AIRLINE TICKET PRICE PREDICTION SYSTEM

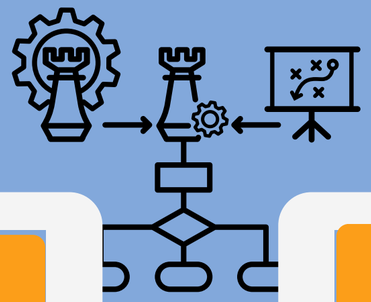
a DSCI Project

Adit Ghorpade - 612210054
Srushti Deshmukh - 642302007

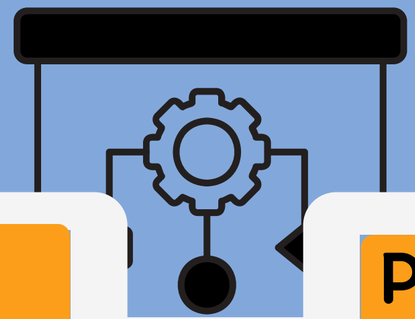
Summary

Introduction

Pre-processing



Model used



Dataset

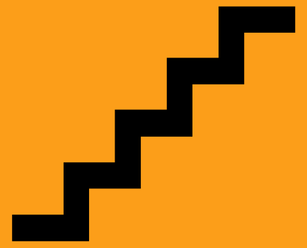
Data Analysis

Performace Evaluation

Introduction

- Airline prices vary due to factors like flight duration, no. of stops, airline, destination and date of travel.
- This project- A machine learning-based system, analyzes previous flight data and builds a predictive model to estimate airline ticket prices based on various flight parameters

Key Steps



Data Pre-processing

Handling missing values, encoding categorical data, and detecting outliers.

Feature Engineering

Extracting and analyzing important factors affecting price.

Exploratory Data Analysis

Handling missing values, encoding categorical data, and detecting outliers.

Model Training

Using RandomForestRegressor to train on flight data.

Evaluation & Prediction

Testing on unseen data and measuring accuracy with R^2 score.

Data Set Description

- Total Records: 10,683
- Total Columns: 11

Airline | Date of Journey | Source | Destination

Route | Dep_Time | Arrival_Time

Duration | Total_Stops

Additional_info

PRICE

Data Set Features

Date / Time

- Date_of_Journey
- Dep_Time
- Arrival_Time

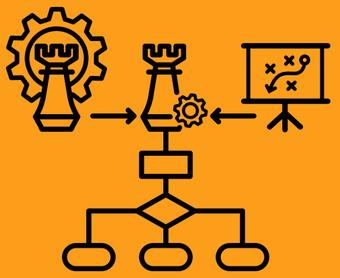
Categorical

- Airline
- Source
- Destination
- Total_Stops
- Additional_Info
- Routes

Numerical

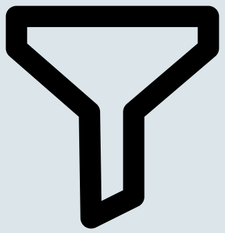
- Duration
- Price
  (Target Variable)

Pre-processing



Handling missing values

Remove or fill missing data using techniques like median

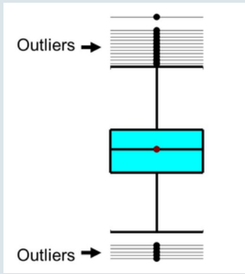


Extracting Useful Features

Transform raw data into meaningful insights

Encoding Methods

One-Hot Encoding
Target Mean Encoding
Label Encoding

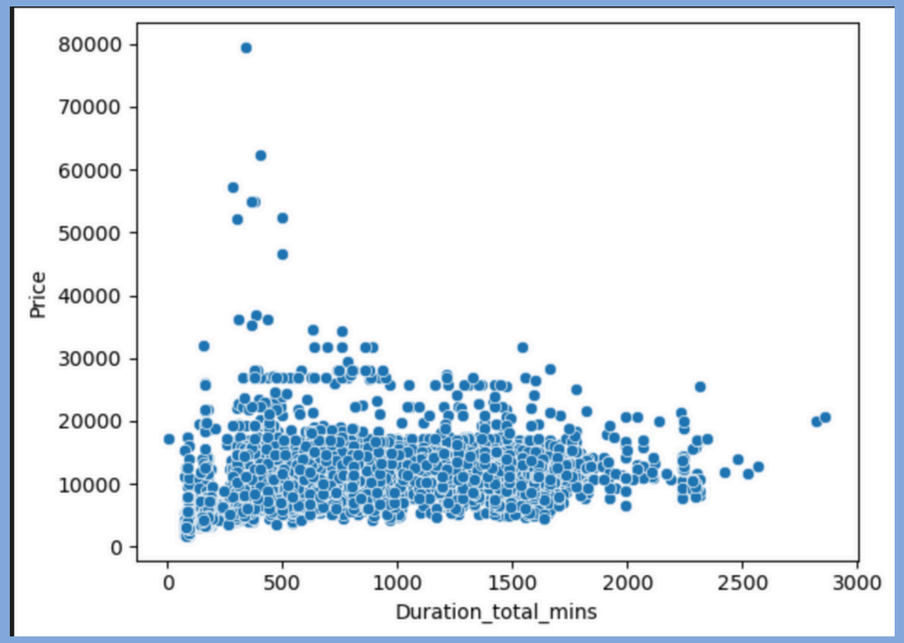


Outlier Detection

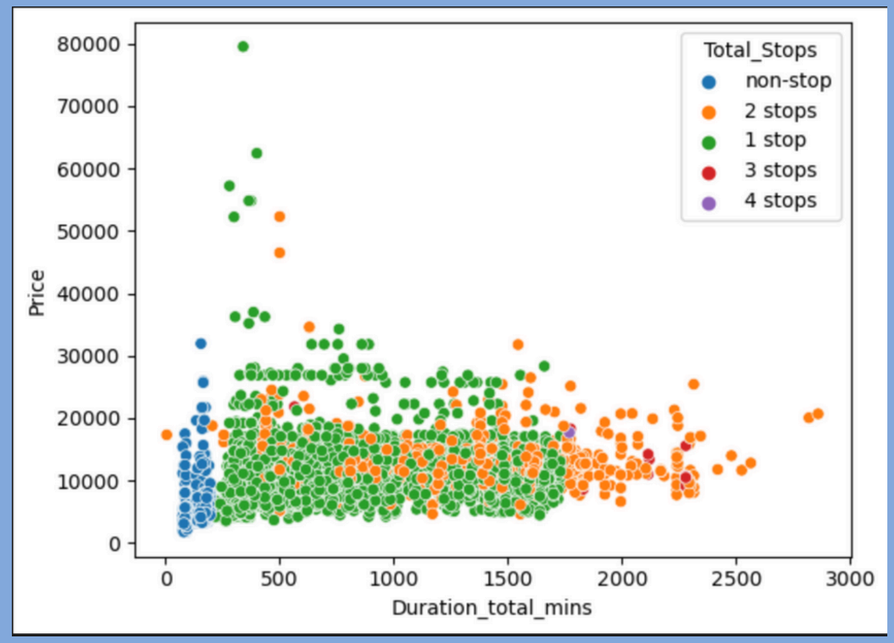
Identifying and handling extreme values using IQR

Exploratory Data Analysis

Scatter Plots- Used to analyze the relationship between two numerical variables.



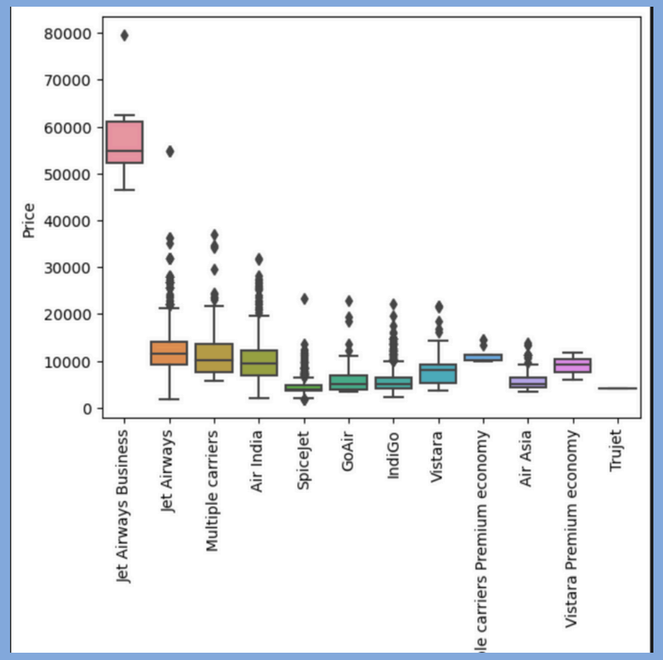
Duration vs Price



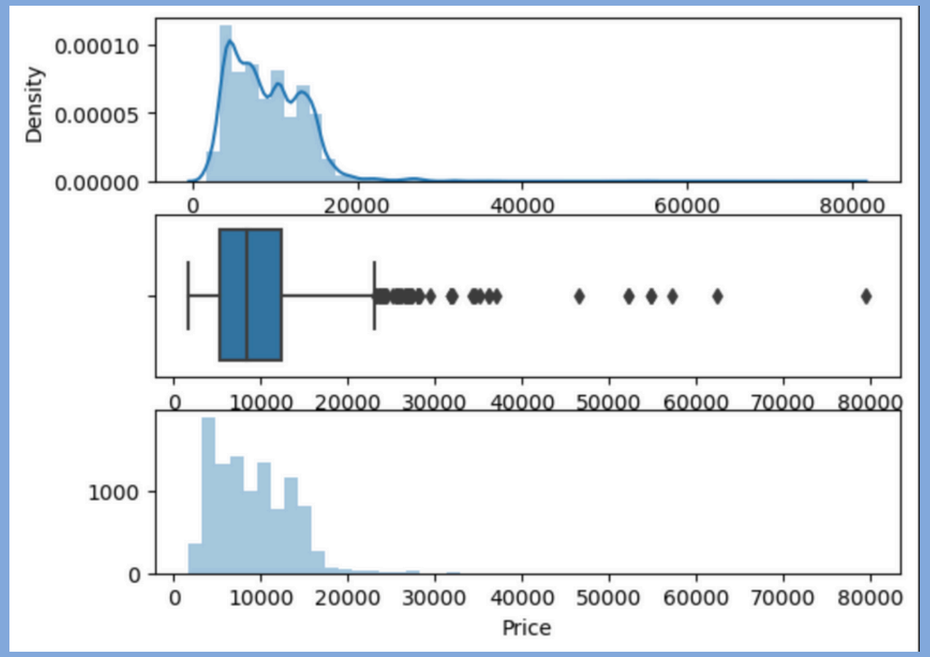
Duration vs Price vs Total_Stops

Exploratory Data Analysis

Box Plot (Whisker Plot)- Shows the distribution of data and detects outliers.



Airline vs Price



Outliers in Price

Model Used

RandomForestRegressor Model-

- A powerful ensemble learning algorithm based on multiple decision trees.
- Reduces overfitting by averaging predictions from multiple trees.
- Handles non-linear relationships between features and the target variable.
- Works well with large datasets and high-dimensional data.

📌 **How It Works?**

1. Creates multiple decision trees on different subsets of data.
2. Averages the predictions from all trees to improve accuracy.

Performance Evaluation

- **R^2 Score**- Measures how well the model explains variance in the target variable.
- **Mean Absolute Error (MAE)**- Average of absolute differences between actual and predicted values.
- **Mean Squared Error (MSE)**- Penalizes larger errors more than smaller ones.
- **Root Mean Squared Error (RMSE)**- Square root of MSE, making it more interpretable.
- **Mean Absolute Percentage Error (MAPE)**- Expresses error as a percentage of actual values.

Visualization

- Residual Plot (sns.distplot) helps check error distribution.

Conclusion

- System accurately estimates ticket prices using machine learning.
- Project involved data preprocessing, feature extraction, EDA, model training, and evaluation.
- RandomForestRegressor was used to learn from previous flight data and predict prices.
- Performance validated using R^2 score, MAE, RMSE, ensuring reliable predictions.
- Key factors affecting prices: Number of stops, airline, destination.
- Therefore, helps travelers plan trips efficiently & assists airlines in optimizing pricing.