Airline Fare Prediction Using Machine Learning Algorithms

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.**

Keywords— Airline Fare, Machine Learning, Linear Regression, Lasso Regression, Ridge Regression, Decision Tree, Stacking Regression, Random Forest, Prediction Model.

I. Introduction

In today's world, airlines attempt to control flight ticket costs in order to maximize profits. Most people who fly regularly know the best times to buy cheap tickets. However, many customers who are not good at booking tickets fall into the discount trap set by the company, causing them to spend their money. The main goal of airline companies is to make a profit, while the customer is looking for the best purchase. Customers frequently aim to purchase tickets far in advance of the departure date in order to prevent price increases as the departure date approaches. Due to the great complexity of the fare models used by airlines, it is very difficult for a customer to buy an airline ticket at a very low price because the price is constantly fluctuating. Airlines can lower their ticket prices when they need to create a market and when tickets are harder to obtain. These tactics consider a number of financial, marketing, commercial, and social factors that are all linked to ultimate flight pricing. They might be able to get the most profit possible. As a result, costs may be influenced by various factors. The price model used by airlines is so complex that prices fluctuate constantly, making it very difficult for customers to buy tickets at very low prices. Surveys of customers and airlines have grown steadily over the last two decades. From a customer point of view, it is an important question to establish a low price or a good time to buy a ticket. In this paper, we will be using the collected data from three different sources to build the models using Machine Learning algorithms. Customers can save millions of rupees by using the proposed method to get the information they need to order tickets at the proper moment.

II. Background And Related Works

There aren't enough publicly available datasets to allow researchers to make accurate predictions. The owner company's ticket price tactics are highly commercially sensitive and remain confidential. So, most airlines do not publish their ticket price plans. Researchers rely on limited datasets obtained through the Web scraping technique. Using these limited resources, a few works have developed various strategies for ticket price prediction. On the customer side, there are two kinds of research: one tries to predict the best time to buy a ticket, and the other tries to estimate the value of the ticket.

The authors of [1] used a linear quantile mixed regression model to predict the cheapest ticket price before departure. To predict the minimum price, one low-cost quintile is used instead of all observations. The data used included 2,271 flights with a total of 126,412 records with a single route collected 60 days prior to shipment: amount, departure date, statement date, list of dates before departure, and weekday (weekend or weekday). One-way lounge tickets with non-stop flights are included in the database. Test results have shown that the model works well for a while before departure, but it does not work as the number of days before the trip increases.

[2] suggested a low-cost model for a specific route (a specific flight on a specific route on a particular day of travel). Three months of data for five international routes were used to train the model. The data has features such as the cost of the same trip, the prices of the most recent pre-day trips, the prices of their trips on the day of the week, their trip rates on the same day of the month. Learn++ is a group-learning algorithm. For example, NSE has been modified and trained to learn from past price fluctuation trends and make predictions about future values over time. When compared to KNN (12.58 percent) and P.A., the model exhibited the lowest mean absolute percentage error (MAPE) of 10.7% (15.41%). A flight's cost could not be predicted, and multi-stop journeys were not considered.

The model of [3] predicts the cost per kilometre for a given distance. The flight must be booked at least 90 days in advance of travel for the model to work. During the 75-day AviaSales time and the 90-day Saber term, a tiny daily fee was collected for each flight. The model was built using the following features: departure city, destination, date of purchase, date of departure, and ticket options and pricing. They used regression analysis to create the model. Ticket prices for domestic and international flights were also examined by the authors. Tickets purchased in advance of departure on international flights were found to be more cost effective, but more research was needed to confirm this for domestic flights. The authors did not evaluate the model's performance. Because they were gathered during a limited period of time and for specific routes, they were confined.

Regression machine learning models for airline ticket price prediction have been developed by [4]. Data from 1814 flights on a single international route was used in the development of this model, including departure and arrival times, bag allowance, and the number of free baggage allowances per flight. They used eight different regression machine learning models, which are Extreme Learning Machine (ELM), Multilayer Perceptron (MLP), Generalized Regression Neural Network, Random Forest Regression Tree, Regression Tree, Linear Regression (LR), Regression SVM (Polynomial and Linear), Bagging Regression Tree. The model produced the following performance results: The

Bagging Regression is accurate to 87.42% and 85.91% accuracy for Random Forest Regression Tree.

A well-known Bayesian estimating method, the Kalman filter, was used by [5] to estimate the ticket prices. Based on the Kalman Filter line model, the study developed an algorithm that estimates the ticket price for a specific aircraft. A matrix of observed data, similar to that of a linear Kalman Filter model, is used to develop model characteristics.

[6] used a mix of the ARMA(Autoregressive–moving-average model) and random forest algorithms to build and implement a ticket price forecasting system. The model is powerful. The top performers were Random Forest and Multilayer Perceptron. As a result, the two most powerful models were given different weightings (Random Forest and Multilayer Perceptron). The model was trained using 51,000 records collected from three domestic airlines 21 days before the departure day of seven non-stop flights. On the other hand, international and international flights are not included in the data. The following information is extracted from the data: airline, flight number, purchase date, departure time, arrival time, fare rate, stopping rate, amount, departure airport, arrival terminal, and arrival date. Based on the R2 test ratings, Random Forest and Multilayer Perceptron performed very well with 4.4 and 7.7 percent, respectively.

Most of the studies didn't include multi-stop flights in the data, and this paper is all about three different datasets from three different sources, including multi-stop flights. This paper is all about finding the insights of the multi-stop flights through data analysis and predicting the prices of the flights using seven different machine learning models and comparing their performances.

III. DATA COLLECTION

Collecting the data is the most important part of this project. To prepare the models, several sources of information from various websites are used. For data about airfare, various sources are available, ranging from APIs to customer travel sites. We have collected three datasets for this project from three different sources.

A. Data Collection

The scraped data from MakeMyTrip's website is used as the first dataset, which comes from Kaggle. A total of ten thousand data points is contained in the report. These include the following: airline, source, and destination; route; number of stops; additional information; price; arrival and departure; date and time; and duration.
The second one is from Data World, a scrapped one from the EaseMyTrip website, and their in-house site scraping teams at Prompt Cloud and Data Stock developed this dataset. This dataset is a subset of a larger dataset. There are around 30K records in this dataset with features such as Source, Destination, Layover1, Layover2, Layover3, Number of Stops, Fare, Departure, and Arrival Day & Time, Airline.
The third data source is from Kaggle. Domestic airlines for New Zealand airlines (controlled by Air New Zealand, a comprehensive service network company, and Jetstar, a less expensive network company) on 12 major and 40-second routes in New Zealand are included in this database. The Web Crawler (Octoparse) collected hourly airfares for two airlines from September 19, 2019 to December 18, 2019, and provided the data for a total of 90 days.

B. Data cleaning and pre-processing

In order to clean the data, the null or missing values have been dropped from the dataset as they were very few and dropping them wouldn't affect the data. Changed the datatypes and had to split some attributes to make them more meaningful. Performed feature encoding to handle categorical data. Usually, one-hot encoding is performed on nominal data and label encoding on ordinal data. Adding and removing some columns was done to make the data more meaningful and finally removed some outliers from the data, as removing outliers makes the data more meaningful. The Make My Trip, Ease My Trip, and New Zealand data are now clean and ready to be analyzed and used to build models.

C. Data Visualisation to find insights

a) MakeMyTrip Data

In the MakeMyTrip dataset, a boxplot figure (1) is plotted to find the relationship between airlines and the fare or how the fare behaves for different airlines present in the MakeMyTrip dataset.

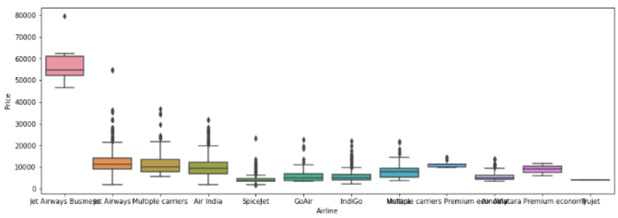

Fig. 1, Boxplot of Airline and Fare

From figure (1) boxplot graph, the only airline whose price range is between 45000 and 80000 is Jet Airways, and the remaining airlines' price range is average, which makes Jet Airways the costliest airline present in the MakeMyTrip dataset.
A box plot is plotted to find the relationship between the source airport and the fare. This boxplot reveals how the source airport affects the prices of the airlines there.

Fig. 2, Boxplot of Source Airport and Fare

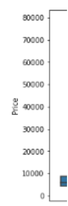

Figure (2) boxplot shows that Bangalore's airport covers all price ranges for flights, from low to very high, making it the only airport in the dataset to include the costliest flights as a source airport. After Bangalore, Delhi has flights that cover the entire spectrum of flight options. Kolkata, Mumbai, and Delhi Chennai airports have affordable flights.
A boxplot similar to the previous one is used to determine the relationship between the ticket and the destination airport.

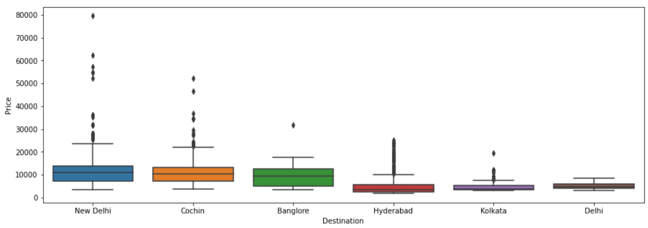

Fig. 3, Boxplot of Destination Airport and Fare

The boxplot figure (3) shows that New Delhi's airport covers all price ranges for flights, from low to very high, making it the only airport in the dataset to include the costliest flights. After New Delhi, Cochin has flights that cover the entire spectrum of flight options. Bangalore, Hyderabad, Kolkata, and Delhi airports have affordable flights.

b) EasyMyTrip Data

A countplot is used to find the observational count of each category. Here, a countplot is used to find the total number of flights departing in the morning, afternoon, evening, and midnight.

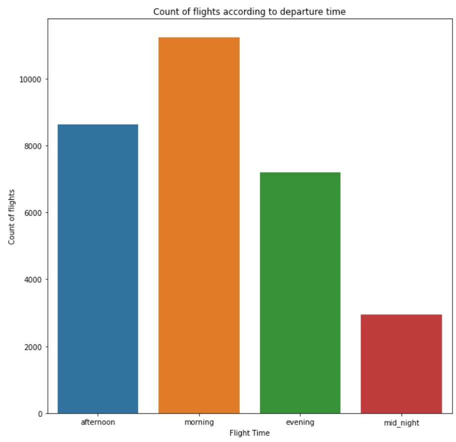

Fig. 4, Number of flights available at different times in a day.

The countplot figure (4) shows that many flights are leaving in the morning, which means the availability of flights in the morning is greater, and the availability of flights in the afternoon and evening is moderate, but the availability of flights in the mid-night is less. This concludes that there are very few flights that travel in the mid-night.

A scatter plot is used to find out how much one variable is affected by another. Here, the scatter plot is used to find out how the flight price is affected by the flight time.

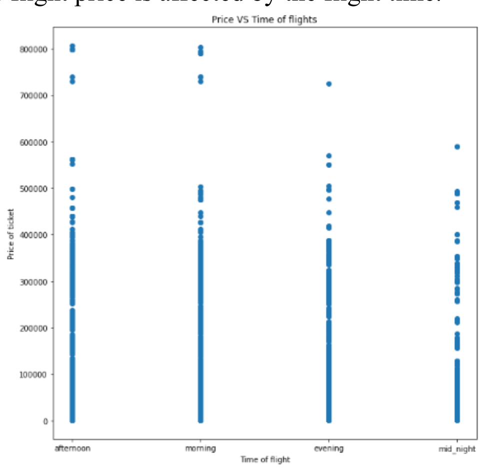

Fig. 5: Scatterplot of Fights at different times and fare price.

The scatter plot in figure (5) says that the prices of the flight tickets available in the morning and afternoon are competitively very high compared to the flight ticket prices of the flights available in the evening and midnight.

A count plot for the Airlines are plotted to find the count of airlines present in the dataset, which helps to gain more insights through the data. The countplot revealed that there are many airlines whose count is less than 500. As it can be difficult to get information from many airlines with such less count. So, placing all the airlines with a very low count into a variable named "other airlines" makes it easier to analyze. In this way, we'll be able to consider all the airlines even if their count is lower. The figure (6) shows the countplot of modified data.

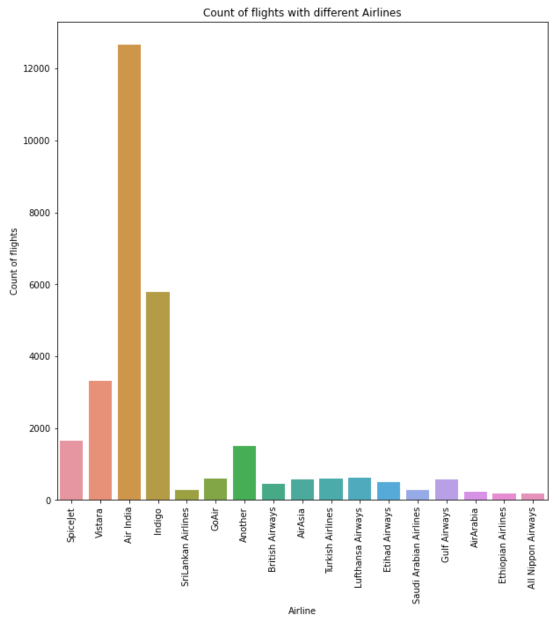

Fig. 6, Countplot of the Airlines to find the number of flights available in an airline

Figure (7) of the scatter plot for airlines and ticket prices shows the behavior of the fare price for the respective airlines. By using the airline vs. price scatter, we can gain insights into how the data behaves. The price ranges from 0 to 800000 and the airlines available are 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.

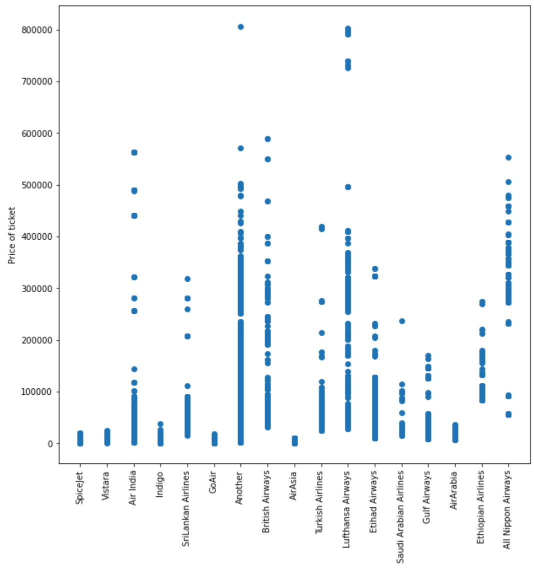

Fig. 7, Scatter plot of the Airlines and the Ticket Fare

A boxplot to find the behavior of the number of stops with the fare is plotted in figure (8), which shows that the flights with two stops have all price range flights, whereas the flights with three stops have very few.

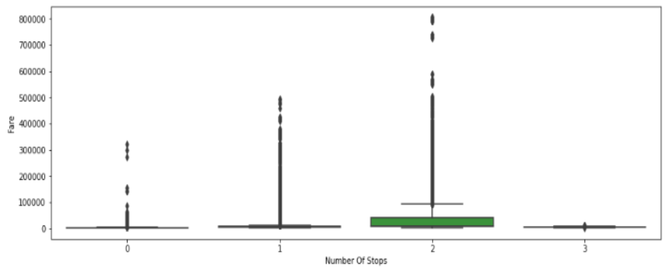

Fig. 8, Boxplot of Number of Stops with Ticket Fare

c) *New Zealand Data*

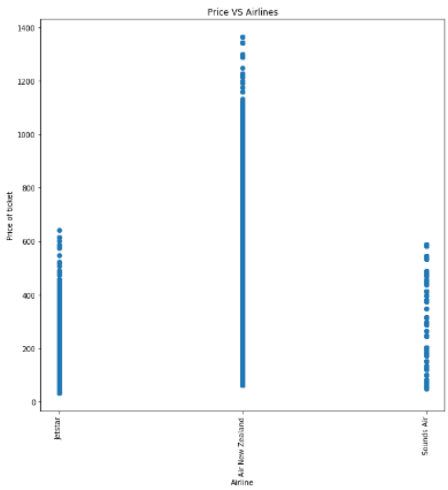

Fig. 9, Scatterplot of Airlines and Ticket Fare

In the New Zealand dataset, a scatter plot is used to find the relationship between the airline and the price of the ticket. The scatter plot in figure (9) shows that the Air New Zealand Airlines has all range flights from lower prices to the highest and the JetStar and Sounds Air prices are affordable and are not as expensive as Air New Zealand Airlines.

A countplot is used here to find the monthly availability of flights. The figure (10) clearly shows that the October month has many flights and the November and September months have fewer flights compared to October. There are very few flights in the first eight months of the year.

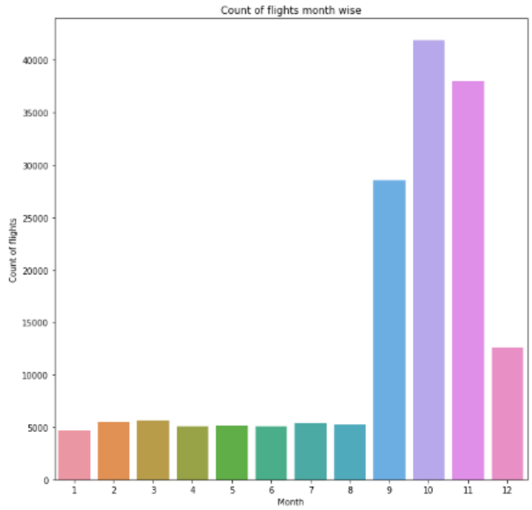

Fig. 10, Number of flights available in each month

IV. MACHINE LEARNING MODELS

Machine learning introduces several techniques for predicting aircraft ticket pricing. Algorithms that we have used include:

- *KNN Regression.*
- *Linear Regression.*
- *Lasso Regression.*
- *Ridge Regression.*
- *Decision Tree Regression.*
- *Stacking Tree Regression.*
- *Random Forest Regression.*

These models have been implemented using the sci-kit learn python library. In order to verify the performance of these models, parameters such as R-square, MAE, MSE, and RMSE are used.

A. *KNN Regression*

A k-neighbor regression analysis gives the average of its k nearest neighbors. Like SVM, this is a non-parametric approach. The results are obtained using only a few values to get the best value. KNN is a supervised classification technique used as a regressor. It adds a new data point to the class. Since no assumptions are made, it is not parametric. It calculates the distance between each training example and a new data set. The model selects K elements from the data set that are near the new data point. The distance is calculated using the Euclidean distance, the Manhattan distance or the Hamilton distance..

B. *Linear Regression*

Linear regression is a supervised learning (ML) technique. It performs regression tasks. It is a linear model, assuming that there is a linear relationship between the input variable (x) and a single output variable (y). Y can be calculated by linear inclusion of input variables, especially (x). Because our data set contains many independent features that prices may depend on, we will use multiple linear regression (MLR) to estimate the relationship between two or more independent variables and a dependent variable.

C. Lasso Regression

Lasso regression takes precedence over other regression approaches for more accurate predictions. This model employs shrinkage. When reduced, the data value is reduced to the center point, the so-called average value. We recommend a simple, sparse model (that is, a model with few parameters). This type of regression is best suited for models with a large number of multicollinearities, or for automating aspects of the model selection process such as variable selection and parameter deletion. The lasso regression employs the L1 regularization method. Since the function is selected automatically, it is used when there are many functions.

D. Ridge Regression

Ridge regression is a data modeling technique used to eliminate multicollinearity. Ridge regression is the most appropriate technique when there are fewer observations than predictor variables. Ridge regression constraint variables have a circular shape when plotted, while the LASSO plot has a diamond shape.

E. Decision Tree Regression

A decision tree is a tree structure used to build regression or classification models. In addition, a decision tree is generated for each data set that is reduced in size. This generates solutions and leaf nodes. The decision tree selects independent variables from the dataset as decision nodes for making a decision. When test data is entered into the model, the result is determined by looking at which segment the data point belongs to. And the decision tree will output the average of all data points in the subsection of the section that the data point belongs to.

F. Stacking Regression

Stacking regression is a technique for improving prediction accuracy by creating linear combinations of multiple predictors. Cross-validation data and the least squares method are used to determine the coefficients of the combination under non-negative requirements. In this case, we used Ridge Regressor, Lasso Regressor, and KNN Regressor as regressors, and Decision Tree Regressor as metal regressors.

G. Random Forest Regression

The random forest algorithm combines less accurate models to create more accurate models. It combines the base model with another model to create a larger model. The features are scanned and passed on to the trees without replacement in order to generate strongly uncorrelated decision trees. It is necessary to have a lower correlation between trees in order to choose the best split. The main principle that distinguishes the random forest from the decision tree is the aggregated uncorrelated trees. A random forest is an ensemble learning technique in which the training model uses a variety of learning algorithms that are then combined to produce a final predicted result. When the output of the random forest model is examined, a random number of features and data sets will average the predicted values, which falls within the bagging area of ensemble learning.

V. RESULT ANALYSIS

Three different datasets from three different sources are used for building the machine learning models, and the cleaned and preprocessed data is used to train and test the machine learning model. Python3 is used to create different machine learning models that would accurately predict flight ticket values.

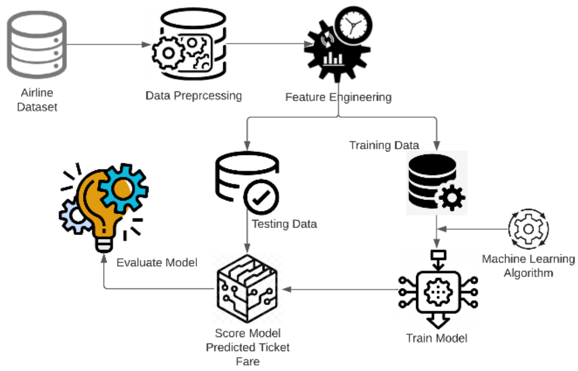

Fig. 11, Architecture Diagram

The machine learning algorithms used are KNN Regression, Linear Regression, Lasso Regression, Ridge Regression, Decision Tree Regression, Stacking Tree Regression, and Random Forest Regression. The Sklearn library's mtual info classif feature selection is used for exploratory data analysis. After that, the data is split up to be used for training and testing models, with training data being used to train models and test models using their accuracy.

```
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

As there are seven machine learning models to be trained and tested. The function in figure (12) simply takes the machine learning model as a parameter and then trains and tests the model. The function then evaluates the model using different performance metrics. But before passing the machine learning model into the function, the models should be hypertuned using GridSearchCV or RandomSearchCV.

A. GridSearchCV

Sklearn's model selection package contains the library function GridSearchCV, which fits our estimator (model) to our training data, which will loop through predefined hyperparameters. Finally, we can choose the best hyperparameter from the list.

B. RandomSearchCV

Both RandomSearchCV and GridSearchCV have the same goal. It is to identify the optimal parameters to improve the model. However, not all parameters are checked in this case. Instead, the search is randomized and all other parameters remain constant, but the parameters under test can be changed.

After performing hypertuning, the optimal features and parameters are determined.

After hyper-tuning the parameters, the machine learning model is trained and tested. After training and testing the model, it is time to check the accuracy of all models and compare them to find the best and optimal one. Performance metrics are used to compare the accuracy of machine learning models trained with different algorithms. The sklearn_metrics module is used to implement functions that measure the errors of each model using regression metrics. The measure of error of each model will be checked using the listed indicators.

C. R^2 or Coefficient of Determination

The performance of the regression model is measured by the coefficient of determination (R2 metric). By measuring the model's improvement over time, you can compare the model's improvement over time to a given baseline. This baseline is established by tracing the average of the data. As a scale-free score, R2 is always less than or equal to 1 regardless of the size of the value.

$$R^2 = 1 - \frac{RSS(Sum\ of\ squares\ of\ residuals)}{TSS(Total\ sum\ of\ squares)} \quad (1)$$

D. Mean Absolute Error

The total value of the difference between the target value and the model prediction (MAE). Unlike MSE, MAE is less susceptible to outliers and penalizes errors less severely than the latter. All individual differences have the same weight in MAE.

$$MAE = \frac{SAE(Sum\ of\ absolute\ errors)}{n(total\ number\ of\ point\ sets\ in\ data)} \quad (2)$$

E. Mean Squared Error

A common regression metric is MSE, or mean square error. This average is calculated using the predicted value of the regression model divided by the target value. As a result of squared the difference, even the slightest error is penalized and the performance of the model is overestimated. It has priority over other metrics because it can be distinguished from other metrics.

$$\frac{1}{n} \sum (y(Predicted) - \check{y}(Actual))^2 \quad (3)$$

F. Root Mean Squared Error

RMSE is calculated by taking the square root of the root-mean-squared difference between the target value and the value predicted by the model. Errors are squared before averaging and large errors are penalized. Therefore, RMSE is useful when large errors are not desired.

$$RMSE = \sqrt{MSE} \quad (4)$$

The comparison of evaluation metrics on 7 different algorithms is shown in table I. The Random Forest Regressor algorithm outperformed all the other 6 machine learning algorithms with an average R2 score of 0.8 in all the three datasets.

TABLE I. COMPARISION OF EVALUATION METRICS

	MakeMyTrip Dataset	**Ease Trip Dataset**	**New Zealand Dataset**
Features	Airline, Source, Destination, Route, Total Stops, Additional Info, Price, Arrival and Departure Date & Time, Duration.	Source, Destination, Layoverl, Layover2, Layover3, Number of Stops, Fare, Departure and Arrival Day & Time, Airline.	Dep.Airport, Arr. Airport, Direct, Airline, Airfare {NZ$}, Travel & Departure & Arrival Date and Time, Duration.
KNN Regressor	R2 Score: 0.642309871372 MAE: 1715.8134456403056 MSE: 6963478.936203401 RMSE: 41.42237856087342	R2 Score: 0.5660481116 MAE: 4577.682833333333 MSE: 115134280.2883889 RMSE: 67.6585754604199	R2 Score: 0.5609824970435516 MAE: 79.74001617476941 MSE: 12597.288351094869 RMSE: 8.929726545352294
Linear Regression	R2 Score: 0.60243222891 MAE: 1954.3389042858266 MSE: 7739813.25756899 RMSE: 44.207905450109564	R2 Score: 0.3515456939648146 MAE: 7495.205929280411 MSE: 172045154.828066 RMSE: 86.5748573737226	R2 Score: 0.39186990171685654 MAE: 101.82702336061142 MSE: 17449.851432944783 RMSE: 10.090937684903787

Lasso Regression	R2 Score: 0.60242854615 MAE: 1954.2738446991762 MSE: 7739884.953319146 RMSE: 44.20716960741974	R2 Score: 0.279515365451395 MAE: 8556.231008890012 MSE: 191155937.04675218 RMSE: 92.49989734529446	R2 Score: 0.3918698731585024 MAE: 101.82705662471288 MSE: 17449.85225240601 RMSE: 10.090939333120227
Ridge Regression	R2 Score: 0.6024463588220 MAE: 1954.291630906175 MSE: 7739538.177853114 RMSE: 44.207370775767416	R2 Score: 0.3515451901720341 MAE: 7495.235092173127 MSE: 172045288.4922365 RMSE: 86.57502579943667	R2 Score: 0.39186432671885874 MAE: 101.82748809842741 MSE: 17450.011403462726 RMSE: 10.090960712361703
Decision Tree Regression	R2 Score: 0.8074816660994 MAE: 1317.0199399874502 MSE: 3747929.438515795 RMSE: 36.29076934962182	R2 Score: 0.697073806192891 MAE: 2656 .87229352892 MSE: 80371096.97625 RMSE: 51.544857100674164	R2 Score: 0.7234778551703466 MAE: 63.297915653863434 MSE: 7934.602084026536 RMSE: 7.955998721333698
Stacking Regressor	R2 Score :0.5076921745774976 MAE: 2069.199134300421 MSE: 9584204.030490173 RMSE: 45.48845055945983	R2 Score: 0.3610610738511121 MAE: 5445.661972222223 MSE: 169520574.46743056 RMSE: 73.79472862083188	R2 Score: 0.3716274733970071 MAE: 96.84403427413163 MSE: 18030.69320976298 RMSE: 9.840936656341794
Random Forest Regressor	R2 Score :0.8382948758883376 MAE: 1113.1461082341784 MSE: 3148060.669017083 RMSE: 33.363844326368906	R2 Score: 0.8445851055962676 MAE: 2507.0445648411983 MSE: 41234022.69276754 RMSE: 50.07039609231385	R2 Score: 0.8188864625563459 MAE: 51.50922520636723 MSE: 5252.370768483116 RMSE: 7.176992769006197

VI. CONCLUSION AND FUTURE SCOPE

To estimate the dynamic fare of flights, three different datasets from three different sources have been used. Many insights have been found while visualizing the dataset. Seven different machine learning algorithms have been used to build the model. Only limited information can be obtained because data is acquired from websites that sell flight tickets. The correctness of the model is determined by the evaluation metrics table I values obtained from the procedure. The Random Forest Regressor outperformed the other algorithms with good accuracy. So, Random Forest Regressor works fine for predicting the airline fare price. If more data, such as actual seat availability, could be obtained in the future, the anticipated results would be more accurate. Prediction-based services are currently employed in a variety of sectors, including stock price predictor programs used by stock brokers and services like Zestimate, which provides an estimate of housing values. As a result, in the aviation business, a service like this is required to assist clients in reserving tickets. There have been numerous studies conducted on this topic using various methodologies, and additional research is required to increase the accuracy of prediction utilizing various algorithms. To acquire more reliable findings, more accurate data with greater features might be employed.

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 Computer Science,* vol. 66, pp. 267-276, 2015.

[5] 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.

[6] A. Boruah, K. Baruah, B. Das, M. Das and 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.

[7] William Groves and Maria Gini, "A regression model for predicting optimal purchase timing for airline tickets.," Technical report, University of Minnesota, Minneapolis, USA, Report number 11-025, 2011.

[8] D. Tanouz, R. R. Subramanian, D. Eswar, G. V. P. Reddy, A. R. Kumar and C. V. N. M. Praneeth, "Credit Card Fraud Detection Using Machine Learning," in *5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2021.

[9] 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.

[10] S. Amara and R. R. Subramanian, "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.

[11] Andi and Hari Kirshnan, "An Accurate Bitcoin Price Prediction using logistic regression with LSTM Machine Learning model," *Journal of Soft Computing Paradigm 3,* pp. 205-217, 2021.

[12] 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.

[13] 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.

[14] 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.

[15] L. Yuling and L. Zhichao, "Design and implementation of ticket price forecasting system," in *AIP Conference Proceedings*, 2018.

[16] 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.

[17] S. Deepa, A. Alli, Sheetac and S. Gokila, "Machine learning regression model for material synthesis prices prediction in agriculture," in *materialstoday*, 2021.

[18] S. Matthew and Lewis, "Identifying airline price discrimination and the effect of competition," *International Journal of Industrial Organization,* vol. 78, 2021.

[19] 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.

[20] 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.