ISSN: 2366-1313

Flight Ticket Price Prediction Using Machine Learning

[1]M. Chinna babu, [2]Eslavath Rahul, [3]Dulam Mohan, [4]Nittu Raj Kumar

[1]Assistant Professor, Dept.of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad,

mchinna64@gmail.com

[2]BTech student, Dept.of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad,
eslavathrahul111@gmail.com

[3]BTech student, Dept.of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad,
mohandulam222@gmail.com

[4]BTech student, Dept.of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad,
nittu.rajkumar111@gmail.com

Abstract: *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 project implements the validations or contradictions towards myths regarding the airline industry, a comparison study among various models in predicting the optimal time to buy the flight ticket and the amount that can be saved if done so. Remarkably, the trends of the prices are highly sensitive to the route, month of departure, day of departure, time of departure, whether the day of departure is a holiday and airline carrier. Highly competitive routes like most business routes (tier 1 to tier 1 cities like Mumbai-Delhi) had a non-decreasing trend where prices increased as days to departure decreased, however other routes (tier 1 to tier 2 cities like Delhi - Guwahati) had a specific time frame where the prices are minimum. Moreover, the data also uncovered two basic categories of airline carriers operating in India – the economical group and the luxurious group, and in most cases, the minimum priced flight was a member of the economical group. The data also validated the fact that, there are certain time-periods of the day where the prices are expected to be maximum. The scope of the project can be extensively extended across the various routes to make significant savings on the purchase of flight prices across the Indian Domestic Airline market.*

Keywords: *Feature selection, Airfare price, Machine learning, Pricing Models, Prediction Model, Random Forest.*

I. INTRODUCTION

The flight ticket buying system is to purchase a ticket many days prior to flight take-off so as to stay away from the effect

ISSN: 2366-1313

of the most extreme charge. Mostly, aviation routes don't agree this procedure. Plane organizations may diminish the cost at the time, they need to build the market and at the time when the tickets are less accessible. They may maximize the costs. So, the cost may rely upon different factors. To foresee the costs this venture uses AI to exhibit the ways of flight tickets after some time. All organizations have the privilege and opportunity to change its ticket costs at any time. Explorer can set aside cash by booking a ticket at the least costs. People who had travelled by flight frequently are aware of price fluctuations. The airlines use complex policies of Revenue Management for execution of distinctive evaluating systems. The evaluating system as a result changes the charge depending on time, season, and festive days to change the header or footer on successive pages. The ultimate aim of the airways is to earn profit whereas the customer searches for the minimum rate. Customers usually try to buy the ticket well in advance of departure date so as to avoid hike in airfare as date comes closer. But actually, this is not the fact. The customer may wind up by giving more than they ought to for the same seat [1].

Recently the airline organizations are giving more attention to the complex tactics and processes to finalize the ticket costs in dynamic manner. Also, with the explosive growth of the net and ecommerce, air passengers today will check transportation and availability of any airlines round the world simply. Once satisfying with associate degree of transportation, these customers should buy their desired tickets online through official airline or agent websites to assist the shoppers to shop for the foremost inexpensive transportation, there are variety of prediction models to predict the transportation costs. Social media these days is an integral part of people's daily routines and therefore this resource as a result, is abundant in user opinions. The analysis of some specific opinions will inform corporations on the amount of satisfaction within customers. Airline price ticket costs modification terribly dynamically and for a similar flight day by day. It is terribly tough for a customer to buy an air ticket within the lowest value since the value changes dynamically. We addressed the matter regarding the market section level airfare ticket cost forecasting by usage of publicly obtainable datasets and completely unique machine learning model to forecast market section level price cost of airline ticket. The purpose of this study is to raise and analyse the options that influence transportation and to develop and tune models to predict the transportation well ahead [2].

ISSN: 2366-1313

Most studies on airfare price prediction have focused on either the national level or a specific market. Research at the market segment level, however, is still very limited. We define the term market segment as the market/airport pair between the flight origin and the destination. Being able to predict the airfare trend at the specific market segment level is crucial for airlines to adjust strategy and resources for a specific route. However, existing studies on market segment price prediction use heuristic-based conventional statistical models, such as linear regression [3], and are based on the assumption that there exists a linear relationship between the dependent and independent variables, which in many cases, may not be true. Recent advances in Artificial Intelligence (AI) and Machine Learning (ML) make it possible to infer rules and model variations on airfare price based on a large number of features, often uncovering hidden relationships amongst the features automatically.

II. LITERATURE SURVEY

Air ticket price prediction is a challenging task since the factors involved in pricing dynamically change over time and make the price fluctuate. In the last decade, researchers have incorporated machine learning algorithms and data mining strategies to better model observed prices. Among them, regression models, such as Linear Regression (LR), Support Vector Machines (SVMs), Random Forests (RF), are frequently used in predicting accurate airfare price [4].

Early work also considered using classification models to predict the trends of the itineraries. Ren et al. [5] proposed using LR, Naive Bayes, Softmax regression, and SVMs to build a prediction model and classify the ticket price into five bins (60% to 80%, 80% to 100%, 100% to 120%, and etc.) to compare the relative values with the overall average price. More than nine thousand data points, including six features (e.g., the departure week begin, price quote date, the number of stops in the itinerary, etc.), were used to build the models. The authors reported the best training error rate close to 22.9% using LR model. Their SVM regression model failed to produce a satisfying result. Instead, an SVM classification model was used to classify the prices into either “higher” or “lower” than the average.

In [6], four LR models were compared to obtain the best fit model, which aims to provide an unbiased information to the passenger whether to buy the ticket or wait longer for a better price. The authors suggested using linear quantile mixed models to predict the lowest ticket prices, which are called the “real bargains”.

ISSN: 2366-1313

However, this work is limited to only one class of tickets, economy, and only on one direction single leg flights from San Francisco Airport to John F. Kennedy Airport.

Wohlfarth et al. [7] integrated clustering as a preliminary stage with multiple state-of the- art supervised learning algorithms (classification tree (CART) and RF) to assist the customers' decision-making process. Their framework uses the K-Means algorithm to group flights with similar behavior in the price series. They then use CART to interpret meaningful rules, and RF to provide information about the importance of each feature. Also, the authors pointed out that one element, the number of seats left, is a key freature for ticket price prediction. Aside from flight-specific features, many other attributes affect the competitive market. Accurately predicting the market demand, for example, can reduce a travel agency's accumulated costs, which are caused by over purchasing or lost orders.

In [8], the author applied Artificial Neural Network (ANN) and Genetic Algorithms (GA) to predict air ticket sales revenue for the travel agency. The input features Included international oil price, Taiwan stock market weighted index, Taiwan's monthly unemployment rate, and so on. Specifically, the GA selects the optimum input features to improve the performance of the ANNs. The model showed good performance with a 9.11% Mean Absolute Percentage Error.

As airline ticket data is not well organized and ready for direct analysis, collecting and processing those data always requires a great deal of effort. For most analyses found in the literature, researchers evaluate their models' performance on different datasets by either crawling the data from the web or requesting private data from collaborative organizations. As a result, it is difficult to replicate the research and conduct comparisons of the models' performance. For U.S. airlines, the fare data is publicly available in the T100 and DB1A/1B databases. However, due to the limited association between the prices and specific flights information, these datasets are seldom used independently to generate scientific research outcomes. However, researchers who are interested in analysing the price dispersion, for example, are more likely to consider investigating the information from those datasets. In Rama-Murthy's dissertation, the Official Airline Guide (OAG) and DB1B data are used to model the airfare prices. The author also incorporates the Sabre AirPrice data, which was provided by SABRE, but they only provide the information of their online users. As this online user data does

ISSN: 2366-1313

not represent the whole consumer market, it can bias the results obtained from the data[9].

Authors have initiated to address for improving the plane estimate favour, also considering the plane estimation as a timeline issue. Concentrated on extracting possible figures of the fare modification by ML techniques. Data model is first presented by the authors to arrange the price values and pull-out attributes. Fitting the flowing way and understanding the idea of floating in the sequence, the authors have presented Learn++. This paper study has still in its early stage. instead of the limited price on a whole travelling place, a perfect view of independent plane will be worth as passengers might get partiality while buying the tickets. Predicting the fight fares in the limited time by improving the study in a present scenario. Improvement is done in a linear model and algorithm used in this has drawn information and the practical data as time, week, day, date etc, are given as input to the forecast value[10].

III. PROPOSED SYSTEM

Our proposed framework utilizes both the DB1B and T- 100 datasets, in combination with macroeconomic data to predict the quarterly average airfare at the market segment level. Figure 1 shows an overview of the major components of the proposed framework. In the data pre-processing step, all datasets are cleaned to exclude possible erroneous samples, transformed and combined based on the market segment. The feature extraction module serves to extract and generate handcrafted features that aim to characterize the market segment. The goal of the feature selection module is to optimize the performance of the prediction model by analysing the effectiveness of the features and remove any irrelevant features. Finally, we use the selected features to build our prediction model, which generates the output value as the predicted air ticket price.

1) Data Pre-processing: In the DB1B and T-100 datasets, many attributes contain the same information. Directly merging the tables creates many duplicate fields. Also, the data reported by the airlines may include erroneous values caused by human error, currency conversion error, etc. Therefore, a properly designed data pre-processing workflow is crucial to generate accurate input data in order to build the machine learning model.

2) Feature Extraction: Several features have been extracted from the DB1B and T-100 dataset to represent a specific aspect of the market segment. Furthermore, to exploit the relationship between the airline industry and the overall economic

ISSN: 2366-1313

conditions, several macroeconomic features are also added to the feature set. Table I describes all the features that are identified during feature extraction.

Table.1 The List of Features Generated During Feature Extraction

Feature Name	Description
Distance	Market distance between the origin and destination airports
Seat Class	Indicator for economy or premium seat type
Passenger Volume	Total number of passengers traveled between the origin and destination airports
Load Factor	The ratio of the total number of passenger to the total number of seat in a market
Competition Factor	The market HHI
LCC Presence	Indicator of LCC operating in the market
Crude Oil Price	Quarterly average of crude oil price
CPI	Quarterly average of Consumer Price Index
Quarter	Indicates the three month period of the year

ALGORITHMS

While we go through the algorithms we employed (XGBoost, Random Forest, and Decision Tree) and also how they operate in our models, please read the discussion below.

A. Decision Tree

The decision tree appears to be the most well-known and commonly employed categorization technique. A decision tree is a collection of nodes that resembles a diagram, for each junction indicating a test on a characteristic and each branch indicating a test outcome, such that each node in a decision tree (terminal node) has a class label. A tree can be "trained" by dividing the resources collection into subgroups depending on a characteristic values test. This procedure is known as partitioning the data because it is performed iteratively on each derived subset. The recursion ends when all subgroups at a node have the same posterior probability, or when the split no longer adds additional value to the predictions. A decision tree.

appropriate for experimental extracting knowledge since it does not need subject matter expertise or parameters configuration Assume S is a collection of cases, A is a property, Sv is the subgroup of S with Such = v, as well as Value (A) is the collection of all number of values of A.

B. Random Forest

A Random Forest is an ensemble approach that can handle simultaneous regression and classification problems by combining many decision trees using a technique known as Bootstrap as well as Aggregation, or bagging. The core idea is to use numerous decision trees to determine the final result instead of depending on personal decision trees. Random Forest's foundation learning methods are numerous decision trees. We arbitrarily choose rows and characteristics from the dataset to create sample datasets for each model. This section is known as Bootstrap. We simply have to understand

ISSN: 2366-1313

the purity in our dataset, and we'll use that characteristic as the root of the tree which has the smallest impurity or, in other words, the smallest Gini index. Mathematically.

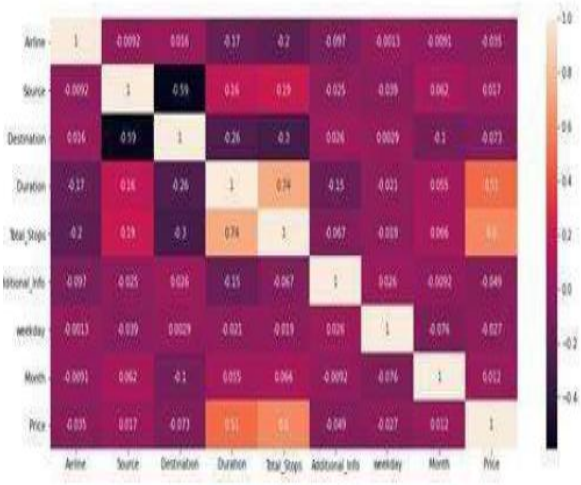

Fig.1 Random forest process

C. XGBoost

XGBoost is an effective method for developing supervised regression models. Knowing as to its (XGBoost) goal function and baseline learners can help determine the truth of this proposition. This optimization problem has both a loss function and regularization component. It makes a distinction between real and theoretical predictions, i.e., how far the model outputs deviatoric the real amounts. In XGBoost, the most used standard error in regression problems is quarantine, whereas: logistic is used for classifications

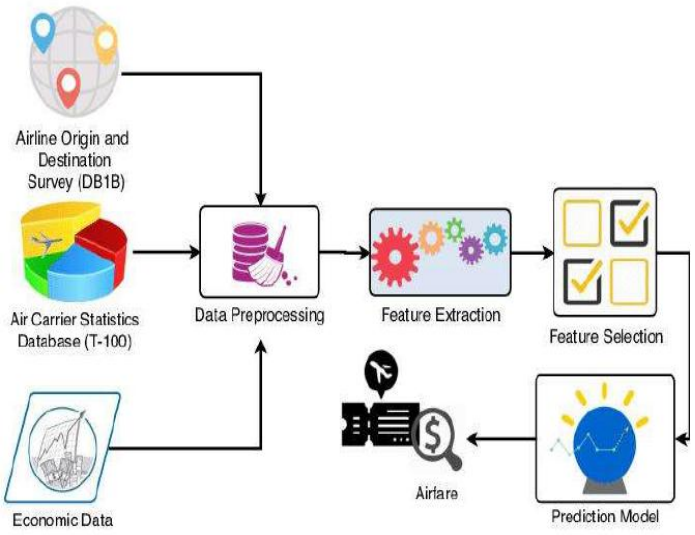

Fig.2 System architecture

System Architecture Description

1. In the first step we are giving all the valued details to the software that we are created using python.

2. In the next step data will be pre-processed with the data which we are given to training module.

3. Coming to next step it will be do the feature extraction and select the feature which match in the training data.

4. After completing all the prediction by system and finally displayed the predicted output to.

Table.1Aaccuracy graph of algorithm

SYSTEM ARCHITECTURE

ISSN: 2366-1313

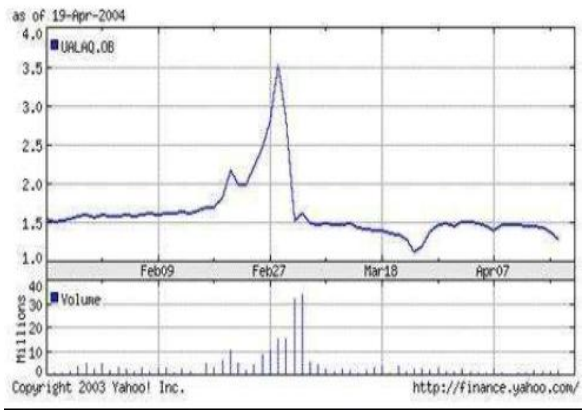

Table.2 Accuracy table of algorithm

Algorithm	Training accuracy	Testing accuracy
XGBoost	0.92	0.77
Random Forest	0.95	0.78
Decision Tree	0.97	0.67

IV. RESULTS

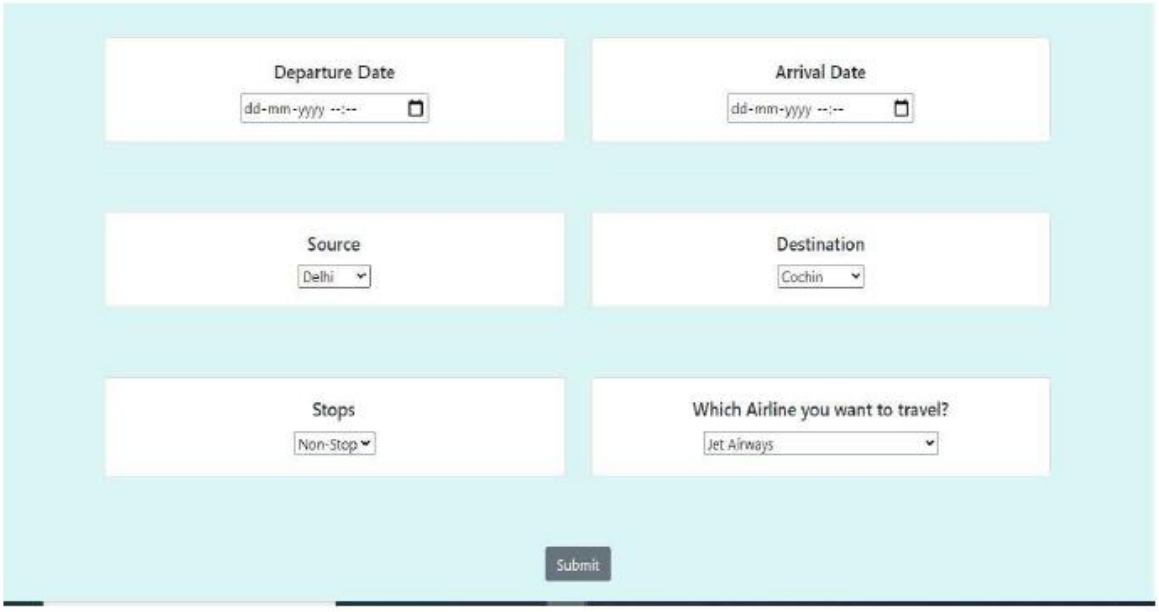

Fig.3Normal web page look like this after updating the details into the colums then we will get the predicted output.

After updating all the values to it then we will get finally the resultant output that is what we are predicting for flight ticket price using machine learning.

The goal in this step is to develop a benchmark model that serves us as a baseline, upon which we will measure the performance of a better and more tuned algorithm. We are using different Regression Technique and comparing them to see which algorithm is giving better performance than other and At the end we will combine all of them using Stacking and see how our model is predicting.

ISSN: 2366-1313

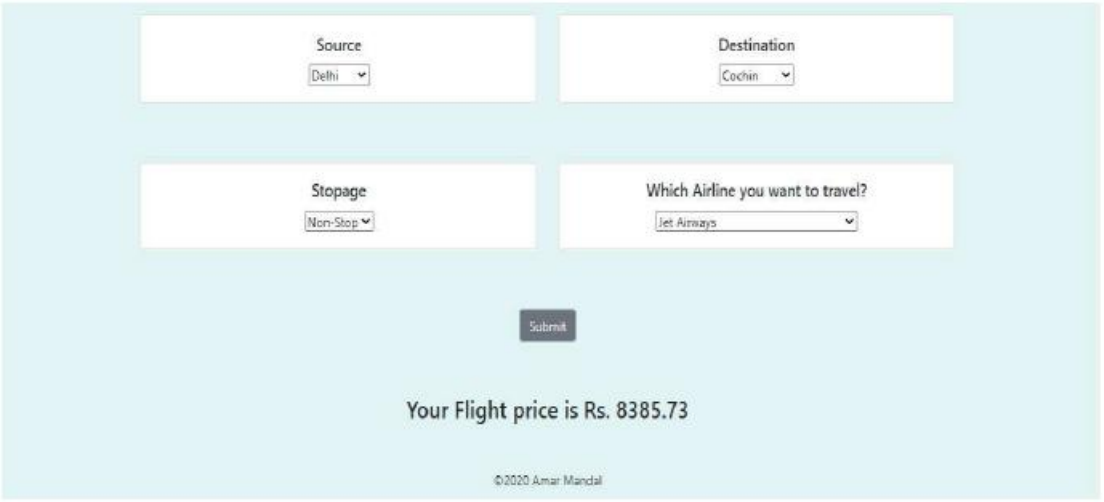

Fig.4 Prediction of flight price

V. CONCLUSION

In the proposed paper the overall survey for the dynamic price changes in the flight tickets is presented. this gives the information about the highs and lows in the airfares according to the days, weekend and time of the day that is morning, evening and night. also, the machine learning models in the computational intelligence field that are evaluated before on different datasets are studied. their accuracy and performances are evaluated and compared in order to get better result. For the prediction of the ticket prices perfectly different prediction models are tested for the better prediction accuracy. As the pricing models of the company are developed in order to maximize the revenue management. So, to get result with maximum accuracy regression analysis is used. From the studies, the feature that influences the prices of the ticket are to be considered. In future the details about number of available seats can improve the performance of the model.

REFERENCES

1. B. Smith, J. Leimkuhler, R. Darrow, and Samuels, Yield managemental American airlines, Interfaces, vol. 22, pp. 831, 1992.

2. 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.

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

ISSN: 2366-1313

4. R. Ren, Y. Yang and S. Yuan, Prediction of airline ticket price, Technical Report, Stanford University, 2015

5. M. Papadakis , Predicting Airfare Prices, 2014.

6. L. Breiman, Random forests, Machine Learning, vol. 45, pp. 5-32, 2001.

7. Viet Hoang Vu, Quang Tran Minh and Phu H. Phung,An Airfare Prediction Model for Developing Markets, IEEE paper 2018.

8. S.B. Kotsiantis, Decision trees: a recent overview, Artificial Intelligence Review, vol. 39, no. 4, pp. 261-283, 2013.

9. L. Breiman, Random forests, Machine Learning, vol. 45, pp. 5-32, 2001.

10. S. Haykin, Neural Networks A Comprehensive Foundation. Prentice Hall, 2nd Edition, 1999.

11. H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola and V. Vapnik, Support vector regression machines, Advances in neural information processing systems, vol. 9, pp. 155-161, 1997.