Flight Fare Prediction

Aswin Shaiju
Computer Science and Engineering
PES University
Bengaluru, India
aswinshaiju2002@gmail.com

Ashrita B Kumar
Computer Science and Engineering
PES University
Bengaluru, India
ashritakum@gmail.com

Shal Ritvik Sinha
Computer Science and Engineering
PES University
Bengaluru,India
shalritvik2001@gmail.com

***Abstract*— The aim of this project is to help users predict fares for flights on particular days. This will help the users to book flights prior to their travelling date and at a much reasonable rate. Considering the amount of flights flying day in and day out from the country, it becomes difficult for people to figure the best and the most reasonable flight to their destination. In order to overcome this problem, we have used various machine learning models such as the KNN, SVM and Random Forest to make predictions about the price of the flight based on various features more accurately and efficiently.**

Keywords—SVM, Random Forest, KNN, ensemble learning, Hyperparameter Tuning,ExtraTreesRegressor

I. Introduction

KNN also called K- nearest neighbour is a supervised machine learning algorithm that can be used for classification and regression problems. K nearest neighbour is one of the simplest algorithms to learn. K nearest neighbour is non-parametric i,e. It does not make any assumptions for underlying data assumptions. K nearest neighbour is also termed as a lazy algorithm as it does not learn during the training phase rather it stores the data points but learns during the testing phase. It is a distance-based algorithm.Ease of Use. Flight fare prediction can be made easier with the application of SVM and has a higher accuracy rate.Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence the algorithm is termed as Support Vector Machine.

A. Random Forest Model

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning,** which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

B. What do we aim at solving

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 flights 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”.

II. IMPLEMENTATION OF THE MODEL

A 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.

.

A. Previous 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.

B.Proposed Solutions

Proposed Solution would be to get the model with better accuracy compared as suggested by our peers in their project. So, we have used KNN which gave us 81% accuracy.

Experimental Results

- In order to demonstrate the importance of each feature for airfare price prediction, we extracted the importance scores generated by the feature selection module.
- Considering the visualisations that we have done, we can observe the clear relation between the attributes that are essential for the solution that we are predicting.
- We have used visualisation methods such as Box Plots, Bar Graphs, Histograms, Heat Maps etc. in order to come up with accurate visualisations.
- We have successfully calculated the correlation between the required attributes and have derived an accurate R^2 value for our dataset.

III. Contribution of each team member

Each team member has contributed equally in this project. We all have divided our responsibilities equally amongst us

A. Team members:

- **Aswin Shaiju**

Onehot encoding, Hyperparameter Tuning using RandomizedSearchCV, Analysis of data by creating various plots.

- **Ashrita B Kumar**

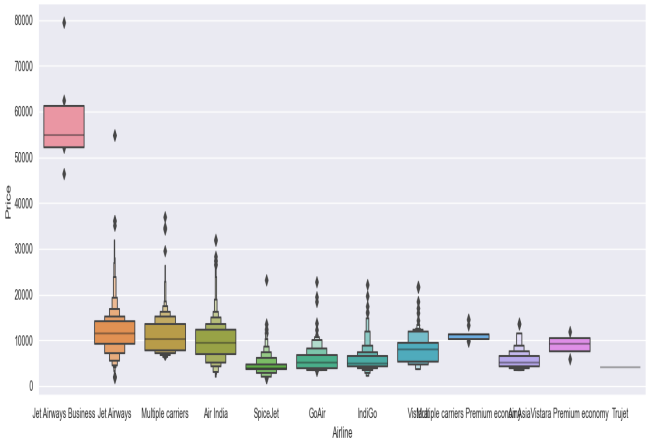
Prediction using visualisations, fitting randomforest model on Xtrain and Ytrain ,Feature Selection parameters, Documentation ,error fixing.

- **Shal Ritvik Sinha**

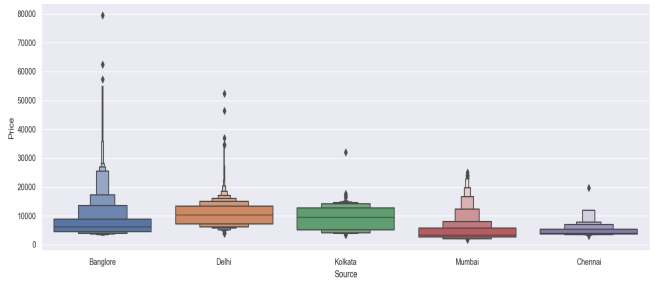
Preprocessing ,splitting the dependent and independent variables into training and testing data ,documentation ,finding correlation between independent and dependent variables,error fixing.

B. Figures and Tables

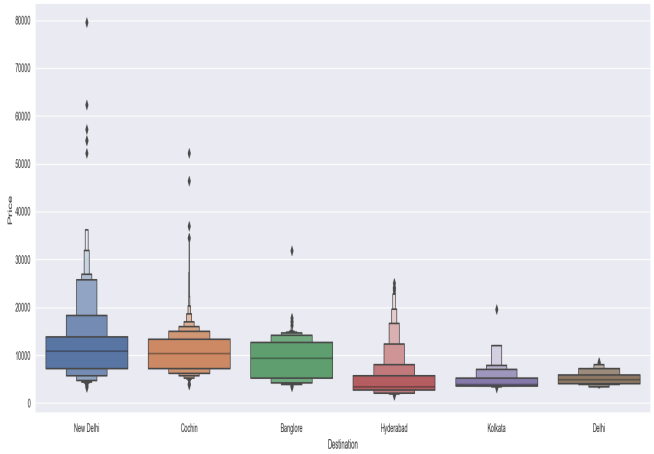
a) Positioning Figures and Tables:



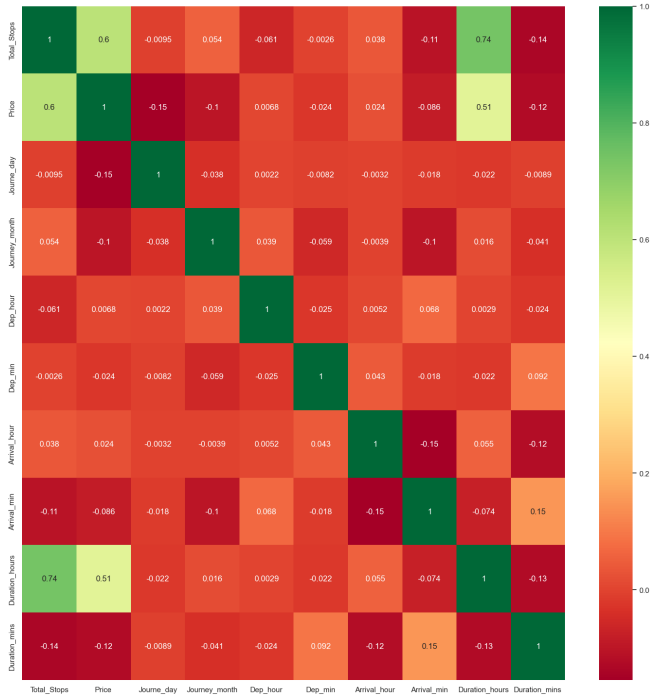
Category plot for Airline vs Price



Category plot for Source vs Price.

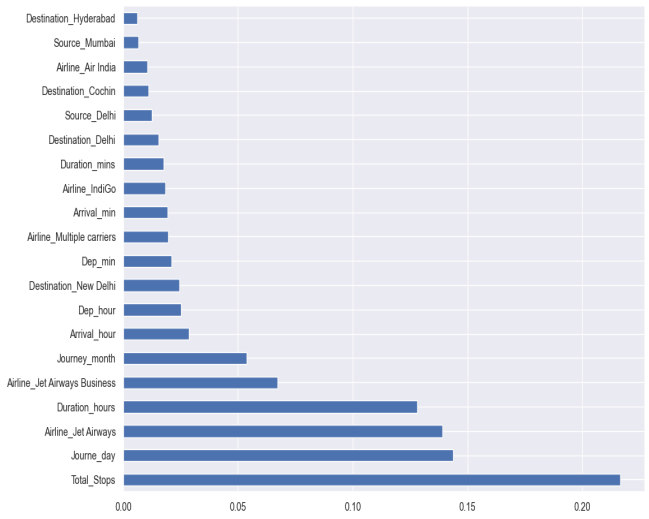


Category plot for Destination vs Price

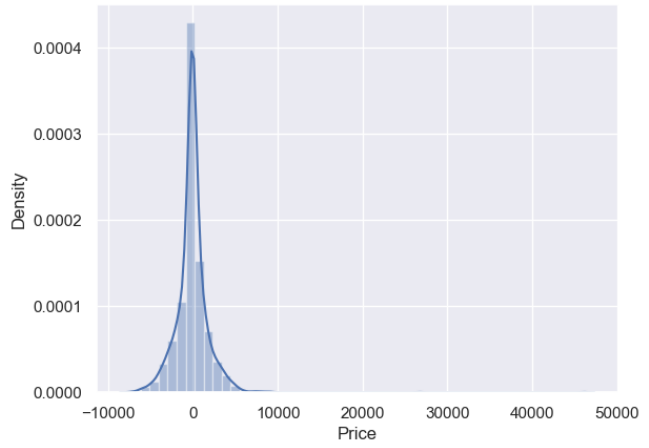


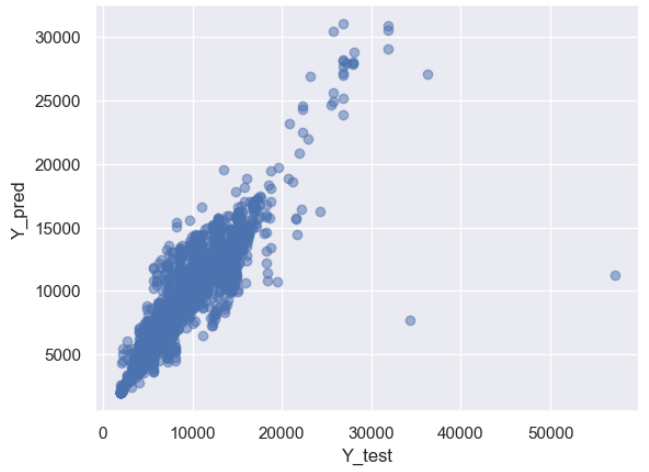
Heatmap for correlation between Independent and dependent attributes

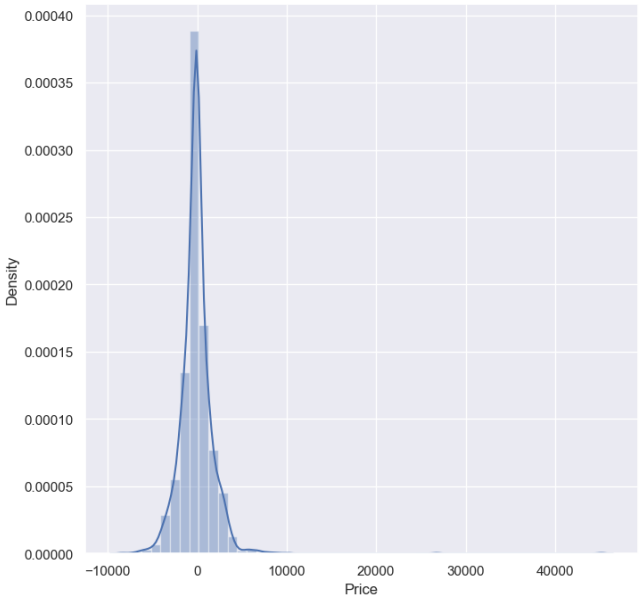
	Airline	Source	Destination	Route	Duration	Total_Stops	Additional_Info	Price	Journe_day	Journey_month	Dep_hour	Dep_min	Arrival_hour	Arrival_min
0	IndiGo	Banglore	New Delhi	BLR → DEL	2h 50m	non-stop	No info	3897	24	3	22	20	1	10
1	Air India	Kolkata	Banglore	CCU → BLR → BBI → BLR	7h 25m	2 stops	No info	7662	1	5	5	50	13	15
2	Jet Airways	Delhi	Cochin	DEL → LKO → BOM → COK	19h	2 stops	No info	13882	9	6	9	25	4	25
3	IndiGo	Kolkata	Banglore	CCU → NAG → BLR	5h 25m	1 stop	No info	6218	12	5	18	5	23	30
4	IndiGo	Banglore	New Delhi	BLR → NAG → DEL	4h 45m	1 stop	No info	13302	1	3	16	50	21	35

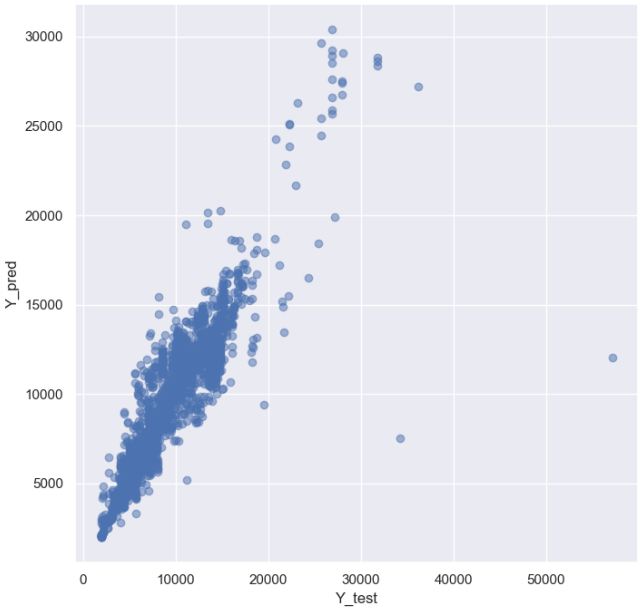


Graph of feature importances for better visualization









References

- https://www.analyticsvidhya.com/blog/2022/01/flight-fare-prediction-using-machine-learning/
- https://www.datascience2000.in/2021/12/flight-fare-prediction-using-machine.html?m=1
- https://360digitmg.com/machine-learning-in-aviation-flight-fare-prediction
- https://www.researchgate.net/publication/335936877_A_Framework_for_Airfare_Price_Prediction_A_Machine_Learning_Approach