Flight Price Prediction

Group 4

INTRODUCTION

AIM

This project aims to analyze a flight booking dataset. The primary goal is to understand price fluctuations and identify key factors influencing these variations.

DESCRIPTION

The key attributes in the 300153 rows × 12 columns dataset are:

- ID: Row number
- Airline: The airline operating the flight.
- Flight: The flight number.
- Source City: Departure city/airport.
- Departure Time: Departure time of the flight.
- Arrival Time: Time of arrival at the destination.
- Stops: Number of stops during the flight.
- Destination City: Destination city/airport.
- Flight Class: Economy/business class.
- Duration: Duration of the flight (hrs)
- Days left: Days left until departure during the time of booking.
- Price: The price of the flight ticket (target variable).

QUESTIONS

- Does the price vary with the airline?
- How is the price affected when tickets are bought a few days before departure?
- Does the ticket price change based on the departure time and arrival time?
- How does the price change with changes in Source City and Destination City?
- How does the ticket price vary between Economy and Business class?

DATA PREPROCESSING

01.

Dropped unnecessary variables like ID and flight. Renamed class to flight_class for clarity.

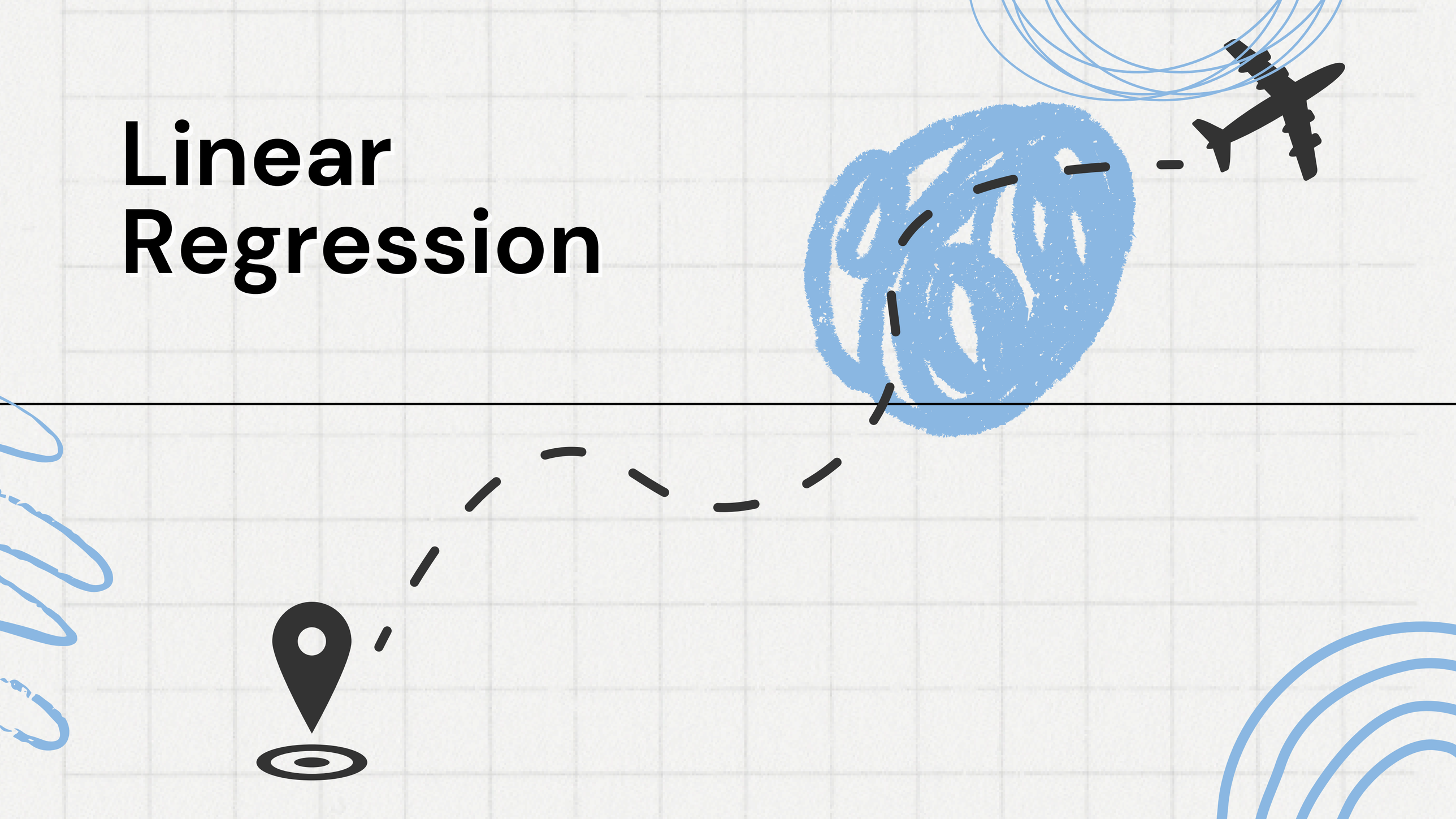
02.

Label encoding was used to replace string values with an integer for categorical variables for KNN and Decision Tree models.

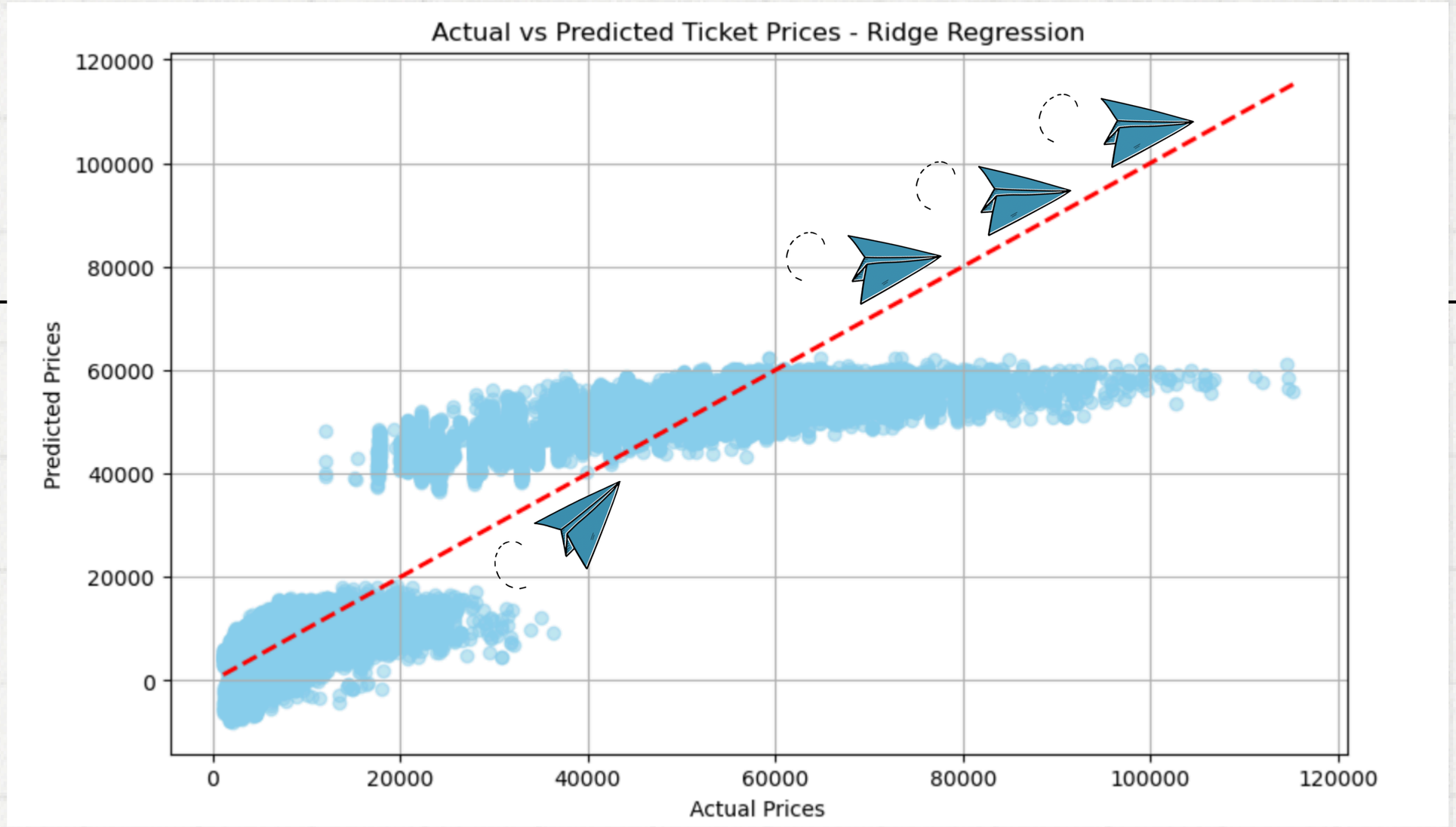
03.

Used one-hot encoding for categorial variables for linear regression.

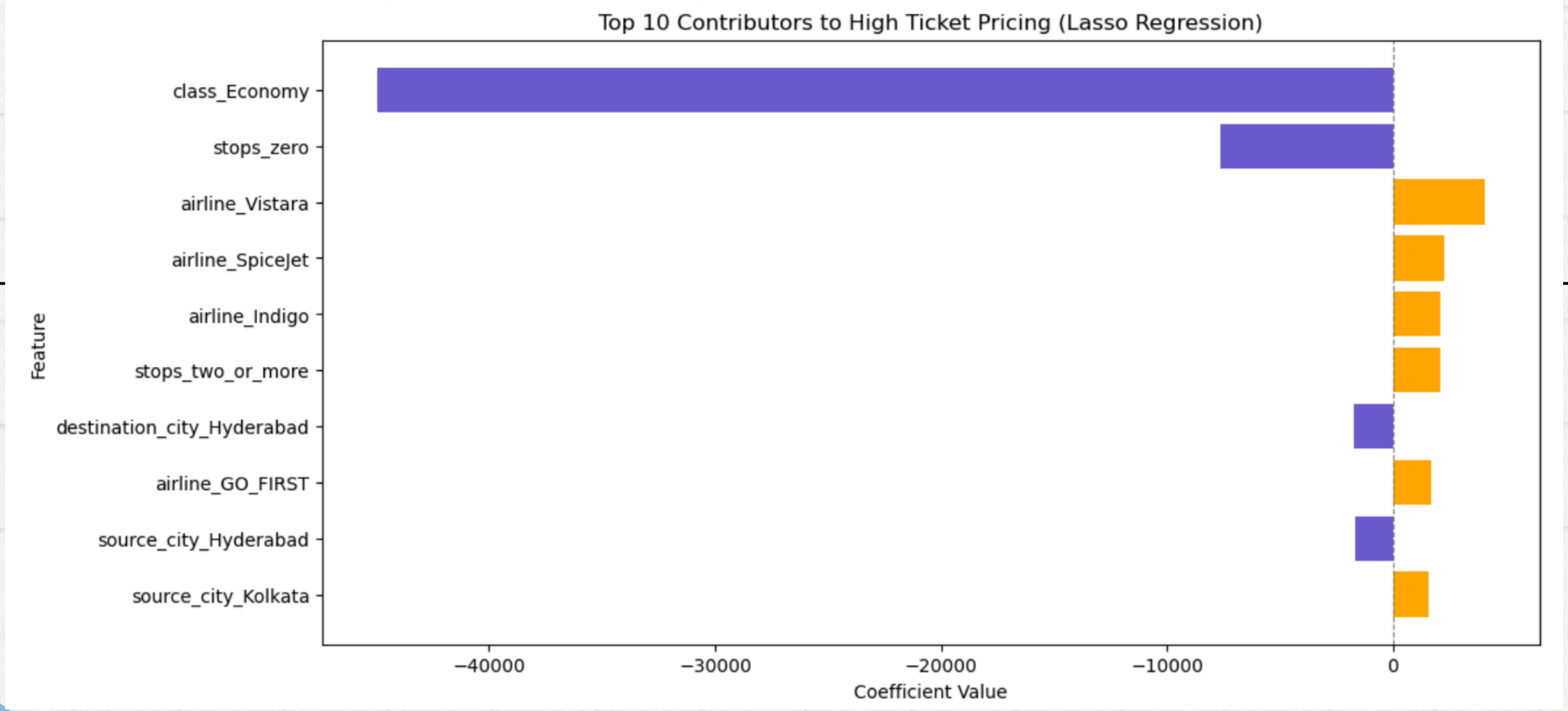
Linear Regression



Actual vs Predicted



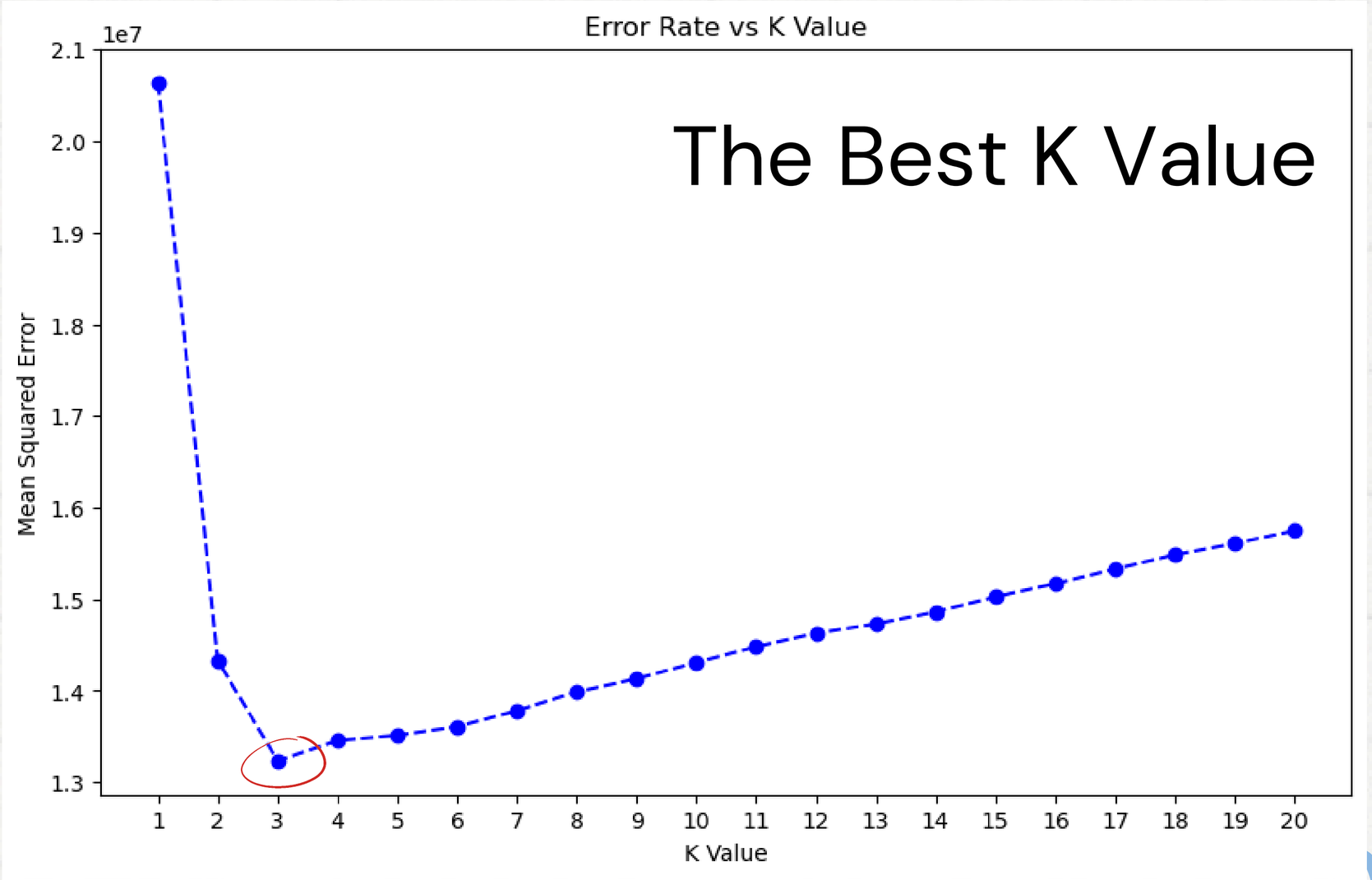
Contributors to high ticket prices

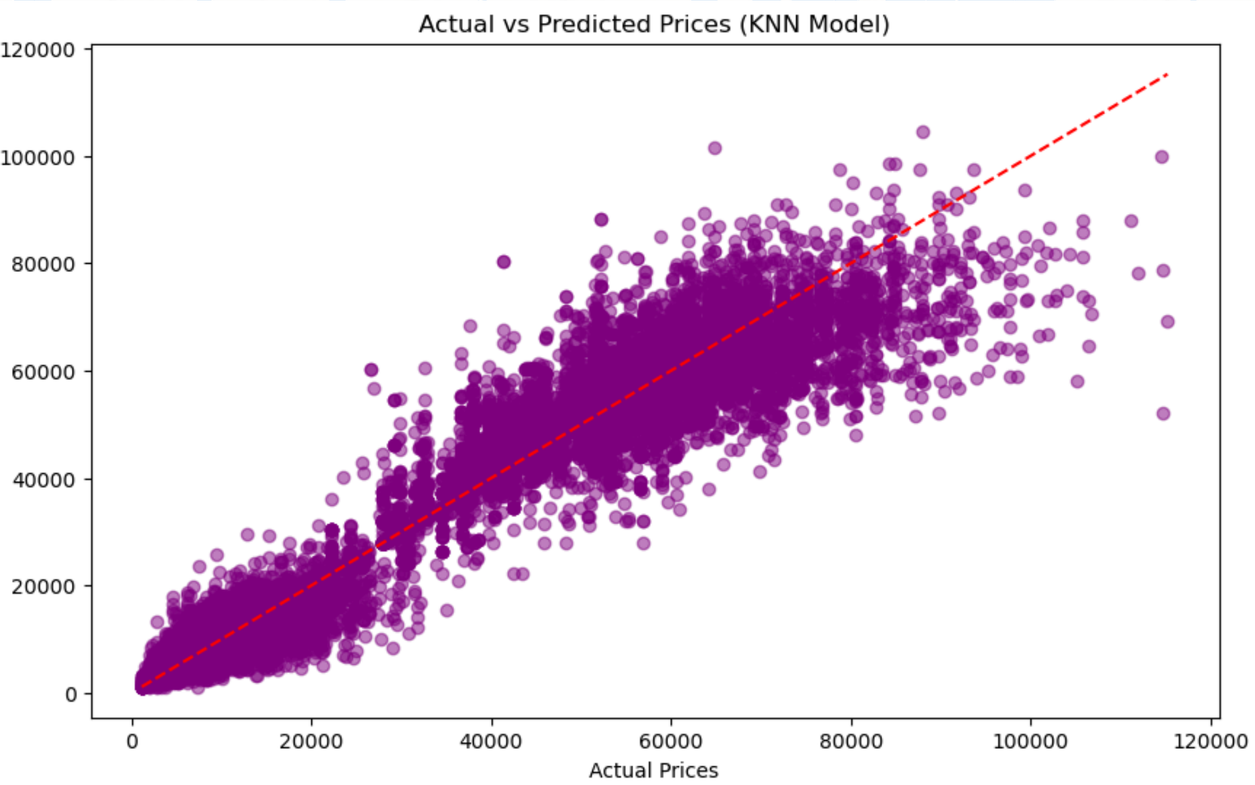


Model Performance

```
Best Ridge Alpha: 1.0
Ridge R²: 0.9113048857987591
Ridge Mean Absolute Error (MAE): 4553.280681349071 5
Ridge Mean Square Error (MSE): 45720759.122906476
Ridge Root Mean Square Error(RMSE): 6761.712735905 4875
```

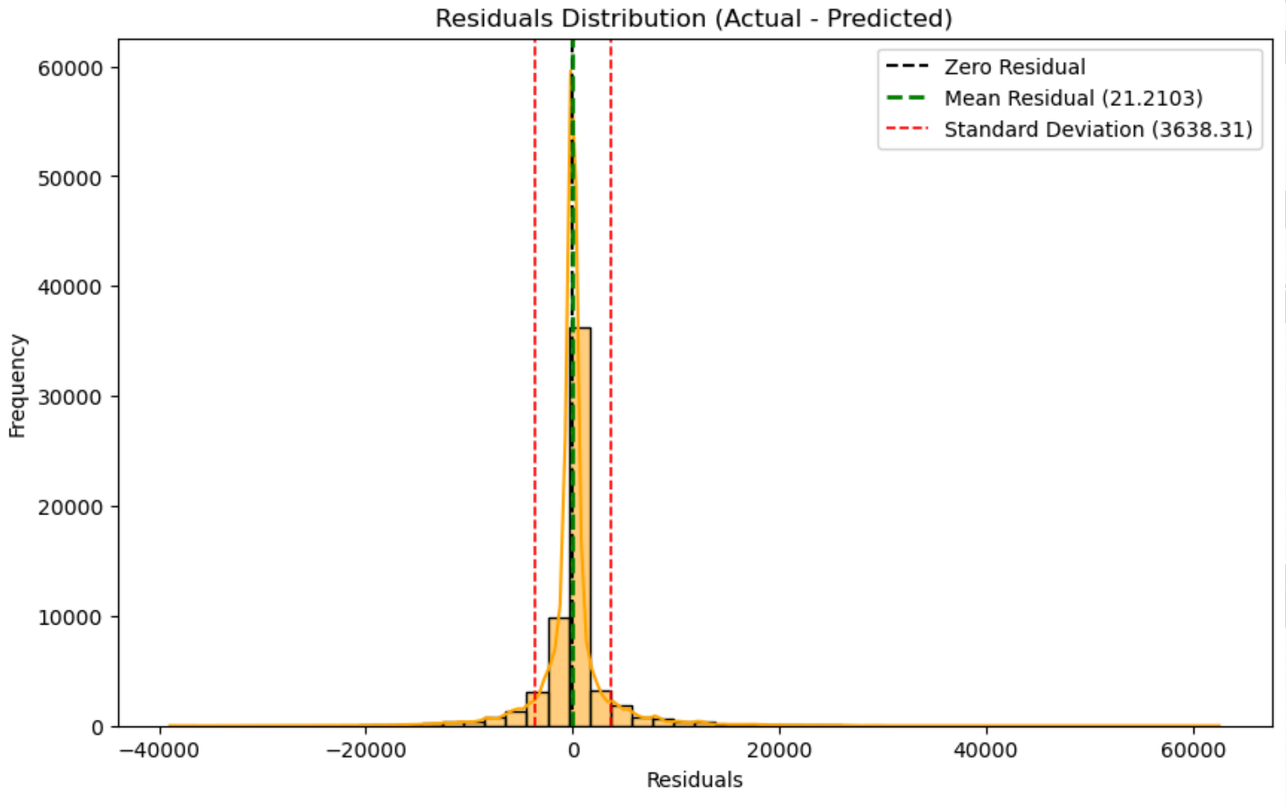
K- Nearest Neighbor





Actual VS Predicted

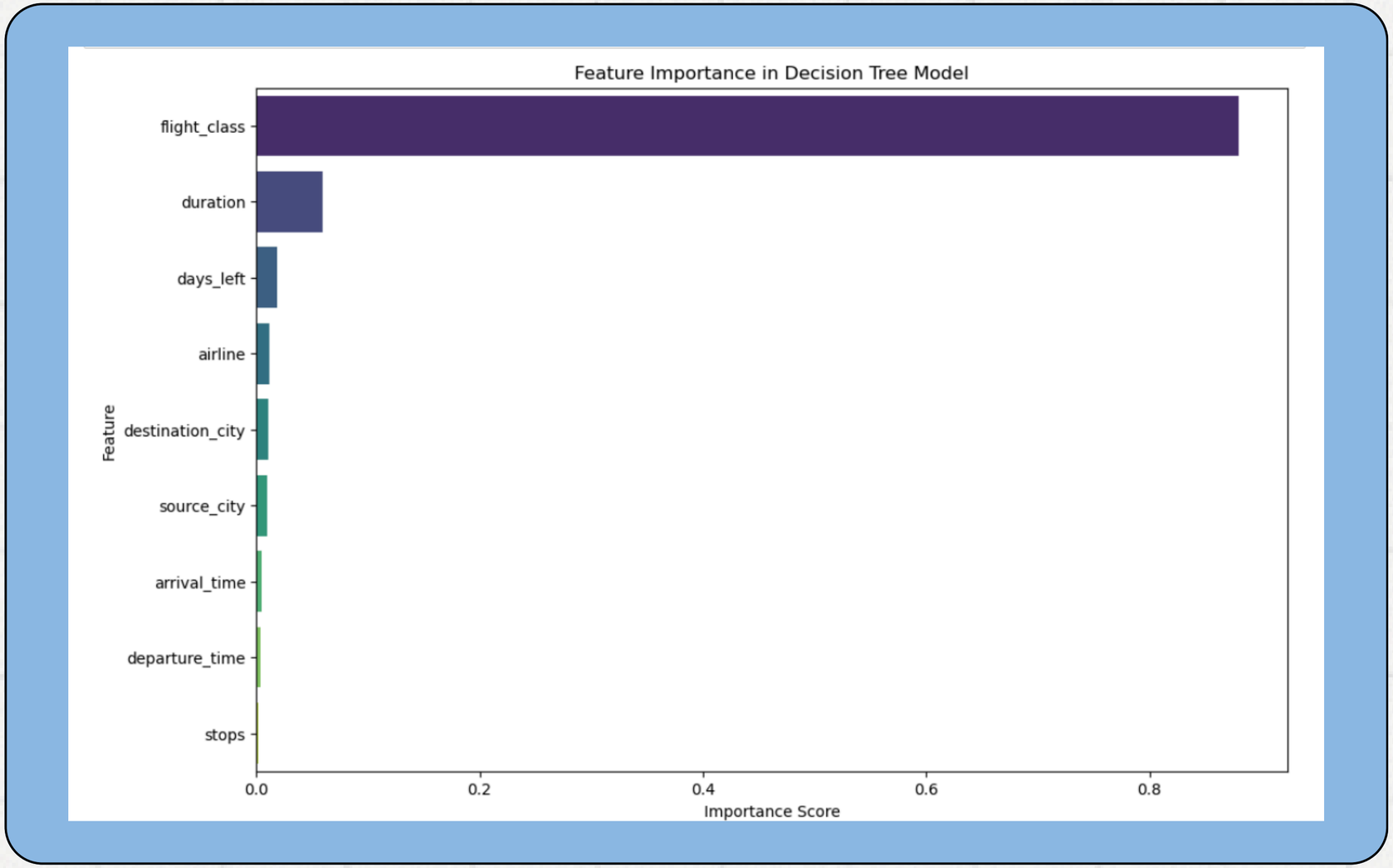
Residual Distribution



Model Performance

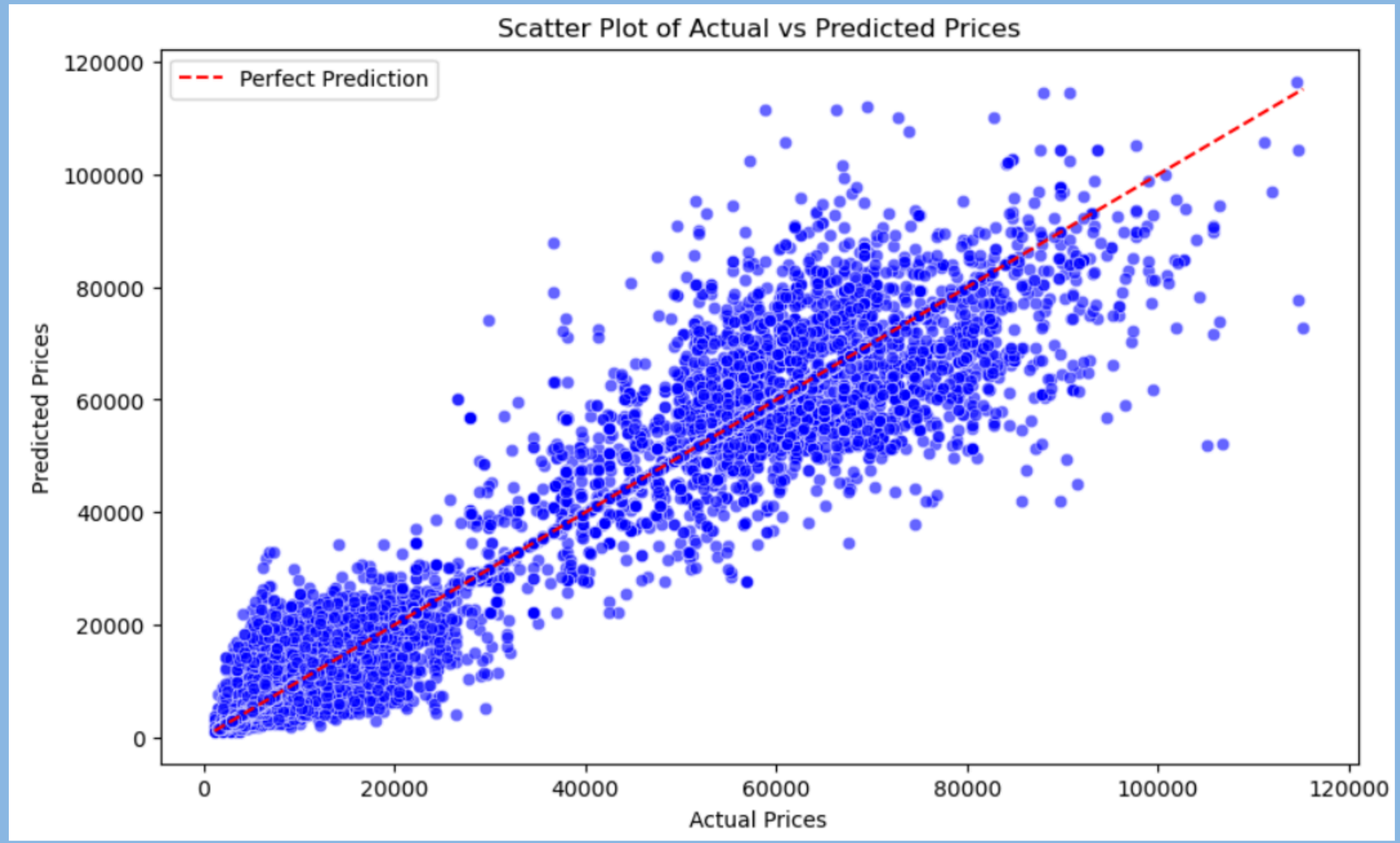
Best K value: 3
Mean Absolute Error (MAE): 1670.66
Mean Squared Error (MSE): 13237540.51
Root Mean Squared Error (RMSE): 3638.34
R^2 Score: 0.97

Decision Tree

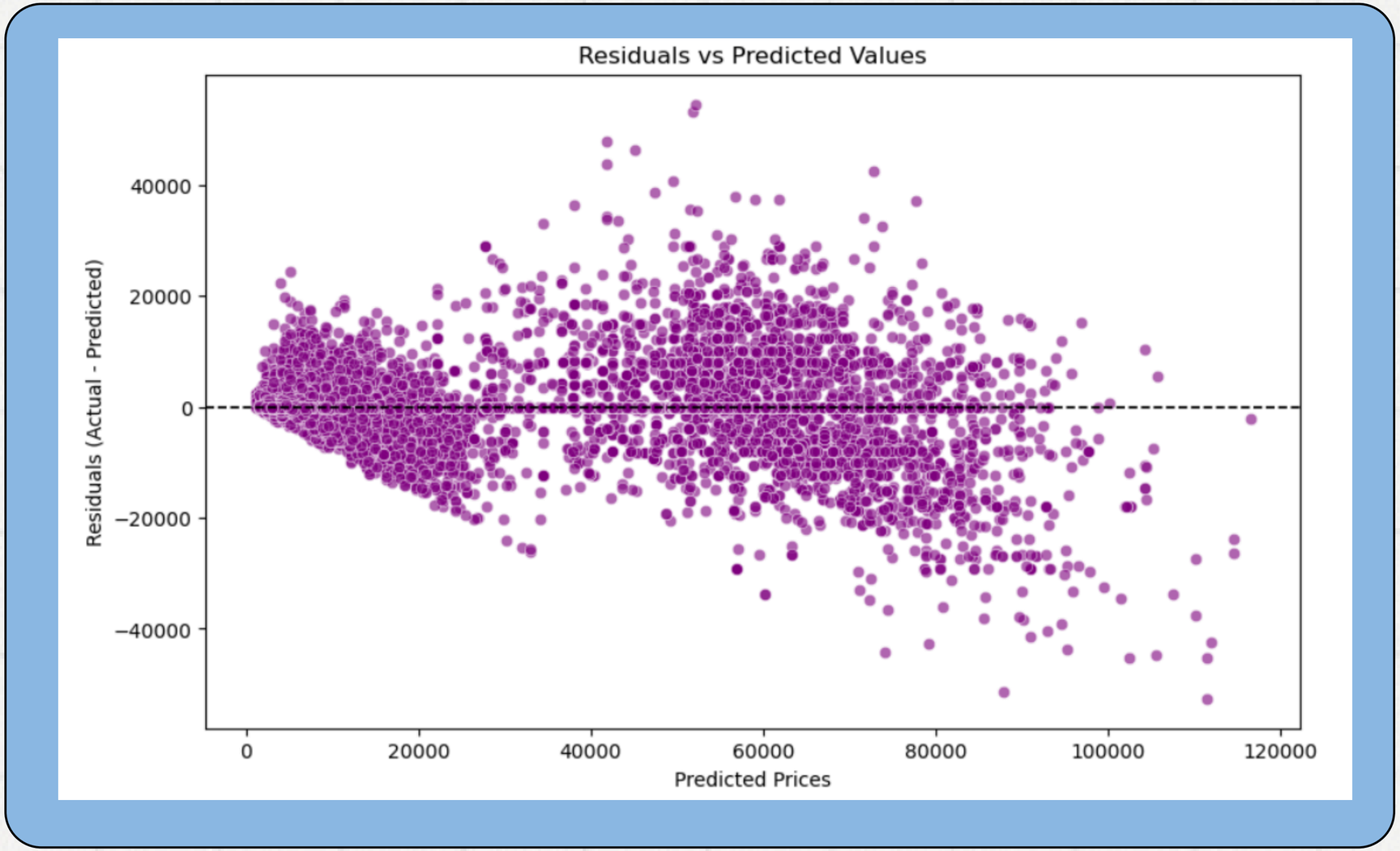


Feature Importance in a Decision Tree Model

Scatter plot to show flight predictions



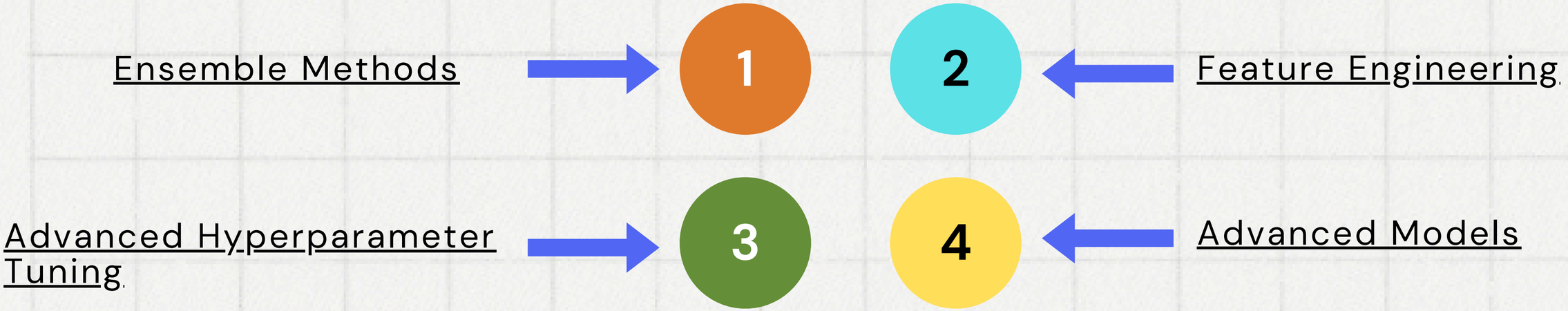
Plot for the Residuals



Model Performances

Metric	Linear Regression	KNN	Decision Tree
MAE	4523.53	1670.66	1166.44
MSE	53,916,434.30	13,237,540.51	12,252,834.70
RMSE	7342.78	3638.34	3500.40
R^2 Score	0.911	0.97	0.98

Future Steps



In conclusion, the Decision Tree model currently provides the best balance of accuracy and interpretability for flight price prediction, but future improvements with ensemble methods and tuning may yield even better results.

THANKYOU