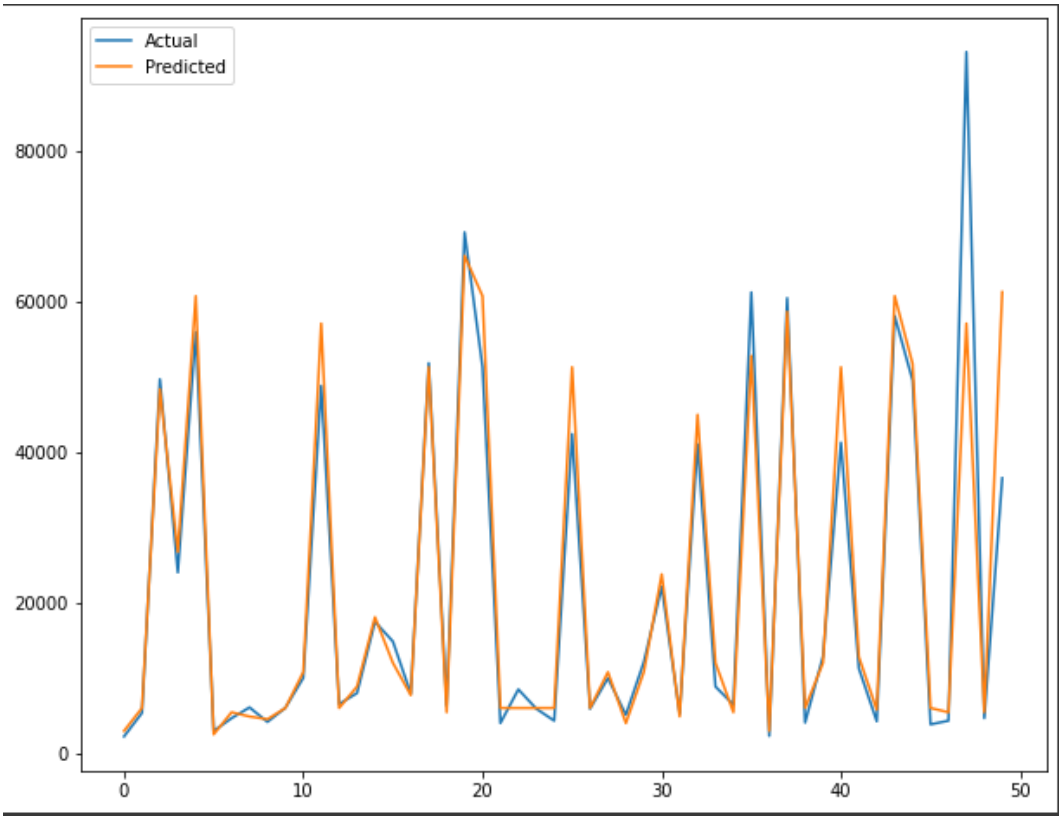

Figure 14. Actual and predicted value graph using Decision Tree.

From the above figure 14. The actual and the predicted values of the data are used to predict the values of the target variable to build a model.

Random Forest Regression

A supervised learning technique called Random Forest Regression leverages the ensemble learning approach for regression. The ensemble learning method combines predictions from various machine learning algorithms to provide predictions that are more accurate than those from a single model.

Prediction values for Random Forest:

	Y-test	Y_pred
0	2278	2994.647110
1	5402	6023.915230
2	49752	48714.461760
3	24056	26789.171200
4	55983	60601.140935
5	3000	2647.433598
6	4697	5486.636727

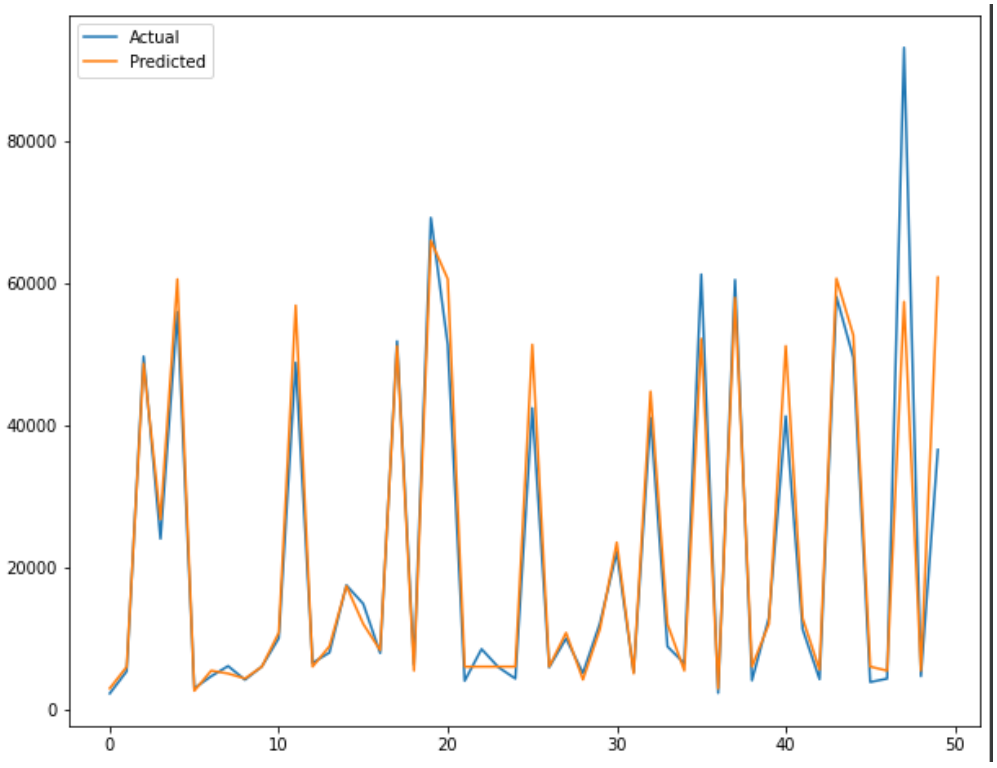

Figure 15. Actual and predicted values graph using Random Forest.

From the above figure 15 The actual and predicted values of the data are used to ensemble learning method which combines predictions from various machine learning algorithms to provide accuracy than those from a single model.

Interpretation of Results:

Linear Regression	
Mean Absolute percentage error	0.43855603
Mean Absolute error	4647.79987
Mean squared error	48924158.69
Root Mean square Error	6994.580665

Lasso Regression	
Mean Absolute percentage error	0.438447852
Mean Absolute error	4647.432576
Mean squared error	48924489.78
Root Mean square Error	6994.604333

From the above results:

The simplest type of regression is linear regression, in which the model is completely unaffected by the weights it chooses. This means that the model may give a feature a lot of weight during the training stage if it thinks the characteristic is particularly essential. Sometimes, this causes small datasets to overfit. Thus, the lasso regression was used to reduce errors.

A variation of linear regression known as lasso penalises the model for the total of the absolute weight values. As a result, the absolute values of weight will generally be lower and likely to be zero. According to our dataset, linear regression had less errors as compared to lasso regression and hence linear regression proved to be more accurate

Decision Tree Regression	
Mean Absolute percentage error	0.193950093
Mean Absolute error	2830.42644
Mean squared error	23833069.97
Root Mean square Error	4881.912531

Random Forest Regression	
Mean Absolute percentage error	0.191895398
Mean Absolute error	2789.196381
Mean squared error	23188011.34
Root Mean square Error	4815.393166

From the above results:

While Random Forest Regression is an ensemble approach of Decision Trees based on randomly divided data, Decision Tree Regression creates its model in the framework of a tree along with decision nodes and leaf nodes. This entire group can be compared to a forest with various independent random samples growing on each tree.

Decision trees are graphs that show all potential outcomes of a decision using a branching technique, which is a key distinction between them and the random forest algorithm. In contrast, the random forest algorithm output is a group of decision trees that work according to the output. Due to their accuracy and the fact that current computers and systems can handle enormous, previously unmanageable datasets, random forest algorithms are frequently used. Hence, according to our dataset, random forest has the highest accuracy proving to be best regression model and thus, the dataset is trained and tested on Random Forest model.

From the above results we can infer that Random Forest Regression has the highest accuracy as the error values are low and hence the research questions are predicted based on the Random Forest regression.

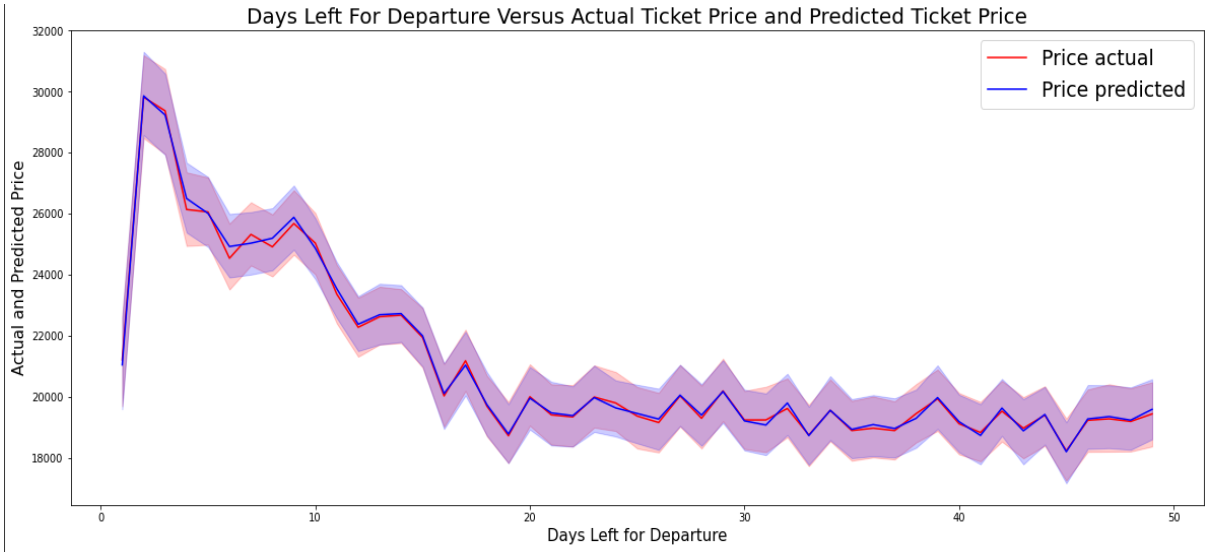

Figure 16. Days left for departure vs actual price and predicted ticket price

From the above plot, we can infer the prediction that the price increases as the days left for departure decreases.

We can see that the graph has two lines in which red represents the actual price before testing the value, whereas the blue line represents the predicted value from the testing variables.

Conclusion:

- The exploratory analysis shows that it has 30353 entries with 12 variables from the dataset.
- The price of the tickets varies depending on the class and the number of days before the ticket is bought.
- The price of the ticket varies depending on the airline.
- There is a sudden rise in the price of actual value which can be an outlier. But in the predicted value there won't be a sudden rise in the price.
- The Economy class seat count is higher than Business class as well as the customer preference based on availability and number of days before it's booked.
- The values of the data are called in the predictive models which is classified as actual and predicted value which is used to find the future values of the data for ticket rate.
- For each regression model the actual and predictive model graph looks the same but the Mean absolute error (MAE) and Root mean squared error (RMSE) will differ.
- The random forest regression model is more accurate in predicting the flight fares as the RMSE is comparatively lesser than the other regression models
- From the random forest regression, we can conclude that,
  a. the predicted price of ticket increases as the days left before departure decreases.
  b. the predicted price of ticket is high when the flight departs at night and arrives in the evening.

Reference:

- (n.d.). What is linear regression? IBM. https://www.ibm.com/topics/linear-regression#:~:text=Resources-,What%20is%20linear%20regression%3F,is%20called%20the%20independent%20variable.
- (n.d.). Lasso Regression: Simple Definition. Statistics How To. https://www.statisticshowto.com/lasso-regression/
- (n.d.). Decision Trees. Scikitlearn. https://scikit-learn.org/stable/modules/tree.html
- Random forest. (2022, October 25). In Wikipedia. https://en.wikipedia.org/wiki/Random_forest