International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 05 | May 2022 www.irjet.net p-ISSN: 2395-0072

AIRLINE FARE PRICE PREDICTION

Kartik Rathi[1], Anubhav Kumar[2], Manish Yadav[3]

[1,2,3] *Students, Computer science and Engineering department, MIET Meerut,Uttar Pradesh,India*

Abstract The price of airline ticket changes frequently nowadays and there's plenty of difference. Price change keeps happening within few hours for the identical flight. The shoppers want to induce the most cost-effective possible price while the airline companies want the utmost profit and revenue possible. To unravel this problem, researchers introduce different models to avoid wasting consumers money- minimum price predicting model and models that tell us an optimal time to shop for a ticket while airlines use techniques like demand prediction and price discrimination to maximize their revenue

1. INTRODUCTION

This project aims to develop an application which can predict the flight prices for various flights using different machine learning techniques. The user will get the expected values and with its reference the user can plan to book their tickets suitably. At this time, carrier ticket costs can shift powerfully and fundamentally for an identical flight, in any event, for accessible seats inside the identical cabin. Clients are attempting to urge the foremost minimal cost while Airlines companies try to stay their general income as high as could reasonably be expected and boost their benefit.

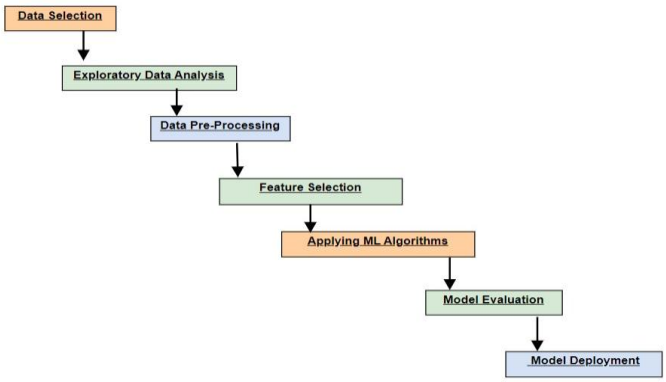
Airlines utilize different computational methods to extend their income, as an example, demand forecast and value segregation. The proposed system can help save immeasurable rupees of shoppers by proving them the knowledge to book tickets at the correct time. Parameters on which fares are calculated-

- Airline
- Date of Journey
- Source
- Destination
- Departure Time
- Duration
- Total Stops
- Weekday/Weekend

Now we can perform Exploratory Data Analysis on the given information. we'll discover correlation between the highlights. At that time a Machine Learning model are made to utilizing those highlights.

1.1 Proposed Methodology

For this project, we've implemented the machine learning life cycle to make a basic web application which is able to predict the flight fare by applying machine learning algorithms on historical flight data using some python libraries like Pandas, NumPy, Matplotlib, seaborn, and sklearn. Below image shows the number of steps that we followed from the life cycle.



Data selection is that the initial step where historical data of flight is assembles for the model to predict prices. Our dataset consists of quite 10,000+ records of information associated with flights and costs. A number of the features of the dataset are source, destination, departure date, point, and number of stops, point in time, prices. Within the exploratory data analysis step, we cleaned the dataset by removing the duplicate values and null values. If the null values aren't removed, the accuracy of the model are affected. Next step is data pre-processing where we observed that almost all of the information was present in string format. Data from each feature is extracted i.e., day and month is extracted from date of journey in integer format, hours and minutes is extracted from time of departure. Features like source and destination needed to be converted into values as they were of categorical type. For this One hot-encoding and label encoding techniques are wont to convert categorical data into the integer data.

Feature selection step is involved in selecting important features that are more correlated to the value. So, some features to be selected and passed to the group of models. Random forest is an ensemble learning method that basically uses group of decision trees as group of models. Random amount of knowledge is passed to decision trees and every decision tree predicts values in line with the dataset given to that. From the predictions made by the choice trees the features like extra information and route which are unnecessary features which can affect the accuracy of the model and so, they have to be removed before getting our model ready for prediction.

After selecting the features which are more correlated to cost the following step involves applying machine algorithm and creating a model. As our dataset incorporates labeled data, we've got to use supervised machine learning algorithms also in supervised we are going to be using regression algorithms as our dataset contains continuous values within the features. Regression models are wont to describe relationship between dependent and independent variables. The machine learning algorithms that we are going to be using in our project are:

1.2 Linear Regression

In simple linear regression there's only 1 independent and one dependent feature but as our dataset consists of the many independent features on which the worth may rely on, we are going to be using multiple linear regression which estimates relationship between two or more independent variables and one dependent variable. The multiple linear regression models are represented by:

$$y = m_1x_1 + m_2x_2 + \cdots + m_nx_n + C$$

Where, y = the predicted value of the dependent variable

x_n= the independent variables

m= independent variables coefficients

C= y-intercept x

1.3 Decision Tree

There are basically of two type's Decision tree i.e., classification and regression tree where classification is employed for categorical values and regression is employed for continuous values. Decision tree chooses independent variable from dataset as decision nodes for decision making.It divides the entire dataset in several sub-section and when test data is passed to the model the output is determined by checking the section to which the info point belongs to. And to whichever section the info point belongs to the choice tree will give output because the average value of all the information points within the sub-section.

1.4 Random Forest

Random Forest is an ensemble learning technique where training model uses multiple learning algorithms then combine individual results to urge a final predicted result. Under ensemble learning random forest falls into bagging category where random number of features and records will average value of the expected values if considered because the output of the random forest model.

2. PERFORMANCE METRICS

Performance metrics are statistical models which is able to be accustomed compare the accuracy of the machine learning models trained by different algorithms. The sklearn.metrics module are accustomed implement the functions to live the errors from each model using the regression metrics. Following metrics are accustomed check the error measure of every model.

2.1 MAE (Mean Absolute Error)

Mean Absolute Error is basically the sum of average of the absolute difference between the expected and actual values.

MAE = 1/n [∑(y-ý)]

y = actual output values,

ý = predicted output values

n = Total number of data points

Lesser the value of MAE better the performance of your model.

2.2 MSE (Mean Square Error)

Mean Square Error squares the difference of actual and predicted output values before summing all rather than using absolute value.

MSE = 1/n [∑ $(y - y')^2$]

y=actual output values

y'=predicted output values

n = Total number of data points

Lower the value of MSE better the performance of the model.

2.2 RMSE (Root Mean Square Error)

RMSE is measured by taking the square root of the average of the squared difference between the prediction and also the actual value.

RMSE = √1/n [∑ $(y - y')^2$]

y=actual output values

y'=predicted output values

n = Total number of data points

RMSE is greater than MAE and lesser the value of RMSE between different models the better the performance of that model.

2.3 R^2 (Coefficient of determination)

It helps you to grasp how well the independent variable adjusted with the variance in your model.

$$R^2 = 1 - \sum (y_i - \bar{y}_i)^2 / \sum (y_i - \bar{y})^2$$

The worth of R-square lies between 0 to 1. The closer its value to at least one, the higher your model is when comparing with other model values. There are different cross-validation techniques like GridsearchCV and RandomizedsearchCV which can be used for improving the accuracy of the model. Parameters of the models like number of trees in random forest or max depth of decision tree will be changed using this method which can help us in further enhancement of the accuracy.

The last three steps of the life cycle model are involved within the deployment of the trained machine learning model. Therefore, after getting the model with the most effective accuracy we store that model in a very file using pickle module. The back-end of the applying are going to be created using Flask Framework where API end-points such and acquire and POST are going to be created to perform operations associated with fetching and displaying data on the front-end of the appliance. The front-end of the applying are going to be created using the bootstrap framework where user will have the functionality of entering their flight data. This data is sent to the back-end service where the model will predict the output consistent with the provided data. the anticipated value is distributed to the front-end and displayed.

3. RESULTS

We had used many algorithms for prepared our ML model i.e., Linear Regression, Decision tree, Random Forest. From all of these, Random Forest gives us the most accurate predictions. Models' performance using a few metrics are given below:

A) LINEAR REGRESSION:

R2 SCORE : 0.6220180540708748

MAE: 1866.1170014687825

MSE: 6974960.920111607

RMSE: 2641.0151306101234

B) DECISION TREE:

R2 SCORE : 0.74098856555086889

MAE : 1270.242692247699

MSE : 4779579.164036292

RMSE : 2186.224865844383

C) RANDOM FOREST:

R2 SCORE : 0.8598264560438941

MAE : 1055.5042333744152

MSE : 2586644.6841106885

RMSE : 1608.3049101804945

By observing all of the performance metrics we can conclude that Random Forest will give us accurate and better results. So, we use Random Forest model for the deployment.

4. CONCLUSION

This project can result in saving money of inexperienced people by providing them the information related to trends of the flight prices and also give them a predicted value of the price which they use to decide whether to book ticket now or later. On working with different models, it was found out that Random Forest algorithm gives the highest accuracy in predicting the output.

REFERENCES

[1] K. Tziridis T. Kalampokas G.Papakostas and K. Diamantaras "Airfare price prediction using machine learning techniques" in European Signal Processing Conference (EUSIPCO), DOI: 10.23919/EUSIPCO .2017.8081365L. Li Y. Chen and Z. Li" Yawning detection for monitoring driver fatigue based on two cameras" Proc. 12th Int. IEEE Conf. Intell. Transp. Syst. pp. 1-6 Oct. 2009.

[2] William Groves and Maria Gini "An agent for optimizing airline ticket purchasing" in proceedings of the 2013 international conference on autonomous agents and multi-agent systems.

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

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

[5] T. Janssen "A linear quantile mixed regression model for prediction of airline ticket prices"

[6] medium.com/analytics-vidhya/mae-mse-rmse-coefficient-ofdetermination-adjusted-r-squared-which-metric-is-bettercd0326a5697e article on performance metrics

[7] www.keboola.com/blog/random-forest-regression article on random forest

[8] https://towardsdatascience.com/machine-learning-basics-decisiontree-regression-1d73ea003fda article on decision tree regression

BIOGRAPHIES

KARTIK RATHI
MIET,MEERUT(STUDENT)
AKTU

ANUBHAV KUMAR
MIET,MEERUT(STUDENT)
AKTU

MANISH YADAV
MIET,MEERUT(STUDENT)
AKTU