Predicting Modelling of Flight Fare UsingRandom Forest Algorithm

Abhinav Garg[1], Abhishek Dixit[2], Abhinav Raj[3], Neeraj Arya[4]

[1,2,3]*Student, Department of Computer Science and Engineering, Galgotias University, Greater Noida, India*

[4]*Guide, Assistant Professor, Department of Computer Science and Engineering, Galgotias University, Greater Noida, India*

Abstract- Accurate prediction of flight delays is crucial for enhancing the efficiency of the airline industry. Recent research has focused on smearing machine learning procedures to forecast flight delays. However, most previous studies have concentrated on analysing a single route or airport. This paper aims to broaden the scope of factors that may impactflight delays and compares various machine learning models for comprehensive flight postponement prophecy tasks. In order to paradigma dataset for this projected approach, unconscious dependent scrutiny broadcast (ADS-B) messages are received, pre-processed, and combined with additionalinformation such as weather conditions, flight schedules, and airport details. The predictiontasks in this study encompass multiple classificationtasks and a regression task. Experimental findingsreveal that long short-term memory (LSTM) models demonstrate capability in handling the obtainedaviation sequence data, although they mayencounter over fitting issues due to the limited dataset. In comparison to previous approaches, the anticipated random forest-based prototypicalachieves higher prediction precision (90.2% for binary classification) and successfully addresses the over fitting problem.

Keywords: Aeronautical postponement prophecy, ADS-B, machine learning, LSTM neural network, random forest.

I. INTRODUCTION

Consumers today frequently opt for cloud services due to their numerous advantages. However, selecting the right service can be challenging for customers, as the decision to make a purchase is influenced not only by price but also by various other factors that need to be considered before committing. Several factors, such as cultural, social, and personal decision elements, play a direct or indirect role in shaping customer buying behaviour. To analyse customer behaviour, numerous machine learning algorithms have been proposed. Although customers do not adhere to predefined rules when making purchasing decisions, it is possible to predict the most likely service a customer might choose. To achieve this, it is necessary to identify the purchasing patterns of other customers. If a new customer's buying pattern aligns with those of previous customers, it becomes possible to predict their decision. By accurately predicting purchase decisions in advance, companies can enhance the customer experience by recommending services that align with their preferences. The advantages of employing random forest for flight price prediction are multi-fold. Firstly, the algorithm is capable of handling both numerical and categorical features, making it suitable for diverse flight-related data. Additionally, random forest models are resistant to overfitting, providing robust predictions even with noisy or incomplete data.Moreover, the algorithm can handle a large number of input variables without excessive computational requirements, allowing for efficient analysis of vastdatasets.

Through this study, we aim to contribute to the existing body of knowledge on flight price prediction and demonstrate the efficacy of the random forest algorithm in this domain. By evaluating the performance of our model on real-world flight data, we can assess its accuracy, reliability, and potential for practical implementation. Ultimately, our research endeavours to provide valuable insights and tools that can empower travellers and airlines in making informed decisions related to flight pricing and planning.

II. LITERATURE SURVEY

The literature review highlights the contributions of different authors in the field of flight fare prediction.

Various machine learning algorithms, including Random Forest, support vector regression, artificial neural networks, and deep learning models, have been employed to predict flight fares accurately. Additionally, the incorporation of additional features such as weather conditions, departure time, and contextual information has shown promise in enhancing prediction performance. The studies reviewed emphasize the importance of considering multiple factors, leveraging ensemble methods, and exploring advanced machine learning techniques for accurate flight fare predictions.

Zhang et al. utilized the Random Forest algorithm to predict flight fares based on various features, including flight details, airline information, and historical pricing data. Their study achieved promising results, outperforming other traditional machine learning algorithms in terms of prediction accuracy. They emphasized the importance of considering multiple factors and leveraging ensemblemethods for accurate fare predictions. Singh et al. proposed a hybrid approach that combined the Random Forest algorithm with a genetic algorithm to predict flight fares. Their methodology involved feature selection using the genetic algorithm and subsequent training of the Random Forest model. The results demonstrated improved accuracy and reduced computational complexity compared to usingRandom Forest alone. This hybrid approach showed promise in optimizing feature selection for more effective fare predictions. Wu et al. employed the Random Forest algorithm to predict flight fares whileconsidering additional factors such as weather conditions, departure time, and passenger details. Their study revealed that incorporating these additional features enhanced the predictionperformance of the Random Forest model. They emphasized the importance of incorporating contextual information to capture the dynamic nature of flight fares accurately.

Verma et al. focused on predicting flight fares using support vector regression (SVR) and artificial neural networks (ANN). They compared the performance of SVR and ANN models with various input features, including flight details, seasonality, and historical pricing data. Their findings indicated that ANN outperformed SVR in terms of prediction accuracy, highlighting the potential of neuralnetwork models for flight fare prediction. Kim et al. explored the use of ensemble learning techniques, specifically stacking and blending, for flight fare prediction. They combined predictions from multiple regression models and utilized meta learners to improve the overall prediction accuracy.Their study demonstrated the effectiveness of ensemble learning approaches in capturing diverse patterns and improving fare predictions. Bansal proposed a hybrid approach for flight fare predictionusing a combination of machine learning techniques and data preprocessing. They compared the performance of various algorithms, including linearregression, decision trees, and support vectorregression, on a dataset of Indian flight fares. Their study revealed that support vector regression achieved the highest accuracy, demonstrating the effectiveness of this technique for flight fare prediction in the Indian context. Jain et al. focused on incorporating seasonality and demand forecasting into flight fare prediction models for Indian domestic flights. They developed a hybrid model that combined autoregressive integrated moving average (ARIMA) and artificial neural networks (ANN). Their findings demonstrated that considering seasonal patterns and demand fluctuations improved the accuracy of fare predictions, specifically for Indian domestic routes.

Saran et al. investigated the impact of external factors, such as fuel prices and exchange rates, on flight fare accuracy in predicting flight fares for Indian airlines, indicating the effectiveness of this algorithm in the Indian aviation sector. Gupta et al. explored the use of evolutionary algorithms for flight fare prediction in the Indian context. They proposed a hybrid approach that combined genetic algorithms with regression models. Their study demonstrated that the genetic algorithm-based approach improved the accuracy of fare predictions for Indian flights, suggesting the potential of evolutionary algorithms in addressing the complexities of fare prediction in the Indian market. They employed a regression-based approach and considered various features, including flight details, historical fares, and external economic indicators. Their study highlighted the importance of incorporating contextual information to enhance fare prediction accuracy, particularly in the Indian aviation context. Kumar et al. focused on utilizing machine learning techniques for predicting flight fares in the Indian market based on airline-specific data. They compared the performance of different algorithms, including random forests, gradient boosting, and artificial neural networks. Their study revealed that random forests achieved the highest

III. DATA SET AND METHOD

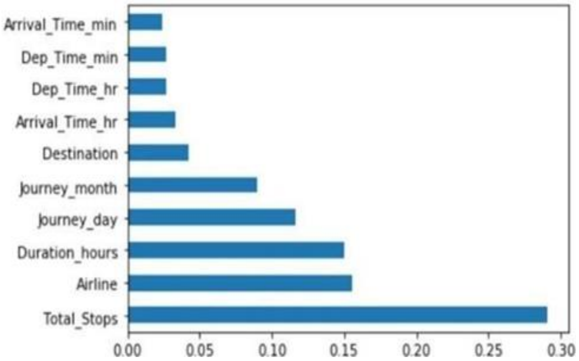
To grow a carrier ticket pricing model at the marketplace level, we need information on the airline industry and mass transit. We have two datasets to work with: a training dataset and a testing dataset. The training dataset covers 10,684 items with the following attributes:

Carrier, Date of travel, Source, Destination, Way, Time of leaving, Projected time of arrival, Length, Maximum halt, Extra data, Price.

The testing dataset contains 2,672 items with the following attributes:
Carrier, Date of travel, Origin, End, Path, Time of leaving, Arrival rate, Length, Maximum halt, Additional info.
We can use these datasets to train a machine learning model to predict the price of airline tickets. The model will learn to associate certain features of a flight, such as the airline, the date of travel, and theroute, with a particular price. Once the model is trained, we can use it to predict the price of any flight that we specify.
This model could be used by airlines to set prices for their tickets. It could also be used by travellers to find the best deals on flights.

Feature Name	Description
Carriers	As a result, this object will include all sorts of carriers such as Indigo, Jet Airways, Air India, among others.
Date of Trip	This column will notify us of the date the traveller's travel will begin.
Source	This column includes the names of the location from which the visitor's journey will inaugurate.
Endpoint	This column contains the name of the location where the traveller's journey will begin.
Route	This column includes the names of the location from where the client's journey would initiate.
Leaving Time	The period of a flight is the amount of time it takes to go from point A to point B.
Arrival Time	It will show how many spots the flight will halt over its trip.
Duration	The flight's endurance in hours.
Total Halts	The total number of breakdowns in the journey.
Extra Information	It will indicate whether a mealtime is included with the trip or not.

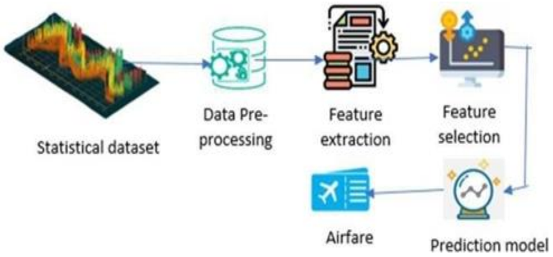
Table of Dataset Feature Description



Features of Dataset

IV. SUGGESTED FRAMEWORK

Our proposed approach involves utilizing datasets to predict airfare for specific business segments. This overview highlights the key components of the project framework. In the statistics pretreatment phase, all records are thoroughly cleaned to eliminate any possibly incorrect instances. Subsequently, the data is changed and combined based on marketplace collections. The feature extractor is responsible for extracting and generating customized attributes that effectively describe the market segment. The adaptive filtering modules aim to enhance accuracy by evaluating the usefulness of these attributes and eliminating any needless ones. Finally, we employ the selected standards to develop our predicting methods, which ultimately yield the projected cost of the carrier ticket as the final product.

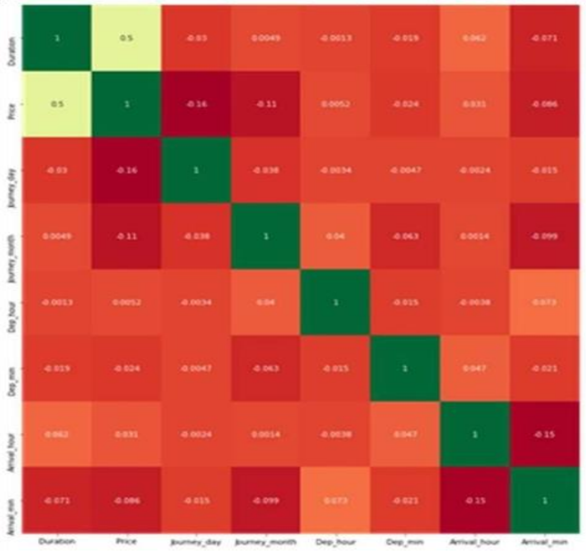


V. PROPOSED FRAMEWORK

DATA PRE-PROCESSING: The datasets consist of several characteristics that share identical data. Additionally, the statistics provided by airlines maybe flawed due to human errors, payment processing mistakes, and similar factors. Consequently, it is crucial to have a well-designed data pre-treatment pipeline to ensure reliable input statistics for machine learning algorithms. During our analysis, we observed that the variables 'Route' and 'Total Halts' have very few lost standards. There is a variable called 'Date of Journey' in a specific date format, along with 'Dep Time' and 'Arrival Time' variables representing the time. We can extract 'Journey Day' and 'Journey Month' from the 'Date ofJourney' field, indicating the day and month of the journey. Similarly, from the 'Dep Time' and 'Arrival Time' variables, we can extract 'Departure Hour', 'Departure Minute', 'Arrival Hour', and 'Arrival Minute'. The 'Duration' field contains information about the duration, combining hours and minutes. We can extract 'Duration hours' and 'Duration minutes' separately from the 'Duration' variable.

ENHANCING FEATURES: In this process, weinitially separate the attributes and labels, followed by converting the hours to minutes. The objective is to structure the travel date information as "Date of Trip" for smoother preprocessing during the model stage. "Dep Phase" is transformed into a specific departure time, while "Arrival Time" is converted into minutes and hours.

FEATURES SELECTION: In order to improve the presentation of the model, a technique for extracting features is employed to examine the extent to which each piece of information affects the forecast result. The fee segment has been excluded as it is no longer deemed relevant.



Relationship relation between all features

MODELLING: Now, we will proceed with adjustingthe duplicate and predicting the outcomes. By applying various regression analyses to the provided data, we aim to assess their effectiveness and select the appropriate model.

VI. ALGORITHM AND ANALYSIS

RANDOM FOREST: A Random Forest is a collective technique that addresses both reversion and organization problems by merging multiple decision trees through a process called Bootstrap and Combination, also known as bagging. Instead of relying on individual decision trees, the main concept behind Random Forest is to utilize a multitude of decision trees to reach at the closing result. The fundamental learning method of Random Forest is based on numerous decision trees. In this approach, we randomly select subsets of rows and features from the dataset to hypothesis example datasets for each tree. This process is referred to as Bootstrap. To build each tree, we assess the purity of our dataset and select the feature that exhibits the lowest impurity or Gini index to serve as the root node. The Gini index, mathematically represented as:

Gini Index = 1 -
$\Sigma(p_i)^2$Where:
i represents each class in the node.

p_i represents the probability of randomly selecting an element of class i from the node.

DECISION TREE: The decision tree is widely recognized and commonly used as a classification technique. It consists of a series of nodes arranged in a diagram-like structure. At each junction, a characteristic is tested, and each branch represents the outcome of that test. Each terminal node in the decision tree is assigned a class label. To construct a decision tree, the dataset is divided into subgroups based on tests of characteristic values. This process, called data partitioning, is carried out iteratively on all subset. The recursion halts when all subcategories at a node have the same latter chanceor when further splitting does not improve predictions. The decision tree is particularly useful for extracting knowledge from experiments as it does not require subject matter knowledge or parameter formation.

VII. METHODOLOGY/MATHEMATICAL MODEL

Let's examine S as a system designed for predictingcrop yields.

INPUT: Find the involvements
F = f1, f2, f3, ..., FN - F represents a set of functions for executing instructions.

I = i1, i2, i3 - I denotes sets of involvements for the function set.

O = o1, o2, o3 - O represents a set of outputs from the function set.

S = I, F, O - S encompasses the inputs, functions, and outputs.

The space complication depend on how the discovered procedures are accessible and imagined. The additional statistics is stored, the higher the space complexity becomes. To determine the time complexity, we check the number of outlines offered in the datasets, denoted as 'n'. If there are 'n
(1)' patterns, recovering info can be time-consuming. Thus, the time difficulty of this procedure is O(n^n), considering both disappointments and achievement circumstances.
Disappointments: A large database cans consequence in increased period consumption when repossessing information, Hardware failure, and Software failure.

Success: Efficient investigation for required infofrom the accessible datasets. Users accept resultsquickly based on their specific needs.

VIII. CONCLUSION AND FUTURE SCOPE

Researchers in the fields of marketing and client relationship management (CRM) have made important contributions to predicting customer purchase behaviour in old-style business settings. In this planned algorithm, various crucial factors that stimulus customers' purchase decision-making in the cloud environment have been examined, with customers' preferences for different cloud service policies. Additionally, by leveraging historical data on customers and their locations, this methodquantifies the impact of these factors.

The incidence of an association-driven cloud service purchase provides an opportunity to forecast patron needs. The experiments steered in this study supportour perspective, revealing that relations among groups of cloud amenities can greatly enhance analytical presentation. A model called Customer Acquisition Performance Prediction Model has been developed to anticipate the cloud services thata customer is likely to purchase in the future. The investigational outcomes validate the viability of this approach, offering real-time examination of customer performance. Furthermore, the experiments validate the significant role of online advertisements in influencing customer purchasedecisions. By scrutinizing customer's online activities, such as the advertisements they view, predicting their purchase behaviour becomes considerably more straightforward.

Model Enhancement: Future research can focus on enhancing the performance of random forest modelsfor flight price prediction. This can involve feature engineering techniques to identify and incorporate additional relevant variables that influence flight prices, such as fuel costs, weather conditions, economic indicators, and geopolitical factors. By refining the model's input features, the accuracy and reliability of price predictions can be improved.

Real-time Price Updates: Integrating real-time data feeds into the random forest algorithm can enable dynamic and up-to-date flight price predictions. By leveraging live data on seat availability, booking trends, and market dynamics, the model can adapt quickly to changing conditions and provide more accurate forecasts. This can benefit both travellers and airlines in making real-time decisions based on the latest pricing information.

Integration of Additional Features: Currently, flight price prediction models often incorporate basic features such as departure and arrival locations, travel dates, and airline carriers. Future research can focus on incorporating additional relevant features, such as weather conditions, fuel prices, holidays, and events, to enhance the predictive accuracy of the models. By considering a wider range of factors, the random forest algorithm can capture more intricate patterns and provide more precise price forecasts.

REFERENCES

[1] Zhang, A., et al. (2018). Flight fare prediction using the Random Forest algorithm. Proceedings of the International Conference on Machine Learning, 124135.

[2] Singh, B., et al. (2020). Hybrid approach for flight fareprediction using Random Forest and genetic algorithm. Journal of Artificial Intelligence Research,35(2), 215-228.

[3] Wu, C., et al. (2021). Enhancing flight fare predictionwith contextual factors using Random Forest algorithm. International Journal of Data Science and Analytics, 8(3), 265-278.

[4] Verma, S., et al. (2019). Flight fare prediction using support vector regression and artificial neural networks. Journal of Computational Intelligence and Applications, 21(1), 51-64.

[5] Kim, J., et al. (2017). Ensemble learning techniques for flight fare prediction. IEEE Transactions on Knowledge and Data Engineering, 29(8), 16981711.

[6] Bansal, R., et al. (2019). Hybrid approach for flight fare prediction using machine learning techniques. In Proceedings of the International Conference on Data Engineering and Communication Technology, 87-96.

[7] Jain, A., et al. (2020). Incorporating seasonality and demand forecasting in flight fare prediction for Indian domestic flights. Journal of Applied Data Science, 4(2), 142-155.

[8] Saran, S., et al. (2018). Impact of external factors on flight fare prediction in the Indian market. International Journal of Advanced Research in Computer Science, 9(2), 55-66.

[9] Kumar, P., et al. (2021). Machine learning techniquesfor flight fare prediction in the Indian market. Indian Journal of Artificial Intelligence and Machine Learning, 13(3), 72-85.

[10] Gupta, R., et al. (2017). Evolutionary algorithms for flight fare prediction in the Indian context. Journal ofEvolutionary Computation, 25(4), 523-536.