INDIAN FLIGHT FARE PREDICTION: A PROPOSAL

Jaywrat Singh Champawat[1], Udhhav Arora[2], Dr. K. Vijaya[3]

[1,2,3]*Department of Computer Science and Engineering, SRMIST, India*

ABSTRACT

The price of a ticket of an airline changes very rapidly these days, and the difference is a lot. It can vary even in a few hours for the same flight or even the some particular. Customers want to get the cheapest price possible while the airline companies want the maximum profit and revenue possible. To solve this problem, researchers have proposed different models to save consumer's money- minimum price predicting model and models that tell us an optimal time to buy a ticket while airlines use techniques such as demand prediction and price discrimination to maximize their revenue. There has been some research on the airfare market but it is seen that a very few of them are done on the Indian market. This paper aims to find an efficient model which will predict the price of a ticket with better accuracy, with features that might have been hind-sighted in the previous models, solely on the Indian market.

I. INTRODUCTION

These days, carrier ticket costs can shift powerfully and fundamentally for a similar flight, in any event, for close by seats inside the same cabin. Clients are trying to get the most minimal cost while carriers are attempting to keep their general income as high as could reasonably be expected and boost their benefit.

Aircrafts utilize different sorts of computational methods to expand their income, for example, demand forecast and value segregation. From the client side, two sorts of models are proposed by various scientists to set aside cash for clients: models that anticipate the ideal time to purchase a ticket and models that predict the minimum ticket cost.

Our initial investigation shows that models on the two sides depend on restricted arrangement of highlights, for example, verifiable ticket value information, ticket buy date and flight date. Parameters on which fares are calculated-

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

We will perform Exploratory Data Analysis on the given information. We will discover correlation between the highlights. After that a Machine Learning model will be made to utilizing those highlights.

II. LITERATURE SURVEY

Flight fare prediction is a very challenging task as a lot of factors depend upon the price of a flight ticket. Many researchers have used different Machine Learning algorithms to get a model with higher accuracy in prediction of the ticket price. Researchers have using various regression model such as Support Vector Machines (SVM), Linear Regression (LR), Decision Tree, Random Forests etc. to predict accurate flight fare.

After further reading it was found that the models are divided into two types- one which predicts the minimum price of an air ticket and one which helps to generate maximum revenue, which can be referred to as customer side models and airline side models respectively. There has been other research besides these categories also, such as research on various factors which lead to the change is ticket prices and how demand changes its price. Those researches found out that customers who travel for leisure are more sensitive to the ticket prices rather than the customers who travel for business purposes. The date of booking and the date of travel is also looked upon by many researchers as how it influences the surge in price. Studies are also done on the effects of delays on the fare.

III. PROPOSED WORK

Machine Learning is the study of algorithms that tend to improve through experience. Formally you improve a task, based on performance measure, based on experience. It is a sub topic of Artificial Intelligence. Where AI deals with intelligent tasks performed by a non-human agent, ML deals with making decisions based on acquired data. Machine Learning is a very vast field in computer science. Machine learning can be supervised, or non-supervised. Problems in ML include-

- Clustering: given some data, we have to figure out how to group them together
- Regression: given some data, can we predict outcome value
- Classification: given some data, can we figure out which category it belongs to?

Data used in ML can be of three types: Categorical Nominal, Categorical Ordinal, and Continuous. We wish to work with the regression models which we suspect would give us more accurate results. The models that have been shortlisted to be worked upon are-

3.1 RIDGE REGRESSION

It is used in models where the data suffers from multicollinearity. It adds a degree of bias to the regression estimates, thus reducing the standard errors. This is done to try to make the estimates more reliable. It performs the L2 regularization. Its cost function is:

Min(||Y – X(theta)||^2 + λ||theta||^2) (1) λ is the penalty term. λ is denoted by an alpha parameter in the ridge function which controls the penalty term when its value is changed. Ridge works well if there are many large parameters of about the same value.

3.2 LASSO REGRESSION

It is a type of linear regression which shrinks the data values towards a central point. Its acronym is Least Absolute Shrinkage and Selection Operator. It uses L1 regularization which adds a penalty which is equal to the absolute value of the magnitude of coefficients. Its aim is to minimize the following function:

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad (2)$$

λ is the tuning parameter which controls the strength of L1 penalty. It is basically the amount of shrinkage. Lasso tends to do well if there are a small number of significant parameters and the others are close to zero.

3.3 K NEIGHBORS REGRESSOR

It is an algorithm that stores all the available cases and predict the output using a similarity function like distance function. It calculates the mean of the numerical target of the K nearest neighbors. It uses the same distance functions as KNN Classifier. Distance function used by KNN Regressor are: Euclidean, Manhattan and Minkowski. In case of categorical variables, it uses Hamming distance function. KNN is a versatile method which is useful for both regression and classification problems. It provided high accuracy and no assumptions about the data are made.

3.4 DECISION TREE REGRESSOR

It uses decision making tools that uses a flowchart like structure which includes all the possible results, their input costs and utility. It works for categorical and continuous output variables. Nodes can either be a decision node or a result node or end node. The outputs from a decision tree regressor are easy to read and interpret. This way, the data can be used generate important insights on costs, probabilities, and alternatives to various strategies. Moreover, decision trees take less effort for data preparation. Once the variables have been created, there is less data cleaning required.

3.5 RANDOM FOREST REGRESSOR

It uses ensemble learning method for regression. It combines predictions of different learning algorithms to make a more accurate prediction as compared to using a single model. The trees run in parallel without interacting with each other. It performs both regression and classification tasks parallelly. Random forest automates the missing values present in the data. It was used to reduce the risk of overfitting the data and to improve the accuracy of the model.

IV. IMPLEMENTATION

This paper will use Python3 to implement the machine learning algorithms to make a model that will predict the airfare with high accuracy. Various python libraries are imported to perform these actions.

There are various steps in making a ML model which starts with importing the dataset and data cleansing. All the null values and duplicate values are removed from the dataset. Then the data is encoded by converting some variables into a specific format. It converts the categorical data into numerical data.

After the dataset is worked upon, feature selection is done. Features or variables that are not so important are removed from the dataset. Exploratory data analysis is performed to provide insights on the data and to identify the important features by using Extra Tress Regressor. Feature Engineering is performed to reduce the computational costs and sometimes to improve accuracy. This is done with the help of a correlation matrix. After this, the data is split into train and test data where train data being used to train the models while the test data is used to check the accuracy of the models. The optimal features and parameters are decided after performing hypertuning. After all the models are trained, their accuracy is checked by their R-squared value.

V. CONCLUSION

This paper proposed to find a machine learning model which gives higher accuracy in predicting the fare of the Indian flights. On working with different models, it was found out that Random Forest algorithm showed the highest accuracy in predicting the output. The paper gives better result than previous looked upon models and aims to improve in the future.

REFERENCES

[1] K. Tziridis, T. Kalampokas, G. A. Papakostas and K. I. Diamantaras, Airfare prices prediction using machine learning techniques, 2017 25th European Signal Processing Conference (EUSIPCO), Kos, Greece, 2017, pp. 1036-1039, https://doi.org/10.23919/EUSIPCO.2017.8081365 .

[2] Juhar Ahmed Abdella, Nazar Zaki, Khaled Shuaib, Fahad Khan, Airline ticket price and demand prediction: A survey, *Journal of King Saud University - Computer and Information Sciences, 2019, ISSN 1319-1578, https://doi.org/10.1016/j.jksuci.2019.02.001* .

[3] Martijn Brons, Eric Pels, Peter Nijkamp, Piet Rietveld, Price elasticities of demand for passenger air travel: a meta-analysis, *Journal of Air Transport Management, Volume 8, Issue 3, 2002, Pages 165-175, ISSN 0969-6997, https://doi.org/10.1016/S0969-6997(01)00050-3* .

[4] Silke J. Forbes, The effect of air traffic delays on airline prices, *International Journal of Industrial Organization, Volume 26, Issue 5, 2008, Pages 1218-1232, ISSN 0167-7187, https://doi.org/10.1016/j.ijindorg.2007.12.004* .

[5] Szopiński, Tomasz and Nowacki, Robert, The Influence of Purchase Date and Flight Duration Over the Dispersion of Airline Ticket Prices, *Contemporary Economics, Vol. 9, No. 3, pp. 253-366, 2015, https://ssrn.com/abstract=2694195* .

[6] María-Encarnación Andrés Martínez, José-Luis Alfaro Navarro, Jean-François Trinquecoste, The effect of destination type and travel period on the behavior of the price of airline tickets, *Research in Transportation Economics, Volume 62, 2017, Pages 37-43, ISSN 0739-8859, https://doi.org/10.1016/j.retrec.2017.03.003* .

[7] Scikit-learn: Machine Learning in Python, *Pedregosa et al., JMLR 12, pp. 2825-2830, 2011*.