International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 11, November-2022

Analysis of Flight Fare Detection using Machine Learning

Arya Karambelkar
Information Technology
K. J. Somaiya College of Engineering

Param Mamania
Information Technology
K. J. Somaiya College of Engineering

Mr. Vaibhav Chunekar
Information Technology
K. J. Somaiya College of Engineering

***Abstract*—Recent years have seen a dramatic rise in air travel for a variety of reasons, including economic expansion and the emergence of low-cost airlines. Fares are very variable because Indian airlines employ a revenue management system to respond to market conditions in real time. The cost of a flight ticket varies based on duration of the flight, destination, route, arrival time, departure time as well as certain events like holidays or vacations. The main aim is to predict the correct flight fare using Machine Learning Techniques and Auto ML. Additionally, the properties of the provided dataset have been examined using a variety of visualisation techniques, including scatter plots, distribution plots, catplots, and ggplots. The results show that the Random Forest Regressor and Randomized Search CV techniques give highest accuracy with the fare prediction dataset.**

***Keywords*— Regression, Machine Learning, Deep Learning, Auto ML, Auto SK Learn, Flight Fare Detection**

I. INTRODUCTION

In today's fast-paced world, everyone expects the fastest way of approach to any problem. This scenario arose in the sphere of transportation, where it is regarded as the most essential platform for the development of various enterprises such as trade, finance, IT personnel, tourism, and so on. In this scenario, it is of the utmost importance to provide the quickest and safest means of transportation, and the solution to this problem is the transportation provided by airlines. Airline transport is the backbone of the tourism industry. Air travel is the most popular mode of international transportation, and India receives a sizable influx of international visitors every year. This results in the creation of jobs, both in India and in the countries that are visited by tourists.

It is in the best interest of airlines to maximise their profits, and there are two main types of customers they serve: leisure travellers, who are more price-conscious because they're paying the cost themselves, and business travellers, who are less price-conscious because they're not paying the cost themselves and often make their travel decisions much closer to the time of travel. Airline seats are a very perishable good; once a plane takes off, the passenger's ability to generate revenue for the airline has ended. If a company isn't careful with its price, they can have to take a trip with empty seats, or they might have a full plane yet lose money since they couldn't charge more. This means that airfares can range greatly since different customers are ready to pay varying sums to meet their diverse requirements. Airline companies use this reality to their advantage.

Distance, flight time, peak season, number of stops, and destination are just a few of the variables that can drive up or down the cost of an airline ticket. The cost of the flight can be lowered to some amount by adjusting the aforementioned variables. In this paper, we use Machine learning approaches, including Auto ML, to the problem of estimating the cost of airline tickets.

II. LITERARY REVIEW

Recent study on the topic of flight fare prediction has aimed at developing data-driven approaches for forecasting future flight prices and their trends.

Ratnakanth G [1] utilised Deep Neural Network that functions same as the human brain. The data is preprocessed, and the Min-Max normalisation approach was used to change the values that are already present in the dataset in order to obtain excellent performance. Randomised Search CV algorithm is used for hyperparameter tuning of the Deep Learning algorithm. Finally, the dataset was visualised using univariate analysis, bivariate analysis, and correlative analysis for all of the features in the dataset.

R. Raja Subramanian et al [2] collected data from MakeMyTrip, Data World and Kaggle to build Machine Learning models. The paper uses KNN Regression, Linear Regression, Lasso Regression, Ridge Regression, and Random Forest Regression. The models have been implemented using the sci-kit learn python library. The research found out the Random Forest Regressor algorithm works the best with high accuracy.

S. Naveen Prasath et al [4] researched and found out the factors that impact the flight fare fluctuations. The paper systematically demonstrates the K-Nearest Neighbours technique to estimate the prices at a particular instance using Machine Learning techniques. After doing a comparison of the highest and lowest levels of airfare for specific days, weekends, and times of the day, such as morning, evening, and night, regression analysis was carried out to predict the flight prices.

Zhichao Zhao et al [5] carried out flight fare prediction in China based on the multi-attribute dual-stage attention (MADA) mechanism. In order to encode and decode the input multi-dimensional fare-related characteristics, a Seq2Seq neural network has been implemented. In addition, effective information variables are extracted by the utilisation of dual-stage attention processes. For the purpose of determining the pattern of fluctuating fares, the mean square error loss function is used to train the real data.

III. METHODOLOGY

A. Importing Libraries and Dataset

The first step is importing different libraries that will help in feature extraction, data analysis and model building. Pandas [15] and NumPy [14] are installed in the Google Colaboratory [23]. Matplotlib [18] and Seaborn [16] libraries are installed for visualisations. Next, two datasets consisting of training data and test data are installed in the Google Colaboratory. The testing data and the training data have been merged to produce a single dataset in order to make the process of Feature Engineering more straightforward. The chosen dataset contains both training and testing data, and it has a total size of 13,354 data rows and 11 columns. These rows and columns contain different features such as the total number of stops, the route, the duration, and the destination.

B. Data Analysis

For this purpose, only the training dataset is used to perform data analysis. First, the relation between the airline and price is found out by using catplot [17] which is present in the seaborn library [16]. Figure 1 displays the results obtained in the form of data visualisation using catplot.

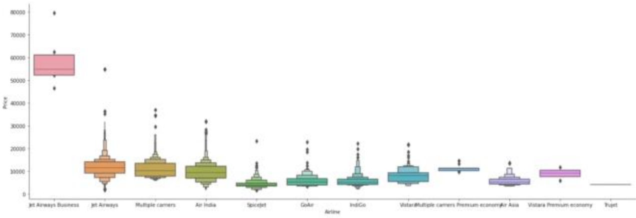


Figure 1. *Relation between airline and price*

Next, the relation between the source and price is calculated using catplot [17]. Figure 2 shows that Bangalore and Delhi had higher quartile ranges than the other cities. In the same manner, a catplot for the relation between destination and price is calculated.

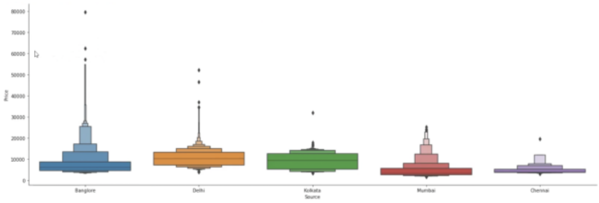


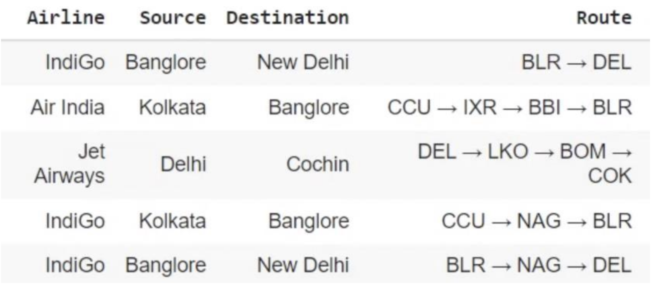
Figure 2. *Relation between source and price*

C. Feature Engineering

The data must now be transformed into a format that the model can comprehend. For the model's interpretation, several forms of alphabetical or continuous data should then be transformed into numeric data. The Date of Journey column in the dataset contains data in the format of DD/MM/YYYY. For simplicity, the Date of Journey column is converted into 3 columns - Date, Month and Year. The original column is dropped. Then, the data type of the previous 3 columns is converted to integer type using the DataFrame.astype() method [24] function.

The arrival time is not of consistent format throughout the dataset and may include both time and date or solely time. For this reason, the column is structured to display only the time in the format HH:MM. The arrival time is then divided into 2 columns - Arrival Hour and Arrival Minute. The departure time is also converted into departure Hour and Departure Time by following the same procedure. Additionally, the data in the column that contains information about the total number of stops has been reformatted such that it will appear as a numeric value with an integer datatype. Similarly, all the other columns will be converted to numeric values.

The dataset includes a column labelled "Route," which displays the path that the flight takes. If it is a flight without any stops in between, then the route will just provide the source and the destination. If the flight, on the other hand, makes many stops, those stops will be shown in the column with the help of an arrow (→). Figure 3 displays the route column in the dataset.



Airline	Source	Destination	Route
IndiGo	Banglore	New Delhi	BLR → DEL
Air India	Kolkata	Banglore	CCU → IXR → BBI → BLR
Jet Airways	Delhi	Cochin	DEL → LKO → BOM → COK
IndiGo	Kolkata	Banglore	CCU → NAG → BLR
IndiGo	Banglore	New Delhi	BLR → NAG → DEL

Figure 3. *Snapshot of dataset*

In the test data, the price column is empty and has null values. This research paper considers both the test data and train data together. For time being, the price column in test data is filled with the mean of the price column in the train data. To use the label encoder [25], it needs to be imported from sklearn.preprocessing. Label encoding is used on all the columns of the dataset. The method fit_transform is used which is the combination of fit method and transform method, it is equivalent to fit().transform(). If we use fit and transform separately when we need both, then it will decrease the efficiency of the model so we use fit_transform() which will do both the work. The data is arranged categorically in a systematic manner. All the columns have been converted into labels.

D. Feature Selection

This section views all the features and selects the features that will be useful for developing the model. The first step for this is to separate the data into independent and dependent variables. Next, Lasso and SelectFromModel are imported from the sklearn library [20]. The SelectFromModel takes a parameter as an input and on the basis of that parameter, it will select the model. The parameter taken here is Lasso. In statistics and machine learning, lasso is a regression analysis approach that uses variable selection and regularisation to improve the predictability and interpretability of the produced statistical model. The data is split into train and test data. The alpha value in Lasso is selected as 0.005. It is preferable to select lower alpha values. The model is fitted with the train data and the column names of the selected features are displayed.

E. Model Building using ML

The model is built using the Random Forest Regression [21] Machine Learning algorithm. A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. It is of two types - Random Forest Classifier and Random Forest Regressor. This paper uses Random Forest Regressor.

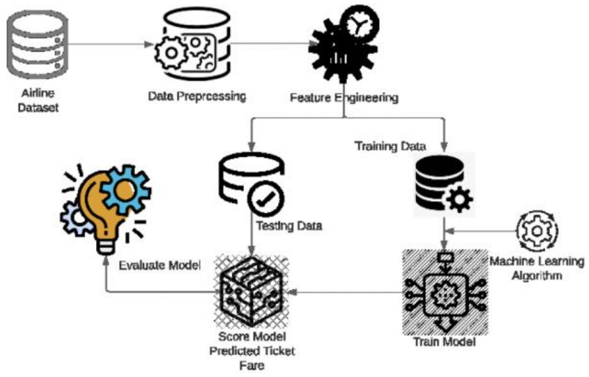


Figure 4. *Architecture Diagram*

This paper discusses the use of Random Forest because this algorithm takes less time to train. It also predicts output with high accuracy and even for the large dataset, it runs efficiently. Random Forest Regressor can maintain accuracy even when a large proportion of the data is missing. The features selected for training are Airline name, Source, Destination, Date, Month, Year, Number of Stops, Arrival time, and so on. Distribution plot of the difference between the actual and predicted values is created. From Figure 5, we observe that zero is in abundance and the plot follows normal distribution. This suggests that the difference between predicted value and actual value is zero. This ensures good accuracy of the project.

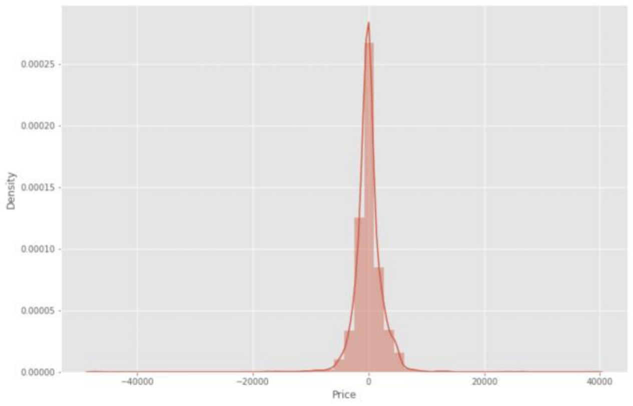


Figure 5. *Distribution plot*

F. Model Building and Prediction using Auto SK Learn

Automated Machine Learning will automate all the machine learning model building and hypertuning parts. Auto SK Learn is used for this purpose in this project. Auto SK Learn only automates the model building part. The preprocessing part needs to be done manually. If raw data is passed in the model without any preprocessing, it will cause the model to fail. This is because there will be different types of data available - unstructured, numerical, object, alphabetical, etc. After performing Auto ML, the top perfectly fitted models are found out. In this case, gradient boosting and random forest were the top models. Figure 5 displays the scatter plot obtained of the predicted and testing data. The cluster in the scatter plot means that the model has high accuracy and the prediction is performed successfully.

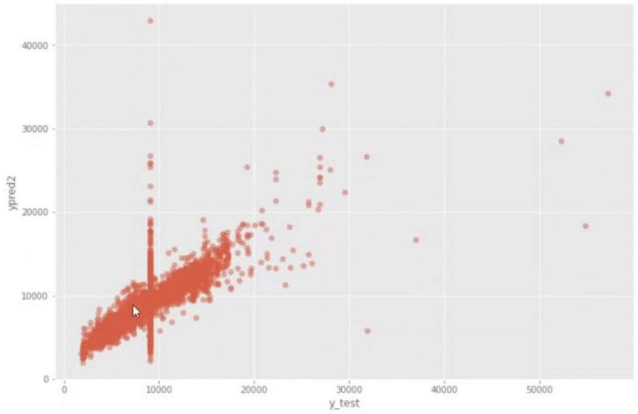


Figure 6. *Scatter Plot*

IV. RESULTS AND INFERENCES

The Random Forest Regressor algorithm has a mean absolute error of 1531.75, mean squared error of 6409856.80 and root mean square error of 2531.77. Coefficient of determination, also called as R^2 score is used to evaluate the performance of a linear regression model. It is the amount of the variation in the output dependent attribute which is predictable from the input independent variable(s). A higher value of R^2 is desirable as it indicates better results. The model built gives R^2 value of 0.61.

One paper utilises Deep Learning techniques such as Deep neural network [1]. The paper suggests that the accuracy obtained by using Deep learning is better than the accuracy obtained using Machine learning models. One of the most common limitations of this project is obtaining information because data is acquired from websites that sell flight tickets.

The paper titled, "Airline Fare Prediction Using Machine Learning Algorithms" predicted a root mean square error of 33.36 when Random Forest Regressor algorithm is used [2]. The model used in the paper is hypertuned so that the error is reduced.

V. CONCLUSION

The research paper depicts how using Automated Machine Learning saves the time of model building but highlights that the data preprocessing part must be done manually and that it cannot be automated. The prediction of the flight rate was carried out successfully using one of the most widely used algorithms - Random Forest Regressor. The accuracy achieved is very high which is seen from the distribution plot and scatter plot obtained from the training data and testing data. The data visualisation techniques have been applied to illustrate the ideology behind the attributes of the dataset. To acquire more reliable findings, more accurate data with greater features might be employed.

REFERENCES

[1] G. Ratnakanth, "Prediction of Flight Fare using Deep Learning Techniques," 2022 International Conference on Computing, Communication and Power Technology (IC3P), 2022, pp. 308-313, doi: 10.1109/IC3P52835.2022.00071.

[2] R. R. Subramanian, M. S. Murali, B. Deepak, P. Deepak, H. N. Reddy and R. R. Sudharsan, "Airline Fare Prediction Using Machine Learning Algorithms," 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), 2022, pp. 877-884, doi: 10.1109/ICSSIT53264.2022.9716563.

[3] C. Chariton and Min-Hyung Choi, "Enhancing usability of flight and fare search functions for airline and travel Web sites," International Conference on Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004., 2004, pp. 320-325 Vol.1, doi: 10.1109/ITCC.2004.1286473.

[4] S. N. Prasath, S. Kumar M and S. Eliyas, "A Prediction of Flight Fare Using K-Nearest Neighbors," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2022, pp. 1347-1351, doi: 10.1109/ICACITE53722.2022.9823876.

[5] Zhao, Z., You, J., Gan, G., Li, X. & Ding, J. 2022, "Civil airline fare prediction with a multi-attribute dual-stage attention mechanism", Applied Intelligence, vol. 52, no. 5, pp. 5047-5062.

[6] Malighetti, P., Paleari, S. & Redondi, R. 2009, "Pricing strategies of low-cost airlines: The Ryanair case study", Journal of Air Transport Management, vol. 15, no. 4, pp. 195-203.

[7] Tziridis, K., Kalampokas, T., Papakostas, G.A. & Diamantaras, K.I. 2017, "Airfare prices prediction using machine learning techniques", 25th European Signal Processing Conference, EUSIPCO 2017, pp. 1036.

[8] Groves, W. & Gini, M. 2013, "An agent for optimizing airline ticket purchasing", 12th International Conference on Autonomous Agents and Multiagent Systems 2013, AAMAS 2013, pp. 1341.

[9] Makridakis, S., Spiliotis, E. & Assimakopoulos, V. 2018, "Statistical and Machine Learning forecasting methods: Concerns and ways forward", PLoS ONE, vol. 13, no. 3.

[10] Lu, M., Zhang, Y. & Lu, C. 2021, "Approach for Dynamic Flight Pricing Based on Strategy Learning", Dianzi Yu Xinxi Xuebao/Journal of Electronics and Information Technology, vol. 43, no. 4, pp. 1022-1028.

[11] Joshi, N., Singh, G., Kumar, S., Jain, R. & Nagrath, P. 2020, Airline Prices Analysis and Prediction Using Decision Tree Regressor.

[12] Boruah, A., Baruah, K., Das, B., Das, M.J., Gohain, N.B. (2019). A Bayesian Approach for Flight Fare Prediction Based on Kalman Filter. In: Panigrahi, C., Pujari, A., Misra, S., Pati, B., Li, KC. (eds) Progress in Advanced Computing and Intelligent Engineering. Advances in Intelligent Systems and Computing, vol 714. Springer, Singapore. https://doi.org/10.1007/978-981-13-0224-4_18

[13] https://www.python.org

[14] https://numpy.org

[15] https://pandas.pydata.org

[16] https://seaborn.pydata.org

[17] https://seaborn.pydata.org/generated/seaborn.catplot.html

[18] https://matplotlib.org

[19] https://en.wikipedia.org/wiki/Machine_learning

[20] https://scikit-learn.org/stable/

[21] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[22] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html

[23] https://colab.research.google.com

[24] https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.astype.html

[25] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html