iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** XI **Month of publication:** November 2022

DOI: https://doi.org/10.22214/ijraset.2022.47475

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 10 Issue XI Nov 2022- Available at www.ijraset.com

A Flight Fare Prediction Using Machine Learning

Ketan Jayatkar[1], Dipak Jagtap[2], Pratik Dengale[3], Aditya Satam[4], Prof. Mahendra Nivangune[5]
[1, 2, 3, 4]*Department of Computer Engineering Sinhgad Academy of Engineering, Pune, India*

Abstract: ***While airlines (the sellers) always work to increase their revenue by changing pricing for the same service, air travellers (the buyers) frequently search for the ideal time of year to buy flights in order to save as much money as possible. The choice to raise or lower tickets at various points leading up to departure dates can be made by the sellers based on all the relevant data, such as historical sales, market demand, consumer profile, and behaviour. The buyers, on the other hand, have limited access to data to help them decide whether to delay or make a quick flight purchase. In this study, we suggest a new model that might assist the purchaser in anticipating price movements even in the absence of official airlines. Our results showed that the suggested model, despite lacking several essential components, such as the number of unsold seats on flights, can forecast trends as well as actual changes in airfare up to the departure dates using public airfare data that is readily available online. We also determined the characteristics that have the biggest effects on changes in airfare.***
Keywords: Atmospheric modelling, predictive models, data Models, prediction algorithms, adaptation models, Tools, indexes

I. INTRODUCTION

Due to the Internet's and e-phenomenal commerce's expansion, travellers may now readily check the prices and availability of all airlines worldwide. These consumers can purchase their chosen tickets online through official airline or agent websites once they are satisfied with an airfare. A number of prediction models have been developed to forecast the airfare before departure in order to assist clients in purchasing the least priced flights. To generate "purchase" or "wait" signals for clients, different data mining techniques as well as time series data analytics were applied [2]. Regression techniques such as partial least square regression [3] and linear quantile mixed regression are also used to build prediction models. The fact that the aforementioned prediction models exclusively concentrate on mature aviation markets, such those in the US, is a recurring theme. Additionally, a model that can forecast the airfare more accurately and ultimately help clients make better decisions to get the best airfare can be built using more of the airfare information that can be collected online. Currently, just a handful of these features are used to predict the price. Waiting to buy the cheapest airfare may result in missing the flight because it is impossible to predict when all of the seats on a given aircraft are sold out. It's also intriguing to learn which characteristics have the biggest effects on changes in airfare before departure dates.

II. LITERATURE REVIEW

1) *K. Tziridis, Th. Kalampokas, G.A. Papakostas K.I. Diamantaras:* The issue of predicting ticket rates is covered in this essay. In order to achieve this, a collection of characteristics that define a typical flight are chosen, presuming that these characteristics have an impact on the cost of an airline ticket. Eight cutting-edge machine learning (ML) models are employed to forecast the pricing of airline tickets using the attributes, and the models' performance is compared to one another. This study examines the relationship between forecast accuracy and the feature set used to represent an airline, in addition to the prediction accuracy of each model.
2) *Tao Liu, Jian Cao Yudong Tan, Quanwu Xiao:* In this study, we offer the ACER context-aware ensemble regression model, which incorporates various context-aware models and adaptively modifies context features. Context characteristics are arbitrarily chosen to efficiently cluster data, and several regression models are trained for data with various contexts. This approach is inspired by bagging and boosting. The context feature list is additionally constantly modified by removing some unnecessary elements. Our model is contrasted in the experiment on the real data set with the baseline regression model, random forest, and traditional time series models. The outcomes demonstrate that ACER outperforms the other models by a wide margin
3) *Viet Hoang Vu, Quang Tran Minh, Phu, H. Phung:* In this article, we provide a brand-new model that might assist the customer in anticipating price trends without relying on official airline information. Our results showed that the suggested model, despite lacking several essential components, such as the number of unsold seats on flights, can forecast trends as well as actual changes in airfare up to the departure dates using public airfare data that is readily available online. We also determined the characteristics that have the biggest effects on changes in airfare.

4) *William Groves, Maria Gini:* However, only sporadically does the earliest buying method result in the ideal lowest cost ticket. This paper suggests a model for determining the best course of action for potential departures. The ultimate use of this concept is to automatically make daily purchases on behalf of purchasers of airline tickets in order to reduce their costs
5) *K.I. Diamantaras, T. Papadimitriou:* In order to handle medium scale data in applications involving pattern classication, this study introduces a parallel version of the kernelized Slackmin method. The fundamental ideas of the serial Slackmin method are first discussed, with emphasis on how easily it may be parallelized due to its parallel nature. Utilizing the parallel processing features of a low-cost NVIDIA GPU card's CUDA architecture, parallelization is made possible.
6) *Hang Zhou, Weicong Li, Ziqi Jiang, Fanger Cai and Yuting Xue,:* In order to identify the elements affecting flight operation, this study first describes the factors impacting flight operation in previous research findings. It next analyses and filters the factors. The GRU neural network model is then created and validated using actual flight data. Finally, the benefits of the model developed in this research are highlighted and contrasted with a number of widely used neural network models and random forest models in machine learning.
7) *Micha Zoutendijk, Mihaela Mitici:* In the present study, we derive probabilistic delay forecasts for flights landing and taking off from a local reference airport. To the best of our knowledge, this marks the first instance in which probabilistic projections for individual flight delays are made. We use the mixture density networks and random forest regression machine learning methods. We take into account features based on the flight schedules that are accessible at the reference airport as well as the weather data gathered at the airports where the flights originated and ended. The performance of the examined machine learning methods, which calculate delay probability density functions, is evaluated using appropriate metrics (pdf). Additionally, the effect of these algorithms' hyperparameter selection is examined.
8) *O. Basturk, C. Cetek:* In this paper, machine learning algorithms are given for predicting aircraft Estimated Time of Arrival (ETA). The management of delays and air traffic flow, runway and gate assignment, collaborative decision-making (CDM), coordination of ground personnel and equipment, and optimization of arrival sequence, among other things, depend on accurate ETA forecast. Machine learning can create predictions with flimsy or no assumptions while learning from past data.

III. PROPOSED SYSTEM

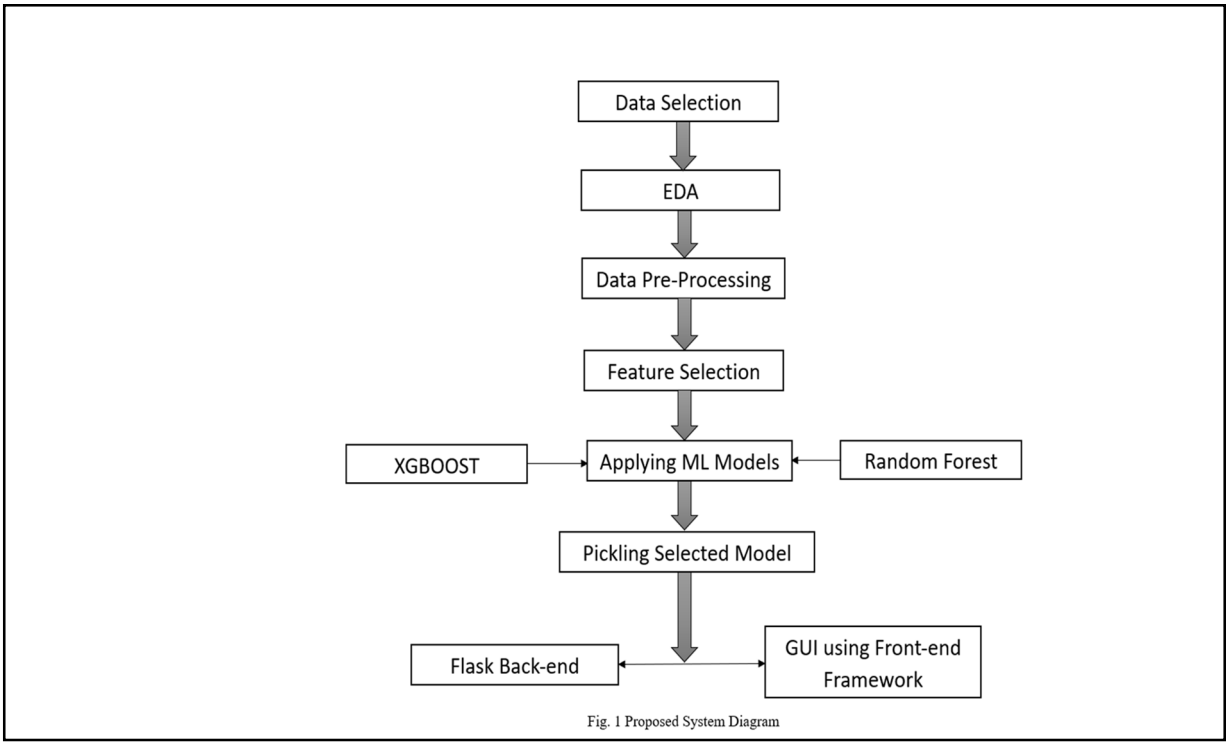


Fig. 1 Proposed System Diagram

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue XI Nov 2022- Available at www.ijraset.com

IV. METHODOLOGY

Following steps were performed while building the system.

A. Data Collection

Both the training and testing datasets have been extracted from Kaggle data repository. They contain categorical as well as nominal data related to the Indian Airlines from the year 2019. The dataset provides vital information about some impacting features to predict the fare of a flight - such as the places of departures and arrivals, time of departure and arrivals, the route of the flight, the number of halts during the journey and the price of the ticket depending on those features. It's an enormous dataset of 10683 rows and 11 columns (each representing one attribute).

B. Data Pre-processing

While pre-processing the data, we converted the date of journey, departure time and the arrival time from string datatype to date-time object and extracted the numeric values from them; the month-date numeric value from the date of journey attribute and hour- minute numeric value from the departure time and arrival time attributes respectively. Later, we have implemented the 'One hot encoding' method for the nominal categorical data and the label encoding method for ordinal categorical data present in both the training as well as the testing dataset. 'One hot encoding' is a process of converting the categorical data variables into numerical values thus making it suitable to use while implementing machine learning algorithms. One hot encoding method was applied to nominal categorical data attributes such as the 'source', the 'destination' and the 'airline company' chosen by the user. 'Label encoding' helps us convert the labels into numeric values in order to make the dataset suitable for use. Label encoding method was applied to the nominal categorical data attributes such as the 'total number of halts in the journey'. The columns were re- arranged at the last step.

C. Data Cleaning

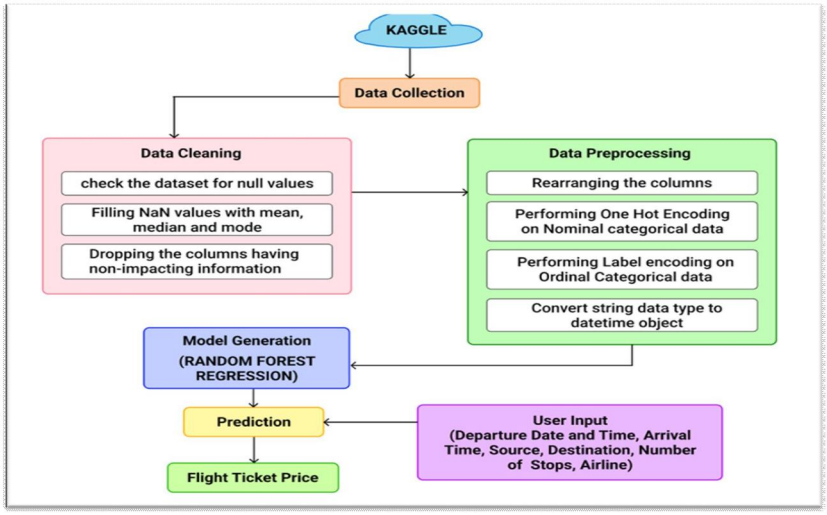
The null values present in the training dataset where removed. A few columns which were of no use for the feature selection processwere deleted from the dataset. The columns of attributes having the categorical data were dropped from the dataset after the new columns containing the numerical values extracted from the pre-processed data were stored for the prediction. Thus, the training dataset suitable for use was obtained and it had the following attribute columns

D. Generating the Model

The model has been generated using the Random Forest Regression.

E. Presenting the Final Prediction

The user input fields will be provided on a webpage developed using the flask framework. The webpage body was built using HTML5 and the same was styled using CSS3. After the user fills all the required input fields and submits the form, the data will be sent to the generate random forest regression model and the predicted value of the ticket price will be displayed



ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue XI Nov 2022- Available at www.ijraset.com

V. ALGORITHMS AND ANALYSIS

While we go through the algorithms we employed (XGBoost, Random Forest, and Decision Tree) and also how they operate in our models, please read the discussion below.

A. Decision Tree

The decision tree appears to be the most well-known and commonly employed categorization technique. A decision tree is a collection of nodes that resembles a diagram, for each junction indicating a test on the a characteristic and each branch indicating a test outcome, such that each node in a decision tree (terminal node) has a class label. A tree can be "trained" by dividing the resources collection into subgroups depending on a characteristic values test. This procedure is known as partitioning the data because it is performed iteratively on each derived subset. The recursion ends when all subgroups at a node have the same posterior probability, or when the split no longer adds additional value to the predictions. A decision tree is appropriate for experimental extracting knowledge since it does not need subject matter expertise or parameters configuration. Assume S is a collection of cases, A is a property, Sv is the subgroup of S with Such a = v, as well as Value (A) is the collection of all number of values of A, then

$$(,) = ( ) - \; | | \; .Entropy( )$$

B. Random Forest

A Random Forest is an ensemble approach that can handle simultaneous regression and classification problems by combining many decision trees using a technique known as Bootstrap as well as Aggregation, or bagging. The core idea is to use numerous decision trees to determine the final result instead of depending on personal decision trees. Random Forest's foundation learning methods are numerous decision trees. We arbitrarily choose rows and characteristics from the dataset to create sample datasets for each model. This section is known as Bootstrap. We simply have to understand the purity in our dataset, and we'll use that characteristic as the root of the tree which has the smallest impurity or, in other words, the smallest Gini index. Mathematically Gini index can be written as:

$$= 1 - \sum ( )$$
$$= 1 - [( ) + ( )]$$

C. XGBoost

XGBoost is an effective method for developing supervised regression models. Knowing as to its (XGBoost) goal function and baseline learners can help determine the truth of this proposition. This optimization problem has both a loss function and a regularization component. It makes a distinction between real and theoretical predictions, i.e. how far the model outputs deviate from the real amounts. In XGBoost, the most used standard error in regression problems is quarantine, whereas reg:logistics is used for classifications. he formula may be used to compute the output value of each model.

VI. IMPLEMENTATION

A. Model

Random Forest Regression: Random Forest Regression is a supervised learning algorithm that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. It operates by building decision trees during training time and outputting the mean of the classes as the prediction of all the trees.[8]

B. UI Development

In this project, Flask framework has been used for the UI development. The main web page of the project takes the required inputs from the user in order to predict the price for the flight. The user inputs required are Departure date and Departure Time, Arrival time of the flight, Source and Destination of the journey, the number of halts during the whole journey and most importantly the airline company which we choose to travel with. After inputting all the fields, the user will click the Submit" button and then the form is submitted. Model enters the scenario at the backend after the submission of the form. The inputs take the help of the historical data and are analysed through supervised machine learning techniques resulting in the prediction of the ticket price. The routing of the pages is done based on the URLs. When the browser finds the '/' in the URL it redirects the user to the home page. After the submission of the form, the user is redirected to the '/result' URL i.e., to the result page where we can see the final result i.e., the prediction of the ticket price. The webpage body was built using HTML5 and the same was styled using CSS3.

VII. EXPERIMENTAL RESULTS

We drew the graph using data visualization to highlight the significance of each attribute for predicting flight prices.

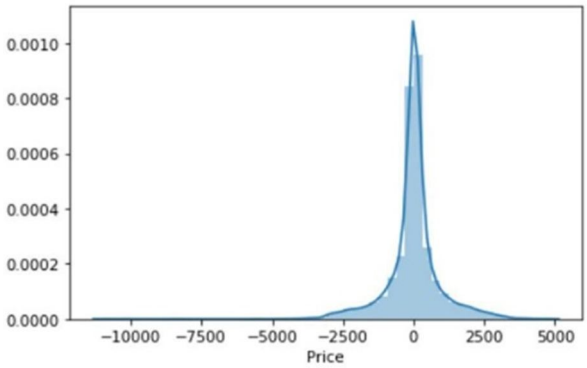


Fig. 4. Normal Distribution Curve between the difference of x-axis and y-axis of training dataset and price.

Meanwhile, depending on our studies, the highest accuracy in the training data for the decision tree method is 97 percent, and the greatest efficiency in Random Forest using testing data is roughly 78 percent.

Table I. Results

Algorithm	Training Accuracy	Testing accuracy
XGBoost	0.92	0.77
Random Forest	0.95	0.78
Decision Tree	0.97	0.67

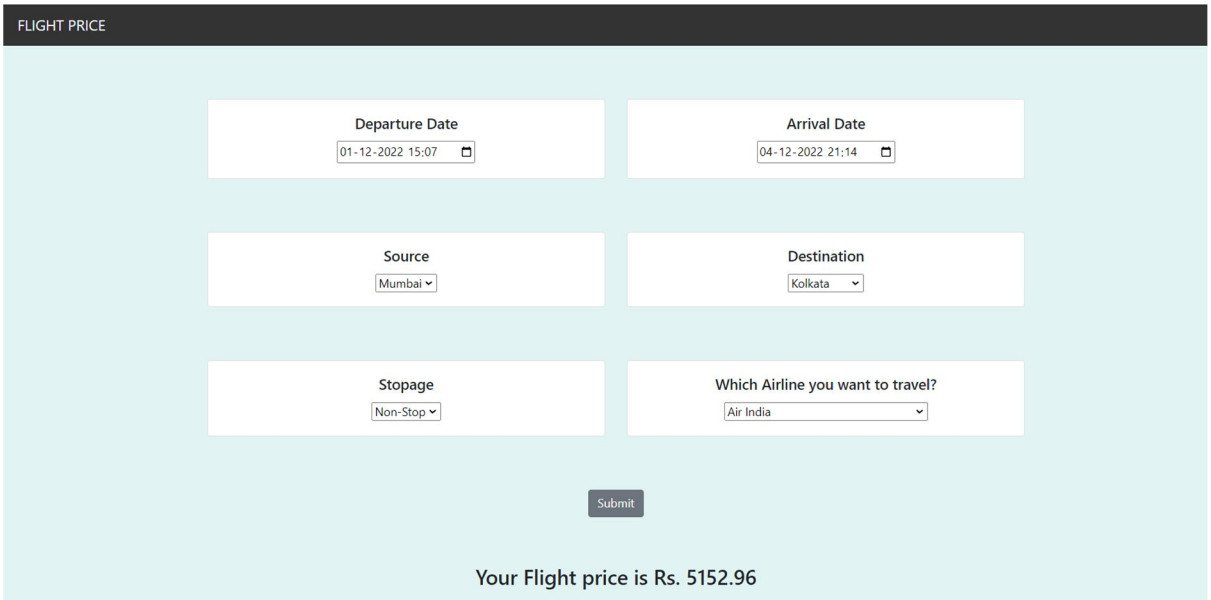


Fig. Prediction of flight price

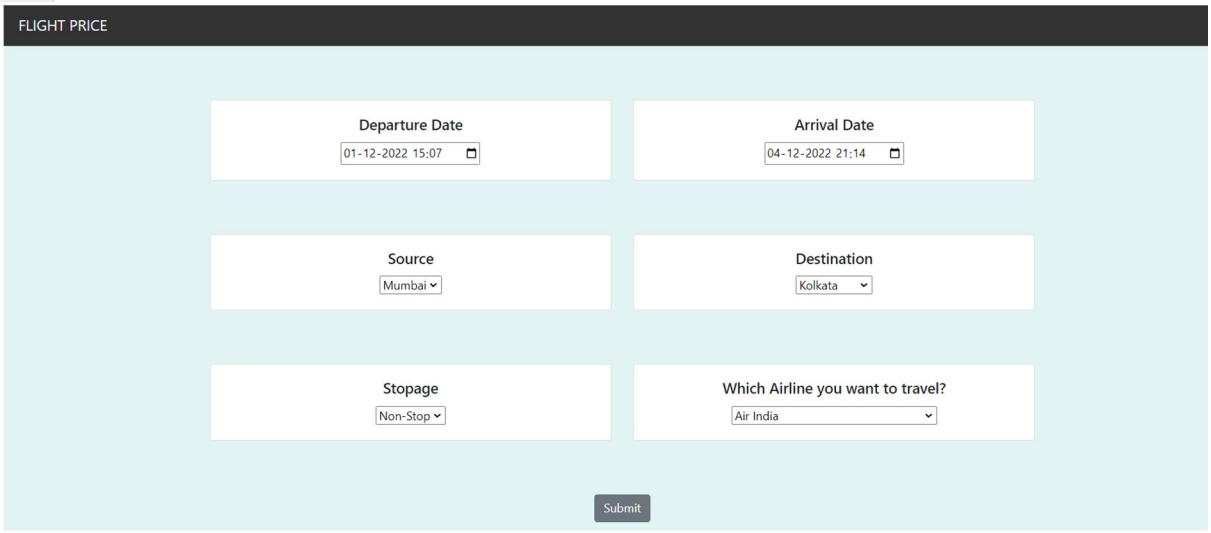


Fig. Taking user input of Day, hour and minute to predict price

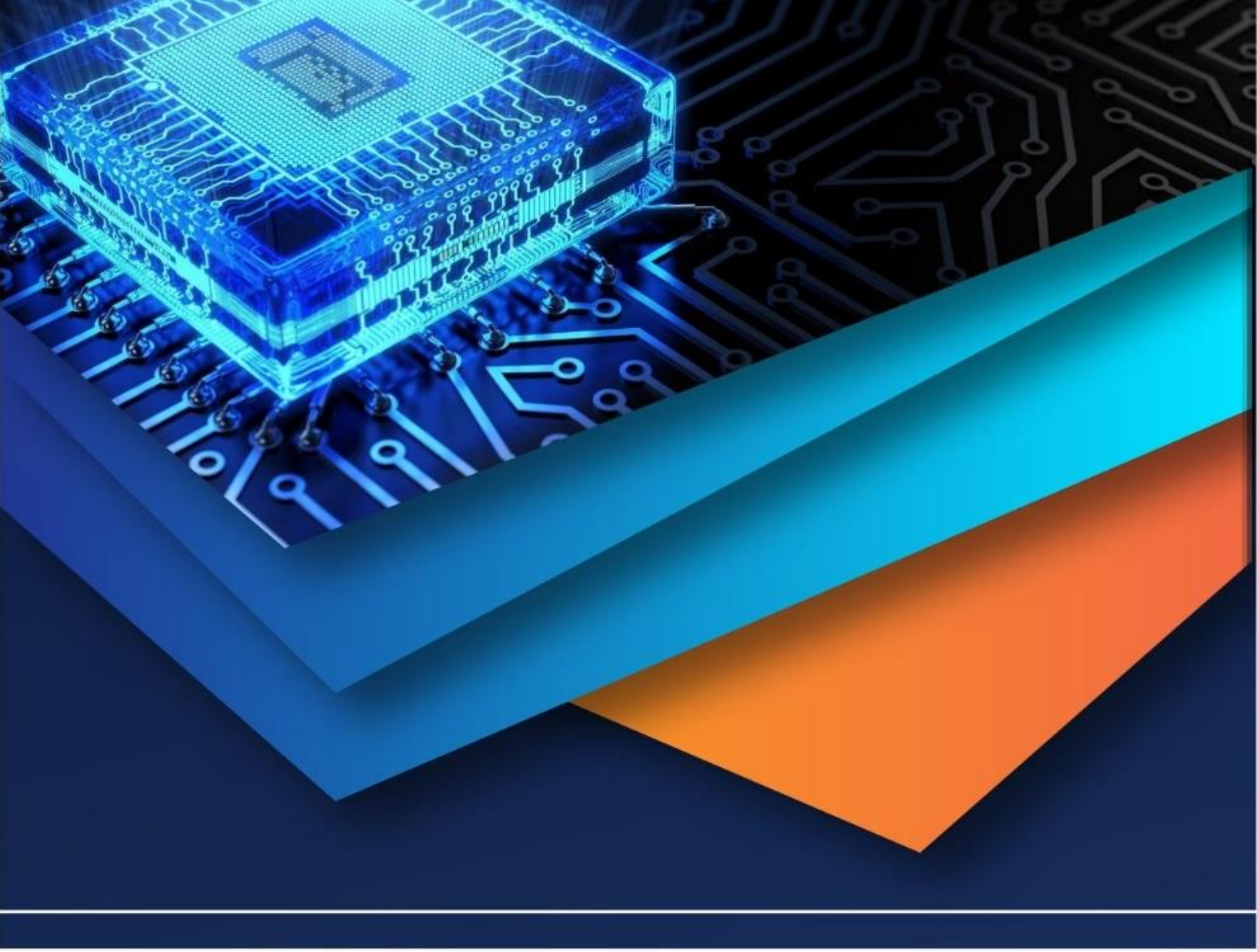
VIII. CONCLUSION AND FUTURE SCOPE

In this study, we evaluated a number of traditional machine learning algorithms on our airfare dataset to create a comprehensible prediction model that can forecast the trend of airfare in a developing aviation market (Vietnam) and assist customers in choosing the best time to book flights in order to maximise savings. Using our own technique, we used information that was gathered from internet travel agency websites. We could only obtain a small amount of publicly available data, missing important details like the number of unfilled seats on a flight. By layering two independent prediction models, Random Forest and Multilayer Perceptron, we may create a final interpretable prediction model. We stack data using fine-tuned weights, with R-squared serving as the primary evaluation metric.

The future aim is to work more on the feature selection and model accuracy. We also plan to extend the study by working with larger datasets and greater number of experimentations on the same to procure more accurate airfares which will in turn help users to get an estimated cost of their next airplane travel and can benefit them to make the best deal. We also plan to level up web applications' user interface to provide a premium user experience. We can also consider various other crucial features that affect airplane ticket prices like public holidays, number of luggage, number of hours till departure, crude oil price, etc. in order to get best results. In the near future, there is also a plan to host the web application.

REFERENCES

[1] K. Tziridis, Th. Kalampokas, G.A. Papakostas K.I. Diamantaras, "Airfare Prices Prediction Using Machine Learning Techniques", IEEE Commun. Mag., vol. 54, no. 5, pp. 138–145, May 2017.
[2] Tao Liu, Jian Cao Yudong Tan, Quanwu Xiao , "ACER:An Adaptive Context-Aware Ensemble Regression Model for Airfare Price Prediction", IEEE Commun.(2017).
[3] Viet Hoang Vu, Quang Tran Minh, Phu, H. Phung, "Airfare Prediction Model for Developing Markets", 2018.
[4] William Groves , Maria Gini, "Optimal Airline Ticket Purchasing Using Automated User-Guided Feature Selection", IEEE, 21-25 July, 2013.
[5] K.I. Diamantaras, T. Papadimitriou, "Parallel pattern classification utilizing GPU-based Kernelized Slackmin algorithm", IEEE, 2016.
[6] Hang Zhou, Weicong Li, Ziqi Jiang, Fanger Cai and Yuting Xue, "Flight Departure Time Prediction Based on Deep Learning",– IEEE, March , 2022.
[7] Micha Zoutendijk, Mihaela Mitici, Probabilistic Flight Delay Predictions Using Machine Learning and Applications to the Flight-to-Gate Assignment Problem", IEEE, JULY 2021.
[8] O. Basturk, C. Cetek, "Prediction of aircraft estimated time of arrival using machine learning methods", IEEE, 2021.,
[9] Guan Gui, Fan Liu, Jinlong Sun, Jie Yang, Ziqi Zhou, "Flight Delay Prediction Based on Aviation Big Data and Machine Learning", IEEE,t 2019
[10] Bin Yua, Zhen Guo, Sobhan Asian , Huaizhu Wang Gang Chen, "Flight delay prediction for commercial air transport: A deep learning approach", IEEE, 2019



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY