MSDS 422
Group 6: Nikhil, Elvis, Reese, Aaron

Machine Learning Strategies for Effective Flight Fare Prediction

Executive Summary:

This project aimed to predict flight fares for domestic flights in India by analyzing factors such as travel duration, airline, source city, departure time, number of stops, arrival time, destination city, and ticket class. Our exploratory data analysis revealed that ticket class and the number of stops were the most significant factors influencing flight prices. We employed Linear Regression, Random Forest, Lasso Regression, and SVM models, with Random Forest achieving the highest accuracy (R-squared score of 0.98). To further improve predictive accuracy, we recommend integrating third-party datasets such as weather data, economic indicators, and customer reviews. These additional variables can provide valuable context and enhance our models, enabling more informed decision-making for airlines and travelers alike.

Introduction:

Our project will discuss the flight fares for domestic flights in India. While exploring, the dataset we set out to address the complex dynamics surrounding airline pricing strategies and the factors influencing the cost of airline tickets. These factors include a multitude of variables such as duration, airline, flight, source_city, departure_time, stops, arrival_time, destination_city, and ticket class. By utilizing all of these variables and the vast data we had we hope to develop an accurate and precise model for forecasting future flight prices.

Literature Review:

The paper "Flight Price Prediction Using Machine Learning" by Ankita Panigrahi et al. investigates the use of machine learning algorithms to predict airline ticket prices which is what we set out to achieve as well. Employing methods such as Artificial Neural Networks, Linear Regression, Decision Trees, and Random Forest, the study found that Decision Trees provided the most accurate predictions. The dataset that was used in the study was sourced from Kaggle and included variables like airline, date, and travel duration. The study concluded that integrating evolutionary algorithms could further enhance prediction accuracy, suggesting potential improvements for future research in flight price forecasting. They came to this conclusion because they found that a huge discrepancy in price was due to when they received their information because flight prices fluctuate heavily on a day to day basis.

Exploratory Data Analysis:

We began the process by going through an exploratory data analysis phase where we charted our variables to see the effect that individual variables had on pricing. Through our charting, we revealed that the class of ticket and the number of stops were among the most significant factors influencing flight pricing. While the higher ticket prices for better ticket classes made sense, we observed a direct correlation between the number of stops and the duration of the flight. Therefore, flights with longer duration, travel distances, and stops tend to incur higher costs and were the reason for the higher priced tickets despite the inconveniences of stops during travel. From here we split the data into test and train datasets allowing us to implement modeling techniques to create accurate forecasts. After our findings, we will be able to inform people when the most cost-efficient time to purchase airline tickets is when they travel.

Data Preparation/Feature Engineering

The dataset was prepared by first handling any missing values and making sure each column had the correct data type. Categorical features like 'airline', 'source_city', 'departure_time', 'stops', 'arrival_time', 'destination_city', and 'class' were converted into a numerical format using one-hot encoding. Numerical features such as 'duration' and 'days_left' were standardized so they would have a mean of zero and a standard deviation of one, which helps the models perform better. The data was then split into training and testing sets in an 80-20 ratio to ensure the models were trained and tested properly. Outliers were identified and removed using Z-scores to make the models more robust. Different regression models, including Linear Regression, Random Forest Regressor, Support Vector Regression (SVR), and Lasso Regression, were used to predict outcomes, and their performances were measured using Mean Squared Error (MSE) and R^2 score. The transformations and scaling were done with the assumption that the data was normally distributed and consistent, while the testing helped verify that the models worked well on new, unseen data.

Methodology and Various Tools:

For our project we used Jupyter Notebook to be able to run our datasets. In addition we had to run the code through a MacBook Pro which has 11 core CPU and a 14 core GPU as the dataset that we used was large. The use of Anaconda, is able to provide a Python Environment to help streamline the development process. We used different models such as Linear Regression, Random Forest, Lasso Regression, and SVM models.

Findings and Conclusions:

Utilizing a Linear Regression model, we were able to predict flight costs based on features such as travel duration, airline, departure and arrival cities, among others. This model provides valuable insights for businesses like United, Southwest, and Delta to make informed pricing decisions. For performance evaluation, we employed metrics like Mean Squared Error (MSE) and R-squared score. The Linear Regression model achieved an MSE of 6608 and an R-squared score of 0.91, indicating that 91% of the variance in the dataset was explained by this model, demonstrating a strong fit.

The Random Forest model further improved prediction accuracy by reducing noise and handling outliers effectively, avoiding overfitting. This model achieved an impressive R-squared score of 0.98, indicating an excellent fit. The SVM model, while computationally intensive, effectively managed high-dimensional feature spaces and non-linear relationships, resulting in an R-squared score of 0.89. Although slightly lower, this still represents a strong model. Lastly, the Lasso Regression model helped in identifying the most significant features for predicting costs. With an R-squared score of 0.90, it also demonstrated a good fit for the data. Each of these models provided unique strengths, contributing to a comprehensive understanding of flight fare prediction.

Lessons Learned and Recommendations:

Through our analysis of flight fare predictions, we gained critical insights into the importance of feature selection and the handling of outliers and noise. Variables such as ticket class and the number of stops emerged as significant predictors of flight prices, underscoring the need for careful feature engineering. The Random Forest model's ability to reduce noise and prevent overfitting was particularly valuable, demonstrating the necessity of robust models in achieving high predictive accuracy. Comparing different models, we found that while Random Forest achieved the highest R-squared score, each model had its strengths and was useful for different aspects of the analysis.

Moving forward, we recommend enhancing our feature set by incorporating additional variables such as weather conditions, economic indicators, and customer reviews. For instance, integrating weather data can provide insights into how adverse weather conditions, like monsoons, impact travel demand and pricing strategies. Economic indicators such as GDP and inflation rates can also influence travel behaviors, while customer reviews can offer sentiment analysis to gauge airline reputation and its effect on pricing. By combining these third-party datasets with our original dataset, we can develop more accurate predictive models. Implementing a real-time prediction system and developing a user-friendly interface for stakeholders will ensure that our model remains relevant and practical. Finally, regular updates and monitoring will help maintain the accuracy and reliability of our predictions, enabling more informed decision-making for airlines and travelers alike.

References:

https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction

https://ceur-ws.org/Vol-3283/Paper90.pdf