CS – 677 (A3)

FINAL PROJECT – DATSCIENCE WITH PYTHON

Project Title – Flight Price Prediction

Presented by – Pratyush Patel (U54611283)

**INTRODUCTION**

The Flight Price Prediction system employs machine learning algorithms and historical flight data to forecast future airfares accurately. By collecting and processing extensive data, the system identifies key features crucial for predicting flight prices. A sophisticated machine learning model is developed and optimized, incorporating real-time data integration to adapt to market fluctuations. The user-friendly interface allows travelers to input details and receive precise predictions, complemented by visualizations and historical trends. Ongoing accuracy assessments and a feedback loop ensure continuous improvement. The system's scalability and deployment on robust infrastructure guarantee reliability, making it a transformative tool for informed travel planning.

**PROJECT GOAL**

The primary objective of this project is to craft a model capable of precisely estimating airline ticket prices, thereby empowering travelers with the information they need to make well-informed decisions and optimize their trip planning. Through a comprehensive analysis of diverse factors, including geographical locations, airline carriers, and departure times, the goal is to create a robust and accurate tool for predicting fluctuations in ticket prices. The model's focus on key parameters ensures a nuanced understanding of the dynamic nature of airfares. By leveraging advanced machine learning techniques, the system aims to unravel complex patterns in historical data, allowing users to anticipate and adapt to changes in the cost of air travel. The project's ultimate aim is to provide a reliable and user-friendly solution that enhances transparency in the often-volatile realm of airline ticket pricing, granting travelers greater control over their travel budgets and preferences.

**Dataset Description**

Flight Dataset Description:

1. Airlines: This column records the names of the airline carriers providing the flights, encompassing a range of international and domestic companies.

2. Departure Time: Denotes the scheduled departure time of the flight, indicating the hour and minute of takeoff.

3. Arrival Time: Represents the anticipated time of arrival at the destination, specifying both the hour and minute of landing.

4. Departure City: Indicates the city from which the flight originates, serving as the starting point of the journey.

5. Source City: Identifies the city from which the flight departs, providing additional information on the departure location.

6. Stops: This column enumerates the number of stops or layovers during the flight, categorizing it as a direct, one-stop, or multi-stop journey.

7. Days Left: Reflects the number of days remaining until the scheduled departure date, assisting travelers in planning and decision-making.

8. Price: Records the cost of the airline ticket, allowing users to evaluate and compare pricing options for their desired flights.

9. Flight Duration: Specifies the total time taken for the flight to reach its destination, encompassing both airborne and layover durations.

10. Class: Differentiates between travel classes, such as Economy, Business, or First Class, providing insights into the available amenities and comfort levels.

This comprehensive dataset combines essential information about airlines, flight timings, cities involved, travel parameters, and pricing details. It serves as a valuable resource for developing predictive models and conducting analyses to enhance the understanding of air travel patterns and pricing dynamics.

**DATASET VISUALISATION:**

1. Plot for majorly used airlines by passengers.

A pie chart with different colored circles

Description automatically generated

1. Plot for major flights taken from which source city

A graph of blue bars with names

Description automatically generated

1. Plot for major flights taken to which destination city

A graph of blue bars with names

Description automatically generated

**FACTORS AFFECTING THE PRICES OF THE FLIGHT**

The prices of flights are influenced by a multitude of factors that collectively shape the dynamics of the airline industry. Firstly, the demand for a particular route plays a pivotal role; during peak travel seasons or high-demand periods, ticket prices tend to escalate. Additionally, the availability of seats and the airline's pricing strategy, which may include dynamic pricing algorithms, impact fare fluctuations. The type of aircraft and its associated operating costs, such as fuel and maintenance, also contribute to ticket pricing variations. Moreover, the inclusion of amenities and the travel class chosen by passengers can significantly affect overall costs. Market competition among airlines servicing the same routes, economic conditions, and geopolitical events further contribute to the volatility of airfare. Seasonal trends, regulatory fees, and external factors like natural disasters or global health crises add additional layers of complexity to the pricing ecosystem. As a result, the intricate interplay of these diverse factors collectively determines the final price of a flight ticket.

1. Dot plot between duration and price

A blue dots on a white background

Description automatically generated

1. Plot between departure time and price

A graph of a line graph

Description automatically generated with medium confidence

1. Plot between days left for the trip and price

A graph of blue dots

Description automatically generated with medium confidence

**BOX PLOTS OF VARIOUS FACTORS AFFECTING THE PRICE**

Box plots are powerful visual tools that effectively illustrate the distribution and central tendency of various factors influencing flight prices. When examining departure times, box plots can reveal whether certain hours or time intervals correlate with higher or lower ticket prices, helping travelers optimize their schedules. The number of stops, a critical factor in flight selection, can be visualized through box plots to identify how layovers impact pricing. Seasonal variations, represented by box plots, offer insights into when prices tend to peak or dip due to demand fluctuations. Class-based box plots showcase the distribution of prices across different travel classes, aiding passengers in making cost-conscious decisions. Additionally, box plots depicting the days left until departure can unveil patterns in pricing dynamics, guiding travelers on optimal booking times. The integration of these visualizations provides a comprehensive understanding of the multifaceted factors affecting flight prices, empowering both researchers and consumers to navigate the intricate landscape of airfare variations.

A diagram of a graph

Description automatically generated with medium confidence

**DATA TRANSFORMATION**

Data transformation from categorical to numerical and scalable formats is a crucial preprocessing step in preparing datasets for machine learning models. This process involves converting categorical variables into numerical representations that can be effectively utilized in algorithms. One common technique is one-hot encoding, which assigns binary values to each category, creating a set of new binary columns. Additionally, ordinal encoding can be employed for variables with intrinsic order, assigning numerical values based on their ordinal relationships. These transformations ensure that the model can interpret and analyze categorical features, facilitating more accurate predictions. Scaling is another vital aspect, as it standardizes numerical variables to a consistent range, preventing certain features from dominating others. Techniques like Min-Max scaling or Z-score normalization transform the data to a comparable scale, enhancing model performance and preventing biases. Overall, these data transformations are essential for harnessing the predictive power of machine learning models on diverse datasets.

FROM THIS

A table with text on it

Description automatically generated

TO THIS

A table with numbers and lines

Description automatically generated

**MODELS USED FOR PREDICTING THE PRICES**

The models chosen for predicting flight prices cover a diverse range of machine learning algorithms, each with its strengths and applications:

1. \*\*Logistic Regression:\*\*

- Logistic Regression is typically used for binary classification tasks, but it can be adapted for regression by predicting probabilities.

- It's suitable for scenarios where the relationship between the features and the target variable is approximately linear.

2. \*\*K Nearest Neighbors (KNN):\*\*

- KNN is a versatile algorithm used for both classification and regression tasks.

- It predicts the target variable by considering the 'k' nearest data points in the feature space.

- KNN is effective when there are local patterns or clusters in the data that can influence flight prices.

3. \*\*Decision Tree:\*\*

- Decision Trees are powerful for both classification and regression tasks, providing interpretable decision rules.

- They can capture non-linear relationships and interactions between features, making them well-suited for complex datasets.

4. \*\*Random Forest:\*\*

- Random Forest is an ensemble method that builds multiple decision trees and combines their predictions.

- It improves upon the limitations of a single decision tree by reducing overfitting and enhancing generalization.

- Random Forest is particularly useful for handling large datasets with many features.

These models collectively contribute to a robust predictive system for flight prices, each bringing unique advantages to the table. The ensemble approach of Random Forest, the simplicity of Logistic Regression, the flexibility of K Nearest Neighbors, and the interpretability of Decision Trees allow for a comprehensive analysis of the intricate factors affecting flight pricing. The selection of these models provides a balance between accuracy, interpretability, and computational efficiency in predicting flight prices.

A line graph with red dots

Description automatically generated

Optimal value of k is 2.

A screenshot of a graph

Description automatically generated

**Conclusion**

* Among the models evaluated, the Random Forest model (Model 4) outperforms the others, demonstrating the highest accuracy and the lowest error metrics.
* The KNN model (Model 2) also performs well, providing high accuracy and low error metrics.
* The Linear Regression model (Model 1) and Decision Tree model (Model 3) show reasonable performance but with some limitations, especially in terms of accuracy.
* Depending on your specific requirements and trade-offs, I can choose the Random Forest or KNN model for flight price prediction due to their overall better performance in this evaluation.
* Overall, this project provides valuable insights into the applicability of different machine learning models for flight price prediction, offering a foundation for further refinement and optimization in future implementations.