Prediction of Flight-fare using machine learning.

Jathusan Uthayaraj
Department of Computer Engineering
University of Ruhuna
Galle, Sri-Lanka
jathurajan@gmail.com

Senevirathna G.D.I.U
Department of Computer Engineering
University of Ruhuna
Galle, Sri-Lanka
imasha@gmail.com

***Abstract*—. The Flight Fare Prediction System is a comprehensive solution aimed at accurately forecasting flight ticket prices, providing travelers with valuable insights for better planning and decision-making as well as the airlines to maximize their revenue. With the exponential growth of the airline industry and the increasing complexity of fare structures, predicting flight fares has become a challenging task. This system leverages machine learning algorithms and historical flight data to generate accurate fare predictions. The system utilizes a vast dataset comprising historical flight fares, including factors such as travel dates, destinations, airlines, departure times, and various other relevant variables. By analyzing this data using two different regression algorithms, the system learns patterns and relationships, enabling it to make reliable predictions about future flight fares. Acknowledging limitations, future research could address bias and improve generalization.**

Keywords — Machine Learning, Flight fare, Regression algorithms, Hyperparameter tuning, historic data.

I. INTRODUCTION

Travel is a universal pursuit driven by the human desire for exploration and new experiences. India, with its unparalleled diversity, stands as a beacon for travelers, offering some of the world's most captivating cities. Recognizing the significance of informed travel planning, this project endeavors to contribute to the travel industry by developing an application that anticipates flight pricing. Using machine learning model, this project intends to produce an application that will anticipate travel pricing for different flights across India.

The Flight Fare Prediction System Is a machine learning initiative using regression algorithms that aims to estimate aircraft ticket costs using relevant features and historic data. This strategy is provided to travelers, travel firms, and airlines to anticipate trip costs for planning, budgeting, and making sensible selections.

With a focus on enhancing the precision of flight fare predictions, our project uniquely delves into a comparative analysis of Linear Regression and Decision Trees, aiming to dissect their performance intricacies in the context of airline pricing. This project not only addresses the core challenge of airfare prediction but also contributes valuable insights into the efficiency of machine learning methods in this domain.

II. METHODOLOGY

A. *Data*

The dataset utilized in this project, sourced from Kaggle's(https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction/) Flight Price Prediction, presents a comprehensive compilation of flight-related information crucial for the development of the predictive model. Spanning 300,153 entries across 12 columns, the dataset encompasses an array of prominent features, including airline details, flight numbers, source and destination cities, departure and arrival times, stopovers, ticket classes, flight durations, and the target variable – price. This extensive dataset offers a diverse and representative set of parameters for analysis, laying a robust foundation for training and evaluating machine learning models to forecast accurate flight ticket prices.

B. *Pre-processing*

b) Handling missing values

In conducting a preliminary analysis of the dataset, several key steps were taken to ensure its integrity and suitability for machine learning model development. The initial focus involved handling missing values and data cleaning. Fortunately, the dataset proved to be complete, devoid of any missing values, yet a thorough examination was carried out to identify and address potential outliers or anomalies. Visualizations, such as box plots and histograms, were employed to assess and mitigate outliers, thereby enhancing the stability of subsequent machine learning models.

b) Unwanted Feature Removal

During the data preprocessing phase for the Flight Ticket Price Prediction project, a crucial measure was taken to identify and eliminate the "Unnamed: 0" column. This column, serving as a mere index, was found to lack substantive information for predicting flight prices. The deliberate removal of this feature during preprocessing aimed at improving the overall efficiency of the modeling process. This strategic elimination not only simplified

machine learning model training but also significantly enhanced computational efficiency by reducing noise in the dataset.It contributed to optimizing the models' predictive capabilities, allowing them to discern meaningful patterns without interference from irrelevant information.

c) Encoding the dataset features

In the process of encoding categorical features for the Flight Ticket Price Prediction project, significant features such as "airline," "source city," "departure time," "stops," "arrival time," "destination city," and "class" were systematically converted into a numerical format. The utilization of Label Encoding from the scikit-learn library ensured each category within these features received a distinct numerical label. This step is imperative as machine learning algorithms generally require numerical inputs for efficient processing. By employing individual applications of the Label Encoder for each category feature, the encoded values accurately represented the original categories without introducing any unnecessary bias, thus enhancing the dataset's compatibility with machine learning algorithms.

d) Normalization

In the data preprocessing phase of the Flight Ticket Price Prediction project, a vital step involved the application of min-max scaling, a normalization technique, to the numerical features within the dataset. This process ensures that numerical values are transformed into a predefined range, typically between 0 and 1. The significance of normalization arises when dealing with numerical features that possess diverse scales or units. By employing min-max scaling, the risk of features with larger scales overshadowing the training process is mitigated, as each feature contributes proportionately to the model's learning. This normalization technique proves especially crucial in the context of flight ticket price prediction, where numerical features like 'length' and 'days_left' may exhibit disparate measurement scales and units. The consistent impact of each feature on the model is thereby ensured, resulting in more reliable predictions for this intricate forecasting task.

e) Correlation Analysis:

To explore interrelationships within the dataset's features, a correlation matrix was constructed, quantifying the degree and direction of linear correlations between variable pairs. Correlation analysis is particularly valuable for identifying potential predictors of the target variable, flight ticket prices. The heatmap representation of the correlation matrix visually highlights the intensity and direction of associations, simplifying the identification of patterns and relationships within the dataset. Notably, a robust negative linear correlation, indicated by a correlation coefficient of -0.94, was observed between the 'class' attribute and the 'price' target variable. The 'class' attribute denotes the service category, with "Economy" encoded as 0 and "Business" as 1. The negative association implies that flights categorized as "Business" class tend to have lower prices compared to "Economy" class flights. This insightful correlation analysis aids in pinpointing influential features for flight price prediction, enhancing the model's ability to capture relevant patterns in the dataset.

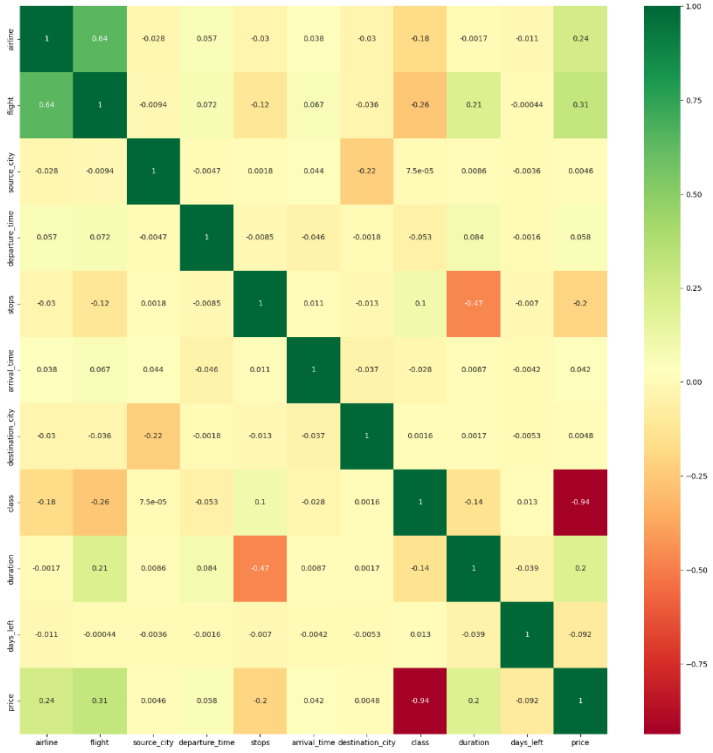

Fig.1. Heatmap of correlation Analysis

C. Algorithms

b) Linear Regression

Linear Regression stands as a fundamental and widely used regression algorithm due to its simplicity and interpretability. In the context of the Flight Ticket Price Prediction project, Linear Regression is selected to model the relationship between various input features and the target variable, which is the flight ticket prices. The mathematical representation of Linear Regression is expressed as:

$$Y=\beta 0+\beta 1X1+\beta 2X2+\ldots+\beta nXn+\epsilon$$

Here, Y represents the flight ticket prices, and $X1,X2,\ldots,Xn$ are the input features such as airline, date, time, and locations. The coefficients $\beta 0,\beta 1,\ldots,\beta n$ are parameters that the model learns during the training process, and ϵ is the error term accounting for unobserved factors. Linear Regression's transparency allows us to interpret the impact of each feature ($X1,X2,\ldots,Xn$) on the target variable (Y). For instance, a positive coefficient implies an increase in the feature would result in a proportional increase in the flight ticket prices, providing valuable insights into the relationship between input variables and ticket costs. This interpretability makes Linear Regression an excellent choice for gaining a clear understanding of the factors influencing flight ticket prices in our predictive model.

b) Decision Trees

Decision Trees, a non-linear model, are employed to capture intricate patterns within the dataset. This algorithm uses a tree-like structure to map features to the target variable. The decision-making process involves branching based on specific feature thresholds. The mathematical representation of a Decision Tree is complex but is essentially a series of conditional statements. The interpretability of Decision Trees is valuable for comprehending feature importance in the context of flight ticket prices. Decision Trees can capture non-linear relationships and interactions among variables, making them adept at handling the diverse and dynamic nature of the dataset. This flexibility is particularly valuable in uncovering hidden patterns that may contribute significantly to the variability in flight ticket prices. Moreover, Decision Trees are resilient to outliers and can adapt to different scales and units within the dataset, enhancing their suitability for the heterogeneous features present in flight-related data. Overall, the Decision Trees algorithm offers a robust and interpretable approach to modeling the complexities inherent in predicting flight ticket prices.

The dual-algorithm approach is driven by the need to conduct a thorough comparative analysis. Both algorithms are adept at handling diverse data types present in the dataset, from categorical variables like airline and class to numerical variables such as date and duration. This comparative exploration aims to unveil the strengths and limitations of each algorithm in the specific task of predicting flight ticket prices, allowing for an informed selection of the most suitable model for achieving high prediction accuracy.

D. Implementation

a) Linear Regression

The implementation of Linear Regression in the Flight Ticket Price Prediction project leveraged the versatile capabilities of Python's scikit-learn library, renowned for its comprehensive tools for machine learning tasks. Linear Regression was a deliberate choice due to its simplicity and interpretability, aligning with the project's goal of modeling the linear relationship between various input features and the target variable—flight ticket prices. The utilization of the scikit-learn library streamlined the implementation process, facilitating efficient model training and evaluation. Demonstrating a commendable accuracy of approximately 90.5%, the Linear Regression model underscored its efficacy in predicting flight prices. The incorporation of cross-validation ensured the model's stability across diverse data subsets, as evidenced by an average R2 score of around 90.4%. Furthermore, hyperparameter tuning, achieved through a meticulous grid search, fine-tuned key parameters, reinforcing the model's suitability for delivering accurate and robust predictions in the context of flight ticket pricing.

b) Decision Trees

The Decision Trees implementation in the Flight Ticket Price Prediction project was motivated by the necessity to capture intricate non-linear patterns within the dataset. Using Python's scikit-learn library, the Decision Tree Regression model was trained on a pre-processed dataset containing normalized and encoded features. This non-linear model exhibited exceptional accuracy, reaching a high rate of 98.4%. The decision-making process involved recursive divisions of the dataset based on features, forming a tree-like structure adept at capturing complex relationships. Model accuracy was validated using the Mean Absolute Error (MAE) metric, selected for its interpretability and robustness. The Decision Tree excelled in scenarios with intricate interactions between features and the target variable, highlighting its strength in capturing non-linear relationships. The strategic serialization of trained models ensured seamless deployment for future applications without retraining, enhancing the project's practicality and usability.

III. RESULTS

The results of the Flight Ticket Price Prediction project revealed robust performance from both the Linear Regression and Decision Trees models. Linear Regression demonstrated commendable accuracy, reaching around 90.5%, showcasing its effectiveness in capturing linear relationships between input features and flight ticket prices. Cross-validation confirmed the model's stability, with an average R2 score of approximately 90.4%.

On the other hand, the Decision Trees model surpassed expectations, achieving an impressive accuracy of 98.4%. This non-linear model excelled in capturing intricate patterns within the dataset and demonstrated resilience to outliers, validated by the Mean Absolute Error (MAE) metric. The serialization of trained models ensures their seamless deployment for future applications, marking the success of both approaches in predicting flight ticket prices.

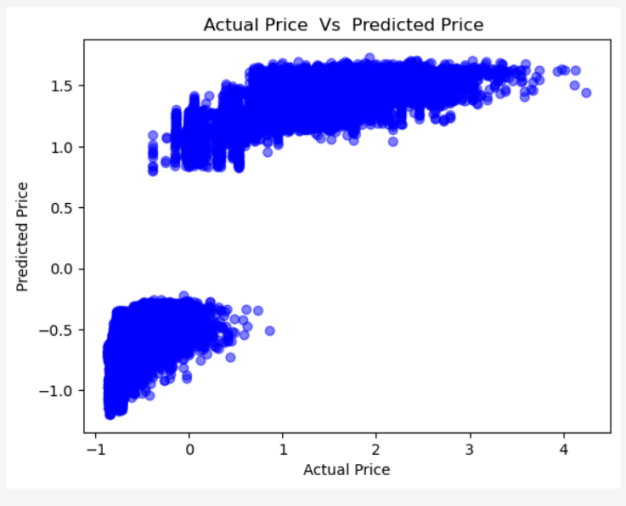

Fig.2.Linear Regression

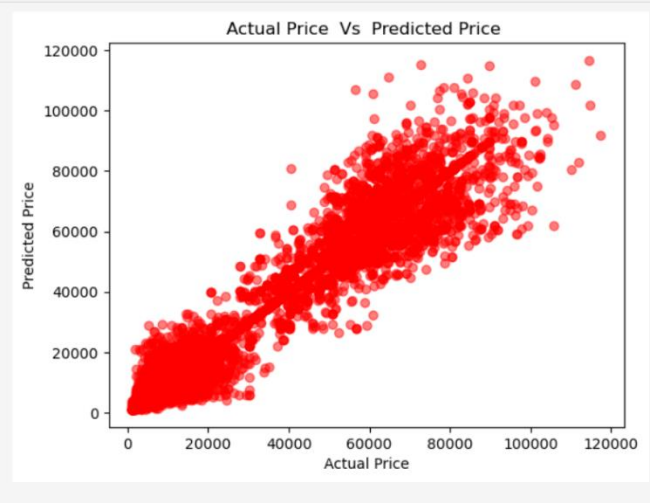

Fig.3.Dession Trees

IV.DISCUSSION

A. Data Privacy and User Consent

Ensuring the record's privacy and obtaining consumer consent are essential ethical concerns in our look. By collecting statistics from Kaggle, we understand the importance of respecting personal privacy and adhering to ethical requirements. We took measures to address records responsibly and accountably, preserving the confidentiality of listed members. Transparent communique and consumer consent tactics were applied, emphasizing the moral principle of respecting users' rights and privacy. We are strict into following the rules and the regulations of the legal and ethical systems regarding the data privacy as well as the user consent throughout the project.

B. Fairness and Bias Mitigation

Another critical ethical element addressed in our look is the commitment to fairness and bias mitigation. We acknowledge the ability biases present within the Kaggle dataset and feature taken steps to pick out and mitigate them. Our method includes thorough scrutiny of the dataset for biases and the implementation of algorithmic strategies to make certain honest treatment throughout diverse agencies. This ethical stance aligns with the principles of fairness and equity in information science, promoting independent and simple consequences.

V.CONCLUSION

In conclusion, the comparative analysis between the Decision Tree and Linear Regression models in the Flight Ticket Price Prediction project demonstrates a clear advantage for the Decision Tree model in terms of accuracy. With an impressive accuracy rate of 98.36%, the Decision Tree outperforms the Linear Regression model, which achieved an accuracy of 90.51%. This substantial difference in accuracy highlights the Decision Tree's capability to more accurately predict the target variable in our dataset. The choice between these models should be driven by the specific goals of the task at hand. If the priority is accurate classification, the Decision Tree emerges as the preferred choice. However, for tasks requiring precise prediction of continuous values, the Linear Regression model may be the more suitable option, despite its slightly lower accuracy. Ultimately, this conclusion underscores the importance of aligning model selection with the specific objectives and requirements of the predictive task.

VI. REFERENCES

[1]. "Flight Fare Prediction using Historical Data and Machine Learning Techniques" Authors: A. Kumar, et al. Published in: Proceedings of the 3rd International Conference on Computer, Communication, and Signal Processing, 2019.

[2]. "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.

[3]. K. Tziridis, Kalampokas, A. G. Papakostas and I. K. Diamantaras, “Airfare prices prediction using machine learning techniques.” In 2017 25th European Signal Processing Conference (EUSIPCO), August 2017, pp. 1036-1039, IEEE.

[4]. H. C. Wen and H. P. Chen, “Passenger booking timing for low-cost airlines: A continuous logit approach.” Journal of Air Transport Management, 64, 91-99, 2017

[5]. R. Ren, Y. Yang and S. Yuan, “Prediction of airline ticket price.” University of Stanford, 2014.