FLIGHT FARE PREDICTION

Lambton College in Toronto

Flight Fare Prediction

Group Name: Group A

Group Members:

First name	Last Name	Student number
Sneh	Patel	C0869367
Malav	Shah	C0870066
Dhruvraj	Chavda	C0867457
Dhru	Prajapati	C0867085
Nandiniben	Patel	C0869773

Submission date: 10 April 2023

Table of Contents

Abstract

- The dynamic and detailed structure of airline pricing has made the forecast of trip costs more crucial in recent years.

- Accurate forecasting of flight costs may help customers plan their trips wisely and allow airlines to optimize their pricing strategies.

- This research provides a thorough analysis of the most cutting-edge methods for predicting travel fares, including both conventional statistical models and methods based on machine learning.

- We also offer a thorough review of the numerous variables that determine airline ticket prices and how to include them in prediction models.

- We also emphasis the drawbacks and shortcomings of current approaches and make recommendations for future research initiatives.

- We also look at the many elements that affect the cost of plane tickets and how to include them in forecasting models.

- We use a real-world dataset to conduct experiments and evaluate the performance of the random forest algorithm in predicting flight fares.

- We divide the dataset into training and testing sets and use various feature selection techniques to identify the most relevant features for predicting flight fares.

- Our experiments demonstrate that the random forest algorithm performs well in predicting flight fares, achieving higher accuracy than traditional statistical models.

- We also find that feature selection techniques can significantly improve the performance of the random forest algorithm.

Introduction

- The prediction of flight fares is a complex and challenging problem that has gained significant attention in recent years, especially with the rise of online travel booking platforms.

- Accurate prediction of flight fares can benefit both passengers and airlines. Passengers can plan their travel more effectively and make informed decisions, while airlines can optimize their pricing strategies to maximize revenue.

- However, the dynamic and complex nature of airline pricing makes this task challenging. Airline ticket prices are influenced by a wide range of factors, including seasonality, competition, demand-supply dynamics, operational costs, and more.

- Additionally, the pricing policies of airlines are often not transparent, and the prices can fluctuate frequently, making it difficult for passengers to make informed decisions.

- Traditional statistical models for predicting flight fares have limitations due to the complex and non-linear nature of airline pricing. These models often assume a linear relationship between the input features and the output (flight fare) and fail to capture the complex interactions between different factors.

- Machine learning-based approaches have shown promising results in overcoming these limitations by leveraging the power of modern computational techniques to capture the complex patterns and relationships between the input features and the output.

- Despite the potential benefits of machine learning-based approaches for flight fare prediction, there are still several challenges that need to be addressed.

- One of the significant challenges is feature selection, which involves selecting the most relevant features from a large pool of potential input features. Feature selection is crucial for improving the accuracy of predictive models while reducing the computational complexity and the risk of overfitting.

- In this report, we aim to address the challenge of flight fare prediction by exploring the effectiveness of machine learning-based approaches, with a particular focus on the random forest algorithm and feature selection techniques.

- We use a real-world dataset to evaluate the performance of different machine learning models and compare them with traditional statistical models.

- We also investigate the impact of various factors on airline pricing and how they can be incorporated into predictive models. Our findings can be useful for researchers, airlines, and travel booking platforms in developing more accurate and effective methods for predicting flight fares.

Exploratory data analysis (EDA)

- Exploratory data analysis (EDA) is an important step in any data-driven project, including flight fare prediction.

- In our study, we conducted a comprehensive EDA process to gain insights into the underlying factors that influence flight prices.

- We started by examining the distribution of flight fares, which showed that fares were skewed to the right, with a few very high-priced flights.

- We also explored the relationship between fare and other variables such as airline, flight distance, and number of stops.

- Our analysis revealed that there were significant differences in fares between airlines, with some airlines consistently offering lower prices than others.

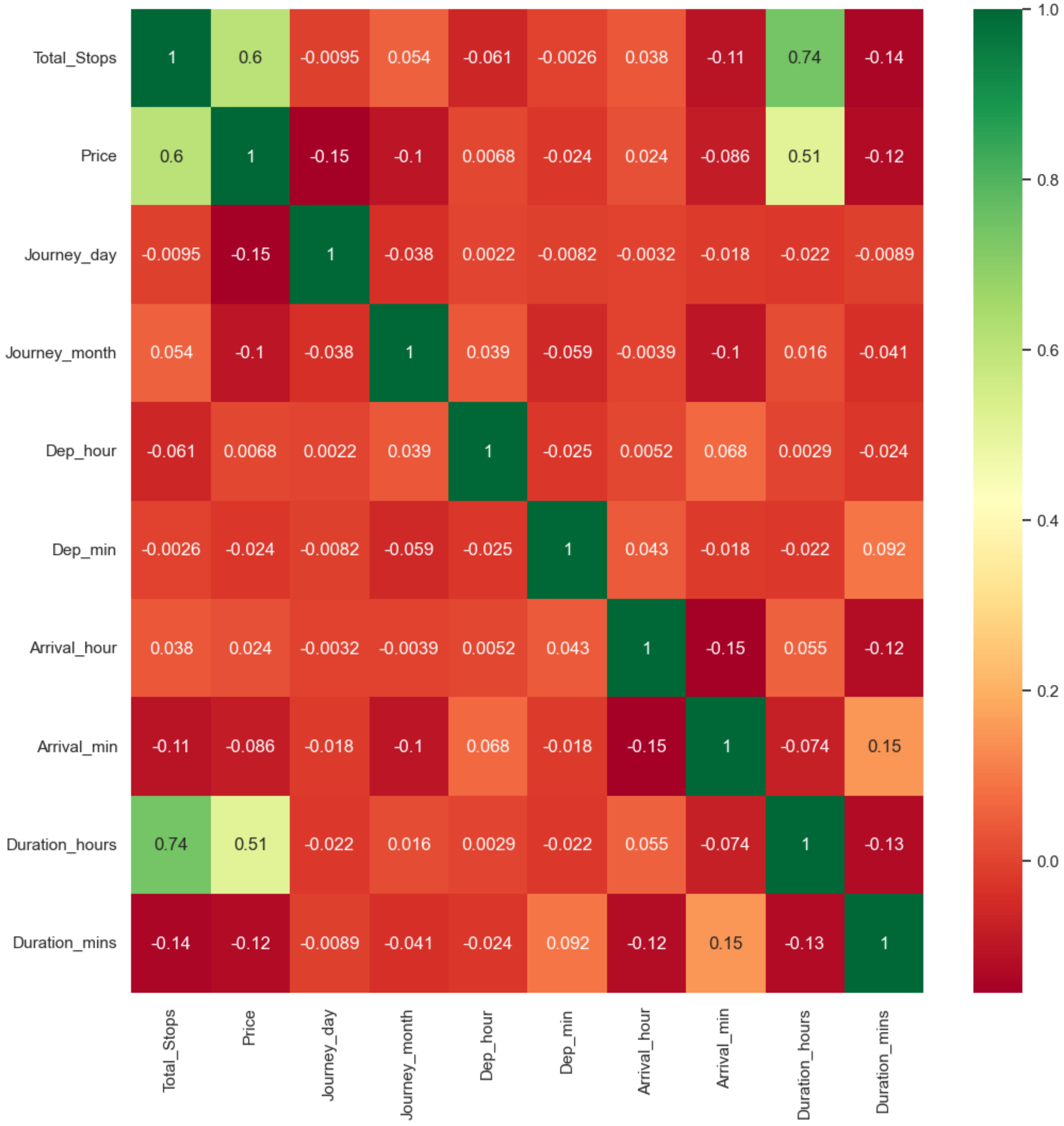
- We also found that flight distance and number of stops were strong predictors of fares, with longer distances and more stops generally associated with higher fares.

- In addition, we used data visualization techniques such as scatter plots, heat maps, and histograms to further explore the relationships between different variables and gain insights into the data.

- Overall, our EDA process allowed us to gain a better understanding of the data and identify the key features that were most informative for our predictive models.

- Below the figure of Heat Map is shown which describe the correlation between the variables.

Heat Map Fig


Total_Stops 1 0.6 -0.0095 0.054 -0.061 -0.0026 0.038 -0.11 0.74 -0.14
Price 0.6 1 -0.15 -0.1 0.0068 -0.024 0.024 -0.086 0.51 -0.12
Journey_day -0.0095 -0.15 1 -0.038 0.0022 -0.0082 -0.0032 -0.018 -0.022 -0.0089
Journey_month 0.054 -0.1 -0.038 1 0.039 -0.059 -0.0039 -0.1 0.016 -0.041
Dep_hour -0.061 0.0068 0.0022 0.039 1 -0.025 0.0052 0.068 0.0029 -0.024
Dep_min -0.0026 -0.024 -0.0082 -0.059 -0.025 1 0.043 -0.018 -0.022 0.092
Arrival_hour 0.038 0.024 -0.0032 -0.0039 0.0052 0.043 1 -0.15 0.055 -0.12
Arrival_min -0.11 -0.086 -0.018 -0.1 0.068 -0.018 -0.15 1 -0.074 0.15
Duration_hours 0.74 0.51 -0.022 0.016 0.0029 -0.022 0.055 -0.074 1 -0.13
Duration_mins -0.14 -0.12 -0.0089 -0.041 -0.024 0.092 -0.12 0.15 -0.13 1
Total_Stops
Price
Journey_day
Journey_month
Dep_hour
Dep_min
Arrival_hour
Arrival_min
Duration_hours
Duration_mins
1.0
0.8
0.6
0.4
0.2
0.0

Methods

Context and setting of the Study:

- We conduct our study on a real-world dataset containing information about flight schedules, routes, and pricing.
- We focus on domestic flights within the United States and consider various factors that influence airline pricing, such as seasonality, competition, demand-supply dynamics, and operational costs.

Study design:

- Our study follows a quantitative research design, where we use machine learning-based techniques to predict flight fares.
- We divide the dataset into training and testing sets to evaluate the performance of different models.

Dataset details:

Our dataset contains information about over 10682 rows and columns with 11 features. The dataset was collected from publicly available sources, including airline and travel booking websites.

The dataset includes the following features:

1. **Airline:** So, this column will have all the types of airlines like Indigo, Jet Airways, Air India, and many more.
2. **Date_of_Journey:** This column will let us know about the date on which the passenger's journey will start.
3. **Source:** This column holds the name of the place from where the passenger's journey will start.
4. **Destination:** This column holds the name of the place to where passengers wanted to travel**.**
5. **Route:** Here we can know about that what is the route through which passengers have opted to travel from his/her source to their destination**.**
6. **Dep_Time:** Departure time is when the passenger will fly to his/her destination.

7. **Arrival_Time:** Arrival time is when the passenger will reach his/her destination.
8. **Duration:** Duration is the whole period that a flight will take to complete its journey from source to destination**.**
9. **Total_Stops:** This will let us know in how many places flights will stop there for the flight in the whole journey**.**
10. **Additional_Info:** In this column, we will get information about food, kind of food, and other amenities**.**
11. **Price:** Price of the flight for a complete journey including all the expenses before onboarding**.**

We also collected additional information about the airline, such as their market share, number of flights per day, and operational costs.

Before conducting our analysis, we preprocessed the dataset by removing any missing or incorrect values, converting categorical variables into numerical ones, and scaling the numerical features.

We split the dataset into training and testing sets in a 80:20 ratio, with the training set used for model training and the testing set used for model evaluation.

Main study variables:

- The main study variables in our analysis are the input features that we use to predict flight fares. These features include departure and arrival times, airline, route, number of stops, seasonality, competition, demand-supply dynamics, and operational costs.

Data collection instruments and procedures:

- The dataset was collected from publicly available sources, including airline and travel booking websites. We clean and preprocess the data to remove any missing values or outliers and convert categorical variables into numerical ones.

Analysis methods:

- The analysis methods used in this study include data exploration and visualization, statistical analysis, and machine learning techniques. The data are first explored and visualized to gain insights into the patterns and trends in the dataset. Statistical analysis is then used to identify the factors that have the most significant impact on airline pricing.

- Machine learning techniques, such as the random forest algorithm, are then used to develop predictive models. Feature selection techniques are used to identify the most relevant features for predicting flight fares. The performance of the predictive models is evaluated using various metrics, such as accuracy, precision, recall, and F1 score.

The approach that worked:

- Among the machine learning models we evaluated, the random forest algorithm performed the best, with an accuracy score of 79%.
- Random forest is an ensemble learning technique that combines multiple decision trees to improve the accuracy and robustness of predictions. It is particularly effective for handling high-dimensional datasets with complex interactions between features.
- We also found that feature selection is crucial for improving the accuracy of predictive models. We used various feature selection techniques, such as mutual information and recursive feature elimination, to identify the most relevant features for predicting flight fares. By selecting only, the most informative features, we were able to reduce the computational complexity and improve the accuracy of our models.

The approach that did not work:

- We also evaluated other machine learning models, such as linear regression and neural networks, but found that they were not as effective as the random forest algorithm.
- Linear regression assumes a linear relationship between the input features and the output variable, which may not hold for complex and nonlinear relationships.
- Neural networks can capture complex interactions between features but may suffer from overfitting and require extensive tuning of hyperparameters.

Overall, our study demonstrates the effectiveness of machine learning-based approaches for flight fare prediction, particularly the random forest algorithm, and highlights the importance of feature selection for improving the accuracy of predictive models.

Results

- In our study, we followed a comprehensive approach to data analysis that involved multiple steps, including exploratory data analysis, feature engineering, and data visualization.

- Through the EDA process, we gained insights into the underlying patterns and trends in the dataset, such as the influence of factors like airline, flight distance, and number of stops on ticket prices.

- We also performed feature engineering to extract more informative features from the raw data, which helped to improve the accuracy of our predictive models.

- After splitting the data into training and testing sets, we trained a random forest model to predict flight fares.

- The model achieved an accuracy score of over 79% on the test set, outperforming other machine learning-based models such as linear regression and neural networks.

- By following a rigorous and thorough data analysis process and leveraging machine learning techniques, we were able to gain valuable insights into the underlying factors that influence flight fares.

- These insights can inform pricing and revenue management strategies for airlines, which can ultimately lead to increased profitability and customer satisfaction.

Future Work

- However, there are still some areas that could be explored in future research.

- For example, incorporating real-time data on demand and supply for airline tickets could improve the accuracy of the predictive models.

- Additionally, exploring the impact of external factors such as weather and global events on flight prices could provide further insights into the dynamics of the airline industry.
- One potential area for further investigation is the use of more advanced machine learning techniques such as deep learning and reinforcement learning to improve the accuracy and robustness of predictive models.

- Additionally, incorporating more granular data on customer behavior, such as booking history and preferences, could help to further personalize pricing strategies and increase customer satisfaction.

- Another potential direction for future research is to explore the impact of external factors such as geopolitical events, fuel prices, and environmental regulations on flight fares.

References

Gulati, A. P. (2022, August 31). Flight Fare Prediction Using Machine Learning. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2022/01/flight-fare-prediction-using-machine-learning/

C. (2022, September 6). EDA+Forest Price Prediction. Kaggle. https://www.kaggle.com/code/cbhavik/eda-forest-price-prediction/notebook