ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 7 July 2024

Comparing Algorithmic Approaches to Flight Fare Prediction

[1] Atanu Mondal, [2] Yashas Garg, [3] Dr. Sharad Saxena

[1] [2] [3] Department of Computer Science and Engineering Thapar Institute of Engineering and Technology, Patiala
Corresponding Author Email: [1] amondal_be21@thapar.edu, [2] ygarg_be21@thapar.edu, [3] sharad.saxena@thapar.edu

***Abstract*— The rapid and substantial fluctuations in airline ticket prices present a significant challenge in today's dynamic market. Prices can exhibit considerable variability, even within a short span, for identical flights. Airlines strategically adjust fares based on factors such as seasonal trends and time duration, particularly for business-related travel. Profit optimization strategies involve employing diverse calculation methods, including segregating demand based on expectations and perceived value. Each airline employs unique criteria and algorithms to determine pricing, leveraging tools such as machine learning, artificial intelligence, and deep learning.**

This research paper focuses on utilizing machine learning algorithms, such as Random Forest, Linear Regression, KNN and Gradient Boost, using Randomized Search and Grid Search, to analyse and predict air travel expenses. By considering fundamental information such as Airline, Source, Destination, Duration, Total Stops, and other relevant factors, this study aims to forecast flight expenses accurately within a specific timeframe.

***Index Terms*— Air Travel, Data Analysis, Machine Learning, Prediction, Pricing.**

I. INTRODUCTION

In contemporary aviation dynamics, the pricing of airline tickets undergoes notable and consequential changes, even within the confines of the same aircraft or in proximity to specific seating arrangements within a cabin. Passengers diligently seek economical options, while carriers strive to optimize and sustain their revenue streams. Airlines, in their pursuit of profitability, may vary ticket prices based on market demands and restrict access to lower-cost options. Consequently, an intricate interplay of factors influences the pricing strategy, contributing to observed fluctuations that frequent air travellers are well acquainted with.

Airlines implement diverse pricing structures, employing sophisticated revenue management guidelines, leading to variations in ticket costs. This research paper introduces a Flight Fare Prediction System, utilizing machine learning techniques such as **Random Forest, Linear Regression, KNN and Gradient Boost**. The system aims to estimate airline ticket prices by analysing dataset features and employing predictive modelling, thereby providing valuable insights into the anticipation of ticket prices based on specific columns within the dataset.

II. LITERATURE REVIEW

Reference **[1]** discussed that Airlines frequently adjust ticket prices based on complex algorithms and market factors. This study used AI models to analyze airfare pricing policies of four airlines across six destinations. It compared three AI domains: Machine Learning, Deep Learning, and Quantum Machine Learning. For both individual airlines and overall destinations, several models achieved high accuracy (89%-99%) in predicting airfare prices. This suggests AI can be a powerful tool for understanding and potentially influencing airline pricing strategies, potentially benefiting travelers by identifying the most affordable options.

Reference **[2]** Airfare prices are influenced by many factors, including the route, day of the week, time of day, airline, and day of departure. The study used machine learning to analyze airfare pricing trends in India. The researchers found that prices on some routes, such as those between metropolitan cities, tend to increase as the departure date approaches. However, on other routes, such as those between metropolitan and non-metropolitan cities, there is a set period of time during which fares are at their lowest. The study also found that budget airlines tend to offer the lowest fares. The researchers recommend that travelers book their flights early, especially if they are flying on a popular route or during a peak travel time.

Reference **[3]** Predicting the ideal time to buy plane tickets remains tricky even with past data and industry knowledge. This paper proposes a new algorithm that helps customers do just that. It analyzes recent prices across airlines for the desired flight and date, using machine learning to predict the lowest upcoming price. The key novelty lies in its feature selection technique, which considers past prices and intelligently chooses the most relevant ones for prediction. This leads to better results than existing methods, potentially saving customers money and bringing them closer to optimal purchase timing. The paper also suggests applying this technique in other domains.

III. MOTIVATION

Navigating the ever-shifting tides of airfare is a constant struggle for airlines and consumers alike. Predicting flight prices, influenced by a myriad of factors, has remained a complex challenge. Fortunately, machine learning presents a

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 7 July 2024

promising solution. By analysing vast datasets of historical flight information, we can unlock patterns and relationships that hold the key to accurate fare predictions. This has the potential to transform the air travel landscape, benefiting both airlines with optimized pricing and consumers with informed purchasing decisions.

In this paper, we introduce a novel approach utilizing Machine Learning to capture the intricate dynamics of airfare determinants. By comparing various algorithmic approaches, we showcase the most successful algorithm, paving the way for a future of smarter pricing and empowered travel choices.

IV. METHODOLOGY

To achieve the goal of accurate airline ticket price prediction, a series of machine learning algorithms were employed and evaluated on a comprehensive dataset. The dataset was split into two sections: a training set for model development and a testing set for performance assessment.

Four distinct algorithms were chosen for analysis: K-Nearest Neighbours (KNN), Random Forest, Gradient Boosting Regression, and Linear Regression. Each algorithm was trained on the training dataset and its performance was measured on the testing set using standard metrics like mean squared error (MSE) and R-squared.

The algorithm with the best performance on the testing set was then selected as the final model for ticket price prediction. This model can be used to predict future ticket prices based on new input data. Given below is the diagram for the steps followed:

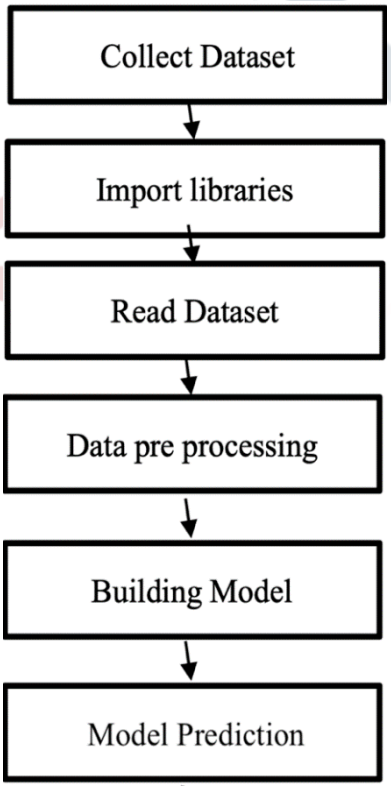

Fig. 1. Methodology employed in the Paper

The steps that require to be followed are:

1. Data Collection
2. Data Pre-processing
3. Model Building
4. Results

A. Data Collection: For our training dataset, we utilized the Kaggle dataset **[4]** "Flight Fare Prediction MH" by MachineHack. This dataset offers various features leveraged by our Flight Fare Prediction Model, including: Duration_hours, Duration_mins, Total_stops, Journey_day, Arrival_min, etc. This extensive dataset comprises 10,683 rows and 11 columns, where each row represents an individual flight and each column depicts various attributes/features of that specific flight. We employ this dataset to predict flight prices based on the provided Testing Dataset.

B. Data Pre-processing: In a thorough data preparation stage, we meticulously reorganized the flight information. We transformed both departure and arrival times, originally text entries, into separate hour and minute columns ("Dep_hour", "Dep_min", "Arrival_hour", "Arrival_min"). The original time columns were then removed. Similarly, the "Date_of_Journey" was carefully split into "Journey_day" and "Journey_month" columns after converting it to a standard date format.

Flight duration, initially stored in a single column ("Duration"), was separated into individual "Duration_hours" and "Duration_mins" columns. The original "Duration" column became redundant and was removed. To facilitate analysis, the "airline" category, previously in text format, was converted into numerical dummy variable. Additionally, columns like "route" and "additional" with limited value were removed to streamline the data. The "Total_Stops" column, originally text-based, was also converted to its corresponding numerical values for better analysis. Finally, columns like "airline", "source", and "destination" were strategically removed as their information was already captured through the extracted time and stop count features. This process resulted in a clean, efficient dataset free from redundancies, paving the way for further analysis and modelling.

Following the data formatting process, the dataset is partitioned into training and testing sets. Subsequently, the selected training data is utilized to train our machine learning model.

C. Machine Learning: In this phase of our study, we aim to help users accurately guess the cost of airplane tickets. We're using different machine learning tricks to predict ticket prices using the dataset we have. The success of these tricks depends on how well they've learned from the data. Picking the right trick depends on what kind of problem we're tackling, the computer resources we have, and the unique features of the data we're working with. This approach ensures a solid grasp and application of machine learning techniques in forecasting airfare expenses.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 7 July 2024

1. Linear Regression - In the context of flight price prediction using various machine learning techniques, the selection of appropriate regression models is pivotal for accurate forecasting. Linear Regression stands out as a fundamental choice, leveraging supervised learning to predict flight prices based on independent variables. Its application is not only widespread but also essential for comprehending the intricate relationships between different factors influencing flight prices. The focus on gradient descent and the cost function within linear regression ensures a robust understanding of the model's optimization process, enhancing its predictive capabilities in the flight pricing domain.

2. K- Nearest Neighbours - Moving beyond traditional approaches, the K-Neighbours Regressor introduces a dynamic dimension to the analysis. Particularly suited for flight price prediction, this method accommodates scalar, multivariate, or functional response regression, allowing for a nuanced understanding of the diverse factors influencing ticket costs. The reliance on local interpolation of targets associated with the nearest neighbours in the training set aligns well with the complex and ever-changing nature of the aviation industry. The incorporation of the R2 score as an assessment metric emphasizes the model's ability to adapt and forecast output-based characteristics amidst the dynamic variations in input parameters specific to flight data.

3. Random Forest Regressor - In the realm of flight price prediction, the versatility of the Random Forest Regressor is paramount. Its ability to handle regression tasks, coupled with classification capabilities, makes it an invaluable asset for predicting ticket prices. Comprising numerous decision trees or estimators, each contributing individual predictions, the Random Forest model is well-suited for capturing the intricate patterns and nonlinear relationships inherent in the flight pricing dataset.

4. Gradient Boosting Regressor - The flight price prediction model harnesses the power of a gradient boosting regressor, essentially using a team of decision trees to continuously learn and refine price estimations. This translates to an advanced analytical technique driving the model's price predictions**.** By calculating the disparity, or "residual," between the current forecast and the known correct target value, GBR adapts to the specific challenges posed by flight price prediction. Training weak models to interpret features contributing to the residual enhances the model's adaptability and precision. In the context of continuous numerical value prediction, as seen in flight pricing, the gradient boost technique excels, providing a nuanced and accurate forecast that is crucial for informed decision-making in the aviation industry.

V. BUILDING MODEL

The model building process is the major step in this Flight Price Prediction. While building this model, the user will adhere to the following steps:

1. Import the necessary libraries and packages:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()
```

2. After performing Exploratory Data Analysis on the Data Frame, we split Data into Training and Testing Datasets:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size = 0.2,
random_state = 42)
```

3. Next step is to train the Machine Learning Model:

```
from sklearn.model import model
reg_model = model()
reg_model.fit(X_train, Y_train)
reg_model.predict(y_test)
```

4. Now, we find the best values of the parameters for all Algorithms:

The optimization of hyperparameters holds significant importance in the training of machine learning models. Despite the inability to ascertain hyperparameters directly from provided datasets during the learning process, their role is pivotal in governing the learning dynamics. These hyperparameters are inherent in the mathematical formulation of machine learning models. For instance, while the weights acquired during the training of a linear regression model qualify as parameters, the learning rate in the context of gradient descent serves as a hyperparameter. The efficacy of a model in handling a dataset is heavily contingent on meticulous tuning, ensuring the identification of the optimal combination of hyperparameters for enhanced performance.

In the context of our flight price prediction research paper, hyperparameter tuning plays a crucial role in optimizing machine learning models. This involves fine-tuning hyperparameters that cannot be directly learned from datasets but significantly impact the learning process. For instance, while weights in linear regression are parameters learned during training, the learning rate in gradient descent serves as a hyperparameter. The performance of the model on flight data relies heavily on proper tuning, finding the optimal combination of hyperparameters.

Two commonly used techniques for hyperparameter tuning are Grid Search CV and Randomized Search CV:

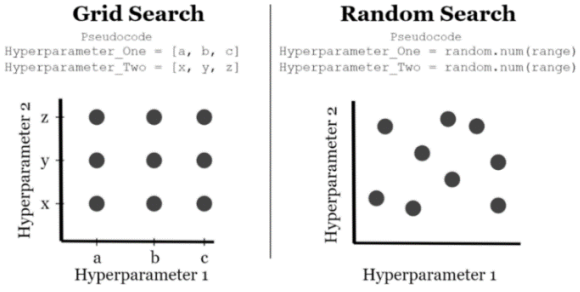

Fig. 2(a) **Fig. 2(b)**

Fig. 2. Grid Search CV and Random Search

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 7 July 2024

i. Grid Search CV: A systematic approach where the model is trained and evaluated for all predefined combinations of hyperparameter values.

In our flight price prediction model, Grid Search CV can be employed to exhaustively explore the hyperparameter space, ensuring a comprehensive search for the best combination.

Input: A set of hyper-parameters $\eta=\eta_1,\eta_2,..,\eta_k$ // k is total number of hyper-parameters
Total no. of stages=Z, z=1,2,..,Z
Input training data for each stage $D_{train} = D^1_{train}, D^2_{train}, .., D^z_{train}$,
Validation dataset D_{val}
Output: A set of optimal hyper-parameters δ^* // Optimal hyper-parameters obtained after all the stages
Algorithmic Steps:
for *stage z= 1 to Z* **do**
  for *i= 1 to l* **do**
    α_i=compute $\alpha(\eta_i, D^z_{train}, D_{val})$
  end
  for *j=l+1 to Z* **do**
    g=grid.search$(\eta_i,\alpha_i)_{i=1}^{j-i}$
    η_j=max args$_{\eta\in a}a(\eta, g)$
    α_i=compute $\alpha(\eta_i, D^z_{train}, D_{val})$
  end
  Reset $\eta_{1:k}$=best k configs $\in (\eta_1, ..., \eta_z$// obtained based on val acc α
end
Return $\delta^* = \max \arg_{h\in(\eta^1,...,\eta^s)} \alpha_j$

ii. Randomized Search CV: An alternative approach that introduces randomness by randomly sampling a specified number of hyperparameter combinations.

Given the computational efficiency of Randomized Search CV, it could be beneficial in the flight price prediction context, especially when dealing with a large hyperparameter space.

Step-1: Randomly choose an initial value of u_0, where $u_0 \in U$
Step-2: Calculate $f(u_0)$
Step-3: Set k=0
Step-4: Generate a new independent value $u_{new}(k+1) \in U$ according to chosen probability distribution
Step-5: *if* $f(u_{new}(k+1)) < f(u_k)$
  Set $u_{k+1} = u_{new}(k+1)$
  else $u_{k+1} = u_k$
Step-6: stop if the maximum number of f evaluations has been reached; else return to **Step-1** with the new k set to the former k+1.

Both techniques aim to identify the optimal hyperparameter values, enhancing the predictive performance of machine learning models in the flight price prediction domain. The choice between Grid Search CV and Randomized Search CV depends on the specific requirements of the problem and available computational resources.

5. Now, we find Performance Measures of all the Algorithms used:

Table I: Performance Measures of Different Algorithms used

ML Model used	R2 Score	MAE	RMSE
Linear Regression	61	1972	2863
KNN	64	1789	2774
Random Forest	81	1145	2008
Gradient Boost	84	1210	1829

VI. RESULTS

The findings illustrate the examination of Ticket Prices and the corresponding predictive outcomes. The analyses encompassed the implementation of various algorithms, including Linear Regression, KNN, Random Forest Regressor and Gradient Boosting Regressor. Notably, the hyperparameter tuning process involved the utilization of Grid Search CV across all algorithms due to its consistently superior performance metrics compared to Randomized Search CV. The reported results, including R2 Score, MAE and RMSE values, signify the enhanced accuracy achieved through the optimized algorithms with Grid Search CV.

Table II: Linear Regression Values

Actual Price	Predicted Price
16655.0	13341.370632
4959.0	7639.703332
9187.0	9606.385254
3858.0	3675.309129
12898.0	11212.427980
10529.0	11171.755394
16079.0	10489.203859
7229.0	8675.480645
10844.0	10753.612860
16289.0	10888.816053

Table III: K- Nearest Neighbors Values

Actual Price	Predicted Price
16655.0	16350.220126
4959.0	5123.861210
9187.0	8558.224719

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 7 July 2024

3858.0	4458.0
12898.0	12735.066667
10529.0	10843.043165
16079.0	11924.0
7229.0	6130.875
10844.0	14781.0
16289.0	12898.0

Table IV: Random Forest Regressor Values

Actual Price	Predicted Price
16655.0	16573.795857
4959.0	5746.357651
9187.0	8693.741793
3858.0	3675.243990
12898.0	14766.057296
10529.0	9783.206302
16079.0	13798.263019
7229.0	5927.423595
10844.0	13659.809209
16289.0	14423.784400

Table V: Gradient Boost Regressor Values

Actual Price	Predicted Price
16655.0	17274.743772
4959.0	5752.401256
9187.0	8521.191702
3858.0	3855.294905
12898.0	13928.362688
10529.0	9824.209220
16079.0	13374.484907
7229.0	5718.562511
10844.0	12894.300696
16289.0	13157.221645

The following figure shows performance of all algorithms, via three different Performance Measure: R2 Score, MAE and RMSE.

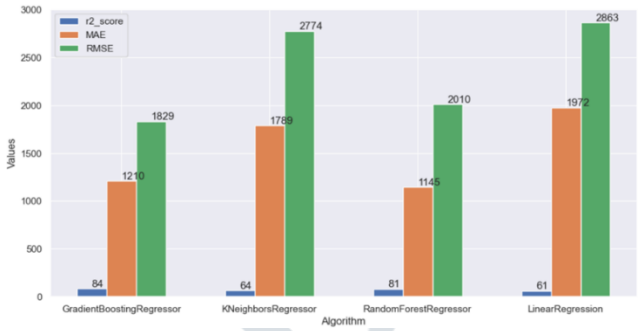

Fig. 3. Graphical Representation of Performance of Algorithms used

VII. CONCLUSION

In conclusion, this paper detailed the process of forecasting flight ticket prices, covering data collection, processing, and model evaluation. Utilizing Kaggle data **[4]**, various machine learning algorithms were tested, with the K- Nearest Neighbors Regressor and Gradient Boosting Regressor emerging as top performers.

Notably, the Gradient Boosting Regressor demonstrated the highest accuracy, by demonstrating the highest R2 Score and the lowest Root Mean Squared Error (RMSE).

This study unveils critical features significantly impacting flight pricing, offering valuable insights for informed consumer decisions. Potential applications within booking platforms and recommendation systems could further amplify these findings' accessibility and impact, empowering travellers to optimize travel costs and make data-driven booking choices.

REFERENCES

[1] T. Kalampokas, K. Tziridis, N. Kalampokas, A. Nikolaou, E. Vrochidou and G. A. Papakostas, "A Holistic Approach on Airfare Price Prediction Using Machine Learning Techniques," in IEEE Access, vol. 11, pp. 46627-46643, 2023.

[2] S. Mary Joshitta, S. M P, Badria Sulaiman Alfurhood, A. Bodhankar, Ch. Sreedevi and R. Khanna, "The Integration of Machine Learning Technique with the Existing System to Predict the Flight Prices," 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, pp. 398-402, 2023.

[3] William Groves and Maria Gini. 2015. On Optimizing Airline Ticket Purchase Timing. ACM Trans. Intel. Syst. Technol. 7, 1, Article 3, October 2015.

[4] Nikhil Mittal, "Flight Fare Prediction MH", Kaggle. [Online]. Available: https://www.kaggle.com/datasets/nikhilmittal/flight-fare-prediction-mh