Flight Price Prediction

Kosaraju Sai Chethan
Department of computer science
Amrita school of Engineering
Bangalore ,India
kschethan2003@gmail.com

N. Sri Lekhya
Department of computer science
Amrita school of Engineering
Bangalore ,India

M. Sree Deepika
Department of computer science
Amrita school of Engineering
Bangalore ,India

K. Vignaj Reddy
Department of computer science
Amrita school of Engineering
Bangalore ,India

***Abstract*— The Flight ticket prices increase or decrease every now and then depending on various factors like timing of the flights, destination, duration of flights. In the proposed system a predictive model will be created by applying machine learning algorithms to the collected historical data of flights. Optimal timing for airline ticket purchasing from the consumer's perspective is challenging principally because buyers have insufficient information for reasoning about future price movements. In this project we majorly targeted to uncover underlying trends of flight prices in India using historical data and also to suggest the best time to buy a flight ticket. The prediction of fares is done based on the route, month of departure, day of departure, time of departure. The paper demonstrated many regression methods: Linear Regression, Decision Tree, Random Forest, XGBoost(XGB), Lasso, Ridge, ElasticNet etc. To build the prediction models. Data preprocessing and feature selection steps were done before building the models. The models were evaluated based on the accuracy, mean absolute error, mean squared error, R2 Score. The random forest regression model performed best with 81.28% accuracy.**

Keywords—Flight Fare prediction, machine learning, Random Forest Regression, XGBoost, Decision Tree, Lasso Regression, Data preprocessing, Feature Selection.

1. INTRODUCTION

To avoid the effects of the most extreme charge, the recommended method for purchasing an airline ticket is to do so far in advance of the flight's departure. Most aviation routes disagree with this procedure. When they need to develop the market and when tickets are more difficult to get, airline companies may lower the price. They may raise prices to the maximum. Therefore, the price may depend on several aspects. This project uses AI to show future aircraft ticket prices in order to forecast costs. Every organisation has the right and ability to adjust the price of its tickets at any time. By purchasing a ticket at the lowest price, an explorer can lay aside money. People who frequently take flights are aware of pricing variations. Airlines implement several evaluating methods using complicated Revenue Management policies. As a result, the fee to modify the header or footer on subsequent pages varies based on the time, season, and holiday. While customers look for the lowest price, the airlines' primary goal is to make a profit. Customers typically aim to purchase their tickets far in advance of the departure date to prevent an increase in price as the departure date approaches. However, this is not the case in reality. For the same seat, the client can end up paying more than they should.

In this paper, we are proposing a machine learning approach to predict the flight ticket prices. This paper is organized asfollow: Section II has the related works where machine learning was used for Flight fare prediction. Section III explains the methodology, where the dataset is described, preprocessed, and split. The evaluation metrics selected to evaluate the performance of the model are described. Section IV discusses the experimental results. Lastly, Section V discusses the remarks and conclusions about this work are presented.

2. RELATED WORK

Flight Fare prediction was addressed in the literature using several methods. In this classification and regression methods random forest gives the best accuracy, i.e.,81.28%. Logistic Regression, KNN, NN, SVM, NB, Decision Tree, and RF, with some feature selection algorithms: Ridge, Extratree Regressor, and Lasso to predict the variation of Flight Fare diseasewith the same dataset used in this work. We have also used tablue to view the data visualization techniques. In this tool we can gethistograms, bar plots, box plots, pie charts etc.

3. METHODOLOGY

A. Data collection:

The dataset for our project was taken from Kaggle. The dataset contains a total of 10683 instances with 11 attributes as detailed in Table

	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662
2	Jet Airways	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops	No info	13882
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218
4	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302

Flight Fare DATASET DESCRIPTION

Data element	Description	Type	Range
Airline	-Name of the Airline	String	Jet Airways, IndiGo, Air India, etc
Date_of_Journey	Flight Journey date-	DateTime	Wide range
Source	From where the flights are starting	String	Delhi, Mumbai, Cochin, Kolkata, Bangalore
Destination	From where the flights are starting	String	Delhi, Mumbai, Cochin, Kolkata, Bangalore
Route	The route each fight takes	String	Wide range
Dep_Time	The time when the flight end it's journey	Nom	12 AM to 12 PM
Arrival_Time	The time when the flight start it's journey	Nom	12 AM to 12 PM
Duration	Total Duration of flights	Time	
Total_Stops	No. of stops each flight makes	Num	0, 1, 2, 3, 4
Additional_Info		Cat	No info
Price	The price taken by each flight to take different routes	Num	Wide range(2000 rupees to 60000 rupees)

[a] Numerical, [b] String, [c] Nominal, [t] Time

B. Data preprocessing

Data preprocessing is crucial because the quality of the data used to build a machine learning model heavily influences how well it performs. In addition to converting, data resampling, and performing feature selection, data pretreatment involves cleaning the data by eliminating outliers, corrupted or missing data points.

1). DATA VISUALIZATION AND CLEANING

Firstly, we checked if there are any missing values and none were found. Second, we checked for outliers and we found some as reported in Table.

List of outliers

Attributes	Outlier values
Price	62345, 2234, 2456, 61967

Here in our dataset the mild outliers contribute to the final diagnosis, only the extreme outliers were removed. The extreme outliers were detected using 1st and 3rd quartile range, where the IQR (interquartile range), and Q_1, Q_3 are the lower and upper quartiles respectively.

$$(75\% \times Q_3) + 3 \times IQR$$

$$(25\% \times Q_1) - 3 \times IQR$$

The data points that are greater than the first expression were removed. Similarly, the data points that are less than the second expression were removed.

2). CONVERTING CATEGORICAL DATA TO NUMERICAL DATA AND DATA PREPROCESSING

We converted Categorical data to numerical data using one hot encoding and converted some nominal and datetime data to usable data.

One Hot Encoding

	Destination_Cochin	Destination_Delhi	Destination_Hyderabad	Destination_Kolkata	Destination_New Delhi
0	0	0	0	0	1
1	0	0	0	0	0
2	1	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	1

Final Dataset after preprocessing has 30 columns

	Total_Stops	Journey_Day	Journey_Month	Dep_hour	Dep_Minute	Arrival_hour	Arrival_Minute	Duration_Hour	Duration_Minute
0	0	24	3	22	20	1	10	2	50
1	2	1	5	5	50	13	15	7	25
2	2	9	6	9	25	4	25	19	0
3	1	12	5	18	5	23	30	5	25
4	1	1	3	16	50	21	35	4	45

3). REMOVED NULL AND DUPLICATES FROM DATASET

Null Values

```
df.dropna(inplace=True)
df.isnull().sum()
```

```
Airline            0
Date_of_Journey    0
Source             0
Destination        0
Route              0
Dep_Time           0
Arrival_Time       0
Duration           0
Total_Stops        0
Additional_Info    0
Price              0
dtype: int64
```

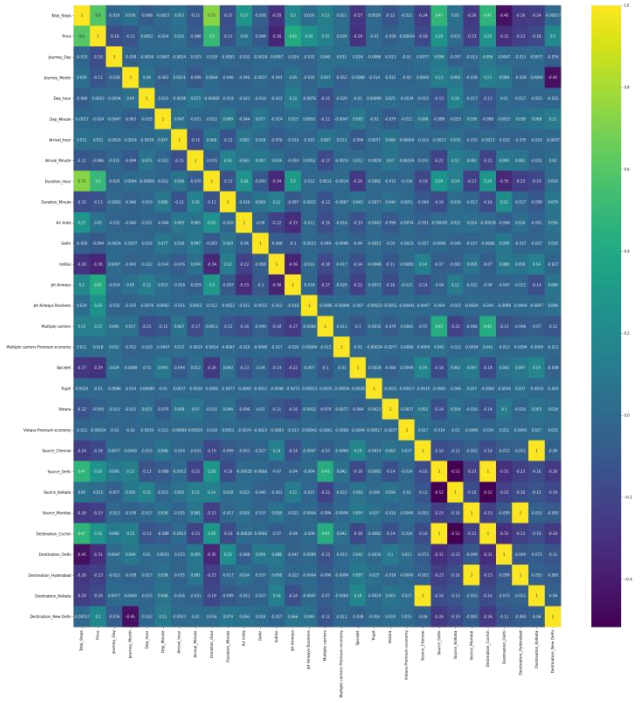
Duplicate Values

```
df.drop_duplicates(keep='first',inplace=True)
df.head()
```

	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662
2	Jet Airways	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops	No info	13882
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218
4	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302

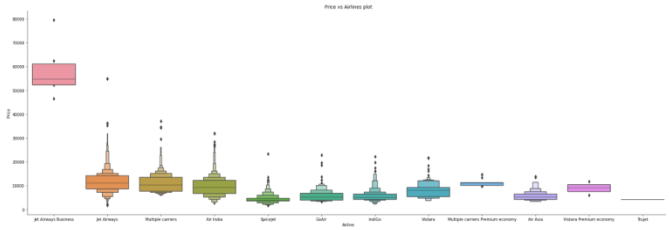
correlation coefficient matrix:

The correlation coefficient matrix was obtained to observe the relation between the different attributes and the output. In the below figure illustrates the correlation matrix where the coefficient indicates both the strength relationship between the variables as well as the direction (whether it is a positive or negative correlation).

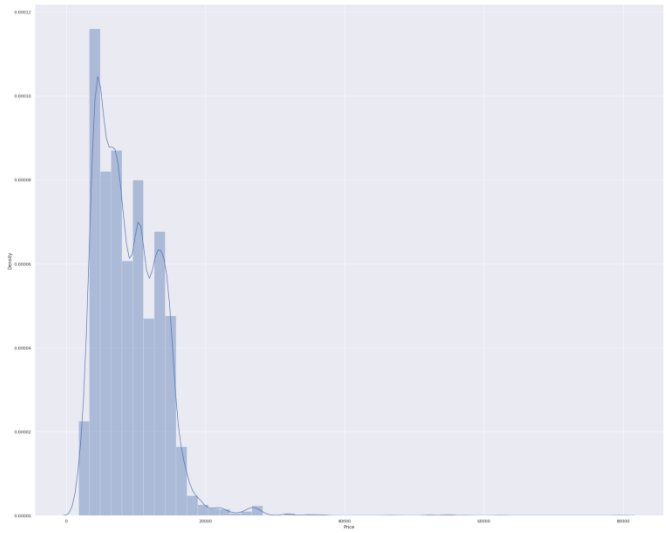


1) Checking Imbalance

Imbalance in the output can distort the prediction accuracy. Therefore, the balance of the output “target” was verified as shown in Figure. After inspection, the data turned out to be balanced with a 9:11 ratio between the two categories. Thus, there was no need to resample the data.



Box plot for Airline vs Price



Distplot for Price

2) DATA TRANSFORMATION

Transformation is applied when the dataset includes data of different formats, or when different datasets are combined. In this case, the categorical data is transformed to numerical data using one hot encoding

	Destination_Cochin	Destination_Delhi	Destination_Hyderabad	Destination_Kolkata	Destination_New Delhi
0	0	0	0	0	1
1	0	0	0	0	0
2	1	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	1

3) DIMENSIONALITY REDUCTION

In machine learning, dimensionality reduction refers to the process of reducing the number of features to decrease the complexity and prevent overfitting, by either feature selection or extraction.

Weka software was used to pick features since it offers a variety of attribute evaluator options that may be tested and used.Feature extraction, which differs slightly from feature selection, is the process of generating a new set of features from the initial set. There is a lot of usage for principal component analysis (PCA). It determines how to project the original data into a space with less dimensions.

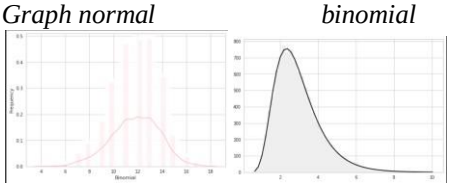
4) DATA SPLITTING

In machine learning, the data is usually split into trainingand testing sets, where the training set is used to train themodel, and the testing set is to test it and predict the output. Hold-out was used in this work with 80% of the data used in training and 20% used for testing.

5) Distributions

We performed normal distribution, Poisson distribution, Bernoulli distribution and binomial distribution.

Graph normal *binomial*



C) Evaluation Metrics

Evaluation metrics are used to test the quality and performance of the machine learning model. In this paper,the best model was chosen based on the following evaluation metrics.

Mean Squared Error: The Mean Squared Error measures how close a regression line is to a set of data points. It is a risk function corresponding to the expected value of the squared error loss. Mean square error is calculated by taking the average, specifically the mean, of errors squared from data as it relates to a function

Mean Absolute Error: Mean Absolute Error is a model evaluation metric used with regression models. The mean absolute error of a model with respect to a test set is the mean of the absolute values of the individual prediction errors on over all instances in the test set.

Root Mean Squared Error: Root mean square error or root mean square deviation is one of the most commonly used measures for evaluating the quality predictions fall from measured true values using Euclidean distance.

R2 Score: Coefficient of determination also called as R2 score is used to evaluate the performance of a linear

For Our model Random Forest Regression gave better results for MES, MAS, and R2 Score.

4. Results and Discussion

These are few selected machine learning techniques were used to build the heart disease prediction model, and the results were obtained after data cleaning. We are future considering for male person getting heart disease arelisted in TABLE III.

Model Metrics	Random forest	Decision Tree	Lasso	NB	KNN	XG Boost	DTR	
Accuracy	0.23	0.51	0.54	0.45	0.81	0.78	0.56	
MAE	0.68	0.86	0.00	0.76	0.71	0.73	0.70	
R2 Score	0.66	0.78	0.00	0.78	0.73	0.73	0.68	

ML Algorithm	MAE	MSE	RMSE
Random Forest	1180.05	4358748.36	2087.76
HyperParameter Tuning Random Forest	1278.10	4277590.50	2068.23
XGBoost	1180.05	4358748.36	2087.76

Here we compared Random Forest and XG Boost.

Proposed System:

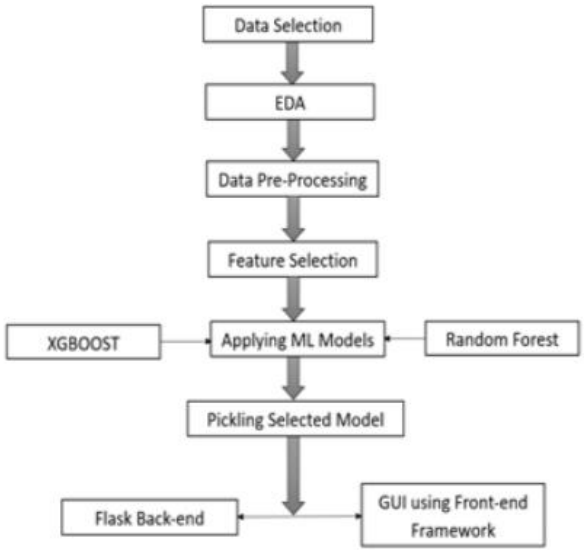


Fig. Proposed System Diagram

IV. CONCLUSION AND FUTURE WORK

Machine Learning algorithms are applied on the dataset to predict the dynamic fare of flights. This gives the predicted values of flight fare to get a flight ticket at minimum cost. The values of R-squared obtained from the algorithm give the accuracy of the model. In the future, if more data could be accessed such as the current availability of seats, the predicted results will be more accurate. Finally, we conclude that this methodology is not preferred for performing this project. We can add more methods, more data for more accurate results.

References

[1] *Tianyi wang, samira Pouyanfar, haiman Tian and Yudong Tao "A Framework for airline price prediction: A machine learning approach"*

[2] *Supriya Rajankar, Neha sakhrakar and Omprakash rajankar "Flight fare prediction using machine learning algorithms" International journal of Engineering Research and Technology (IJERT) June 2019.*

[3] *J. Santos Dominguez-Menchero, Javier Rivera and Emilio Torres Manzanera "Optimal purchase timing in the airline market"..*