International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 01 | January - 2023 Impact Factor: 7.185 ISSN: 2582-3930

# A STUDY ON FLIGHT FARE PREDICTION USING ML ALGORITHM

**Shahrukh Khan[1], Vishal Prakash Maurya[2], Dr. Anurag Tiwari[3], Sarvesh Kumar Patel[4], Vikas Rawat[5]**

*[1,2,4,5]Student, CSE ,BBDITM,LUCKNOW*
*[3]Associate Professor, CSE, BBDITM,LUCKNOW*

---

**Abstract** - There are several differences in today's continuously fluctuating airline ticket prices. Within a few hours, the same flight's price continues changing. The shoppers want to get the best deal possible, while the airline corporations want to make as much money as they can. Researchers have developed many models to address this issue, including ones that estimate the minimal price and the best time to buy a ticket, while airlines use strategies like demand forecasting and price discrimination to increase their profits.

## 1. INTRODUCTION

This project seeks to create an application that uses multiple machine learning methods to forecast flight costs for various flights. The user will receive the anticipated values and, using them as a guide, may make an informed decision about how to purchase their tickets. Currently, airline ticket prices are subject to significant and fundamental changes for the same flight and available seats in the same cabin. While airlines endeavor to maintain their overall income as high as is reasonably possible and increase their profit, customers are seeking to demand the lowest possible price.

It might be challenging to choose the best time to buy an airline ticket from the perspective of the traveller because they have very little knowledge about future business price rates. Different models forecast future air travel costs and classify the ideal window for booking tickets. Airlines use a variety of pricing systems for their tickets, making price decisions later since order displays a larger value for approximation models. Each of these factors is what makes the system challenging. Airlines must control demand since there are only so many seats that can be filled in planes. Assume that the airline may raise prices to slow down the rate at which seats fill when demand exceeds capacity.

Airlines use a variety of computational techniques, such as value segmentation and demand forecasting, to increase their revenue. By demonstrating them, the proposed approach can help consumers save an enormous amount of money. The following factors are used to compute fares are:

- Airline
- Source
- Destination
- Route
- Total stops
- Additional information
- Journey date
- Journey month
- Departure hour
- Departure minute
- Arrival hour
- Arrival minute

We can now use the provided data for an exploratory data analysis. The connections between the highlights will be revealed. A machine learning model was then created using those highlights.

## 2. LITERATURE SURVEY

The customer finds it quite challenging to purchase an airline ticket at the lowest cost. Several methods are employed to do this. Find the day when airline tickets will cost the least. The majority of these methods rely on cutting-edge artificial intelligence (AI) research, often referred to as machine learning.

AI models were used by to connect the PLSR (Partial Least Square Regression) model in order to provide the best presentation for obtaining the cheapest airline tickets, with a 75.3% accuracy. To forecast the price of cheap tickets many days before flight, Janssen developed a straight quantile mixed relapse model. Ren, Yuan, and Yang considered the demonstration of Linear Regression's (77.06% precision), Naive Bayes' (73.06% exactness), Softmax Regression's (76.84% accuracy), and SVM's (80.6% exactness) models in predicting the cost of airline tickets. By accepting the problem as a grouping problem with the help of the machine learning models Ripple Down Rule Learner (74.5% exactness),

Logistic Regression (69.9% precision), and Linear SVM (69.4% exactness), Papadakis predicted that the price of the ticket would decrease subsequently.

In order to create a model that forecasts when to buy airline tickets, Gini and Groves used partial least square regression (PLSR). From 22 February 2011 to 23 June 2011, information was gathered from popular travel booking websites. In order to compare the results of the final model's performance comparisons, supplementary data were also collected.

With current daily airfares provided by Janssen for the San Francisco to New York route, he constructed an expectation model using the Linear Quantile Blended Regression method. The number of days till departure and whether the flight is on a weekday or the end of the week were used as two highlights by the model. The programme accurately forecasts airfare for days that are far from the departure date, but for a significant amount of time near to the departure day, the prediction isn't convincing.

A technique for improving the speed of ticket purchases by Wohlfarth was based on a remarkable pre-planning step known as macked point processors and information mining systems (arrangement and bunching) and a quantifiable investigation strategy. It is suggested that this system convert heterogeneous worth arrangement data into added value arrangement direction that can support unsupervised grouping calculation. Based on comparison estimation practises, the value direction is grouped into a collection. Advancement models evaluate value-change strategies. To select the best coordinating group and then compare the advancement model, a tree-based order computation was employed.

According to Dominguez-research, Menchero's the best moment to buy depends on the nonparametric isotonic relapse approach for a specific course, carriers, and duration. The model suggests the maximum amount of days before purchasing a ticket. For the expectation, there are two different types of variables. The first is the purchase date and passage.

## 3. PROPOSED METHODOLOGY

We have the machine learning life cycle to create a simple web application for this project that can: By leveraging machine learning algorithms on historical flight data using python libraries like Pandas, Numpy, Matplotlib, seaborn and sklearn, it is possible to anticipate the cost of a flight. The amount of steps we followed from the life cycle are depicted in the graphic figure below(Figure.1.).

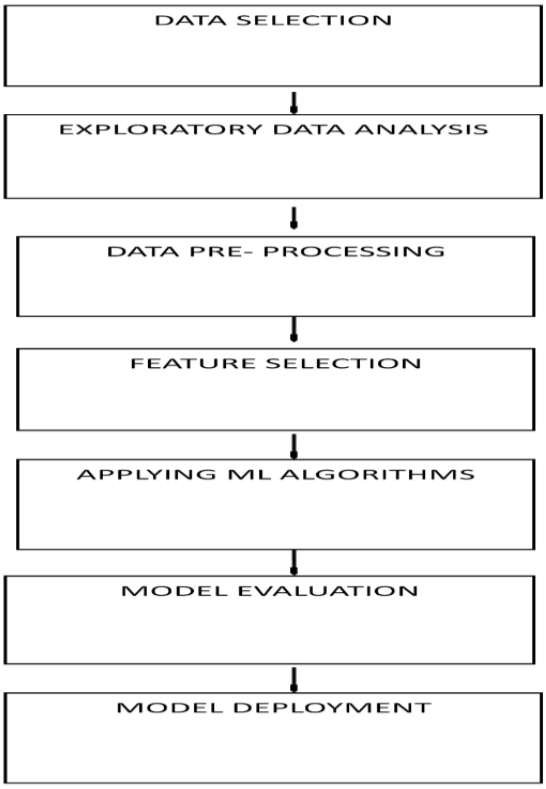


**Figure.1.**

The first step in compiling historical flight data for the model to anticipate pricing is data selection. Over 10,000 entries of data about flights and expenses are included in our dataset. The dataset's source, destination, departure date, point, number of stops, and point in time pricing are only a few of its properties. We cleansed the dataset during the exploratory data analysis process by eliminating duplicate and null values. The model's accuracy will be impacted if the null values are left in.

The following phase is data pre-processing, during which we noticed that practically all of the data was already in string format. Data is taken from each feature, such as the day and month from the journey's date in integer format and the hours and minutes from the departure time. Source and destination were categorical features, so they needed to be turned into values. Hot-encoding and label encoding techniques are frequently used to transform category data into numeric data for this purpose.

In the feature selection process, significant features that are more associated to value are chosen. Therefore, a few features should be chosen and given to the group of models. A set of decision trees are essentially used as a group of models in the ensemble learning technique known as random forest. Decision trees get a random amount of knowledge, and each one makes predictions based on the dataset it has been given. Before preparing our model for prediction, it is vital to eliminate features like extra information and route that are unneeded

features that can damage the model's accuracy from the predictions provided by choice trees.

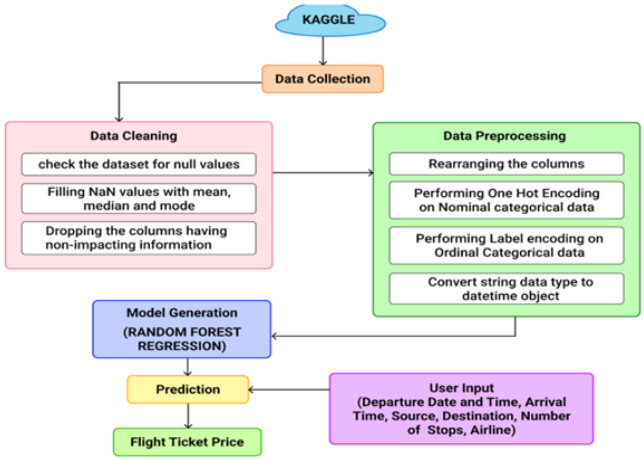


**Figure.2.Methodology block diagram**

The next phase entails using a machine algorithm and building a model after choosing the features that are more associated to cost. We must use supervised machine learning algorithms since our dataset contains labelled data. We must also use supervised machine learning techniques because our dataset contains continuous values for the features. It is common practice for regression models to explain the relationship between dependent and independent variables. We will use the following machine learning algorithms in our project:

**Linear Regression:** We will use multiple linear regressions, which estimates relationships between two or more independent variables and one dependent variable. In simple linear regression, there is only one independent and one dependent feature. However, our dataset contains numerous independent features on which the worth may depend. The following are examples of the multiple linear regression models:

$$y=m_1x_1 + m_2x_2 + m_3x_3+....m_nx_n + C$$

Where,

$y$ = the predicted value of the dependent variable

$x_n$= the independent variables

$m$= independent variables coefficients

$C$= y-intercept x

**Decision Tree:** Classification and regression trees are the two main types of decision trees, with classification being used for categorical values and regression being used for continuous values. A decision tree selects independent variables from a dataset as its decision nodes. When test data is provided to the model, it separates the entire dataset into numerous sub-sections, and when this happens, the output is determined by examining which part the info point belongs to. The output of the decision tree will be the average value of all the information points in the sub-section to which the information point belongs.

**Random Forest:** In an ensemble learning technique called Random Forest, the training model employs a number of learning algorithms, and the separate outputs are then combined to produce the final anticipated outcome. Random Forest belongs to the bagging category of ensemble learning, where a random sample of features and records will average out the predicted values when taken into account as the model's output.

## 4. PERFORMANCE METRICS

Performance metrics are statistical models that can be used to compare the precision of machine learning models developed using various techniques. The routines to live the mistakes from each model using regression metrics are often implemented in the sklearn.metrics module. The following metrics are frequently used to check each model's error measure.

**MAE (Mean Absolute Error):**

The average of the absolute difference between the expected and actual numbers makes up mean absolute error, essentially.

MAE = 1/n [Σ(y-ý)]

Where,

y = actual output values

ý = predicted output values

n = The more data points there are, the lower the value of MAE enhanced functionality of your model.

**MSE (Mean Square Error):**

Instead of utilising an absolute value, mean square error squares the difference between the output values that were expected and those that were actually obtained.

$MSE = 1/n [\Sigma (y-y')^2]$

Where, y=actual output values

y'=predicted output values

n = Total number of data points

Lower the value of MSE enhanced functionality of the model.

**RMSE (Root Mean Square Error):**

The square root of the average of the squared difference between the prediction and the actual value is used to calculate RMSE.

$RMSE = \sqrt{1/n [\Sigma (y-y')^2]}$

Where, y=actual output values

y'=predicted output values

n = Total number of data points

RMSE is bigger than MAE, and the lower the RMSE value comparing different models, the more effectively the model performs.

**$R^2$(Coefficient of determination):**

You can have a better understanding of how well the independent variable handled the variation in your model through $R^2$ .

$R^2 = 1- \sum [ ( yi - \underline{y}_i) / ( yi - \underline{y} )^2 ]$

R-square values range from 0 to 1. When compared to other model values, your model performs better the closer its value is to at least one. To increase the model's accuracy, various cross-validation methods can be utilised, such as GridsearchCV and RandomizedsearchCV. We may further improve the accuracy of the models by changing their parameters, such as the number of trees in the random forest or the maximum depth of the decision tree.

The deployment of the trained machine learning model involves the final three stages of the life cycle model. As a result, after obtaining the model with the highest level of accuracy, we use the pickle module to store that model in a specific file. In order to accomplish actions related to acquiring and displaying data on the front end of the application, API end-points like acquire and POST will be constructed on the back end of the application using the Flask Framework. The bootstrap framework will be used to develop the application's front end, which will allow users to enter flight information. This information is transmitted to the back-end service, where the model predicts the results in accordance with the supplied information. The front-end receives and displays the projected value.

## 5. RESULTS

For the preparation of the ML model, variety of methods, including Linear Regression, Decision Trees, and Random Forest are available. Random Forest provides us with the most precise forecasts out of all of these. We get to the conclusion that Random Forest will provide us with accurate and superior outcomes by looking at all of the performance measures. Therefore, we implement the Random Forest model.

## 6. CONCLUSION

By giving customers information on flight pricing patterns and a predicted price that they can use to decide whether it should book a ticket now or later, this effort may enable inexperienced people to save money. A dataset is gathered, pre-processed, data modelling is done, and a value difference for the number of restricted days for travel by the passengers is analysed in order to evaluate the algorithmic rule. Machine learning algorithms are used to predict airline fares accurately, and they provide accurate estimates of the cost of a plane ticket at the lowest possible price. Data that is often accessible is restricted since it is obtained from Kaggle websites that sell airline tickets. Although decision tree and random forest algorithms produce outcomes with higher accuracy, the analysis above demonstrates that after experimenting with various models, it was discovered that the Random Forest method provides the highest level of output prediction accuracy .

## 7. FUTURE SCOPE

The anticipated findings in the future are quite accurate when a large amount of information is accessed as in-depth information in the dataset in the coming days. If someone wants to conduct further research on the topic, they should look for additional historical data sources or be very organised while gathering information manually over an extended period of time, since many possible combinations of plane will be travelled. There is a good chance that different planes will execute their plans differently depending on the plane's individual qualities. Finally, it is intriguing to compare the correctness of our model to that of the contemporary business models being presented.

## 8. 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

[9] Boruah A., Baruah K., Das B., Das M.J., Gohain N.B. (2019) "A Bayesian Approach for Flight Fare Prediction Based on Kalman Filter," https://doi.org/10.1007/978-981-13-0224-4_18

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

[11] G.A. Papakostas, K.I. Diamantaras and T. Papadimitriou, "Parallel pattern classification utilizing GPU-Based kernelized slackmin algorithm," doi:10.1016/j.jpdc.2016.09.001

[12] G. Francis, A. Fidato, and I. Humphreys, "Airport–airline interaction: the impact of low-cost carriers on two european airports," doi:10.1016/s0969-6997(03)00004-8

[13] C. Koopmans and R. Lieshout, "Airline cost changes: To what extent are they passed through to the passenger?" doi:10.1016/j.jairtraman.2015.12.013

[14] S. Lee, K. Seo, and A. Sharma, "Corporate social responsibility and firm performance in the airline industry: The moderating role of oil prices," doi:10.1016/j.tourman.2013.02.002

[15] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," Available: http://arxiv.org/abs/1412.6980

[16] L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5-32, 2001. doi:10.1023/a:1010933404324

[17] H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola and V. Vapnik, "Support vector regression machines," doi:10.1016/s0165-0114(02)00514-6 [18] K. S. Gerardi and A. H. Shapiro, "Does competition reduce price dispersion? new evidence from the airline industry," doi:10.1086/597328

[18] W. Groves and M. Gini, "An agent for optimizing airline ticket purchasing," 12th International Conference on Autonomous Agents
and Multiagent Systems (AAMAS 2013), St. Paul, MN, May 06 -10, 2013 , pp. 1341-1342.

[19] T. Janssen, "A linear quantile mixed regression model for prediction of airline ticket prices," Bachelor Thesis, Radboud University, 2014.

[20] R. Ren, Y. Yang and S. Yuan, "Prediction of airline ticket price,"Technical Report, Stanford University, 2015.

[21] Wohlfarth, T. Clemencon, S.Roueff, "A Dat mining approach to travel price forecasting", 10 th international conference on machine learning Honolulu 2011.