iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** XI **Month of publication:** November 2023

DOI: https://doi.org/10.22214/ijraset.2023.57139

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue XI Nov 2023- Available at www.ijraset.com

Flight Ticket Fare Prediction using Supervised Learning

Vansh Sethi
Department of Electrical Engineering, National Institute of Technology, Jamshedpur, India

Abstract: ***With the increasing interconnectivity of air routes worldwide, air travel has become a widespread and expeditious means of transportation. Forecasting airline fares poses a significant and intricate challenge due to their continual fluctation, influenced by a diverse array of factores. Extensive research suggests that the application of Machine Learning, and Deep Learing techniques enables the swift estimation of flight fares at specific intervals. This study employs a Machine Learning Regression methodology to predict flight fares based on essential parameters such as departure and arrival times, departure location, destination, stopovers, and airline provider. This research employs two distinct datasets for training and testing purposes, evaluating a range of machine learning approaches to forecast flight ticket prices.***
Keywords: Prediction, Accuracy, Regression Algorithm, Supervised Learning, Exploratory Data Analysis, Optimization

I. INTRODUCTION

As the holiday season approaches, the process of curating an ideal vacation itinerary can present itself as an arduous endeavour. The exponential expansion of the internet and E-commerce on a global scale has substantially impacted the commercial aviation sector, reshaping it into a meticulously regulated yet burgeoning marketplace.[1] Consequently, in the realm of Airline Revenue Management, diverse strategies encompassing customer profiling, financial marketing, and social factors are employed to establish ticket fares. This practice commonly manifests in lower airfares when tickets are reserved well in advance, followed by escalated prices for last-minute bookings.[2] Hence, time until departure is one of the main factors which decides flight fare, but not the only one. This study undertakes the task of utilizing supervised Machine Learning algorithms to accurately estimate airline ticket prices while considering a multitude of parameters. The research encompasses as in-depth examination of diverse algorithms, aiming to discern the most suitable model for precise price prediction within the aviation domain. Through meticulous analysis and comparison of various supervised learning methodologies, this research endeavours to identify and validate the optimal algorithm that best encapsulates the intricate relationships between pertinent factors influencing ticket prices. The primary objective is to construct a robust predictive framework capable of accommodating the complexities of airfare dynamics, thereby empowering stakeholders and travelers with enhanced insights for informed decision-making in the realm of air travel.

II. LITERATURE REVIEW

In the study by Juhar Ahmed Abdella, Nazar Zaki, and colleagues as outlined in reference [3], an examination of a predictive model for airline ticket prices utilizing deep learning techniques and social media data is presented. The research provides a comprehensive overview of the current landscape of airline ticket pricing, eluciading the multifaceted factors influencing ticket costs. Additionally, the paper delved into airline strategies aimed at revenue maximization. The predictive model offers users guidance on optimal ticket purchase timing, utilizing a blend of data mining approaches such as Rule Learning, Reinforcement Learning, and time-series methods to enhance predictive accuracy. Noteworthy features incorporated in the study encompass flight details, time until departure, prevailing ticket prices, airline carriers, and flight routes. The culmination of these methodologies resulted in achieving a peak accuracy rate of 61.9% when employing a combination of the aforementioned techniques.

The proposed investigation as outlined in reference [4], focuses on employing machine learning methodologies for predicting airfare prices. This study utilizes a dataset comprising 1814 flights records from Aegean Airlines, serving as the foundation to train a machine learning model. Notably, diverse sets of features were employed to train distinct models, highlighting the impact of feature selection on model accuracy within the research. The research titled “Airline Fare Prediction Using Machine Learning” authored by A.L. Rodrigues [5], and colleagues (2020) centers on the application of machine learning methodologies to forecast airline fares. This study encompasses the analysis of multiple factors, such as airline preferences, travel distances, and past fare records, to develop and refine a predictive model. The authors delve into the evaluation of diverse algorithms’ performance and offer insights into the implication of the discoveries concerning the accuracy of fare prediction.

In the study titled "Predicting Airfare Using Machine Learning Techniques" authored by S. Aruna, and colleagues (2020), a comparative assessment of various machine learning algorithms for airfare prediction is presented. This research encompasses an exploration of factors like season variations, booking timings, and flight class to formulate a predictive model. The authors conduct an evaluation of regression algorithms, encompassing linear regression, support vector regression, and random forest regression, to gauge their predictive performance.

In study cited as reference [6] conducted by William Groves, an intelligent agent is introduced to optimize the timing of purchase for customers. The study employed Partial Least Squares Regression technique in developing a predictive model. Initial stages involved employing various methods for feature selectuon, encompassing Feature Extraction, Lagged Feature Computation, Regression Model Construction, and Optimal Model Selection. The experiments were specifically structured to evaluate the real-world implications of implementing the prediction models. While the lag scheme approach exhibited favourable performance across several machine learning algorithms, the study identified PLS regression as the most effective technique within this domain. This improved efficacy is attributed to its inherent ability to mitigate the impact of correlated and extraneous variables.

In an instance detailed in reference [7] by Tianya Wang, diverse attributes were extracted from datasets and integrated into the data framework to delineate segments within the air transport sector. Utilizing feature selection methodologies, our proposed model demonstrates the capability to forecast the quarterly average ticket price.

III. PROPOSED FRAMEWORK

Our proposed methodology leverages datasets to forecast airfare within distinct business segments. Figure 1 outlines the primary components of our project framework at a high level. The initial phase involves meticulous data pretreatment, encompassing the cleaning of databases to eliminate potential errors, followed by their conversion and integration based on market groupings. Subsequently, the feature extractor identifies and generates specific attributes tailored to describe each market segment effectively. Adaptive filtering modules then refine these attributes, enhancing accuracy by assessing their utility and removing unnecessary features. Finally, employing the selected criteria, we construct forecasting techniques that culminate in the anticipated output—the projected cost of airline tickets.

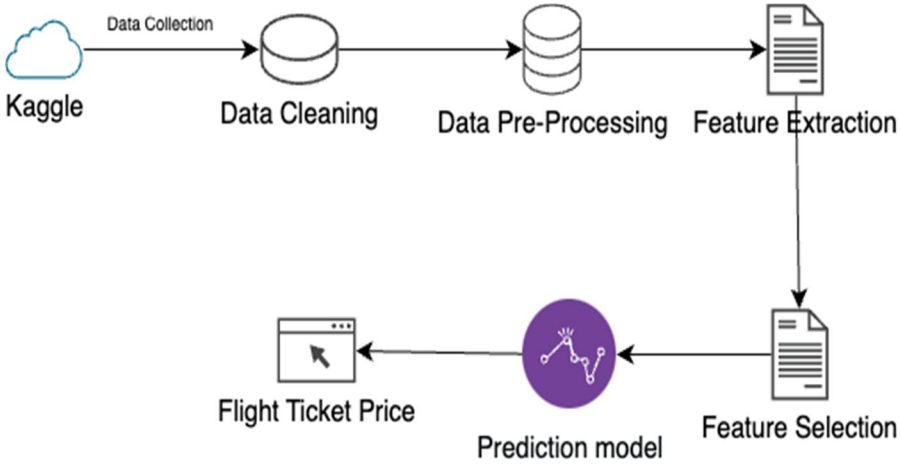


Fig 1. Proposed System for flight ticket price

IV. METHODOLOGY

Following steps were perfomed while building the system.

A. Dataset

The training and testing datasets utilized in this study have been sourced from the Kaggle data repository, encompassing categorical and nominal data associated with Indian Airlines in the year 2019. This dataset offers crucial insights into influential factors used for flight fare prediction, including departure and arrival locations, timings, flight routes, halts during the journey, and corresponding ticket prices. Comprising 10,682 rows and 11 columns, each representing distinct attributes, this extensive dataset serves as a significant resource for analysis and modeling in our research.

Feature Name	Description
Airline	Airline company name
Date_of_Journey	Date of departure of the passenger
Source	Departure airport code
Destination	Arrival airport code
Route	Route from source to destination including stops if any
Dep_Time	Time of departure from source
Arrival_Time	Time of arrival to final destination
Duration	Duration of the journey including time in flight and waiting time for connecting flights if any
Total_Stops	Number of breaks in the journey
Additional_Info	Info of complimentary services like meals etc

Table 1. Description of the attributes

B. Data Cleaning

Many columns in the data had null value or missing data, removed the rows from the dataset which contains null or missing data in any of the columns. Additionaly column 'Additional_Info' was removed as it had value 'no info' only. This ensures the clean data for data pre-processing and prediction.

C. Data Pre-Processing

During the data pre-processing phase, transformations were applied to enhance the dataset's suitability for analysis. String-based attributes like the date of the journey, departure time, and arrival time were converted from string data type to datetime objects. Subsequently, numeric values were extracted, including the month-date numeric value from the journey date attribute and the hour-minute numeric value from departure and arrival time attributes.

D. Feature Engineering

Categorical data underwent encoding techniques tailored to their nature. Nominal categorical attributes like 'source', 'destination', and 'airline' were processed using the 'One hot encoding' method, converting them into numerical values conducive for machine learning algorithms. Concurrently, ordinal categorical data such as the 'total stops' underwent 'Label encoding' to transform labels into numeric representations, optimizing dataset usability. Finally, column rearrangement concluded this pre-processing step, ensuring data readiness for subsequent analysis.

E. Feature Selection

In an effort to mitigate overfitting, streamline model complexity, and enhance overall accuracy, a feature selection process was executed. This involved employing mutual info regression to assess the mutual dependency between individual features and the target variable. Subsequently, the features were sorted based on their dependency, enabling the identification of the most strongly correlated features, which were retained, while less dependent features were omitted.

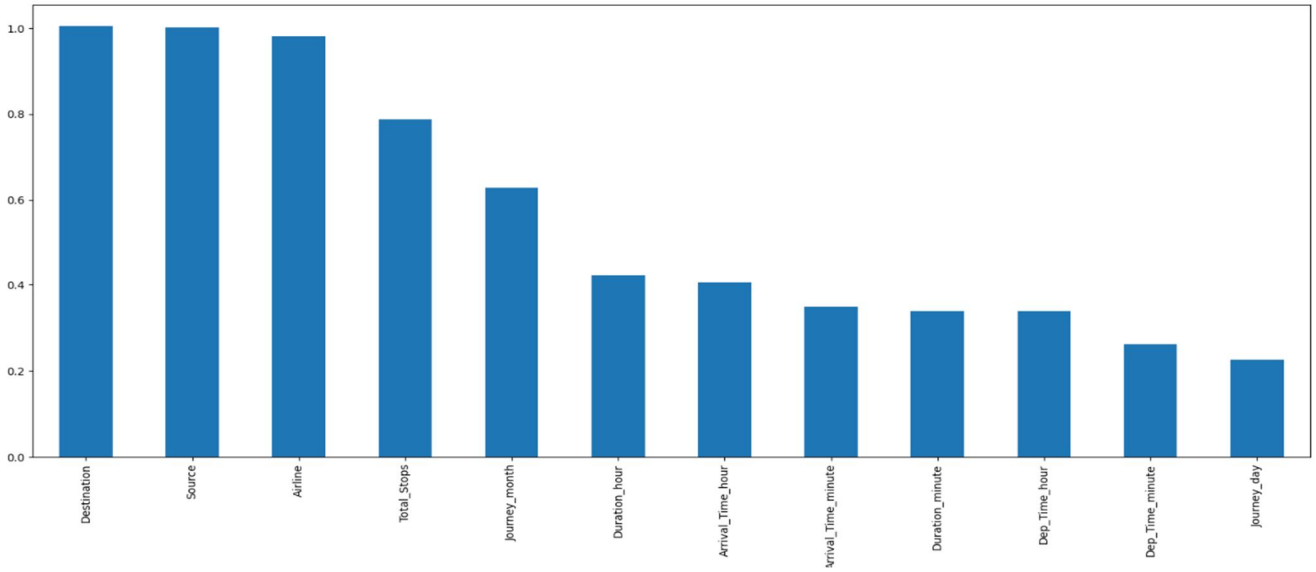


Fig 2. Importance of individaul feature in dataset

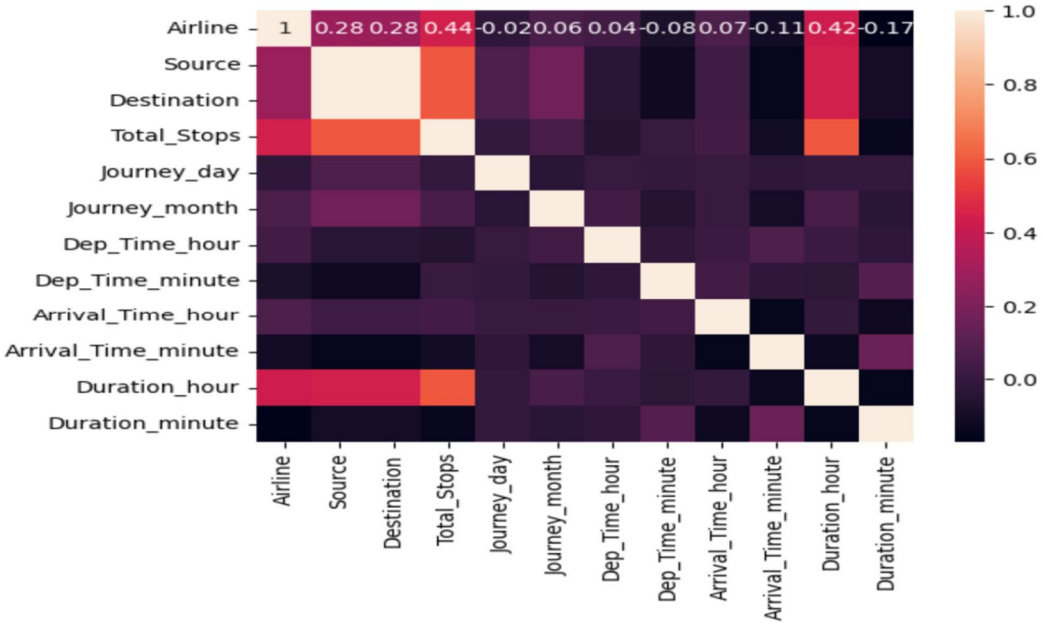


Fig 3. Correlation between all features

F. Modeling

The aforementioned dataset will undergo fitting with various regression algorithms to facilitate a comparative analysis of predictions generated by each model, ultimately determining and selecting the most suitable model.

V. ALGORITHMS ANALYSIS

This section outlines diverse algorithms and their application within our analysis.

A. Decision Tree

The Decision Tree serves as a decision-making tool employing a structured tree format or model to delineate decisions and their potential outcomes, encompassing input costs and utility. Belonging to the domain of supervised learning algorithms, the decision-tree algorithm caters to both continuous and categorical output variables [8].

Its branches or edges symbolize the validity of statements, influencing decisions based on this evaluation. A tree undergoes "training" by segmenting the resource collection into subgroups based on tests of characteristic values. This process, termed data partitioning, iteratively operates on each resulting subset. The recursive procedure concludes when all subgroups within a node share identical posterior probability or when further splits cease to contribute additional predictive value. Renowned for its ability to extract experimental knowledge, a decision tree proves beneficial as it does not necessitate subject matter expertise or parameter configuration.

B. Random Forest

Random forests, also known as random decision forests, represent an ensemble learning technique used for a variety of tasks, including classification and regression. This method involves creating numerous decision trees during the training phase. In classification tasks, the output provided by the random forest corresponds to the class that the majority of trees select. Conversely, in regression tasks, it returns the mean or average prediction derived from individual trees [9][10]. One of the key advantages of random decision forests is their ability to counteract the tendency of decision trees to overfit their training data.

C. Extra Trees Regressor

The ExtraTreesRegressor, an abbreviation for Extra Trees Regressor, stands as an ensemble learning variation rooted in decision tree regression. Its functionality lies in forming an ensemble of decision trees while injecting randomness into the tree construction phase. This randomness encompasses the consideration of random feature subsets and thresholds for node splits within each tree. The core objective of the ExtraTreesRegressor is to mitigate overfitting by harnessing this additional randomness, thereby fostering a model that is more resilient and less prone to sensitivity compared to conventional decision tree regression.

D. Gradient Boosting

Gradient boosting represents a prominent machine learning approach applicable to regression and classification tasks. This technique constructs a predictive model by amalgamating several weak prediction models, typically simple decision trees [11][12], thereby forming an ensemble. Specifically termed as gradient-boosted trees when utilizing decision trees as weak learners, this algorithm commonly demonstrates superior performance in contrast to the random forest method. The construction of a gradient-boosted trees model follows a stage-wise progression akin to other boosting methodologies. However, its distinctive aspect lies in its capability to optimize diverse differentiable loss functions, thus extending beyond the constraints of conventional boosting techniques.

VI. RESULTS AND DISCUSSIONS

In our assessment, we employ various evaluation metrics including MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), and R-squared value to evaluate the performance of all four models.

1) The Mean Absolute Error (MAE) represents the mean value of the absolute variance between the predicted and actual data values [13].

$$MAE = (1/N) * \Sigma|Y[I] - X[I]|$$

where:

N – Number of observations
Y[i] – Predicted value for i^{th} observation
X[i] – actual value for i^{th} observation

2) The Mean Squared Error (MSE) refers to the average of the squared variances between the predicted values and the actual values in the dataset [14].

$$MSE = (1/N) * \Sigma(|Y[I] - X[I]|)^2$$

where:
N – Number of observations
Y[i] – Predicted value for i^{th} observation
X[i] – actual value for i^{th} observation

3) Root Mean Squared Error is the root of MSE [15].
RSME = $(MSE)^{1/2}$

4) R-squared value is used for measuring the accuracy of the model [16].
R^2 = 1- (SSR/SST)
where:
SSR = Σ|Y[I] – X[I]|
SST = Σ|Y[I] – $\bar{Y}$[I]|

Model	MAE	MSE	RSME	R Squared Value	Accuracy
Decision Tree	1349.3683	5563043.1348	36.7337	0.7142	71.4%
Random Forest	1175.9425	3615444.6734	34.2920	0.8142	81.4%
Extra Trees Regressor	1213.3594	4142855.1899	34.8333	0.7871	78.7%
Gradient Boosting	1545.4106	4867491.0440	39.3117	0.7499	74.9%

Table 2. Values of Evaluation Metrics

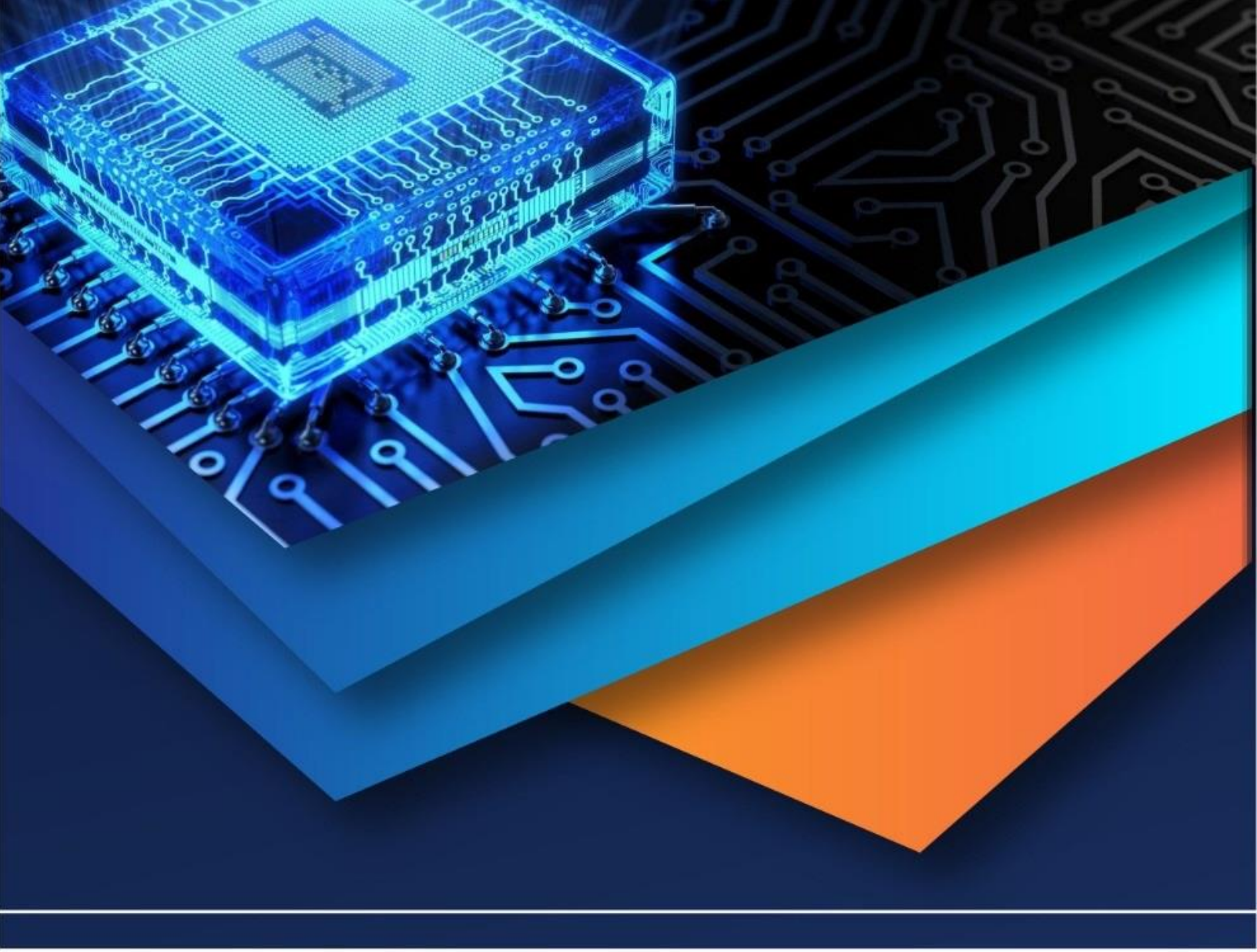
From the above table, conclusion is that Random Forest algorithms gives better results than other decision-based algorithms, it provides the solid accuracy of ~82%.

VII. CONCLUSION AND FUTURE SCOPE

For this paper, an extensive study was carried out with dataset collection from Kaggle. Using visualisation determined the importance of every feature on the outcome variable i.e airfare price. Multiple decision trees based supervised algorithm were compared to find the model with best accuracy, Random Forest performed the best on various parameters like MAE, MSE, and accuracy. The forthcoming objective involves enhancing the model's accuracy through expanded data collection, incorporation of additional features, and implementing more refined feature selection techniques.

REFERENCES

[1] Tom Chitty, CMBC Business News, “This is how airplanes price tickets”, August 3, 2018. Available: https://www.cnbc.com/2018/08/03/how-do-airlines-priceseat-tickets.html.
[2] Moira McCormick, BlackCurce, “Behind the Scenes of Airline Pricing Strategies”, September 19, 2017. Available: https://blog.blackcurve.com/behind-thescenes-of-airline-pricing-strategies.
[3] Juhar Ahmed Abdella, Nazar Zaki and Khaled Shuaib, “Automatic Detection of Airline Ticket Price and Demand: A Review”, 13th International Conference on Innovations in Information technology (IIT), January 10, 2019.
[4] Joshi, Achyut & Sikaria, Himanshu & Devireddy, Tarun. Predicting Flight Prices in India.
[5] “Airline Fare Prediction Using Machine Learning” Authors: A. L. Rodrigues, et al. Published in: Proceedings of the International Conference on Data Engineering and Communication Technology, 2020.
[6] William Groves and Maria Gini “An agent for optimizing airline ticket purchasing” in proceedings of the 2013 international conference on autonomous agents and multi-agent systems.
[7] Tianyi Wang, Samira Pouyanfar, Haiman Tian, Yudong Tao, Miguel Alonso Jr., Steven Luis and Shu-Ching Chen, “A Framework for Airfare Price Prediction: A Machine Learning Approach”, 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI), September 9, 2019.
[8] von Winterfeldt, Detlof; Edwards, Ward (1986). “Decision trees”. Decision Analysis and Behavioral Research. Cambridge University Press. pp. 63–89. ISBN 0-521-27304-8.
[9] Ho, Tin Kam (1995). Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282. Archived from the original (PDF) on 17 April 2016. Retrieved 5 June 2016.
[10] Ho TK (1998). The Random Subspace Method for Constructing Decision Forests IEEE Transactions on Pattern Analysis and Machine Intelligence. 20 (8): 832–844. doi:10.1109/34.709601. S2CID 206420153.
[11] Piryonesi, S. Madeh; El-Diraby, Tamer E. (2020-03-01). “Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index”. Journal of Infrastructure Systems.
[12] Hastie, T.; Tibshirani, R.; Friedman, J. H. (2009). “10. Boosting and Additive Trees”. The Elements of Statistical Learning (2nd ed.). New York: Springer. pp. 337–384.
[13] Zach, Statology, “How to calculate mean Absolute Error in Python”, January 8. 2021, Available: https://www.statology.org/mean-absolute-error-python/
[14] Wikipedia, “Mean Squared error”, Available: https://en.wikipedia.org/wiki/Mean_squared_error
[15] Science Direct, “Root-Mean Squared Error”, Available: https://www.sciencedirect.com/topics/engineering/root-mean-squared-error/
[16] “Coefficient of Determination, R-squared”, Available: https://www.ncl.ac.uk/webtemplate/ask-assets/external/mathsresources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.html



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY