Flight Ticket Price Prediction using Machine Learning

Midhun Krishna
Department of Computer Applications
Amal Jyothi College of Engineering Kanjirappally, India
midhunkrishna2022@mca.ajce.in

Ms Jetty Benjamin
Department of Computer Applications
Amal Jyothi College of Engineering Kanjirappally, India
jettybenjamin@amaljyothi.ac.in

***Abstract* — The airline ticket price changes very quickly these days, and the difference is huge. It can vary even within a few hours for the same flight. For business purposes, many airlines change fares according to seasons or duration of time. Airlines use a variety of calculation methods to increase their profits, for example, separating demand between expectations and value. Each carrier uses its own set of criteria and algorithms to determine the price. Machine learning, artificial intelligence, and deep learning are all clear and important tools. In a particular amount of time, it is possible to obtain the amount of air travel expenses. In this paper, we use machine learning algorithms. KNN, Random Forest, gradient-enhanced regression, SVR, and linear regression are examples of algorithms. Provide basic information such as airline, source, destination, route, total stops, and so on to forecast flight expenses.**

Keywords—Price, Flight, Regressor, Prediction, Accuracy, Random Forest, Machine Learning

I.INTRODUCTION

Nowadays, the cost of a carrier ticket can change significantly and essentially on the same plane, in any case, near the seats within one cabin. Customers try to get very low cost while carriers try to keep their income high as expected and growing. their earnings. Aircraft organizations can reduce the cost in the time required to build a market, making access to tickets difficult. You can increase the cost. Therefore, the cost can depend on various factors. People who travel a lot by plane are aware of price fluctuations. Airlines operate different rating systems using complex revenue management guidelines. This paper highlights a Flight fare prediction system based on machine learning that uses KNN, RandomForest, GradientBoostingRegression, SVR and Linear Regression algorithm to estimate airline ticket prices and analyze this data set using machine learning techniques in order to anticipate the price of an airline ticket based on the columns data set's features

II.LITERATURE REVIEW

K. Tziridis, Th. Kalampokas, et.al in [1] created a technique for predicting airline ticket prices. The authors begin with some basic machine learning material before moving on to the approach, which comprises four distinct steps of feature selection that influence flight costs, data collecting from Greek Aegean Airlines, model selection, and evaluation. Many of the following characteristics were included in the airline dataset. Eight state-of-the-art regression Machine Learning models were employed to forecast: MLP, GRNN, ELM, Random Forest Regression Tree, Regression Tree, Bagging Tree, Regression SVM, and Linear Regression. These machine learning models' outcomes were also compared and analyzed. Other regression tree methods are outperformed by the Bagging Regression Tree model.

Tianyi Wang, Samira Pouyanfar, et. al in [2] states using a Machine Learning technique, the problem of market segment level is stated. DBIB and T-100, two available datasets with basic features, were acquired for training and evaluation of the proposed model. Data cleaning, data transformation, data preprocessing, feature selection, and ML model deployment are all part of the methodology. The Random Forest Model is utilized for development since it outperforms other models, such as LR SVM and Neural Networks, in terms of data performance. With a R squared score of 0.869, this prediction framework has a good level of accuracy.

Gini and Groves [3] For creating a model for predicting the best purchase time for aircraft tickets, we used the Partial Least Square Regression (PLSR). From February 22nd to June 23rd, 2011, data was collected from trip booking websites.

Wohlfarth proposed a ticket-buying speed-up model [4] that relied on a novel pre-processing step called macked point processors and information mining systems, as well as a measurable research technique. This system's purpose is to convert heterogeneous value arrangement input into added value arrangement direction using unsupervised grouping computations.

Supriya Rajankar, Neha Sakharkar, [5] proposed Methods for forecasting the price of an airline ticket at a certain point in time using Machine Learning Regression. The study begins with data gathering, which was done via makemytrip.com. The date of departure, time of departure, place of departure, time of arrival, place of destination, airlines, and total fare are the seven components of this dataset. The data is then cleansed and pre-processed before being analysed with a variety of AI models. The authors analyse the performance of numerous Machine Learning models on data, including LR, Decision Tree, SVM, KNN, and Random Forest, and find that KNN produces R-squared values close to 1, indicating great accuracy.

DOI: 10.5281/zenodo.6906451
ISBN: 978-93-5607-317-3@2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

III.MOTIVATION

Everyone knows that the holidays are a time when people are looking for a much-needed vacation, and that finishing a trip can be a difficult effort. As a result of the global emergence of the Internet and E-commerce, the commercial aviation industry has experienced amazing growth and has become a controlled market. Customers often try to buy the ticket properly before the day of departure to avoid rising airfare as the day is near. But in reality this is not true. The customer may end up giving more than they should same seat. For the provided model, regression analysis visualization and various techniques are used. This model assists the user in accurately predicting the price of an airline ticket. This model will help the common man to easily predict the future fare of plane tickets.

IV.METHODOLOGY

The goal of this work aims to use the provided dataset to create a Machine Learning model that can accurately anticipate the price of a plane ticket. There are two training and testing data sets in the dataset. To increase learning accuracy, the model should be trained with more data. This model's output can be used to forecast airline ticket prices. The ticket prices are forecasted using the KNN algorithm, Random Forest, Gradient Boosting Regression, SVR, and Linear Regression.The structure is:

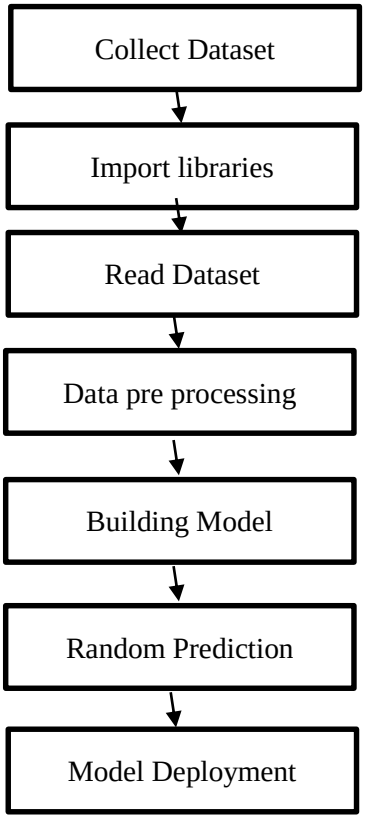


Figure 1

The steps that require to be followed are:

1. Data Collection
2. Data Pre-processing
3. Model Building
4. Analyzing
5. Result

A. **Data Collection:** The training and testing datasets were from the Kaggle data pool. They now include both category and nominal information for Indian Airlines as of 2019. The dataset contains crucial information about some of the elements that determine flight pricing, such as departures and arrivals, time of departure and arrival, flight path, number of halts along the way, and ticket price based on those variables, all of which are used to anticipate flight pricing. There are 10683 rows and 11 columns in this massive dataset (each representing one attribute)

B. **Data Pre-processing:** This is the first stage in any machine learning algorithm. Data cleansing, data transformation, and data minimization are all part of this process. All of this is done to improve the data's effectiveness. The data can be analyzed to improve the accuracy of our model. In order for the categorization to be correct.

a. Cleaning Data – In the training dataset, the null values were deleted. Because they were unnecessary for the feature selection technique, a few columns in the dataset were eliminated. After the new columns with numerical values derived from the preprocessed data were stored for the prediction, the columns of attributes with categorical data were removed from the dataset. As a result, an appropriate training dataset with the following attribute columns was obtained.

b. Formatting the Data – We add a new column week day 1 mean week day 0 mean weekend while pre-processing the data. Format the arrival and departure times, and add an extra two columns to indicate whether the flight is taking place at night or early in the morning. Some flights are less expensive early in the morning and more expensive late at night, indicating a clear correlation. Converting length hour and minute into separate columns is also done, as well as labelling and encoding to convert category data to unique int values.

c. Splitting of Data – After formatting the data, the data is then split into training and testing datasets. After this the data chosen for training is used to train our model.

C. **Machine Learning:** This is used to help the user to anticipate the price of an aeroplane ticket with the greatest degree of precision The machine learning algorithms are used to predict fares , that will use the dataset given. There are different learning algorithm used to predict the airfares. The machine learning algorithms relies on how it is trained. Which algorithm

DOI: 10.5281/zenodo.6906451
ISBN: 978-93-5607-317-3@2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

works best depends on the type of problem you are solving, the computer resources available, and the type of data.

1. **Linear Regression -** Linear regression is a supervised learning machine learning algorithm. It carries out a regression task. Regression models a goal prediction value based on independent variables. It's generally used for forecasting and figuring out how variables are related. Different regression models have different types of relationships between dependent and independent variables. Gradient descent and cost function are the two most important factors in comprehending linear regression. The equation for linear regression is : y(pred) = b0+b1 ∗ x
2. **Support Vector Regression (SVR) –** SVR (Support Vector Regression) is a regression approach that works in the same way as SVM. It's a type of Machine Learning model that's used to solve classification problems or sort data into categories. The R2 score is used to evaluate the performance of the regression model.
3. **K-Neighbors Regressor –** The k-nearest neighbours approach is used to perform regression. Response regression might be scalar, multivariate, or functional. Local interpolation of the targets associated with the training set's nearest neighbours is used to predict the target.. The coefficient of determination, often known as the R2 score, is used to assess the regression model's performance. The independent variation of the input can be used to forecast the value of the difference in the output-based characteristic.
4. **Random Forest Regressor -** Random Forest is an useful machine learning technique for a variety of tasks, including regression and classification. A random forest model is made up of many little decision trees called estimators, each of which produces its own predictions.
5. **Gradient boosting Regression -** The GBR uses regression to calculate the difference between the current forecast and the known correct target value. "Residual" is the term for this disparity. Gradient boosting regression is then used to train the weak model that translates features to the residual. Gradient boost is a technique for forecasting a continuous number.

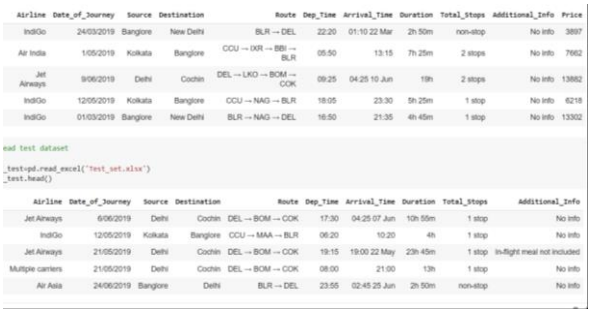
V.BUILD MODEL

The model building is the main step in the Flight Price Prediction. While building the model user use the algorithms

1. Import the packages that are necessary.

```
#import libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

2. Add the data into a Data Frame, then get the shape of data.



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

```
ead test dataset

_test=pd.read_excel("test_set.xlsx")
_test.head()
```

Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info
Jet Airways	6/06/2019	Delhi	Cochin	DEL → BOM → COK	17:30	04:25 07 Jun	10h 55m	1 stop	No info
IndiGo	12/05/2019	Kolkata	Banglore	CCU → MAA → BLR	06:20	10:20	4h	1 stop	No info
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
Multiple carriers	21/05/2019	Delhi	Cochin	DEL → BOM → COK	08:00	21:00	13h	1 stop	No info
Air Asia	24/06/2019	Banglore	Delhi	BLR → DEL	23:55	02:45 25 Jun	2h 50m	non-stop	No info

3. Then split the dataset into training and testing datasets.

```
X=df_train.drop(columns=['Price'])
y=df_train['Price']

from sklearn.model_selection import train_test_split

X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2,random_state=10)
X_train.shape,X_test.shape,y_train.shape,y_test.shape
```

```
((8544, 26), (2137, 26), (8544,), (2137,))
```

4. Cross Validate to validating the model efficiency by training it on the subset of input data

```
#cross validation
from sklearn.model_selection import cross_val_score

#function declaration
def cross_validation(reg_model,X,y):

    score=cross_val_score(reg_model,X,y,scoring='neg_mean_squared_error',cv=10)
    rmse_score=np.sqrt(-score)
    print("\nScores ",rmse_score)
    print("Mean ",rmse_score.mean())
    print("Standard Deviation ",rmse_score.std())
```

5. Use Different Algorithms to find R-square, MSE and MAE values which helps to find Accuracy.

Machine Learning(ML	R squared	MAE	MSE
Kneighbours Regressor	2468.73705	1438.629	6094662.62985
Random forest Regressor	1573.08845	635.6535	2474607.2856
Gradient Boosting Regressor	2466.54433	1563.642	6083830.9535

6. Then identify on test set and calculate the accuracy of the model.

VI.RESULT

The result shows that the table represents study of Price of Tickets and also the prediction of results. The outcomes obtained by the analysis are KN Regressor, Random Forest Regressor, Gradient Boosting Regression, SVR, and Linear Regression. Along with R-square, MSE, and MAE values, the algorithm's accuracy is improved.

DOI: 10.5281/zenodo.6906451
ISBN: 978-93-5607-317-3@2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

Random Forest Regressor

	Price	Price Predicted
2389	6224	8608.6700
5411	14151	13566.0000
2674	10539	10602.4700
970	7934	7008.7500
5845	16754	11771.9600
10268	18799	18363.4875
3476	11522	11368.3300
9502	4823	4823.0000
4033	4030	4826.5000
121	3100	3178.9500

```
Mean Absolute Error: 635.6535495932991
Mean Squared Error: 2474607.285600617
Root Mean Squared Error: 1573.0884544743874
Accuracy: 92.66 %.
```

K Neighbors Regressor

	Price	Price Predicted
2389	6224	10592.823816
5411	14151	13426.959618
2674	10539	10592.823816
970	7934	7408.579375
5845	16754	7968.360124
10268	18799	13986.740368
3476	11522	11257.356400
9502	4823	4506.379680
4033	4030	5009.452588
121	3100	3640.613627

```
Mean Absolute Error: 1563.6427528665088
Mean Squared Error: 6083840.953638745
Root Mean Squared Error: 2466.5443344158125
Accuracy: 80.88 %.
```

Gradient Boosting Regressor

	Price	Price Predicted
2389	6224	10418.725791
5411	14151	11027.535142
2674	10539	10903.094636
970	7934	10706.540060
5845	16754	11234.119814
10268	18799	12760.863783
3476	11522	9931.209000
9502	4823	3741.172839
4033	4030	5321.065318
121	3100	4955.756596

```
Mean Absolute Error: 1438.6298549368273
Mean Squared Error: 6094662.629854937
Root Mean Squared Error: 2468.737051582233
Accuracy: 83.76 %.
```

SVR

	Price	Price Predicted
2389	6224	9701.541982
5411	14151	11490.940451
2674	10539	11252.143399
970	7934	9769.740536
5845	16754	10561.944073
10268	18799	12923.692647
3476	11522	12559.523676
9502	4823	4289.433777
4033	4030	5144.508207
121	3100	5070.058694

	Price	Price Predicted
7733	4145	4302.960000
3521	13817	13827.390000
4059	4823	4991.740000
8987	22270	20168.950000
10265	10231	10332.945000
1056	6144	5012.580000
1548	4649	4658.120000
9606	9709	8638.810000
6626	13377	8719.240833
8453	5241	6213.340000

Linear Regression

DOI: 10.5281/zenodo.6906451
ISBN: 978-93-5607-317-3@2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

VII. CONCLUSION

This paper explains how to forecast flight ticket prices. A set of data is collected, pre-processed, modelled, and investigated in order to test algorithmic rule. Machine Learning methods with square measure for predicting accurate airline fares and providing accurate value of aircraft ticket price at both limited and maximum value. On Kaggle, data is obtained from websites that sell aircraft tickets. As indicated in the above analysis, the KNeighbors Regressor and Gradient Boosting Regressor yield better results, while the Random Forest Regressor forecasts the highest accuracy. The R-squared value predicts the model's accuracy as well. They are frequently attained.

VIII. REFERENCE

[1]. Konstantinos Tziridis , Theofanis Kalampokas ” Airfare Prices Prediction Using Machine Learning Techniques” DOI:10.23919/EUSIPCO.2017.8081365

[2]. T. Wang et al., "A Framework for Airfare Price Prediction: A Machine Learning Approach," doi: 10.1109/IRI.2019.00041.

[3]. Groves, W. and Gini, M., 2021. “A Regression Model For Predicting Optimal Purchase Timing For Airline Tickets.”.262172314

[4]JuharAhmedAbdellaa,NMZakibKhaled,ShuaibaFahadKh ”an Airline ticket price and demand prediction: A survey” https://doi.org/10.1016/j.jksuci.2019.02.001

[5] Supriya Rajankar, Neha Sakharkar, Omprakash Rajankar “Predicting The Price Of A Flight Ticket With The Use Of Machine Learning Algorithms” ISSN 2277-861

DOI: 10.5281/zenodo.6906451
ISBN: 978-93-5607-317-3@2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam