FLIP ROBO

FLIGHT PRICE PREDICTION

Submitted by:

SARANYA M

ACKNOWLEDGMENT

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

I have referred most of the things data trained class notes and some Machine learning articles from Towards Data science

Selenium issues , data cleaning -- Stack overflow and Kaggle

INTRODUCTION

- ## Business Problem Framing

Describe the business problem and how this problem can be related to the real world.

Anyone who has booked a flight ticket knows how unexpectedly the prices vary. The cheapest available ticket on a given flight gets more and 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.

- ## 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.

- Review of Literature

Since the airline company's privatization, the airfare pricing scheme has evolved into a complex framework of sophisticated regulations including numerical simulations that determine airfare marketing strategies . Even though these principles are still mostly unknown, research has revealed that they are influenced by a range of circumstances . Conventional characteristics such as distance, while still important, are no longer the only determinants of pricing structure. Economic, marketing, and sociological factors have all played a growing influence in determining flight pricing.

The majority of research on airfare forecasting has concentrated either on state scale or a single market. However, analysis at the business segment level is still relatively scarce. The marketing strategy is defined as the market/airport pairing in between aircraft sender and receiver.

- Motivation for the Problem Undertaken

Air ticket price prediction is a challenging task since the factors involved in pricing dynamically change overtime and make the price fluctuate. In the last decade, researchers have incorporated machine learning algorithms and data mining strategies to better model observed prices. Among them, regression models, such as Linear Regression(LR), Support Vector Machines (SVMs), Random Forests(RF), are frequently used in predicting accurate airfare price. Early work also considered using classification models top redict the trends of the itineraries. Ren et al. proposed using LR, Referenc e Name Work Description Problem Found Any other criteria 1 IndiGo IndiGo is one of the flight booking app which can predict the flight ticket as well as book the flight ticket. This app predict ticket price for only of Indigo flights not for other flights. Registrati on is required 2 Ixigo Ixigo is one of the flight booking app which can predict the flight ticket as well as book that ticket This app only predict some Popular airlines like SpiceJet, AirAsiaIndia, GoFIRST, Vistara, AirIndia, IndiGo Compared to the current and recent work, our proposed framework manages to handle the price prediction task only using public data sources with minimal features. Also, not restricted by any specific market segment that usually limits the existing work, this proposed framework can be applied to predict the airfare price for any market.

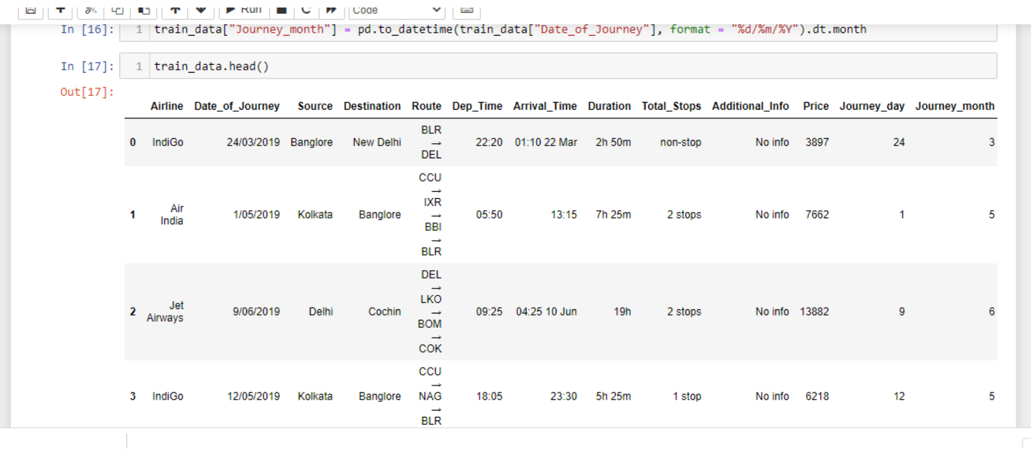
Analytical Problem Framing

- It is critical for airlines to be capable of predicting airfare trends at the market segment level in order to alter strategies and resources for a given route. Scientific literatures on business segment price prediction, on the other hand, use biased conventional predictive methods, including such linear regression , and thus are founded on the supposition that the selected variables have a linear relationship, that might not be true in most cases. Proposed research [8] Prediction of airfare prices utilizing machine learning approach, A dataset of 1814 Aegean Airways data flights was gathered and utilized to develop the machine learning technique for the study effort. Various figures of variables have been used to train the classifiers to demonstrate how feature extraction might affect validity of the model.

- A study by William Groves' shows that an operator can be introduced who can maximize purchase time on behalf of consumers. A model is constructed using the partial least squares approach
- Supriya Rajankar's survey report on aircraft fare forecasting using machine learning models employs a tiny dataset consisting of flights between Delhi and Bombay. K-nearest neighbours (KNN), linear regression, and support vector machine (SVM) algorithms are used.
- Over the course of several months, Santos conducts research on airline routes between Madrid via London, Frankfurt, New York, as well as Paris. The figure depicts the acceptable number of days before purchasing an airline ticket.
- Tianyi Wang suggested a system in which two databases with macroeconomic data are integrated and machine learning methods including such support vector machines and XGBoost are being used to estimate the average price of tickets based on input and output pairings. With the updated R squared performance indicators, the framework obtains a high accuracy of 0.869.
- Data Sources and their formats

Kaggle lets in customers to discover and put up data sets, discover and construct fashions in a web-primarily based totally factstechnology data science environment, paintings with different facts scientists and device mastering engineers, and input competitions to resolve data science challenges.Data collection done through scraping different airline websites, like Yatra.com, EsayMyTrip.com

Data stored in .CSV file format



```
In [16]: train_data["Journey_month"] = pd.to_datetime(train_data["Date_of_Journey"], format = "%d/%m/%Y").dt.month
```

```
In [17]: train_data.head()
```

Out[17]:

	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price	Journey_day	Journey_month
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897	24	3
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662	1	5
2	Jet Airways	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops	No info	13882	9	6
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218	12	5

- Data Preprocessing Done

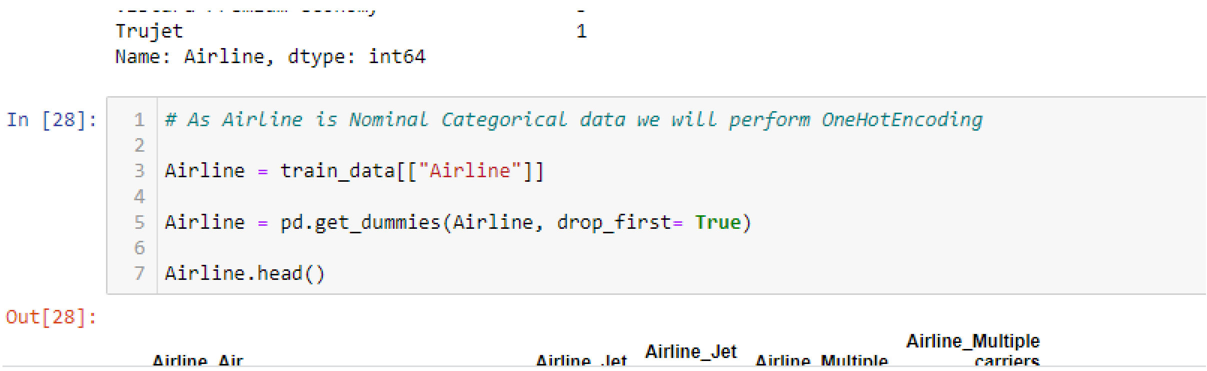
Data preprocessing is a data mining technique.It is used to convert the raw records in a beneficial and useful format. In the dataset, many attributes include the identical information. Directly merging the tables creates many replica fields. Also, the records pronounced via way of means of the airways can also additionally consist of faulty values because of human error, foreign money conversion error, etc. Hence, as it should be designed records preprocessing workflow is vital to generate correct enter records for you to build the machine learning model.

Feature Extraction goals to lessen the wide variety of functions or features in a dataset through developing new functions from the prevailing ones (after which discarding the unique functions). These new decreased set of functions must then be capable of summarize maximum of the statistics contained withinside the unique set of functions.

- Data Inputs- Logic- Output Relationships

Feature selection is the system of decreasing the wide variety of input variables while making a predictive model. It is ideal to lessen the wide variety of input variables to each lessen the computational value of modeling and, in a few cases, to enhance the overall performance of the model. The problem belongs to the Airline domain so booking ticket date of the flight and date of journey is more important, also flight duration is more impact on the out put

Any change in the flight duration and ticket booking and date of journey change in output (flight ticket price)



```
Trujet                                   1
Name: Airline, dtype: int64
```

In [28]:

```
# As Airline is Nominal Categorical data we will perform OneHotEncoding

Airline = train_data[["Airline"]]

Airline = pd.get_dummies(Airline, drop_first= True)

Airline.head()
```

Out[28]:

```
       Airline_Air                     Airline_Jet  Airline_Jet  Airline_Multiple  Airline_Multiple
                                                                                   carriers
```

- State the set of assumptions (if any) related to the problem under consideration

Algorithm used in our model is Random forest. Random forest is supervised machine learning algorithm. It is a collection of multiple decision trees whose results are aggregated into one final result. As the name suggests, "Random Forest is a classifier that contains a multiple number of decision trees on various subsets of the given dataset and takes the average to improve the predictive or final accuracy of that dataset.

Hardware and Software Requirements and Tools Used

Listing down the hardware and software requirements along with the tools, libraries and packages used. Describe all the software tools used along with a detailed description of tasks done with those tools.

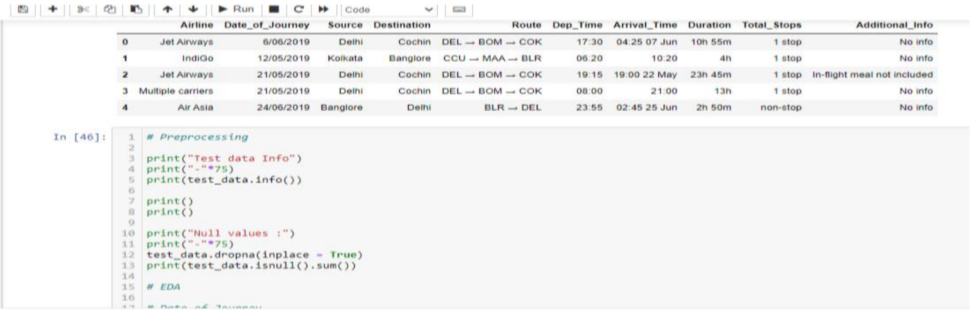
Hardware: Sony vaio laptop ,i5 processor , 4GB Ram

Software : window 10 , Jupyter notebook

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  These techniques are taking into consideration numerous financial, advertising, business and social elements intently linked with the very last airfare charges. It may be tough to betthe airfare price tag fee whilst. We test it nowadays as compared to the alternative day. The vacationers who need to go to a brand new location need to realize the price tag fee to be able to get the most inexpensive and positive price tag fee with their needs. This whole thing brings the concept to make a prediction approximately the flight tickets to be able to make the vacationers simpler to book their tickets with their needs. Due to the excessive complexity of the pricing fashions implemented with the aid of using the airlines, it's miles very tough a client to buy an air price tag with inside the lowest fee, for the reason that fee modifications dynamically.



	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info
0	Jet Airways	6/06/2019	Delhi	Cochin	DEL → BOM → COK	17:30	04:25 07 Jun	10h 55m	1 stop	No info
1	IndiGo	12/05/2019	Kolkata	Banglore	CCU → MAA → BLR	06:20	10:20	4h	1 stop	No info
2	Jet Airways	21/05/2019	Delhi	Cochin	DEL → BOM → COK	19:15	19:00 22 May	23h 45m	1 stop	In-flight meal not included
3	Multiple carriers	21/05/2019	Delhi	Cochin	DEL → BOM → COK	08:00	21:00	13h	1 stop	No info
4	Air Asia	24/06/2019	Banglore	Delhi	BLR → DEL	23:55	02:45 25 Jun	2h 50m	non-stop	No info

In [46]:

```
# Preprocessing

print("Test data Info")
print("-"*75)
print(test_data.info())

print()
print()

print("Null values :")
print("-"*75)
test_data.dropna(inplace = True)
print(test_data.isnull().sum())

# EDA

```

Testing of Identified Approaches (Algorithms)

While we go through the algorithms we employed (XGBoost, Random Forest, and Decision Tree) and also how they operate in our models, please read the discussion below.

A. *Decision Tree*

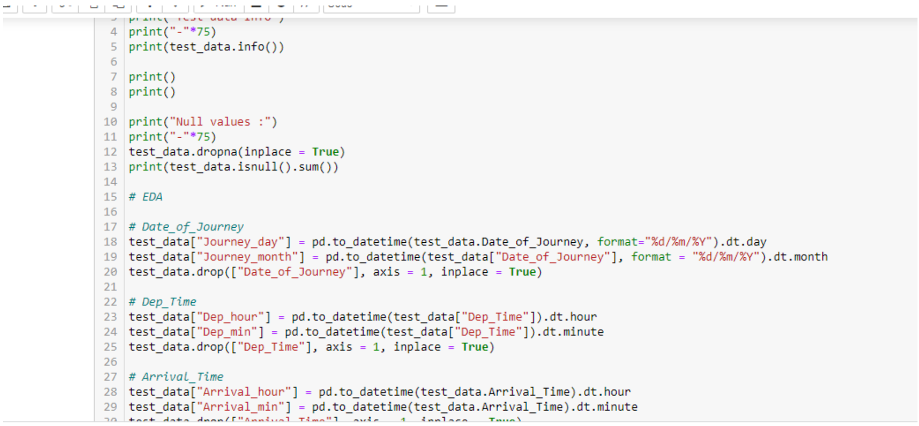
The decision tree appears to be the most well-known and commonly employed categorization technique. A decision tree is a collection of nodes that resembles a diagram, for each junction indicating a test on the a characteristic and each branch indicating a test outcome, such that each node in a decision tree (terminal node) has a class label. A tree can be "trained" by dividing the resources collection into subgroups depending on a characteristic values test. This procedure is known as partitioning the data because it is performed iteratively on each derived subset. The recursion ends when all subgroups at a node have the same posterior probability, or when the split no longer adds additional value to the predictions. A decision tree is appropriate for experimental extracting knowledge since it does not need subject matter expertise or parameters configuration.

A Random Forest is an ensemble approach that can handle simultaneous regression and classification problems by combining many decision trees using a technique known as Bootstrap as well as Aggregation, or bagging. The core idea is to use numerous decision trees to determine the final result instead of depending on personal decision trees. Random Forest's foundation learning methods are numerous decision trees. We arbitrarily choose rows and characteristics from the dataset to create sample datasets for each model. This section is known as Bootstrap. We simply have to understand the purity in our dataset, and we'll use that characteristic as the root of the tree which has the smallest impurity or, in other words, the smallest Gini index. Mathematically Gini index can be written as:

$$Gini\ Index = 1 - \sum_{i=1}^{n} (P_i)^2$$

$$= 1 - [(P_+)^2 + (P_-)^2]$$

- Run and Evaluate selected models



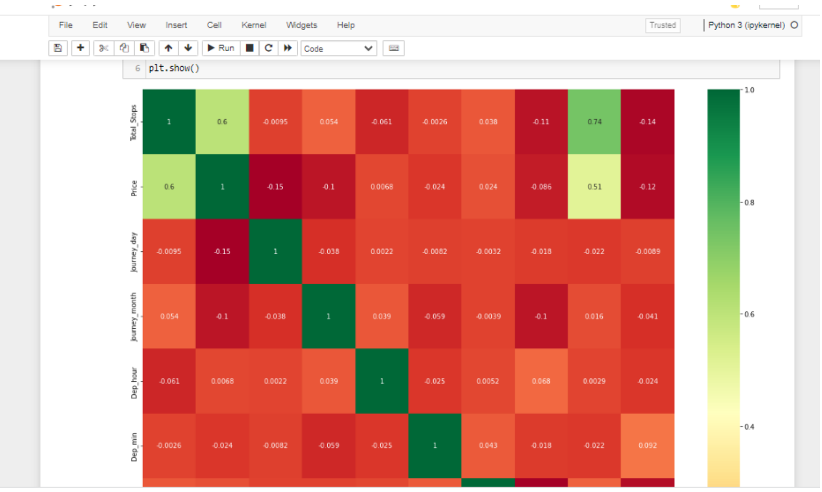
```
print("Test data info")
print("-"*75)
print(test_data.info())

print()
print()

print("Null values :")
print("-"*75)
test_data.dropna(inplace = True)
print(test_data.isnull().sum())

# EDA

# Date_of_Journey
test_data["Journey_day"] = pd.to_datetime(test_data.Date_of_Journey, format="%d/%m/%Y").dt.day
test_data["Journey_month"] = pd.to_datetime(test_data["Date_of_Journey"], format = "%d/%m/%Y").dt.month
test_data.drop(["Date_of_Journey"], axis = 1, inplace = True)

# Dep_Time
test_data["Dep_hour"] = pd.to_datetime(test_data["Dep_Time"]).dt.hour
test_data["Dep_min"] = pd.to_datetime(test_data["Dep_Time"]).dt.minute
test_data.drop(["Dep_Time"], axis = 1, inplace = True)

# Arrival_Time
test_data["Arrival_hour"] = pd.to_datetime(test_data.Arrival_Time).dt.hour
test_data["Arrival_min"] = pd.to_datetime(test_data.Arrival_Time).dt.minute
test_data.drop(["Arrival_Time"], axis = 1, inplace = True)
```

- Key Metrics for success in solving problem under consideration

  Airline, Source, Destination, Route, Total_Stops, Additional_info are the categorical variables we have in our data. Let’s handle each one by one.

  Nominal data → are not in any order → OneHotEncoder is used in this case

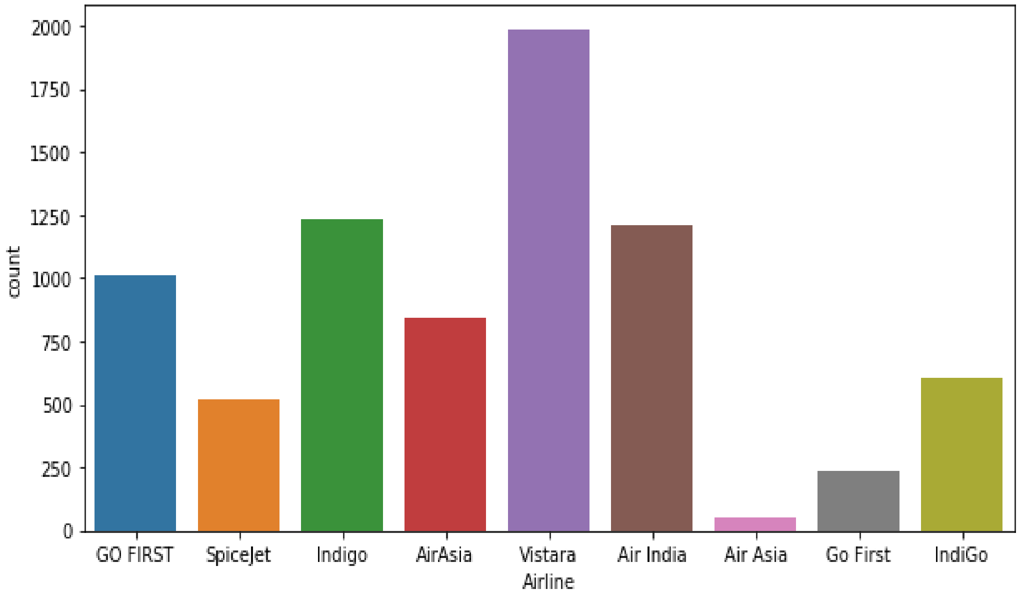
  Ordinal data → are in order → LabelEncoder is used in this case



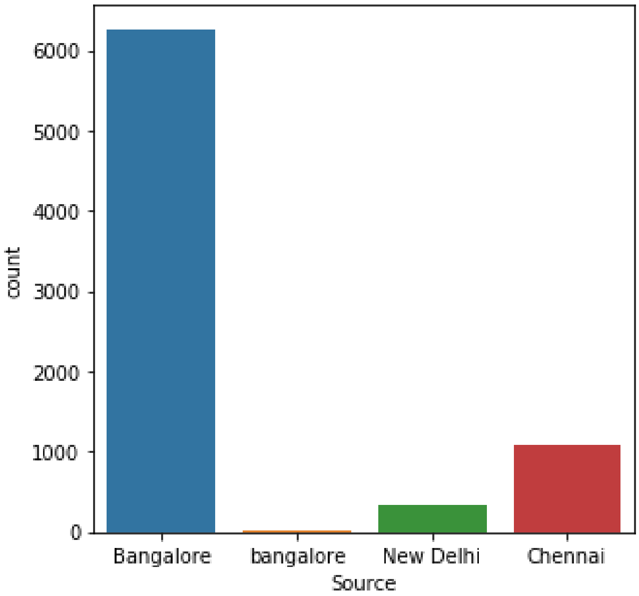
```
plt.show()
```

Visualizations

Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail.

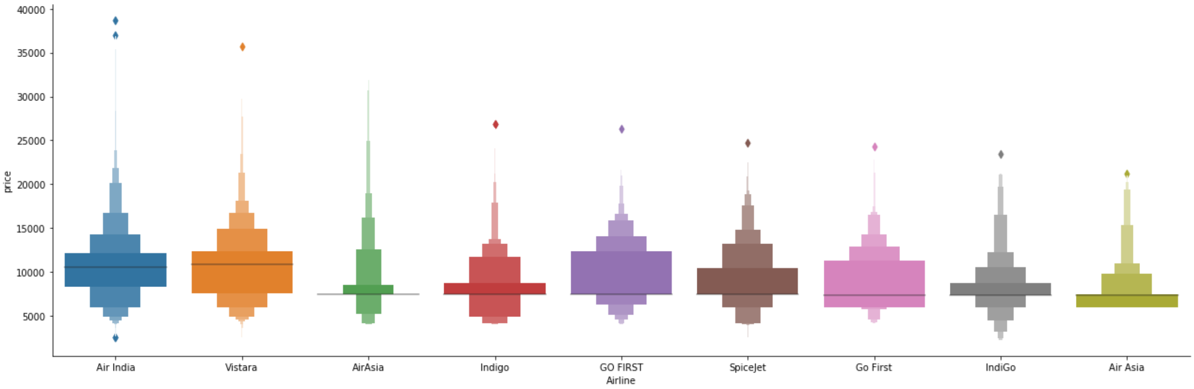


Observation: Airline vistara have highest number of flights

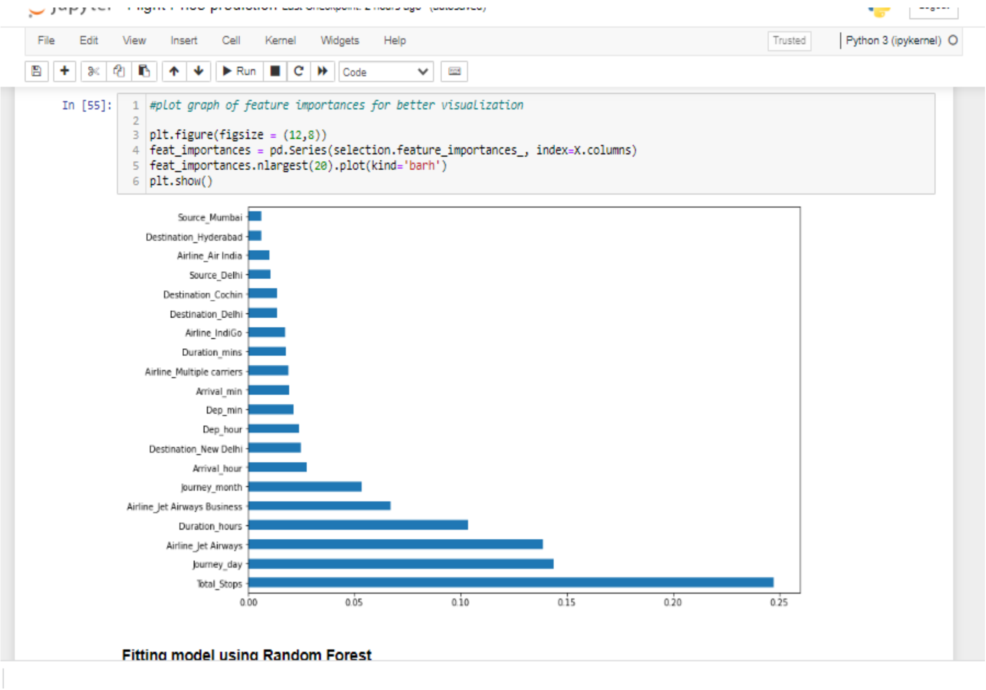


Observation : Data set contains the highest number of flight data from Bangalore

Catplot:

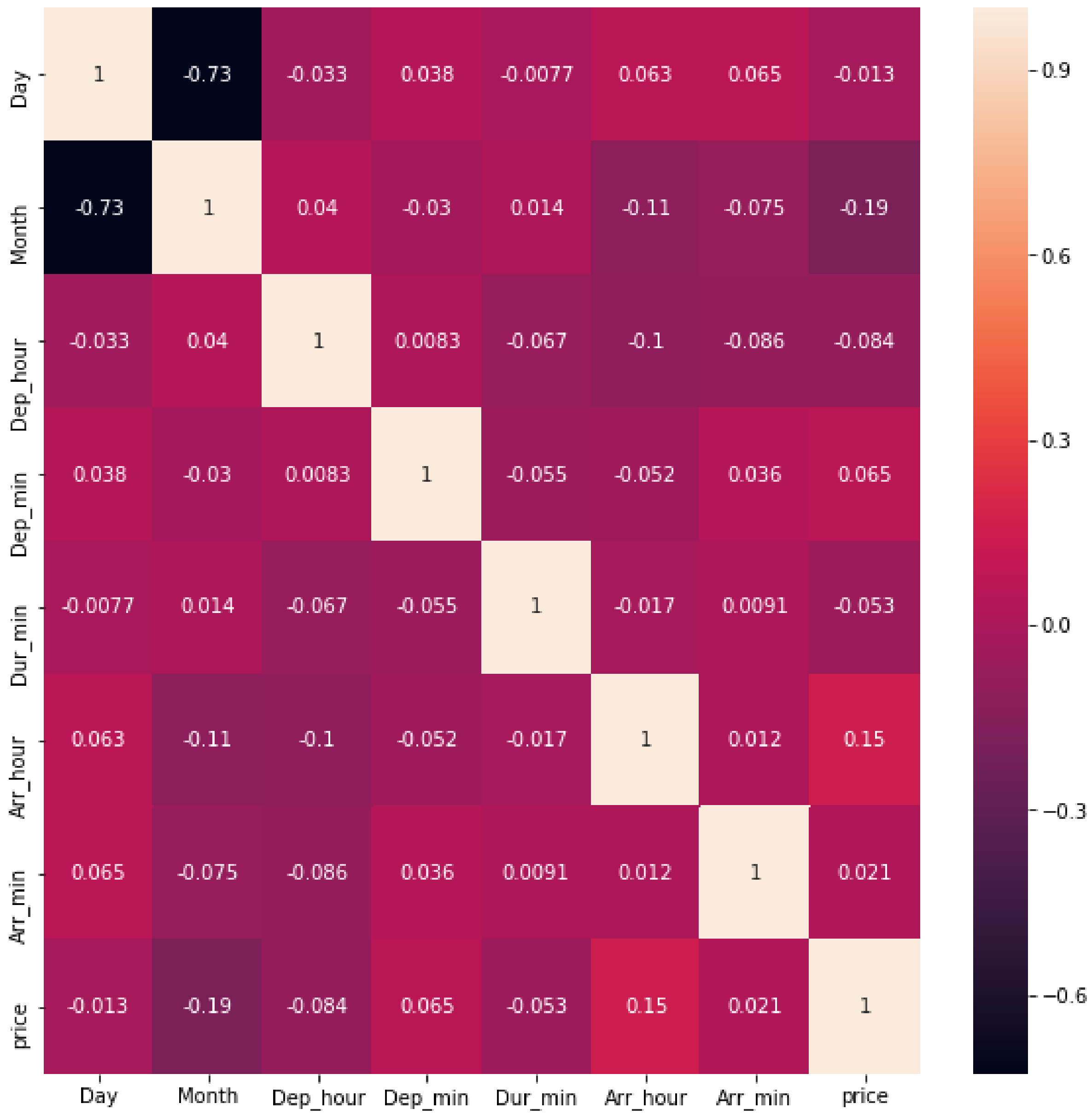


Observation: Air India have highest Air fare compare to other airlines and cheapest airline Air aisa

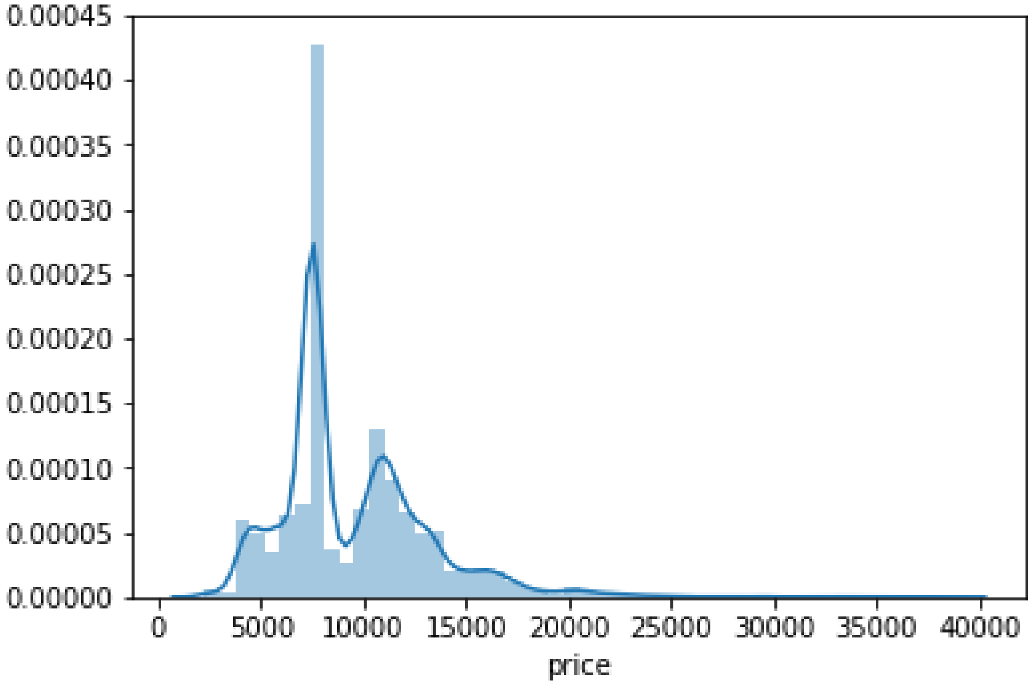


```
In [55]:
#plot graph of feature importances for better visualization

plt.figure(figsize = (12,8))
feat_importances = pd.Series(selection.feature_importances_, index=X.columns)
feat_importances.nlargest(20).plot(kind='barh')
plt.show()
```

Fitting model using Random Forest



Observation : Correlation of the heat map , Arrival hour and dur_hours have positive relation between Feature variable and target variables



Observation: Highly Price distribution happened between 5000 to 10000 Rupees

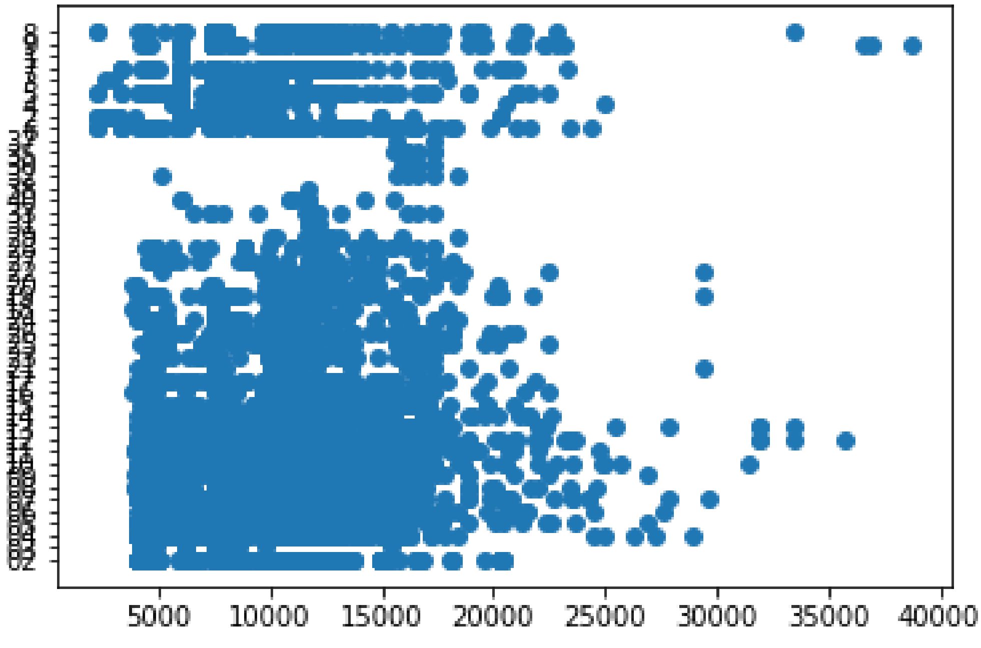


Fig: Price and Hour scatter plot

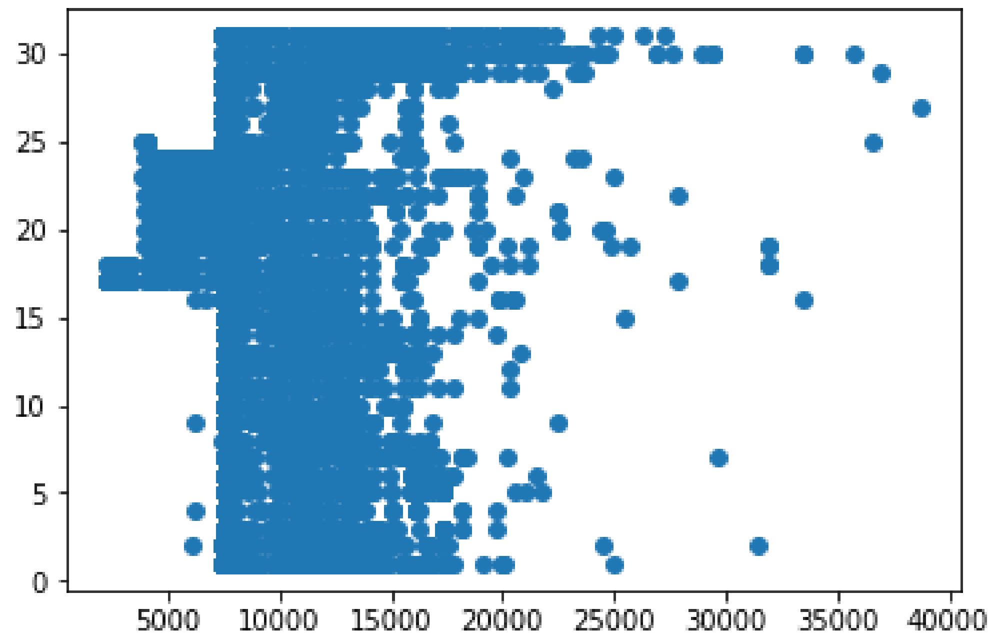
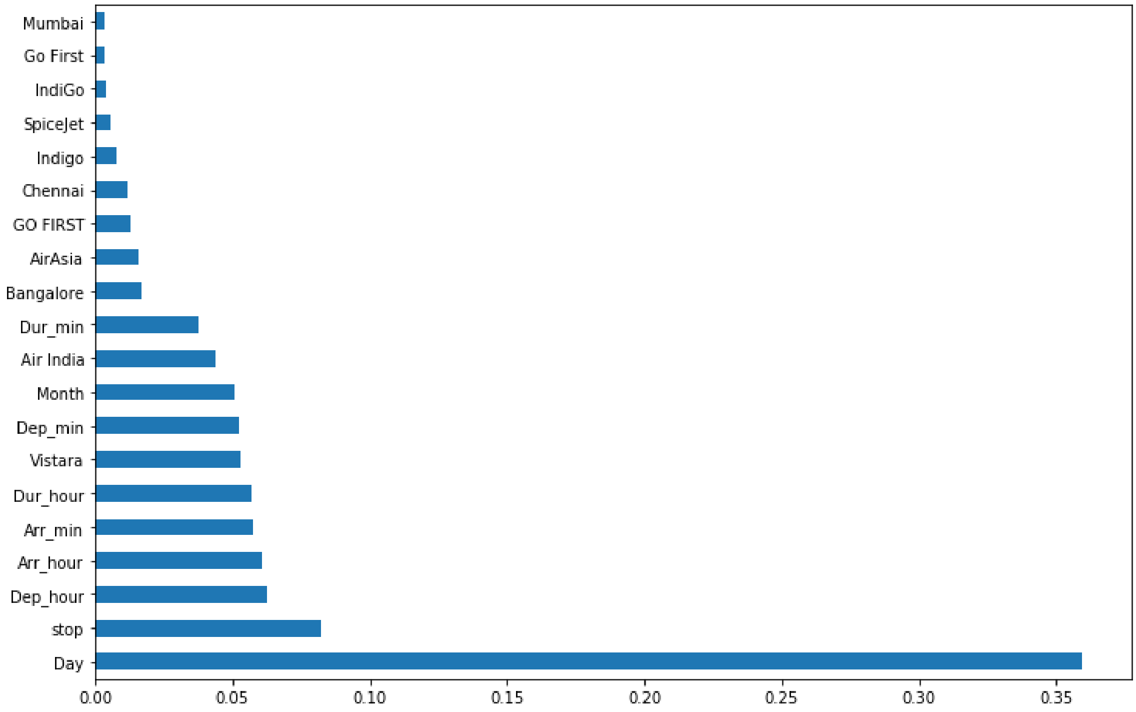


Fig: Price and Day scatter plot



Observation: Bar plots show the clearly feature variables Day and stop ,Dep_hour, dur_hour, Arr_hour have high correlation between the target variables

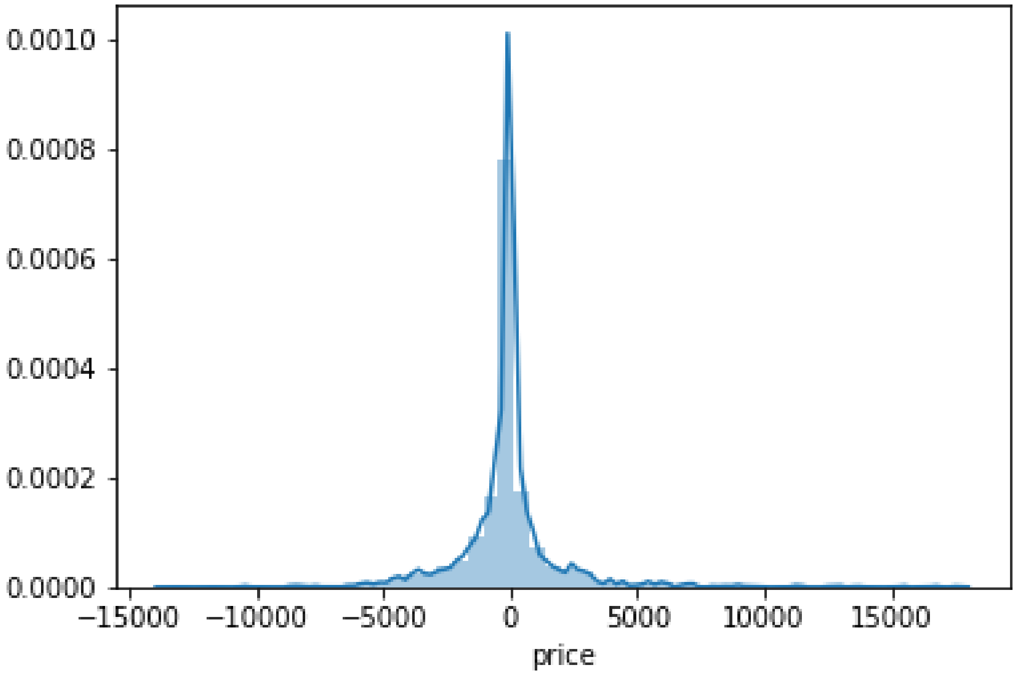


Fig: Test score of the y_test, and y_pred

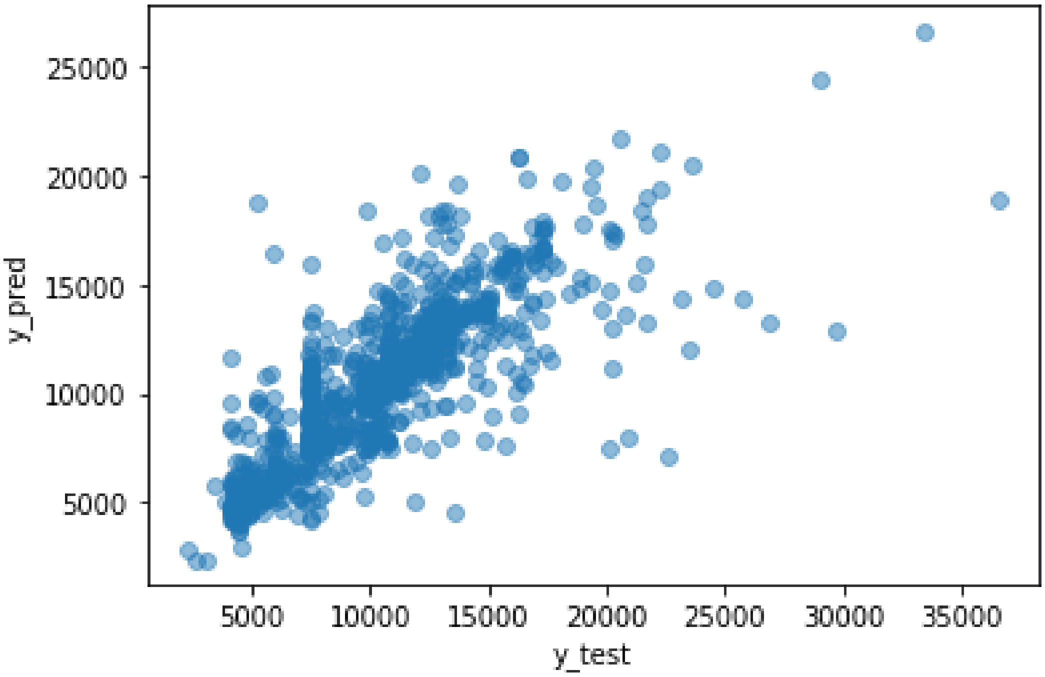


Fig: Scatter plot of y_test and y_pred

If different platforms were used, mention that as well.

- Interpretation of the Results

Give a summary of what results were interpreted from the visualizations, preprocessing and modelling.

CONCLUSION

- Key Findings and Conclusions of the Study

Three machine learning models were examined in this case study to forecast the average flight price at the business segment level. We used training data to train the training data and test data to test it. These records were used to extract a number of characteristics. Our suggested model can estimate the quarterly average flight price using attribute selection strategies.To the highest possible standard, much prior studies into flight price prediction using the large dataset depended on standard statistical approaches, which have their own limitations in terms of underlying issue estimates and hypotheses. To our knowledge, no other research have included statistics from holidays, celebrations, stock market price fluctuations, depression, fuel price, and socioeconomic information to estimate the air transport market sector; nonetheless, there are numerous restrictions.As example, neither of the databases provide precise information about ticket revenue, including such departing and arrival times and days of the week. This framework may be expanded in the future to also include airline tickets payment details, that can offer more detail about each area, such as timestamp of entry and exit, seat placement, covered auxiliary items, and so on. By merging such data, it is feasible to create a more robust and complete daily and even daily flight price forecast model. Furthermore, a huge surge of big commuters triggered by some unique events might alter flight costs in a market sector. Thus, incident data will be gathered from a variety of sources, including social media sites and media organizations, to supplement our forecasting models. We will also examine specific technological Models, such as Deeper Learning methods, meanwhile striving to enhance existing models by modifying their hyper-parameters to get the optimum design for airline price prediction.

Learning Outcomes of the Study in respect of Data Science

This paper reported on a preliminary study in “airfare prices prediction”. We gathered airfare data from a Kaggle website and showed that it is feasible to predict prices for flights based on historical fare data. The experimental results show that ML models are a satisfactory tool for predicting airfare prices. Other important factors in airfare prediction are the data collection and feature selection from which we drew some useful conclusions. From the experiments we concluded which features influence the airfare prediction at most. Limitations of this work and Scope for

Future Work

selection is the system of decreasing the wide variety of input variables while making a predictive model. It is ideal to lessen the wide variety of input variables to each lessen the computational value of modeling and, in a few cases, to enhance the overall performance of the model.