International Journal for Multidisciplinary Research (IJFMR)

E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Airline Ticket Price Prediction Using Machine Learning Algorithms

A.Krishna Kumar[1], G.L.D.Prasanna[2], B.Ramesh Kumar[3], K.Nanya Sri[4], K.Lasya[5], M.Rakshak[6],

[1]Assistant Professor, CSE Department, Vsm College Of Engineering, Ramachandrapuram, Dr.B.R.Ambhedkar Konaseema Dt, Andhra Pradesh - 533255

[2,3,4,5,6]Student, CSE Department, Vsm College Of Engineering, Ramachandrapuram, Dr.B.R.Ambhedkar Konaseema Dt, Andhra Pradesh - 533255

Abstract

This paper discusses the issue of airfare. A set of characteristics defining a typical flight is chosen for this purpose, with the assumption that these characteristics influence the price of an airline ticket. Flight ticket prices fluctuate depending on different parameters such as flight schedule, destination, and duration, a variety of occasions such as vacations or the holiday season. As a result, having a basic understanding of flight rates before booking a vacation will undoubtedly save many individuals money and time. Analysing 3 datasets to get insights about the airline fare and the features of the three datasets are applied to the seven different machine learning (ML) models which are used to predict airline ticket prices, and their performance is compared. The goal is to investigate the factors that determine the cost of a flight. The data can then be used to create a system that predicts flight prices.Examine the elements that influence flight prices. Subsequently, a flight price prediction system can be developed using the data.

Keyword: Decision trees, stacked regression, random forests, regression, Lasso regression, ridge regression, and prediction models.

SECTION I: INTRODUCTION

Airlines try to keep aeroplane ticket prices under control these days in order to increase their earnings. Most frequent travellers are aware of the ideal times to find discounted tickets. But a lot of inexperienced ticket buyers are sucked into the company's discount trap and end up spending their money. While customers search for the greatest deal, airline firms prioritise making a profit. In order to avoid price rises as the departure day draws near, customers usually try to buy tickets far in advance of the departure date. It is quite difficult for a customer to purchase an aircraft ticket at a very low price because of the vast complexity of the rate models employed by airlines.pricing due to the ongoing fluctuations in price. When there is a need to develop a market and when tickets are more difficult to get, airlines can reduce the price of their tickets. These strategies take into account a variety of economic, commercial, marketing, and social aspects that are all connected to the final cost of a flight. They might be able to maximise their earnings. So, a number of things may have an impact on expenses. Because of the intricacy of the pricing scheme that airlines employ, it is exceedingly difficult for customers to purchase tickets at extremely low costs because prices are always changing. Over the past 20 years, customer and airline surveys have increased steadily. From the perspective of the client In our opinion, determining a favourable time or low price for a ticket purchase is crucial. In this work, we will use machine learning methods to build the models utilising the data that we have obtained from three different sources. Using the suggested approach, customers can save millions of rupees by obtaining the information they require to purchase tickets at the appropriate time.

SECTION II:PROPOSED MODEL

Data Collection-The data collection process is the most crucial aspect of this endeavour. A number of information sources from different websites are used to prepare the models. There are a number of sources available for data regarding airfare, including customer travel sites and APIs. For this project, we have gathered three datasets from various sources.

A. Information Gathering

The first dataset is sourced from Kaggle and consists of data that has been scraped from the MakeMyTrip website. The report has ten thousand data points in total. These consist of the airline, the source and the destination; the itinerary; the number of stops; any extra information; the cost; the arrival and departure times; the date and time; and the duration.

The second one comes from Data World, and their internal site scraping teams used one that was scraped from the EaseMyTrip website.This dataset was created at Prompt Cloud and Data Stock. A broader dataset is comprised of subsets like this one. This dataset has about 30K entries with attributes like Arrival Day & Time, Airline, Source, Destination, Layover1, Layover2, Layover3, Number of Stops, and Fare.Kaggle is the third source of information. This database includes domestic airlines for New Zealand airlines on 12 major and 40-second routes in New Zealand (managed by Air New Zealand, a full service network company, and Jetstar, a less expensive network firm). From September 19, 2019 to December 18, 2019, the Web Crawler (Octoparse) gathered hourly airfares for two airlines and made the data available for a total of ninety days.

B. Pre-processing and data cleaning

To ensure that clean the data: as there were so few null or missing values, their removal from the dataset had little impact on the data. To make some qualities more meaningful, it was necessary to separate them and change the datatypes. Feature encoding was used to manage categorical data. For nominal data, one-hot encoding is typically used, while for ordinal data, label encoding. In order to increase the meaning of the data, some columns were added and removed. Ultimately, several outliers were eliminated from the data because doing so increases the meaning of the data. Now that they are clean, the data from Make My Trip, Ease My Trip, and New Zealand may be examined and utilised to create models.

C. Visualising data to gain insights

a) CreateMyTripData

A boxplot figure (1) in the MakeMyTrip dataset is plotted to determine the behaviour of the fare for each airline included in the MakeMyTrip dataset, as well as the relationship between the airlines and the fare.

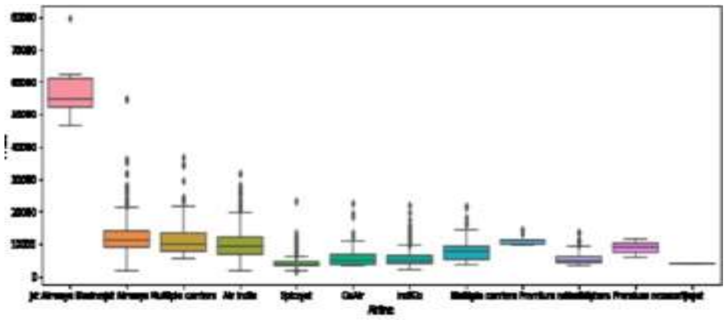


Fig. 1, Boxplot of Airline and Fare

According to the figure (1) boxplot graph, Jet Airways is the most expensive airline in the MakeMyTrip dataset since it is the only airline with a price range between $45,000 and $80,000,000. The price ranges of the other airlines are average.To determine the relationship between the fare and the source airport, a box plot is drawn. This boxplot illustrates how the source airport influences the airline's pricing there.

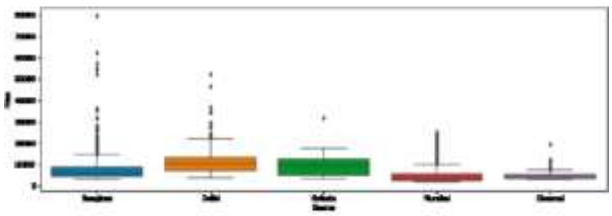


Fig. 2, Boxplot of Source Airport and Fare

Bangalore's airport is the only one in the dataset to include the most expensive flights as a source airport, as seen by the boxplot in Figure (2), which illustrates that it covers all flight price ranges, from extremely low to very high. Delhi has flights that span the whole range of flight possibilities, second only to Bangalore. There are cheap flights from Delhi, Chennai, Mumbai, and Kolkata airports.

To ascertain the association between the ticket and the destination airport, a boxplot akin to the preceding one is employed.

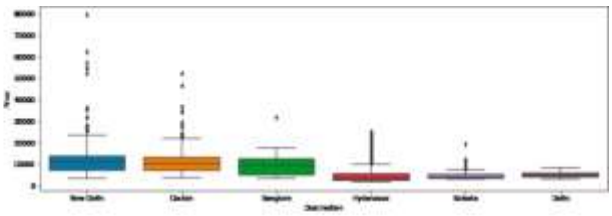


Fig. 3, Boxplot of Destination Airport and Fare

The boxplot figure (3) illustrates that the airport in New Delhi is the only one in the dataset to feature the most expensive flights, covering all flight price ranges from extremely cheap to very expensive. Cochin has flights that span the whole range of flight possibilities, second only to New Delhi. There are cheap flights from Delhi, Bangalore, Hyderabad, Kolkata, and other airports.

b) Easy MyTrip Data

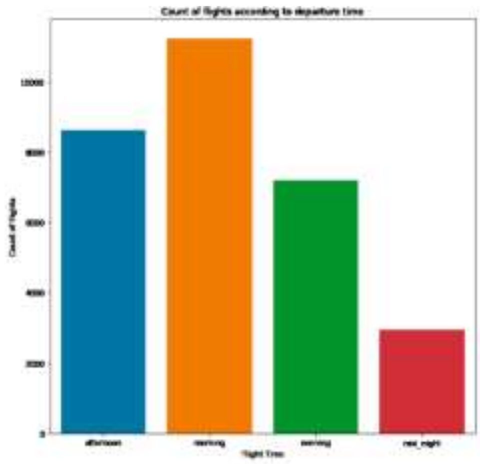


Fig. 4, Number of flights available at different times in a day.

To get the observational count for each category, a countplot is employed. Here, the total number of flights leaving in the morning, afternoon, evening, and midnight is found using a countplot.

There are more flights departing in the morning, which indicates that there are more flights available in the morning, moderate flights available in the afternoon and evening, and less flights available in the midnight hours, according to the countplot figure (4). This indicates that there aren't many flights that take off in the middle of the night.

To determine the degree to which one variable influences another, utilise a scatter plot. In this case, the scatter plot is utilised to determine the impact of flight time on flight pricing.

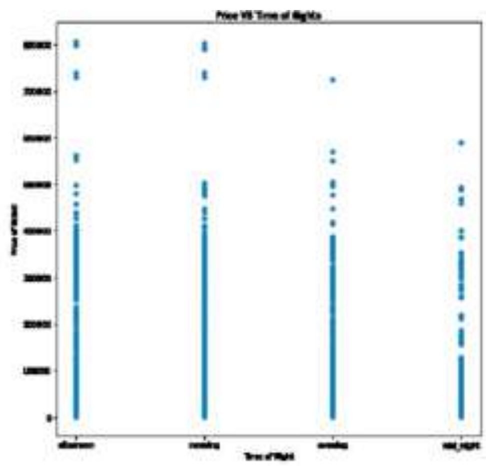


Fig. 5: Scatterplot of Fights at different times and fare price.

According to the scatter plot in Figure 5, the costs of flight tickets for flights available in the morning and afternoon are substantially higher than those of flights available in the evening and at midnight.

In order to obtain more insights from the data, a count plot for the airlines is produced to determine the count of airlines included in the dataset. The countplot showed that a large number of airlines have a count of fewer than

500. Given the small number, it can be challenging to obtain information from numerous airlines. Therefore, it is simpler to include all the airlines with a relatively low count in a variable called "other

airlines"Examine. We will be able to take into account any airline in this manner, even if their number is smaller. The countplot of the changed data is displayed in figure (6).

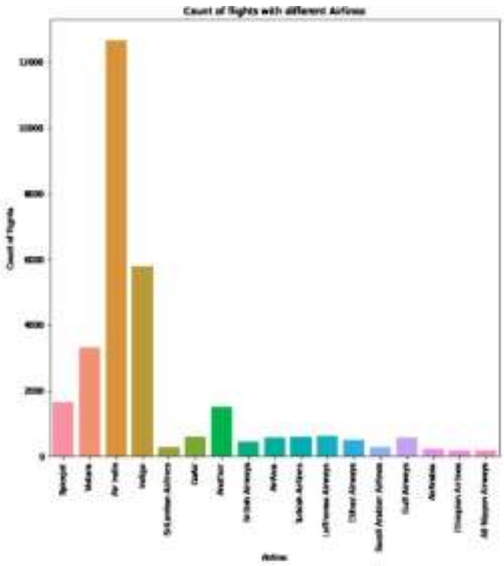


Fig. 6, Countplot of the Airlines to find the number of flights available in an airline

The behaviour of the fare price for each airline is displayed in Figure (7) of the scatter plot for airlines and ticket prices. The airline vs. price scatter helps us understand the behaviour of the data. SpiceJet, Vistara, Air India, Indigo, Sri Lanka Airlines, GoAir, Another, British Airways, AirAsia, Turkish Airlines,Lufthansa Airlines, Etihad Airlines, Saudi Arabian Airlines, Gulf Airways, Air Arabia, Ethiopian Airlines, and All Nippon Airlines are among the airlines that are offered at varying price points between 0 and 800,000.

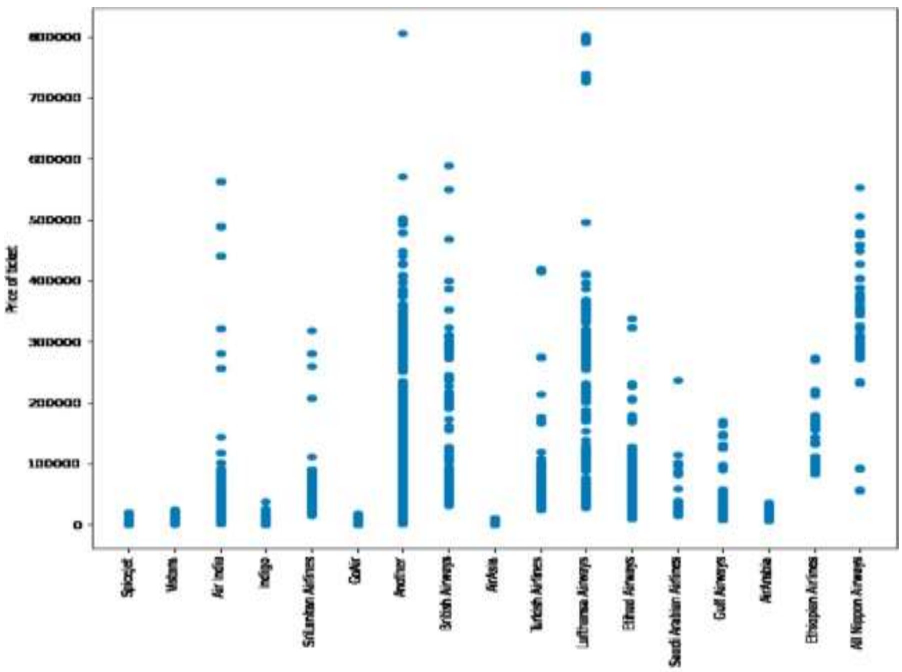


Fig. 7, Scatter plot of the Airlines and the Ticket Fare

Figure (8) plots a boxplot to determine the behaviour of the number of stops with the fare and demonstrates that the flights flights with two stops feature flights in every price category, while there are very few flights with three stops.

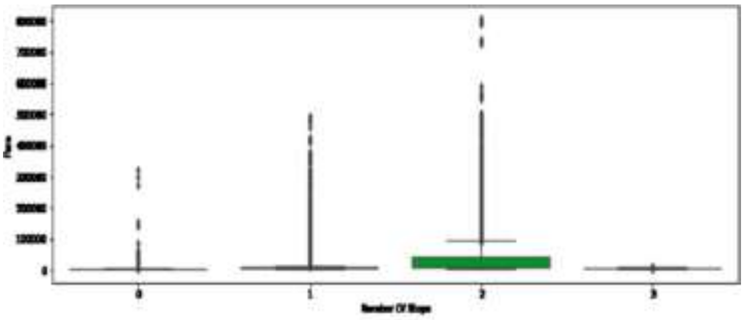


Fig. 8, Boxplot of Number of Stops with Ticket Fare

c) NewZealandData

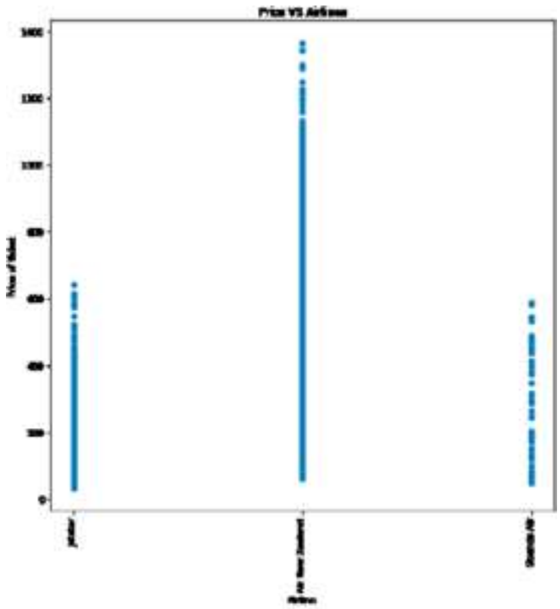


Fig. 9, Scatterplot of Airlines and Ticket Fare

A scatter plot is utilised in the New Zealand dataset to determine the link between the airline and the ticket price. Figure 9's scatter plot demonstrates that while JetStar and Sounds Air have reasonable and less expensive pricing than Air New Zealand Airlines, the airline offers flights across the whole price range.

In this case, the monthly flight availability is determined using a countplot. In contrast to November and September, which have fewer flights than October, figure (10) amply illustrates this point. The first eight months of the year see virtually few flights.

International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

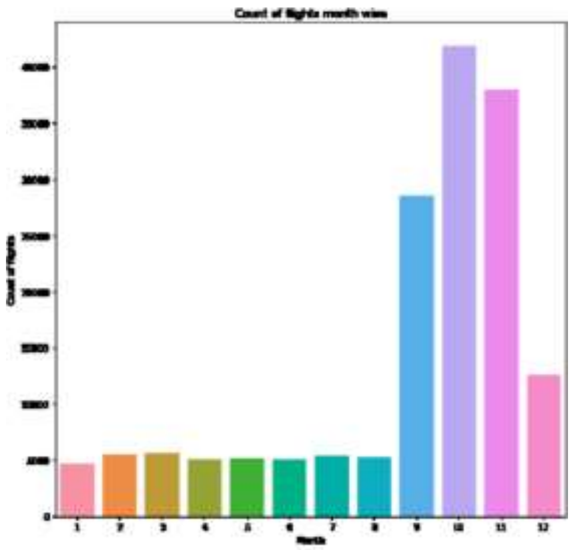


Fig. 10, Number of flights available in each month

SECTION III:IMPLEMENTED

Machine Learning Models-The field of machine learning presents multiple methods for forecasting the cost of airline tickets. We have employed the following algorithms:

- KNN Regression.
- Logistic Regression.
- Lasso Regression.
- Ridge, Decision Tree,
- Stacking Tree,
- Random Forest regressions are among the types of regressions.

The Sci-Kit Learn Python Library has been utilised in the implementation of these models. Parameters like R-square, MAE, MSE, and RMSE are used to confirm the models' performance.

1. KNN Regression

The average of its k closest neighbours is obtained by a k-neighbor regression analysis. This method is non-parametric, just as SVM. To get the best value, the results are obtained using a small number of values. KNN is a regressor that uses supervised classification techniques. It updates the class with a new piece of data. Since there are no presumptions, it is not in a parametric manner. The distance between each training example and a fresh data set is computed. K items from the data set that are close to the new data point are chosen by the model. The Euclidean, Manhattan, or Hamilton distances are used to compute the distance.

B. Regression in Linear Form

One method of supervised learning (ML) is linear regression. It carries out tasks involving regression. Assuming a linear relationship between the input variable (x) and a single output variable (y), the model is linear. By using linear inclusion of the input variables, particularly (x), Y can be computed. We will use multiple linear regression (MLR) to assess the relationship between two or more independent factors and a dependent variable because our data set contains a large number of independent elements that could affect prices changeable.

C. Regression of Lasso

For more accurate predictions, Lasso regression is preferred over other regression techniques. In this model, shrinkage is used. The so-called average value, or centre point, is reached when the data value is lowered. We suggest using a sparse, basic model—one with few parameters. Regression of this kind works well for models that have a lot of multicollinearities or for automating steps in the model selection process, like parameter deletion and variable selection. The L1 regularisation method is used in the lasso regression. When there are multiple functions available, the function that is automatically picked is used.

D) Ridge Regression

A data modelling method called ridge regression is utilised to get rid of multicollinearity. The best method is ridge regression in cases where there areless data than the predictor variables. Plotting the ridge regression constraint variables results in a diamond-shaped figure, whereas the LASSO plot shows a circular shape.

E. Regression Using Decision Trees

Regression or classification models are constructed using decision trees as tree structures. Furthermore, for every smaller data set, a decision tree is created. This produces leaves nodes and solutions. In order to reach a judgement, the decision tree chooses independent variables from the dataset to serve as decision nodes. Upon entering test data into the model, the outcome is ascertained by examining the segment to which the data point pertains. The average of all the data points in the subsection of the section to which the data point belongs will also be output by the decision tree.

F. Regression Stacking

By generating linear combinations of multiple variables, stacking regression is a method for increasing prediction accuracy. The coefficients of the combination are found using the least squares approach using cross-validation data under non-negative constraints. Here, we employed Decision Tree Regressor as the metal regressor and Ridge, Lasso, and KNN regressors as the regressors.

G. Regression with Random Forest

In order to produce more accurate models, the random forest algorithm merges less accurate models. To make a bigger model, it merges the base model with an additional model. To create highly uncorrelated decision trees, the characteristics are scanned and passed on to the trees without replacement. A lower correlation between trees is required in order toto select the optimal split. The aggregated uncorrelated trees are the primary feature that sets the random forest apart from the decision tree. A random forest is an ensemble learning method where a number of learning algorithms are used by the training model and then merged to yield a final predicted result. A random subset of features and data sets will average the projected values when the random forest model's output is analysed; this is within the bagging domain of ensemble learning.

IV. RESULTS

Result Analysis-

The machine learning models are constructed using three distinct datasets from three different sources, and the cleaned and preprocessed data is used for the model's training and testing. Several machine learning models that may precisely forecast the value of airline tickets are developed using Python.

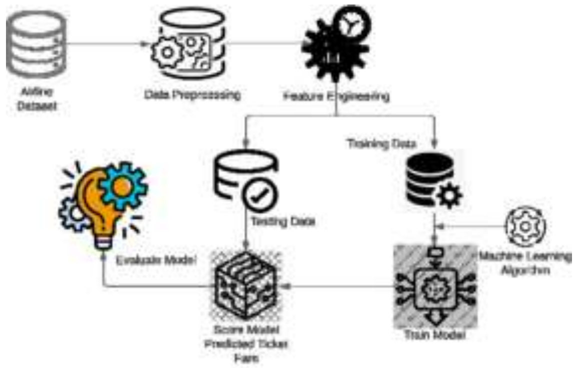


Fig. 11, Architecture Diagram

KNN Regression, Linear Regression, Lasso Regression, Ridge Regression, Decision Tree Regression, Stacking Tree Regression, and Random Forest Regression are the machine learning techniques employed.

For exploratory data analysis, the mtual info classif feature selection function of the Spark library is utilised. Subsequently, the data is partitioned for the purpose of training and testing models. Training data is utilised for model training, while model accuracy is used for testing.

```
def predict(model):
    trained_model = model.fit(x_train,y_train)
    print("Training Score : {}".format(trained_model.score(x_train,y_train)))
    y_prediction = trained_model.predict(x_test)
    print("Predictions are : {}".format(y_prediction))
    print('\n')
    print("Testing Score : {}".format(trained_model.score(x_test, y_test)))
    r2_score = metrics.r2_score(y_test,y_prediction)

    print("R2 Score : {}".format(r2_score))

    print("MAE : ",metrics.mean_absolute_error(y_test,y_prediction))
    print("MSE : ",metrics.mean_squared_error(y_test,y_prediction))
    print("RMSE : ",np.sqrt(metrics.mean_absolute_error(y_test,y_prediction)))
```

Fig. 12, Prediction Function

Seven machine learning models need to be tested and trained. The function in figure (12) merely trains and tests the machine learning model after receiving it as a parameter. Subsequently, the function assesses the model through various performance criteria. However, prior to passing the 1. GridSearchCV

The library function GridSearchCV, which fits our estimator (model) to our training data by looping over preset hyperparameters, is part of Sklearn's model selection package.

Lastly, we can select the hyperparameter that works best from the list.

B. CV RandomSearch

The objective of GridSearchCV and RandomSearchCV is the same. The goal is to determine the best parameters to make the model better. All parameters aren't examined in this instance, though.Rather, the parameters under test can be altered, but the search is random and all other parameters stay same. The best features and settings are identified following hypertuning.

The machine learning model is tested and trained once the parameters are fine-tuned. Once the model has been trained and tested, it is time to compare and assess each model's accuracy.

them to choose which is ideal and best. To compare the accuracy of machine learning models trained using various techniques, performance metrics are employed. Regression metrics-based functions that quantify each model's mistakes are implemented using the sklearn_metrics module.

The specified indications will be used to verify each model's measure of error.

The Coefficient of Determination, or C. R'

The coefficient of determination (R2 metric) is used to gauge how well the regression model performs. You can compare the model's improvement over time to a specified baseline by tracking the model's progress over time. Tracing the average of the data yields this baseline. R2 is a scale-free score, therefore regardless of size, it is always less than or equal to 1.of the worth.

Sum of squares of residuals, or RSS, equals R2 = 1.

Total sum of squares, or TSS, (1)

D. Absolute Mean Error

The entire amount of the model prediction's (MAE) difference from the target value. In contrast to MSE, MAE penalises errors less severely and is less vulnerable to outliers. In MAE, each individual difference carries the same weight.

Total Absolute Error (SAE)

(Total number of point sets in the data) / MAE = n

(2)

E. Squared Mean Error

Mean square error, or MSE, is a popular regression statistic.

The regression model's projected value is divided by the goal value to determine this average. Squared the difference means that even the smallest mistake is penalised and the model's performance is exaggerated. Because it is distinguishable from other measures, it has precedence over them.

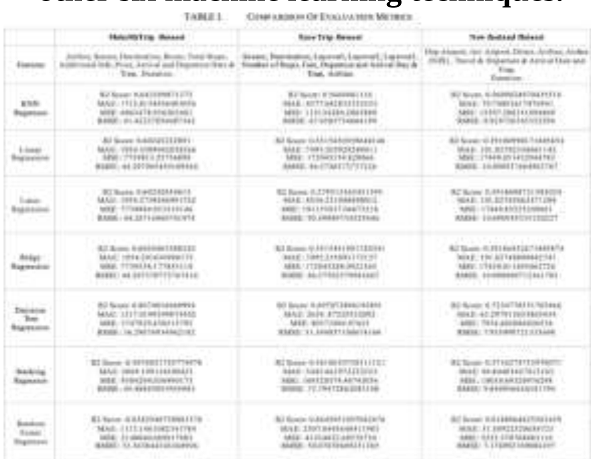
n IC(Assumed) -ỹ(Real)2. (3)

F. Squared Root Mean Error

The root-mean-squared difference between the target value and the value the model predicts is taken as the square root to determine the root mean square error, or RMSE. Mistakes are squared prior to averaging, and significant errors are punished. Consequently, when significant mistakes are not desired, RMSE is helpful.

VMSE = RMSE (4)

Table I displays a comparison of evaluation metrics across seven distinct algorithms. With an average R2 score of 0.8 in all three cases, the Random Forest Regressor method outscored the other six machine learning techniques.



V. CONCLUSION AND FUTURE SCOPE

Three distinct datasets from three different sources have been utilised to estimate the dynamic fare of flights. The process of visualising the dataset has yielded numerous insights. To construct the model, seven distinct machine learning methods were employed. Since the information is gathered from websites that sell airline tickets, only a restricted amount of information is available. The evaluation metrics table I values derived from the process are used to assess the model's accuracy. With respect to accuracy, the Random Forest Regressor performed better than the other algorithms. Thus, the Random Forest Regressor is a good predictor of airline ticket prices. The expected outcomes would be more accurate if further information, such as real seat availability, could be acquired in the future. Currently, prediction-based services are used in a improve forecast accuracy by applying different methods. More precise data with more attributes may be used to obtain more trustworthy results.

VI. REFERENCES

1. T. Janssen, "A linear quantile mized regression model for prediction of airline ticket prices," in A Treatise on Electricity and Magnetism 3rd ed., vol. 2, 2014, pp. 68- 73.
2. Yiwei Chen and F. Vivek Farias, "Robust Dynamic Pricing With Strategic Customers," Mathematics of Operations Research 43, pp. 1119-1142, 2018.
3. Juhar Ahmed Abdella, Nazar Zaki, Khaled Shuaib and Fahad Khan, "Airline ticket price and demand prediction: A survey.," Journal od King Saud University - Computer and Information Sciences, vol. 33, no. 4, pp. 375-391, 2021.
4. Lantseva, Anastasia, Mukhina, Ksenia, Nikishova, Anna, Ivanov, Sergey, Knyazkov and Konstantin, "Data-driven Modeling of Airlines Pricing,"Procedia ComputerScience,
5. vol.66,pp.267-276, 2015.
6. K. Tziridis, T. Kalampokas, G. A. Papakostas and K. I. Diamantaras, "Airfare prices predictiono using machine learning techniques," in 25th European Signal Processing Conference (EUSIPCO). Kos 2017, 2017.
7. A. Boruah, K. Baruah, B. Das, M. Das and
8. N. Gohain, "A Bayesian Approach for Flight Fare Prediction Based on Kalman Filter," in Progress in Advanced Computing and Intelligent Engineering, Singapore, 2019, pp. 191-203.
9. William Groves and Maria Gini, "A regression model for predicting optimal purchase timing for airline tickets.," Technical report, University of Minnesota, Minneapolis, USA, Report number11-025, 2011.
10. D. Tanouz, R. R. Subramanian, D. Eswar, G.
11. V. P. Reddy, A. R. Ku mar and C. V. N. M. Praneeth, "Cred it Card Fraud Detection Using Machine Learning," in 5th International Conference on Intelligent Computing and Control Systems ICICCS), 2021.
12. R. R. Subramanian, N. Akshith, G. N. Murthy, M. Vikas, S. Amara and K. Balaji, "A Survey on Sentiment Analysis," in 11th International Conference on Cloud Computing, Data Science &Engineering (Confluence),2021,2021.
13. S. A mara and R. R. Subra manian, " Collaborating personalized recommender system and content-based recommender system using TextCorpus," in 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India 2020.
14. Andi and Hari Kirshnan, "An Accurate Bitcoin Price Prediction using logistic regression with LSTM

Machine Learning model," Journal of Soft Computing Paradigm3,pp.205-217, 2021.
15. Manoharan and J. Samuel, "Study of Variants of Extreme Learning Machine (ELM) Brands and its Performance Measure on Classification Algorithm," Journal of Soft Computing Paradigm (JSCP) 3, pp. 83- 95, 2021.
16. V. Suma and Shavige Malleshwara Hills, "Data Mining based Prediction of Demand in Indian Market for Refurbished Electronics," Journal of Soft Computing Paradigm(JSCP)2,pp.101-110, 2020.
17. W. K. Michael and A. G. Thomas, "A Framework for the Evaluation of Statistical Prediction Models," CHEST, vol. 158, no. 1, pp. S29-S38, 2020.
18. L. Yuling and L. Zhichao, "Design and implementation of ticket price forecasting system," in AIP Conference Proceedings, 2018.
19. Elizaveta Stavinova, Petr Chunaev and Klavdiya Bochenina, "Forecasting railway ticket dynamic price with Google Trends open data," Procedia Computer Science, vol. 193, pp. 333- 342, 2021.
20. S. Deepa, A. Alli, Sheetac and S. Gokila, "Machine learning regression model for material synthesis prices prediction in agriculture," in materialstoday, 2021.
21. S. Matthew and Lewis, "Identifying airline price discrimination and the effect of competition," International Journal of Industrial Organization, vol. 78, 2021.
22. Ismail Koc and Emel Arslan, "Dynamic ticket pricing of airlines using variant batch size interpretable multi- variable long short-term memory," Expert Systems with Applications,vol.175, 2021.
23. Rian Mehta, Stephen Rice, John Deaton and Scott R. Winter, "Creating a prediction model of passenger preference between low cost and legacy airlines,"Transportation Research Interdisciplinary Perspectives, vol. 3, 2019.