Data Analytics- Mini Project (UE20CS312)
Team Trios

Ashrita B Kumar	PES1UG20CS099
Aswin Shaiju	PES1UG20CS086
Shal Ritvik Sinha	PES1UG20CS717

LITERATURE REVIEW - FLIGHT FARE PREDICTION

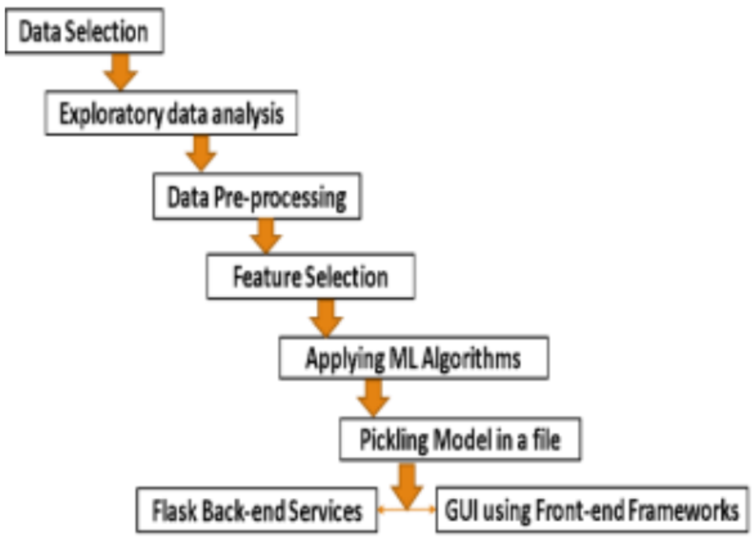
1) What have others done to solve this problem?

- Here, we will be analyzing the flight fare prediction using various Machine Learning datasets using essential data analysis techniques, then we will draw some predictions about the price of the flight based on various features. Techniques such as K- Nearest Neighbours and Support Vector Machines.

Some of the related work-

- Tianyi Wang proposed a framework where two databases are combined with macroeconomic data and machine learning algorithms such as a support vector machine, and XGBoost is used to model the average ticket price based on source and destination pairs. The framework achieves a high prediction accuracy of 0.869 with the adjusted R squared performance metric.
- In a survey paper by Supriya rajankar a survey on flight fare prediction using a machine learning algorithm uses a small dataset consisting of flights between Delhi and Bombay. Algorithms such as K-nearest neighbors (KNN), linear regression, and support vector machine (SVM) are applied.
- Research done by Santos analysis is done on airfare routes from Madrid to London, Frankfurt, New York, and Paris over a few months. The model provides the accepted number of days before buying the flight ticket.

The implementation is done in this format-



2) How have others solved a similar problem? Can we apply any of those solution strategies to the problem we have selected?
Exception: If you are working on a problem for which there is no ready precedent, but know the kind of approaches you want to use, then look for papers that talk about those approaches.

The detailed implementation of the model-

- **A basic web application that will predict the flight prices by applying a machine learning algorithm to historical flight data using python libraries like Pandas, NumPy, Matplotlib, seaborn, and sklearn.**
- **Our dataset consists of more than 10,000 records of data related to flights and their prices. Some of the features of the dataset are the source, destination, departure date, departure time, number of stops, arrival time, prices, and a few more.**

- **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 the distribution of data. We have to remove the redundant data and focus on the attributes that we are trying to extract from this,**

- **The 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 extracted from the date of the journey in integer format, and hours and minutes are extracted from departure time. Features such as source and destination need to be converted into values as they were of categorical type. This makes the classification and visualization of the data easier.**

- **The feature selection step is involved in selecting important features that are more correlated to the price. Random forest uses a group of decision models.**

- **After selecting the features which are more correlated to price the next step involves applying a machine algorithm and creating a model. As our dataset consists of labeled data, we will be using supervised machine learning algorithms and regression algorithms as our dataset contains continuous values in the features. Regression models are used to describe the relationship between dependent and independent variables.**

The algorithms used are- Decision Tree and Random Forest

Refine your problem statement

3) What is the specific problem we are going to solve?

This project aims to develop an application that will predict the flight prices for various flights using machine learning models. The user will get the predicted values and with its reference, the user can decide to book their tickets accordingly. In the current day scenario, flight companies try to manipulate flight ticket prices to maximize their profits. Many people travel regularly by flight and so they have an idea about the best deals that they can get at a given time. But many people are inexperienced in booking tickets and end up falling into discount traps made by the companies where they end up spending more than they should have. The proposed system can help save millions of rupees for customers by providing them with the information to book tickets at the right time. The proposed problem statement is “Flight Fare prediction system”.

4) What are the questions we are going to attempt to answer?

Proper implementation of this project can result in saving money for inexperienced people by providing them the information related to trends that flight prices follow and also give them a predicted value of the price which they use to decide whether to book a ticket whenever they decide to. In conclusion, this type of service can be implemented with good accuracy of prediction. As the predicted value is not fully accurate there is huge scope for improvement of these kinds of services.

5) What are the challenges with this data set (based on the initial exploratory analysis +coarse solution approach (trying library functions, etc., to build a simple model)

Additional info has a lot of no info. Some of the data in the Arrival_Time has the date mentioned while some don’t have it mentioned. This may lead to confusion.

6) What solution approaches would be reasonable to attempt?

With the amount of knowledge and the data we have, we can use the most optimal machine learning way of approaching this solution. The techniques that are easily feasible for this problem are the KNN algorithm and the decision tree algorithm. We also need to perform the performance metrics to compare the accuracy of the machine learning models implemented.

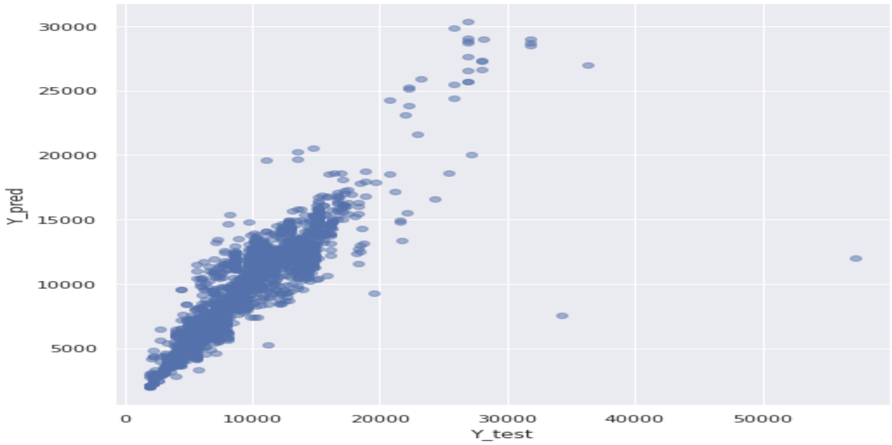
7) What is the use of solving this problem?

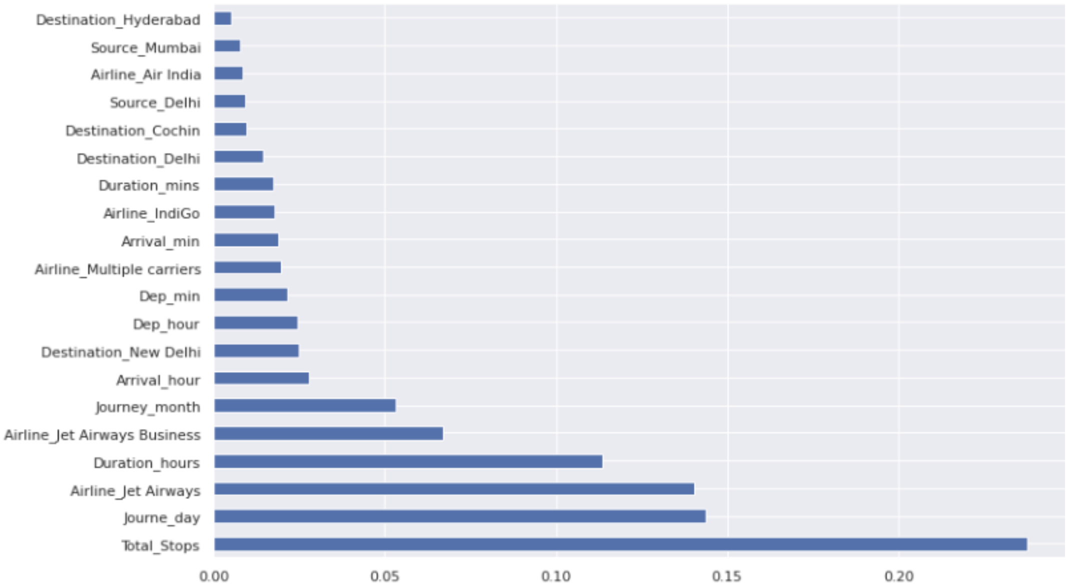
We can predict at what time we can find the lowest fares and also understand the trend.

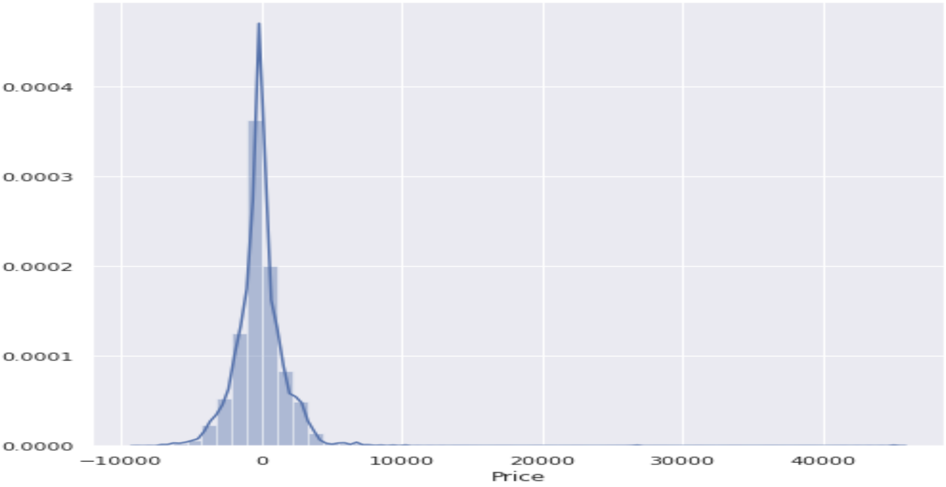
GitHub link-

https://github.com/AshritaKumar/Data-Analytics-Mini-Project--Team-Trios

Visualization


Y_pred
Y_test


Destination_Hyderabad
Source_Mumbai
Airline_Air India
Source_Delhi
Destination_Cochin
Destination_Delhi
Duration_mins
Airline_IndiGo
Arrival_min
Airline_Multiple carriers
Dep_min
Dep_hour
Destination_New Delhi
Arrival_hour
Journey_month
Airline_Jet Airways Business
Duration_hours
Airline_Jet Airways
Journe_day
Total_Stops


Price