IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol. 10, Issue.2, February 2022 E-ISSN: 2347-2693

Airlines Ticket Price Prediction Using Machine learning approach

Kusam Bhargavi[1*], A. Lahari Sai[2], K. Thirupathi Rao[3]

[1,2,3]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, AP, India

DOI: https://doi.org/10.26438/ijcse/v10i2.4953 | Available online at: www.ijcseonline.org

Received: 10/Feb/2022, Accepted: 20/Feb/2022, Published: 28/Feb/2022

***Abstract*—** The main theme of this paper deals with the prediction of airlines prices as they may vary many times based on different constraints or attributes of the dataset which affect the fare of airlines. There are many ML(machine learning) models which help us to predict mainly two types of customer side models. The main focus of them is to provide the optimal time to buy a ticket and the fare of tickets should be as cheap as possible. We have considered a dataset which is consisting of ten thousand six hundred and eighty-three entries and eleven columns or attributes. For this paper, we have plotted dist plots and used some of the algorithms/methods to get the accuracy. This project helps the customers/buyers to tell the best time to purchase the ticket and that too with the lowest price. Out of all the methods used, We found random forest gives the most accuracy than the other models.

***Keywords*—** Python,Pandas,ml: prediction models.

I. INTRODUCTION

As any individual who booked a ticket for airways would know that prices/costs of tickets changes atleast for about seven times a day. It is because Airway industry people uses some advanced strategy known as Revenue Manage- ment in order to execute a special and distinctive valuing techniques/strategy[1],[2]. It is said that the Airways industry is one of the best sophisticated its usage of dynamic pricing techniques or strategies to maximize revenue, based on using suitable usecase algorithms and hidden variables that contribute more to target variables. Therefore, it is said to be challenging for consumers for predicting the price changes in the future. The low expensive available ticket changes over a certain amount of time and the fare of a ticket may be cheap or high. The suitable methods will naturally modify the price according to the time of booking tickets like early morning or afternoon ,evening or night. Price may as usual change with the seasons like summer, winter, weekends and celebration seasons. The ultimate aim of carrier is to build an effective models such that its revenue can be maximized and also prices should be less such that we can attract more customers yet on the other side buyers are searching for least expensive cost. Since many buyers thought that airfare would be the high when the date of purchasing a ticket is most closer to the take off date, yet it is not always

/generally true according to researches [3],[4],[5]. And most of the Buyer may end up with paying higher amount than they ought to pay for the similar seat. Dynamic costing is one of the most commonly used strategy for pricing tickets and is implemented by many of different airline industries to adjust ticket fare in accordance to the various external and internal factors such as changes in demand and also frequently changing the costs of tickets in a way to fetch more customers, promotions done by competitors, to keep the prices as low as possible for customer satisfaction, more availability of seats in the compartments and others[6],[7]. We have used a dataset which is not available in all websites. From the buyer's point of view, predicting

/determining the least price or the finest time to buy a ticket is one of the key issue. The conception of most of the customers is that tickets that are bought in advance are cheaper. It is mostly possible scenario that buyers who will buy a ticket way more earlier end up paying more than those who would be buying the same ticket later with low price. [8] proposed using Naive Bayes, regression, lr and SVM to build a model and classify the ticket fare into bins. Furthermore , early purchasing implies that risk of committing to a specific schedule that might need to be enhanced/changed basically for a fee[9]. The ticket pricing might get affected by various factors thus accuracy may change continuously.

The dataset has been utilized in various studies that assess the determinants of aircraft characteristics and frequency of flights and demand-prediction[10].

In [11], four LR models were compared to get the most effective fit model, that aims to produce AN unbiased data to the rider whether or not to shop for the price ticket or wait longer for a better value.

II. RELATED WORK

In the referred area, we examine,summaraize and distin- guish the qualities, shortcomings of prev work done by manyresearchers and recommend directions of future.

A. Complete evaluation of previous work

Energetic estimating is the foremost common estimating procedures executed by aircraft industry; to alter ticket/fare costs in reaction to different inside and outside components such as changes in request, competitor advancements, capacity of clients to purchase, accessibility of seats of others. Aircrafts got foresee changes factor's to actualize a energetic estimating plot that powerfully alters ticket costs to extend their profit. Other hand the clients are moreover interest to estimate how ticket costs would alter within the future to be able to purchase tickets at lower costs. Subsequently, analysts have created different forecast models both for aircrafts and clients to assist them bargain with energetic estimating. The 2/3 most common strategies proposed for aircrafts are request forecast and cost segregation which was collectively allude to as Aircrafts fig/diagrams/models. Client models include ideal ticket buy time expectation models and ticket cost expectation models. There a trade b/w cash sparing by client and expanding income by factories/ companies. As clients gotten to be more vital by utilizing client side apparatuses, it gets to be more troublesome for the carriers to apply energetic estimating and to create benefit. In this manner, there's a require for a expectation show that can anticipate the ideal ticket costs that can bring shared advantage both for clients and carriers. Existing room for advancements in a few ranges counting anticipating correct esteem of ticket prices/demand, dataset issues, restricted the number of highlights, missing sweeping statement, way better forecast methods and execution and issues of complexity. Be that because it may, the utilized models in these considers around persevere from overhead computational since it is more com- putationally genuinely than identify the ideal time of buying. Inside the locale of ask desire, the first striking work (ref4) predicts quarterly course ask but cannot work for brief term estimate. The other models in ref3,4 recommended for ask desire because it were gage the rate increment or decrement in ask for a flight based on fetched flexibility. Another vital theme that's not however investigated well it was related to the advancement of a cost separation show. None of the past ponders propose a procedure for cost segregation but they or maybe center on demonstrating the presence of cost segregation in carriers estimating procedures. Lack of simplification is additionally one of the shortcomings taken note among existing thinks about. The forecast models proposed so distant work at either flight level or level of course and don't bolster expectation at both levels at the same time. Additionally, a demonstrate that combines forecast for diverse sorts of flights like as continuous flights, multi-stop flights, circular trips and one way trips etc. isn't proposed however. On side of others, issues of dataset which was been provided, features limitations/drawbacks and methods utilized are likely the foremost critical issues and have to be be talked about in subtle elements. Subsequently, we see everything from of those in a isolated area.

Collecting data from various important sources/by using web scraping is the one of the most crucial aspect of our project. There were several sources of the data available on various famous websites which can be used to train the ml models. Most of the Websites that give data consisting about the timings of the flight, multiple routes, fare and airlines . Several sources from API's(it is An app program interface that acts as a connection between pc or between pc programs)to customer travel websites are available for data scraping(it is one of the tool to collect data from websites or web applications) and convert it into structured format . In this project details of the various/several sources and parameters that were collected are considered. To build any model data is collected from a website known as "7 Makemytrip.com" and technology used by us to5 implement and collection of this models is python.

The script excerpt the details from various specified websites and creates an csvfile as output. This file contain the data w26ith different features and their details. Main important aspect is selection the features/parameters that might be needed/used for the airline prediction algorithms. Output which was collected from the various web contains more than 22000 rows in the (object format) for each and every flight but not all are required(duplicated data can be removed).

- Source
- Destination
- Departure- date
- Arrival Time
- Price
- Airways name
- Ticket booked date
- Multi route variable
- Stops
- Additional-info
- Departure- Time

In this study, the main focus is only on reducing airline fares according to certain charges so that more routes are considered without return. This data collection is considered to be one of the busiest routes in India (BOM to DEL) for a period of 3 months i.e., from Feb to April. In the data of each aircraft with all the features collected in person.

A. Pre-processing (data cleaning and preparing data)

All the data that we have collected might not be in structured format in order to apply models. so, first make the format into structured. Now the data is in desired format our next task is to clean the data which can be achieved by removing/drop the rows that contain empty cells or to use some central tendency calculators such as mean , mode , median and replace in place of empty cells it is considered as effective as we are not droping any data. once all primary cleaning is done next step is to convert the parameters in the dataset into numerical format because, the ml models does not understand any object kind of data it only understands numerical features in order to implement models. python

is used to clean and prepare the data. For example, the Source was character type, not an integer so, converting into numerical format by using one hot encoding technique .similarly all the other object types are converted into numerical using desired methods such as label encoding.

B. Analyzing data

Analysis of data is next to data preparation, which uncovers certain hidden patterns/trends and then apply-ing/implementing various machine learning models. And Also, some features can be computed by using existing features. departure date can be obtained by computing the difference between the departure-date and the date on which airway is booked. This parameter difference is considered to be within 45 days. Also, the day in which we are departing plays an crutial role in whether if it is holiday or a weekday. Comparatively the flights scheduled during weekends said to be having more price compared to the flights on normal days. Furthermore, time seems to play an important factor. So that is the reason why the time is been divided into 4 categories: evening, night, morning, afternoon.

III. METHODOLOGY

Since the cost of the flight ticket may differs due to the airport popularity, difference in distance from source to destination, weekdays or normal days, festive seasons and other factors, it is comparatively hard to implement a model which can perform fine and fair for all the flights. So, We decided to train different models for every airline route and the implement a method called predict which takes ml model as one of the inyput and the other inyput is dump. If dump=0 it just calculates all the performance metrics and prediction with out saving for further use and if dump=1 then the results can be reused at any time as it gets stored in the specified file destination. For any continuous model(output in numerical format) the target variable is simply varies according to time.

A. Linear Regression Model

Firstly we need to separate dependent and independent features from dataset as x, y. Where x is the independent variable that changes according to the dependent variable y. In our use case the price of ticket may vary according to various factors. The graph that we got by implementing linear regression might be positively/negatively correlated .It is basically used to find the relation between 2 variables. linear regression is one of a type of analysis where the number x is one and the relationship between the dependent(y) and independent (y) variables vary linearly. Gradient decent and cost function are the important concepts to understand linear regressions. continuous variable the output variable in linear regression.[8] y(prediction)=b0+b1 * x

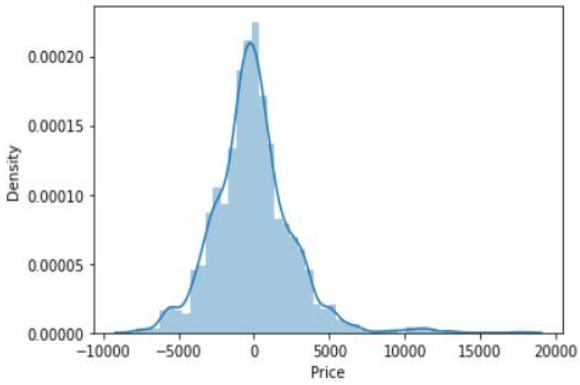


figure III.A

B. Decision tree

Decision tree calculation separates all data collected into smaller sets, simultaneously constructed gradually. Effects of keeping the tree with decision nodules and leaf nodules. The decision tree can contain at least two branches. First, consider a collection of details as one root. Highlighting the respect required to be lowered. If it is possible that the skills are permanent, they are separated before planning the model. Based on feature quality records that are repeatedly dispersed. The calculation of the decision tree can be done using two main features .One Information gain (IG) and the other Gini indicator. the portion of the change in entropy is calculated with the benefit of the details. Entropy is very high in character value, where entropy is part of the (vulnerable) risk of random (random) variance. The Gini Index is a component that covers how a randomly selected segment can be broken down by mistake. Specifies whether a feature with a low Gini reference should / is said to be popular or widely used. The root of the decision tree plays an important role. decision trees are constructed using rule based models.

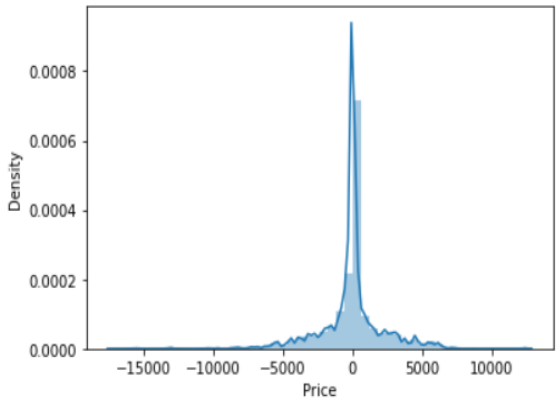
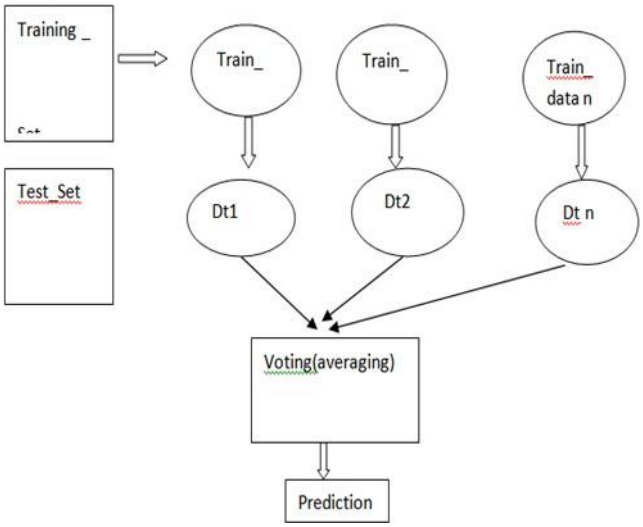


Figure-III.B

C. Random Forest

It is a one of the supervised learning algorithm. It uses a method/concept called ensemble Learning which learns from multiple decision trees to solve a complex problems by combining multiple classifiers that finally improves the performance of this computed model . It's a classifier that has different no of decision trees on several subsets of the complete data that is given and considers

the average to increase the accuracy of prediction model. The random forest considers prediction from each and every tree and also by their majority of predicted votes. And majority is considered as the final output. The higher no of trees in this model leads to greater accuracy and stop the overfitting problem in this model.



Output:Figure-III.C

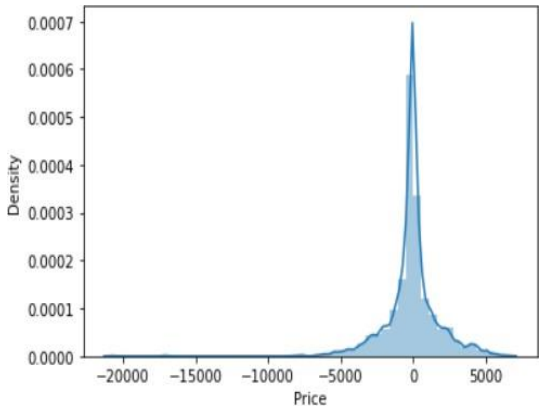


Figure-IV

IV. RESULTS AND DISCUSSION

In this project we have performed 3 models such as random Forest,Linear regression and decision tree. The accuracy of randoom forest was 83 precent where as accuracyof 2nd model(linear regression) is 65.3 percent and 3rd model(Decision Tree) is 75.4 percent. after hypertuning the regression model accuracy got increased to 86 percent.so hypertuning increases the accuracy of a model. Regression model gives the highest accuracy here in our project.

Table 1

Method	RMSE	R2
LR	2582.80	0.653
DECISION TREE	2172.79	0.754
RANDOM FOREST	1817.81	0.828

V. CONCLUSION AND FUTURE SCOPE

In this Project , We have Presented the Previous Study on Ticket Price Prediction and the strengths and weaknesses and also the Drawbacks in their Study. Our main focus is to know which algorithm is best suited for this use case and also the accuracy. There are 2 Kinds of Models one dynamic pricing strategies used by Airways to increase their revenue. And the other is to find the correct time to buy a ticket and optimal cost of ticket(which saves money for the customer).

In this study, a machine learning framework was developed to predict the quarterly average fare value on the market phase level. many options were extracted from the datasets and combined along with economics information, to model the aviation market segments. With the assistance of the feature choice techniques, our planned model is in a position to predict the quarterly average airfare value with associate adjusted R square score of zero.869.Thus, our study demonstrates the effectiveness of machine learning algorithms and techniques, yet as compares the performance of varied machine learning classifiers and finds the simplest one for the airfare value prediction task.

In the future, our framework will be extended to incorporate air ticket dealing info, which may give additional detail about a specific itinerary, like time and date of departure and arrival, seat location, coated adjuvant product, etc. By combining such information with the prevailing market phase and economics options within the current framework, it is possible to make a additional powerful and comprehensive airfare value prediction model on the daily or maybe hourly level. moreover, fare value in a very market phase will be affected by a sudden inflow of huge volume of passengers caused by some special events. Thus, events info will also be collected from numerous sources, that embody social platforms and news agencies, on complement our prediction model. in addition, we'll investigate alternative advanced mil models, like Deep Learning models, while working to boost the prevailing models by standardisation their hyper-parameters to achieve the simplest design for fare price prediction.

REFERENCES

[1] B. Mantin and B. Koo, “Dynamic price dispersion in airline markets,” Transportation Research Part E: Logistics and Transportation Review, vol. 45, no. 6, pp. 1020–1029, 2009.

[2] S. Lee, K. Seo, and A. Sharma, “Corporate social responsibility and firm performance in the airline industry: The moderating role of oil prices,” Tourism Management, vol. 38, pp. 20–30, 2013.

[3] 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.

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

[5] Aegean Airlines, https://en.aegeanair.com.

[6] https://github.com/humain-lab/airfare_prediction.
[7] C. Koopmans and R. Lieshout, "Airline cost changes: To what extent are they passed through to the passenger?" Journal of Air Transport Management, vol. 53, pp. 1–11, 2016.
[8] R. Ren, Y. Yang and S. Yuan, "Prediction of airline ticket price," Technical Report, Stanford Univerisy, 2015.
[9] B. Derudder and F. Witlox, "An appraisal of the use of airline data in assessing the world city network: a research note on data," Urban Studies, vol. 42, no. 13, pp. 2371–2388, 2005.
[10] H. Baik, A. A. Trani, N. Hinze, H. Swingle, S. Ashiabor, and A. Seshadri, "Forecasting model for air taxi, commercial airline, and automobile demand in the united states," Transportation Research Record, vol. 2052, no. 1, pp. 9–20, 2008.
[11] T. Janssen, T. Dijkstra, S. Abbas, and A. C. van Riel, "A linear quantile mixed regression model for prediction of airline ticket prices," Radboud University, 2014.

