Airfare Price Prediction Using Machine Learning Algorithms

Mohit Vyavhare
Student, BSIOTR
Pune, Maharashtra
411014

Komal Wable
Student, BSIOTR
Pune, Maharashtra
411014

Ayush chothe
Student, BSIOTR
Pune, Maharashtra
411014

Shreya Kute
Student, BSIOTR
Pune, Maharashtra
411014

Prof. Ashwini Taskar
Project Guide, BSIOTR
Pune, Maharashtra,
411014

ABSTRACT

There is rapid development in Computer Science or Technology, it has become a major problem for the users how to find useful or needed information. Data mining can be seen as an area of AI that seeks to extract pieces of information and patterns from enormous amounts of data stored in databases. Recent research on features selection has been conducted to find efficient methods for the selection of features. Feature choice involves a mix of search and attribute utility estimation and analysis with relation to specific learning schemes. There is square measure many ways to pick out options in an exceedingly system and genetic algorithms (GA) is one in all the foremost used ways. This paper tells us about an overview of the feature selection Algorithm which is going to search the feature space using the idea of evolutionary computation, to find the optimal feature subset.

Keyword: - *Data mining, K Feature selection, Genetic algorithm, etc.*

1. INTRODUCTION

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 analyze the options that influence transportation and to develop and tune models to predict the transportation well ahead.

The rest of this paper is organized as follows: Section II, presents connected work regarding this paper. Section III describes Methodology and Section IV discusses their results. Finally, Section V concludes conclusion

2. RELATED WORK

Guessing the price of airline tickets has been a daunting task ever since Features involved in variable values time and make the price flexible. For the past ten years, Researchers have included machine learning algorithms and data mining techniques to have a better model of visual values.

Among them are retrospective models, such as Linear Regression (LR), Vector Support Machines (SVMs), Random Forests (RF), which is often used to predict the exact cost of a plane price.

Preliminary work was also considered using the separation models in order predict travel styles. ren et al. proposed using LR, Naive Bayes, Softmax regression, and SVM (Support Vector Machine) to create a prediction model and divide the ticket value into five barrels (60% to 80%, 80% to 100%, 100% to 120%, etc.) to compare values related to the total value. More than nine thousand data points, covering six features (e.g., departure week begins, price date, number stops in the travel system, etc.), were used to build models.[12] Their SVM retrieval model failed to produce a satisfactory result. Instead, the SVM fragmentation model was used to distinguish values "up" or “Below” is a measure.

In, four LR models were compared to obtain the best results an appropriate model, which aims to provide impartial information to the passenger whether to buy a ticket or wait a long-time best price. The authors have suggested the use of line quantile mixed models predict low ticket prices, which is so-called "real agreements". However, this work is limited only one class of tickets, savings, and only one one-legged flight guide from San Francisco Airport to John F. Kennedy Airport. Wohlfarth et al. included integration as a first class with advanced supervised learning algorithms (split tree (CART) and RF) to help make customer decisions process.[14] Their framework uses the K-Means formula to group planes with similar behavior within the value varies. See then use CART to translate logical rules, yet as RF to provide data on the importance of every facet. Also, the authors point out that one factor, a number for the remaining seats, is a crucial factor in guessing the price of tickets. Apart from the flight-related features, there are many other features that affect the competitive market. Accurate prediction I market demand, for example, could undermine a travel agency accumulated costs, caused by excessive purchases, or lost orders. In, the author used Artificial Neural Network (ANN) and Genetic Algorithms (GA) for prediction.[6]

Revenue from the sale of airline tickets to a travel agency. Statistics included include the international oil price, a weighty Taiwan stock market index, Taiwan's monthly unemployment rate, and so on. Specifically, GA selects the best input features to improve ANN performance. The model showed efficiency with 9.11% Mean Absolute Percentage error.

Since 2017, machine learning has improved models are considered to improve pre-draft aircraft prices. Tziridis et al. use eight machines learning models, including ANN, RF, SVM, and LR, to predict ticket prices and compare their performance.

The best retrospective model obtained 88% accuracy. In comparison, the Bagging Regression Tree is visible as the best, most robust and untouchable model using a set of different input features. In, Deep Repressors The overlap was proposed to achieve more accurate predictions.

The proposed method is a novel intended for multiple modes RF and SVM as regressors and can be easily applied to other domains of similar problems. As flight ticket data is not well organized and ready direct analysis, collection and processing of such data on a regular basis it takes a lot of effort. For more analysis available in books, researchers tested the effectiveness of their model's different data sets specify data from the web or requesting confidential data from partner organizations. As a result, it is difficult to replicate research and behavior performance measurement models. For U.S. airlines, fare data is publicly available on T100 and DB1A / 1B database. However, due to the limited interaction between prices and specific flight information, these are such databases it is rarely used independently for scientific research results.[8] However, researchers are interested in the analysis of the dispersion price; for example, there is many which you may consider investigating information from those details. The author also includes Saber Air-Price data, provided by SABRE, but only for themselves provide information for their online users. As this is online user data does not represent the entire consumer market, either can override results obtained from data.

Compared to the present and recent work, our proposed framework can handle only the pricing function using public data sources with fewer features. And, it is not limited by any segment of the market that normally limits your existing work, this proposed framework can be used to predict the flight price of any market.

3. METHODOLOGY

3.1 Feature Selection

Phase one (Feature Selection) - throughout this part the foremost informative options of a flight verify the costs of the air tickets area unit set. This part is extremely vital since it defines the matter underneath determination. For every flight, the subsequent options were considered:

F1: Feature one - time of departure.
F2: Feature two-point.
F3: Feature three - ranges of free baggage (0, 1, or 2).
F4: Feature four - days left till departure.
F5: Feature five - ranges of intermediate stops.
F6: Feature a half-dozen - vacation days (yes or no).
F7: Feature a seven-night long flight (yes or no).
F8: Feature eight - days of the week.

It is a price to notice that the influence of some essential features from the higher than list is examined through AN "one leave-out" rule. we tend to additionally wish to clarify that the feature F4 indicates the number of days between the price ticket purchase and the day of the flight.

Phase two (Data Collection) - during this study, our interest is focused on the prediction of one transportation value while not return. For the sake of the experiments, a group of flights to the same destination for the period between December and July is collected. For each flight, the eight options (F1:F8) were manually collected from the Web, 1814 flights were recorded entirely and are out there in [13].

Phase three (ML Models Selection) - Eight state of the art regression metric capacity unit models [8], [10], [14], [15], [16] were designated for the present study and applied to identical information of flights. The metric capacity unit models compared during this work square measure the following: Multilayer Perceptron (MLP), Generalized Regression Neural Network, Extreme Learning Machine (ELM), Random Forest Regression Tree, Regression Tree, Bagging Regression Tree, Regression SVM (Polynomial and Linear), Linear Regression (LR).

Phase four (Evaluation) - The 1814 flights collected in part 2, were employed in a 10-fold cross-validation procedure to coach the previously mentioned milliliter models. The performance indices used to compare the models square measure the prediction accuracy (% - MSE (Mean Squared Error) between the specified and expected prices) and therefore the time in seconds required to coach every model

3.2. Grid Search Tuning of Hyperparameters

3.2.1. Grid Search Method:

Machine the learning model has many limitations tuning too to adjust these parameters; the performance of the model can better. Hyper parameter tuning is the best way to do it different number of parameter compounds to be tested an editor function. Checking the separator by using training data will create basic machine learning a problem called over fitting.

Overfitting is the situation we are in now, which is a model that performs poorly in test data and is remarkably effective rain data. Therefore, the opposite validation is applied to the grid Find out how to do hyper parameter optimization. The grid search method is a method used to identify correct separator parameters for model accurately predict other unlabeled data. Grid Search The method is used to tune some potential hyper parameters learn directly from the training program. Separation the model has more hyper parameters and gets the best. The combination of these parameters is a challenging process. One of the best methods used for this purpose is Grid Search way. Grid Search Method defines the range of values in each parameter h1, h2 and h3. It will build as many X versions as possible a combination of h1, h2 and h3. This is a type of hyper parameter values are known as grid.[16] Input data

is divided into training set, test set and a confirmation set. The tuning process is done by sorting data sets into distinct parts. Then, the Random Forest Classifier trained with parts of each n-2 personalized solution selected by tuning system. I am verification set is used to validate the advanced model and the latter part is used to test the model. The accuracy of the test once the accuracy of verification is assessed using a model. Then the model is motivated by a set of training and the hyper parameter value determined by the tuning method.

These steps are repeated N times. Directing the search process, central accuracy of verification is used as an appropriate number. Eventually, it will bring back the highest person the accuracy and efficiency of the method is a measure to check the accuracy of that person.

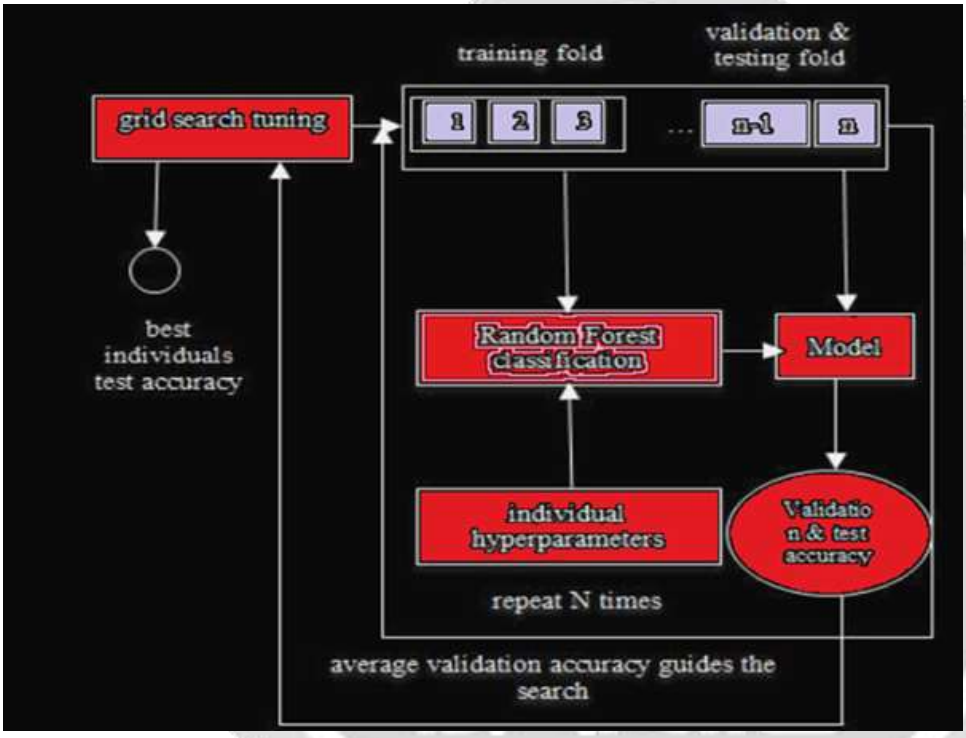

Fig. 3.2.1 Architecture of Hyper-parameter Tuning [17]

3.2.2. Random Forest Classifier

Random Forest could be a versatile, easy-to-use machine learning formula that produces, even while not hyper-parameter standardisation, a superb result most of the time. It is also one of the most widely used algorithms, due to its simplicity and versatility.

Random Forest is a supervised learning algorithm. The "forest" is constructive, a collection of tropical trees, usually trained in the "enclosing" method. A common idea of how to pack a bag is that a combination of learning models enhances the overall effect.

The informal forest forms many decision-making trees and combines them together to obtain the most accurate and stable prediction.

The main unit of random forest sections is the decision tree. Decision tree is a hierarchical structure constructed using the elements (or independent variants) of a data set. Each node of the decision tree is divided into a scale associated with a small set of features. Random Forest could be a assortment of call trees related to a collection of bootstrap samples generated from an explicit information set. The nodes are categorized based on the entropy (or Gini index) of the selected subset of features. The subset sets created from the original data set, using bootstrapping, are the same size as the original data set. Detailed information on random forest categories can be found in the papers by Breiman (Breiman, 1996, 2001).[10][13] In the normal course of a random forest, the bootstrapping method facilitates the development of the random forest with a set of the required number of deciduous trees to improve the accuracy of the sections in the sense of minimizing fragmentation as stated in Suthaharan (2015). Then a method called bagging (bootstrap aggregate) is used to select the best trees with a voting system.[14] This common random forest approach is the one used in the construction of the proposed brain computer.

Another excellent quality of the random forest algorithm is that it is quite easy to measure the relative value of each element in the prediction. Sklearn provides a useful tool for measuring the value of a feature by looking at how the nodes of the tree using that feature reduce pollution in all trees in the forest.[11] It automatically calculates this feature for each element after training and measures the results so that your total value is equal to one.

If you do not know how a decision tree works or what a leaf or place is, here is a good description from Wikipedia: "'In a tree each inner place represents a' check 'on the attribute (e.g., that a coin comes out with heads or tails), each branch represents the result test, and each leaf node represents a class label (a decision taken after counting all the factors).'"

By looking at the value of the feature you can decide which features you might omit because they do not provide enough (or sometimes not at all) in the forecast process. This is important because the general rule in machine learning is that many of the features you have are likely to suffer from overuse and vice versa.

3.2.3. Linear Regression

Linear regression refers to the mathematical technique of fitting given information to a performance of an explicit kind. It is best proverbial for fitting straight lines. During this paper, we tend to make a case for the idea behind regression and illustrate this system with a true world information set.

"The horizontal axis is the miles, and the vertical axis is the cost sold. This figure also shows the plot of the line y = ax + b in red. The reader will agree that this line approximates the data well. In fact, this line is the line that best fits the data. As we will see in this article, Linear Regression makes the meaning of best fits the data precise" [1].

Linear Regression could be a widely used technique employed in several branches of science and technology. It is a core topic in Machine Learning and information Science, two very talked-about fields that have found a large variety of applications. This information could be a self-contained description of linear regression, the specified algebra, and the needed Python for its Implementation.

Linear regression could be a quiet and straightforward regression statistical procedure method used for prognosticative analysis and shows the link between the continual variables. rectilinear regression shows the linear relationship between the variable (X-axis) and the variable (Y-axis), consequently known as rectilinear regression. If there is one input variable (x), such rectilinear regression is termed straightforward rectilinear regression. And if there is quite one input variable, such rectilinear regression is termed multiple rectilinear regression. The rectilinear regression model provides a sloping line describing the link inside the variables.

The top of graph presents the linear relationship between the variable quantity and freelance variables. Once the worth of x (independent variable) increases, the worth of y (dependent variable) is likewise increasing. The line is cited because of the best match line. supported the given knowledge points, we tend to attempt to plot a line that models the simplest points.

To calculate the best-fit line simple regression uses a conventional slope-intercept type.

$$y = mx+b \quad \Longrightarrow \quad y = a_0+a_1x$$

y= Dependent Variable.

x= Independent Variable.

a0= intercept of the line.

a1 = Linear regression coefficient.

4. CONCLUSIONS AND FUTURE SCOPE

According to the outcome of this paper reported on a basic analysis on "airfare price prediction". We gathered transportation information from a selected Indian airline from the online and showed that it is possible to predict costs for flights supported historical fare information.

We have known the compliments and complaints of shoppers, variations in sentiment over an amount of your time. This analysis provides a general opinion of passengers towards airlines. XG-boost gives better results compared to other machine learning models. The dataset we have collected from the Kaggle consisting of more than 2000 Indian airlines flight data and deprived that it is easy to forecast the prices of airlines based on previous price data.

The outcome of the experiments derives from the fact that machine learning models satisfies the need of forecasting the airfare costs. Other vital parts in airfare prediction are the feature selection and data collection from which we have derived some helpful conclusions. We have derived some features which affect the airfare forecasting at most using experiments. Apart from the options elect, there are different options that would improve the price forecasting accuracy. In the future, this work could be extended to predict the airfare prices for the entire flight map of the airline. We need to test these machine learning algorithms on the huge airline datasets, but the preliminary studies remark on the capabilities of ML Models to help the end users to give an idea when to buy the tickets in what period so they will be in profit.

Features took from external factors such as social media data and search engine query are not taken. Therefore, we will introduce and discuss the concept of using social media data for ticket/demand prediction (i.e., twitter sentiment analysis). In today's date, social media sentiment analysis has become an enough library of knowledge for varied data processing models. For example, social media data has been used for event prediction, price prediction and tourist traffic flow prediction. A similar approach might be followed to extract helpful social media info associated with various external factors poignant airline rider demand and price tag value. For example, Analysis of various twitter hash tags might provide valuable information concerning the presence of an occurrence at an origin/destination city, competitors' promotions, volume of traveler traffic flow, weather condition, economic activity etc. This in turns may allow us to predict the amendment in price tag price/demand. It is expected that a data processing model that utilizes info ensuing from social media information would provide higher results than existing ad forecasting route demand and our price ticket worth. But there are no facilities or studies available on the Web that use social media data to forecast the demand of path or cost value. There is room for improvements in several areas including predicting exact value of ticket prices/demand, dataset issues, limited number of features, lacking generality, better prediction techniques and performance and complexity issues

.

5. REFERENCES

[1]. R P. Malighetti, S. Paleari and R. Redondi, "Pricing strategies of low-cost airlines: The Ryanair case study," Journal of Air Transport Management, vol. 15, no. 4, pp. 195-203, 2009.
[2]. P. Malighetti, S. Paleari and R. Redondi, "Has Ryanair's pricing strategy changed over time? An empirical analysis of its 2006–2007 flights," Tourism Management, vol. 31, no. 1, pp. 36-44, 2010.
[3]. W. Groves and M. Gini, "A regression model for predicting optimal purchase timing for airline tickets," Technical Report 11-025, University of Minnesota, Minneapolis, 2011.
[4]. 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.
[5]. M. Papadakis, "Predicting Airfare Prices," 2014.
[6]. T. Janssen, "A linear quantile mixed regression model for prediction of airline ticket prices," Bachelor Thesis, Radboud University, 2014.
[7]. R. Ren, Y. Yang and S. Yuan, "Prediction of airline ticket price," Technical Report, Stanford Univerisy, 2015.
[8]. S. Haykin, Neural Networks – A Comprehensive Foundation. Prentice Hall, 2nd Edition, 1999.
[9]. S.B. Kotsiantis, "Decision trees: a recent overview," Artificial Intelligence Review, vol. 39, no. 4, pp. 261-283, 2013.
[10]. G.B. Huang, Q.Y. Zhu and C.K. Siew, "Extreme learning machine: Theory and applications," Neurocomputing, vol. 70, no. 1-3, pp. 489- 501, 2009.
[11]. G.A. Papakostas, K.I. Diamantaras and T. Papadimitriou, "Parallel pattern classification utilizing GPU-Based kernelized slackmin algorithm," Journal of Parallel and Distributed Computing, vol. 99, pp. 90-99, 2017.
[12]. Aegean Airlines, https://en.aegeanair.com.
[13]. https://github.com/humain-lab/airfare_prediction.
[14]. L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5-32, 2001.
[15]. L. Breiman, J. Friedman, R. Olshen and C. Stone, Classification and Regression Trees. Boca Raton, FL: CRC Press, 1984.
[16]. 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.
[17]. Siji George C G, B. Sumathi2, "Grid Search Tuning of Hyperparameters in Random Forest Classifier for Customer Feedback Sentiment Prediction", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 9, 2020
[18].Mengyu Huang, Ningbo Xiaoshi High School," Theory and Implementation of linear regression", 2020 International Conference on Computer Vision, Image, and Deep Learning (CVIDL).