Sky High Savings: Predicting Indian Airfare Trends

By Lupe Covarrubias

The Problem

- Prices influenced by seasonality, demand, destinations, and competitor pricing.
- Overpaying due to unpredictable fare fluctuations.

What factors might affect airfare?

- Route
- Arrival time
- First Class
- Airline
- Duration
- Day of the week
- Planned vs Last-minute Booking

How can we help?

Objective: Build a predictive model to highlight cost-driving factors and help raise customer satisfaction through saving money.

Empower millions of travelers and enhance travel platforms with accurate airfare prediction tools.

Expedia

Google

hopper

Data Overview

- Dataset sourced from Kaggle, containing ~445k rows and 13 columns.
- Initial cleaning included:

Splitting the Date_of_journey column into year, month, and day (year was dropped since data spanned only 3 months).

Removing duplicates based on flight_code, Destination, Fare, Arrival, Duration_in_hours, and other features.

Initial Observations

- **Dataset was predominantly categorical with limited continuous features (e.g., Fare).**
- **Categorical features had 3-7 values each, critical for creating dummy variables later.**

- **Aggregation by Fare revealed:**

Ahmedabad as the most expensive destination.

The longest flights were also to Ahmedabad.

Preliminary Insights

- Outliers in airfare data were observed due to natural variability.
- Fare distribution showed bimodal behavior with a right skew (Figure 1).
- Median was chosen as the key metric for further analysis to handle skewness effectively.

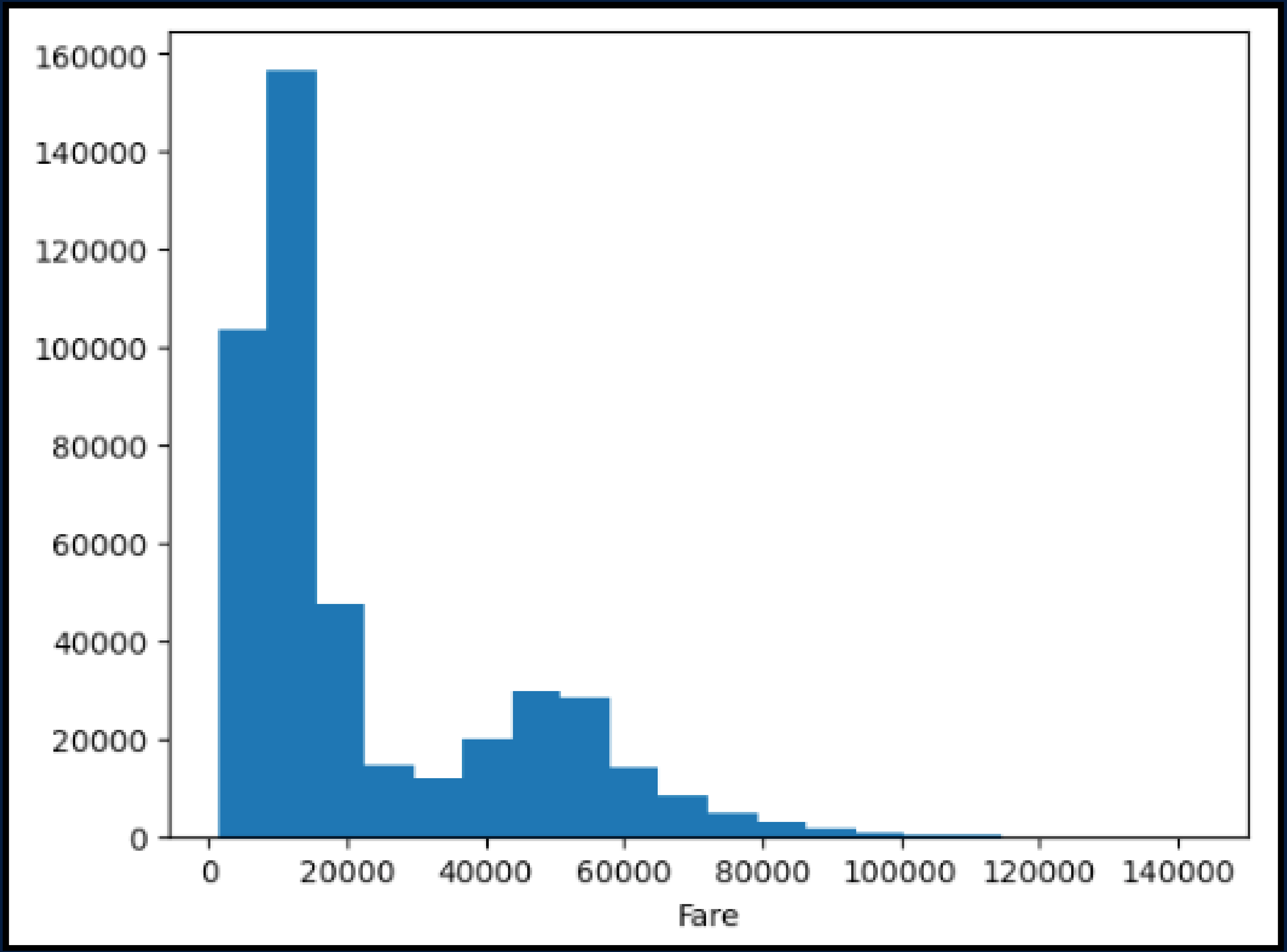

Figure 1

Key Metrics and Initial Observations

- Median Fare = 13,362 rupees < mean = 22,920 rupees
- Median chosen as it is robust to outliers.
- No correlation between numerical features (e.g., Duration_in_hours, Day, Days_left) and airfare.
- Focus shifted to categorical features (e.g., Class, Airlines).

Initial Exploration

- **Only Vistara and Air India offered premium classes (e.g., Business, First Class).**
- **Combined all non-economy values into a single "Premium" category (Figure 2).**

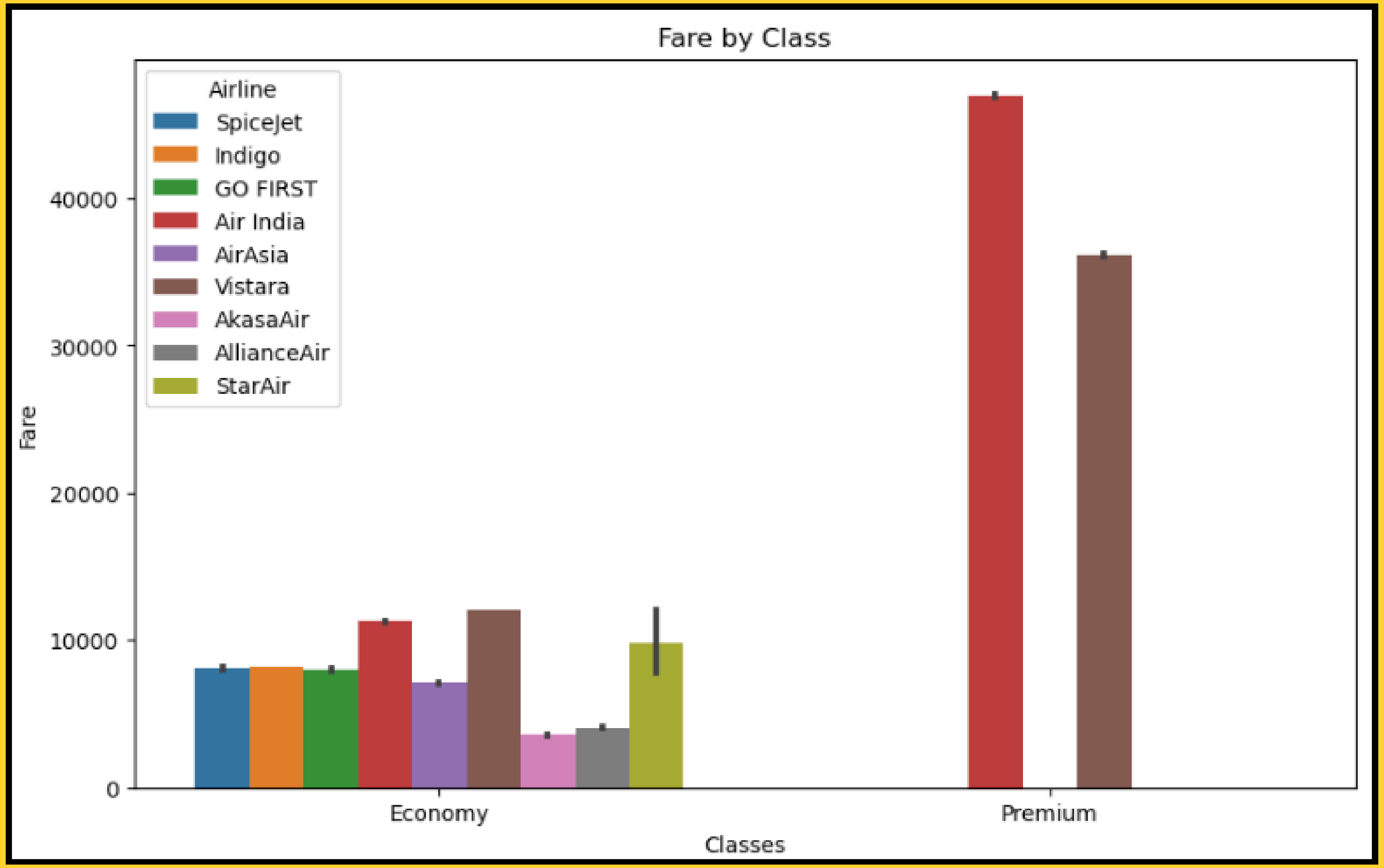

Figure 2

Frequency Patterns from EDA

- Mondays for most airlines
- Economy for most airlines; Vistara's frequent class is Premium.
- Arrivals - After 6 PM, due to full-day travel.
- Departures - Noon to midnight, aligning with travelers’ schedules.
- Stops: One-stop flights dominate; nonstop flights are rare for smaller airlines.
- Ahmedabad to Mumbai - Most expensive, ₹18,712.
- Bangalore to Delhi - Cheapest ₹10,338, frequently operated by budget airline AirAsia.

Fare Trends

- Last-minute bookings - Most expensive
- Planned bookings - Cheapest
- Short-notice bookings - Most frequent, driven by business travel or emergencies
- Sunday (most expensive), Thursday (cheapest).
- Departure Times: Before 6 AM (cheapest), 6 AM–Noon (most expensive).
- More stops = higher fares.
- End of Month Flights higher fares.

Visualized Patterns

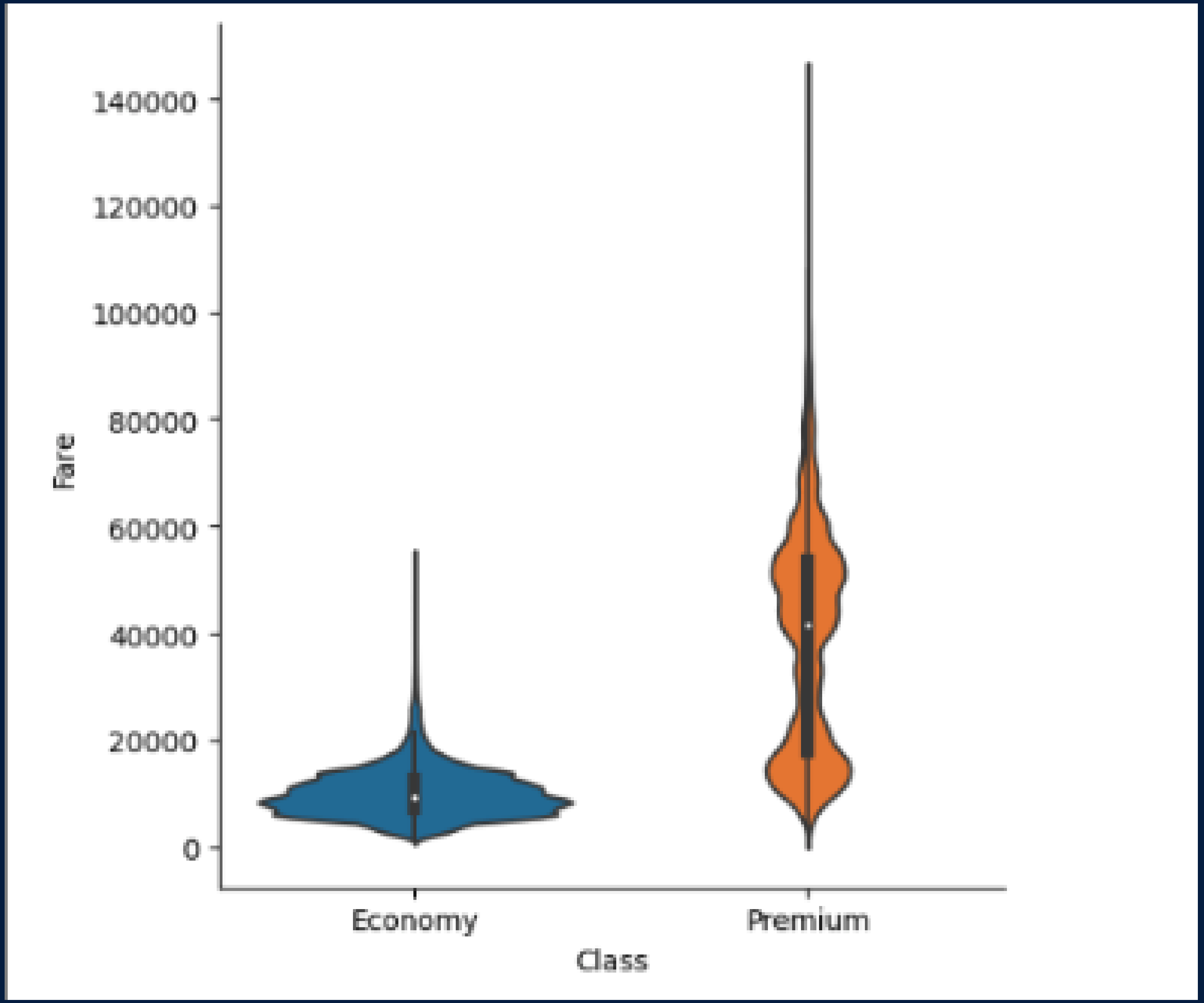

Figure 3

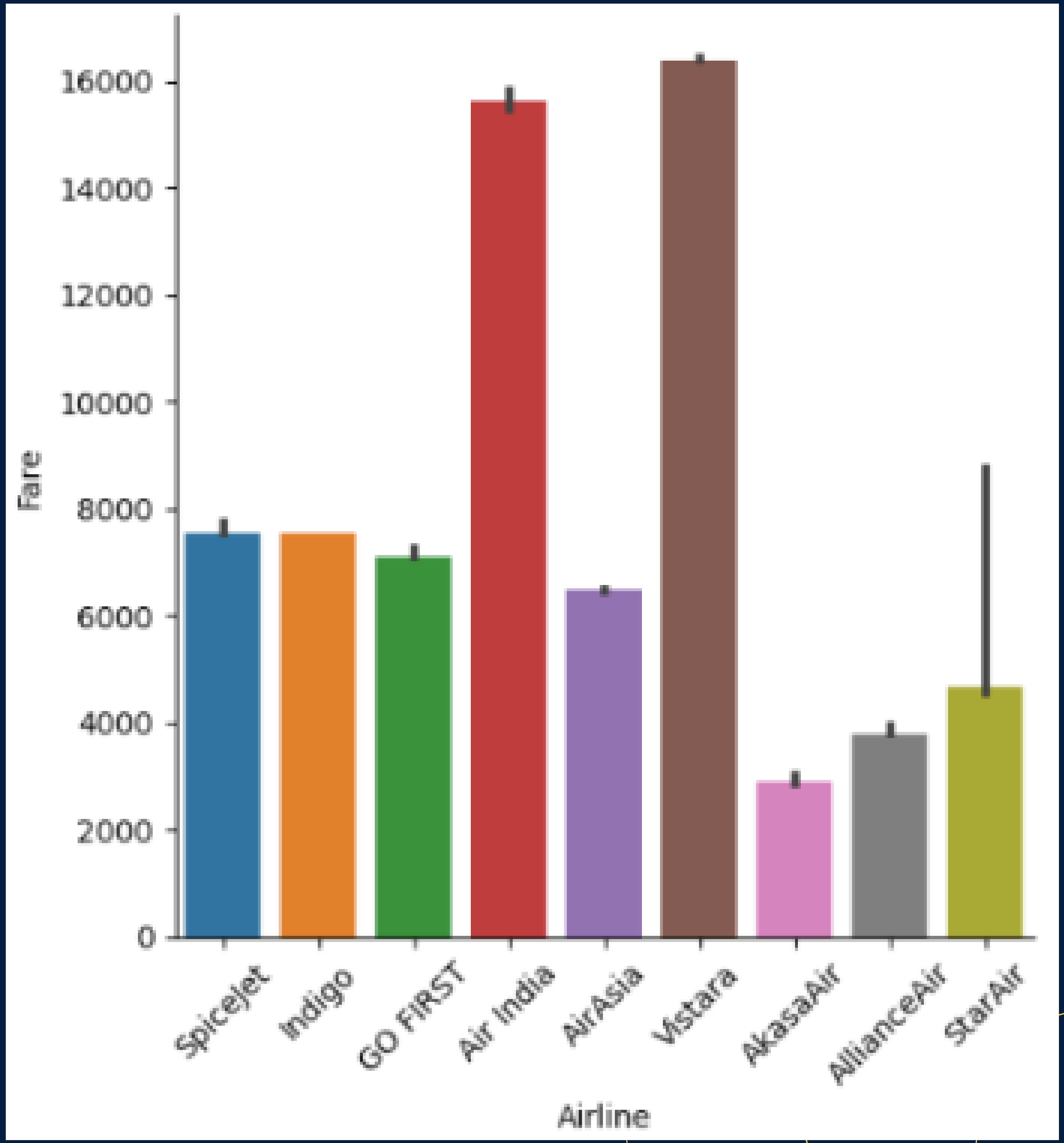

Figure 4

Preprocessing

- **Predominantly categorical with limited continuous features (e.g., Fare).**
  - **3-7 values per feature, critical for creating dummy variables later.**
- **Methods used to reduce multicollinearity**
  - **63 dimensions total**
- **All binary variables, no need for scaling**

Preliminary Modeling

- Dummy Regressor: median = 13,379 rupees for baseline performance.

	Train set	Test set
R-squared	-0.2195	-0.2172
MAE	14289	
MSE	506460343	508088827
RMSE	22505	22541

- Similar metrics for train and test data suggest good generalization
- The dummy regressor highlights the data's complexity

Linear Regression Model

- Similar metrics for train and test data suggest good generalization
- Decent model but room for improvement
- best k = 68

	Train set	Test set
R-squared	0.56	0.56
MAE	9424	9431

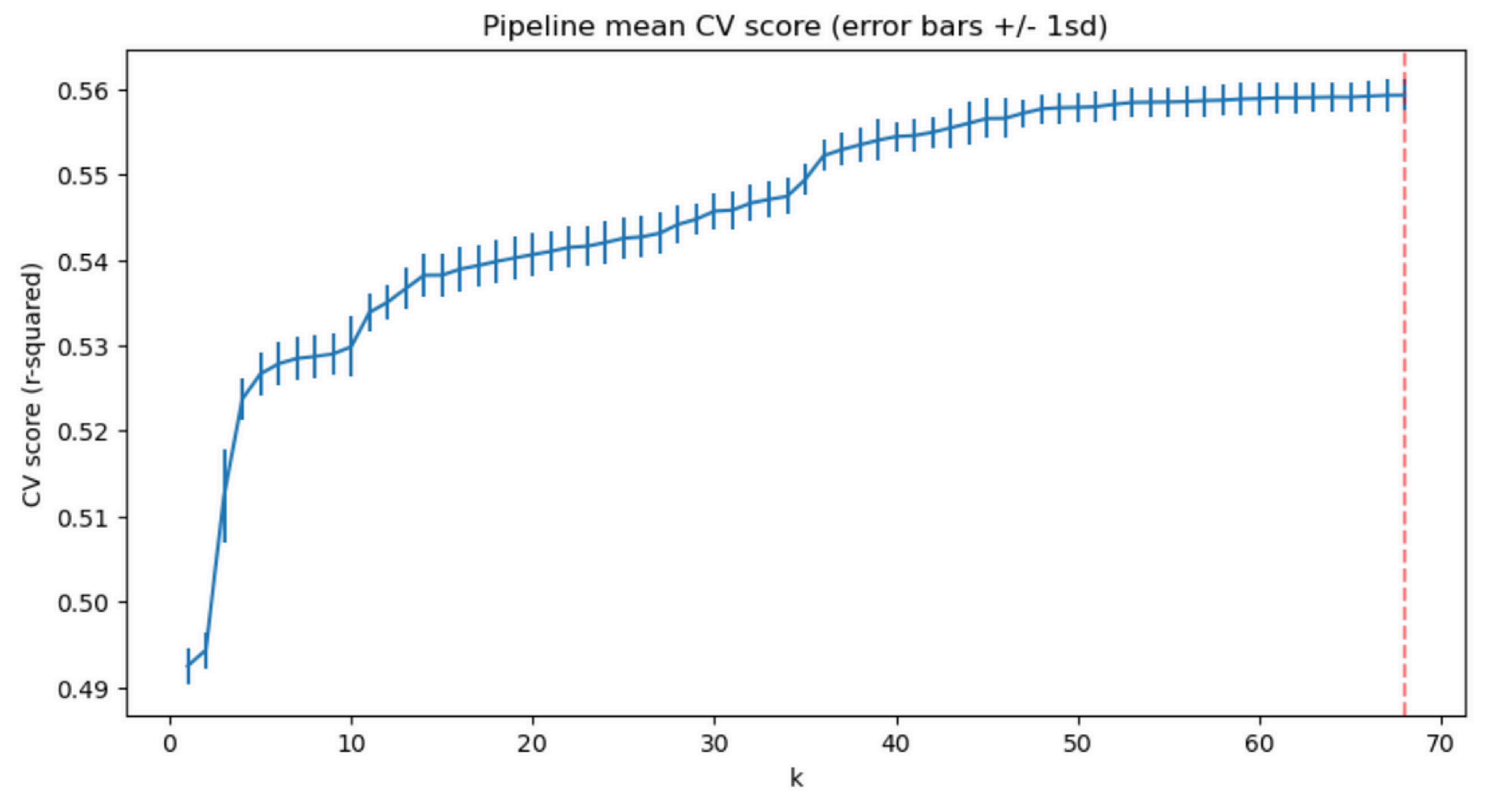

Figure 5

Gradient Boosting Regressor Model

Advantages:

- Handles complex datasets and robust to outliers.
- Random Search CV used for efficient hyperparameter tuning due to large search space.

	Train set	Test set
R-squared	0.66	0.66
MAE	8151	8176

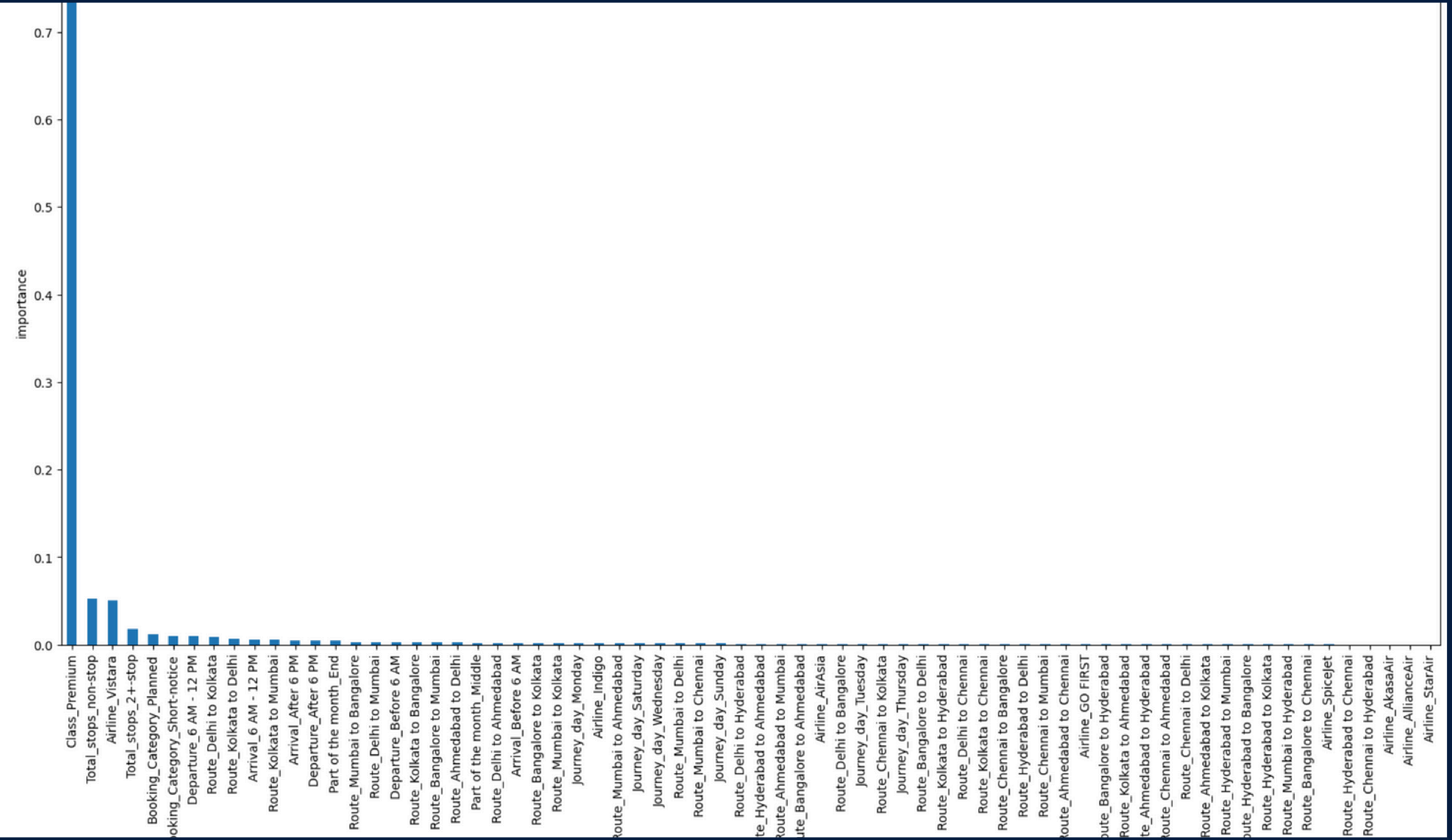

Figure 6

Histogram Gradient Boosting Regressor Model

Advantages:

- **Chosen for its efficiency with high-dimensional categorical data and reduced computation time.**

	Train set	Test set
R-squared	0.65	0.65
MAE	7514	7558

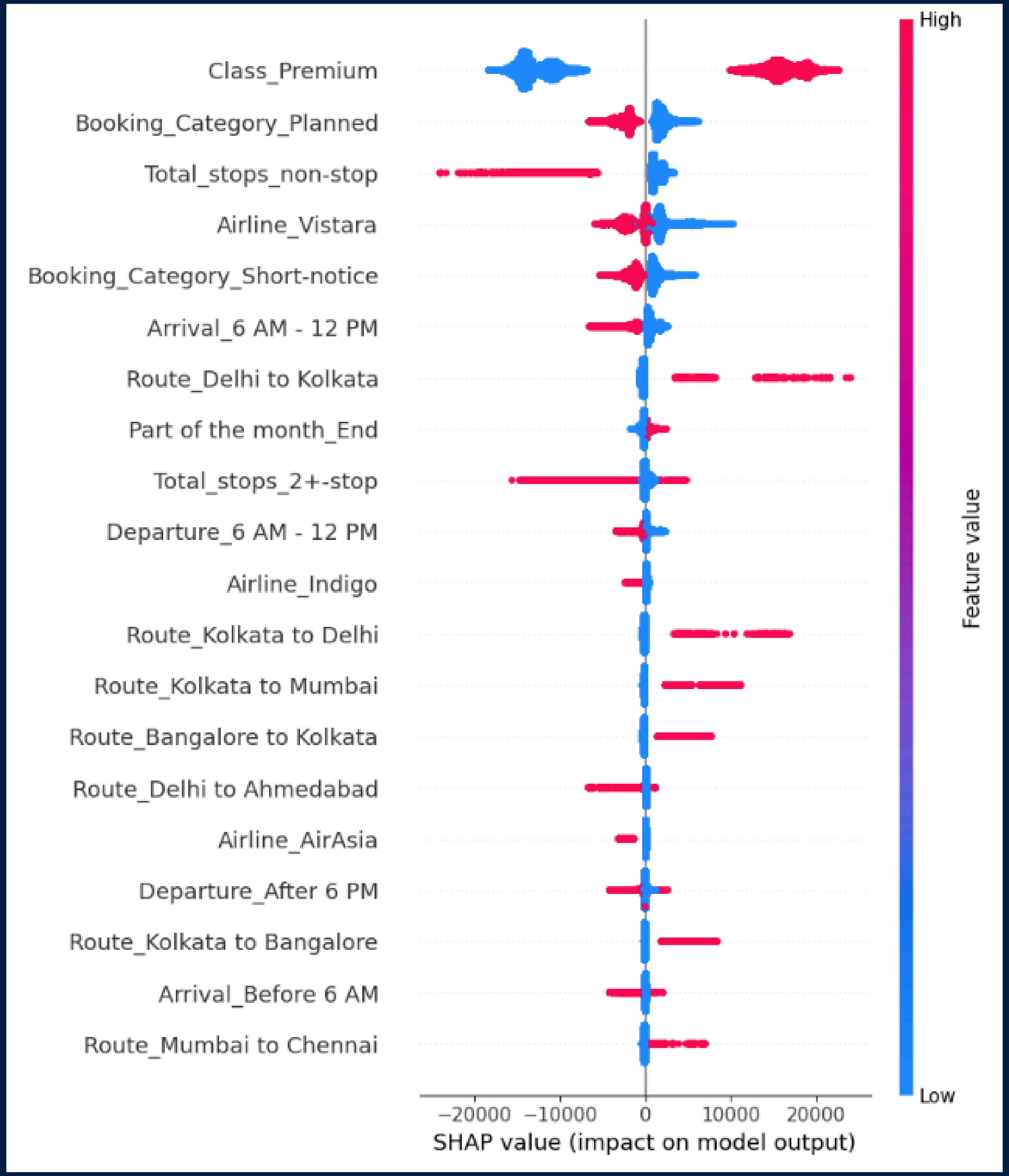

Figure 7

Model Selection

- **Histogram Gradient Boosting Regressor (HGBR)**
- **Best predictive power for fare**
- **Scatter increases as fare values rise**
- **Underprediction for fares above 80,000 rupees**
- **Noticeable outliers due to data noise and model limitations**

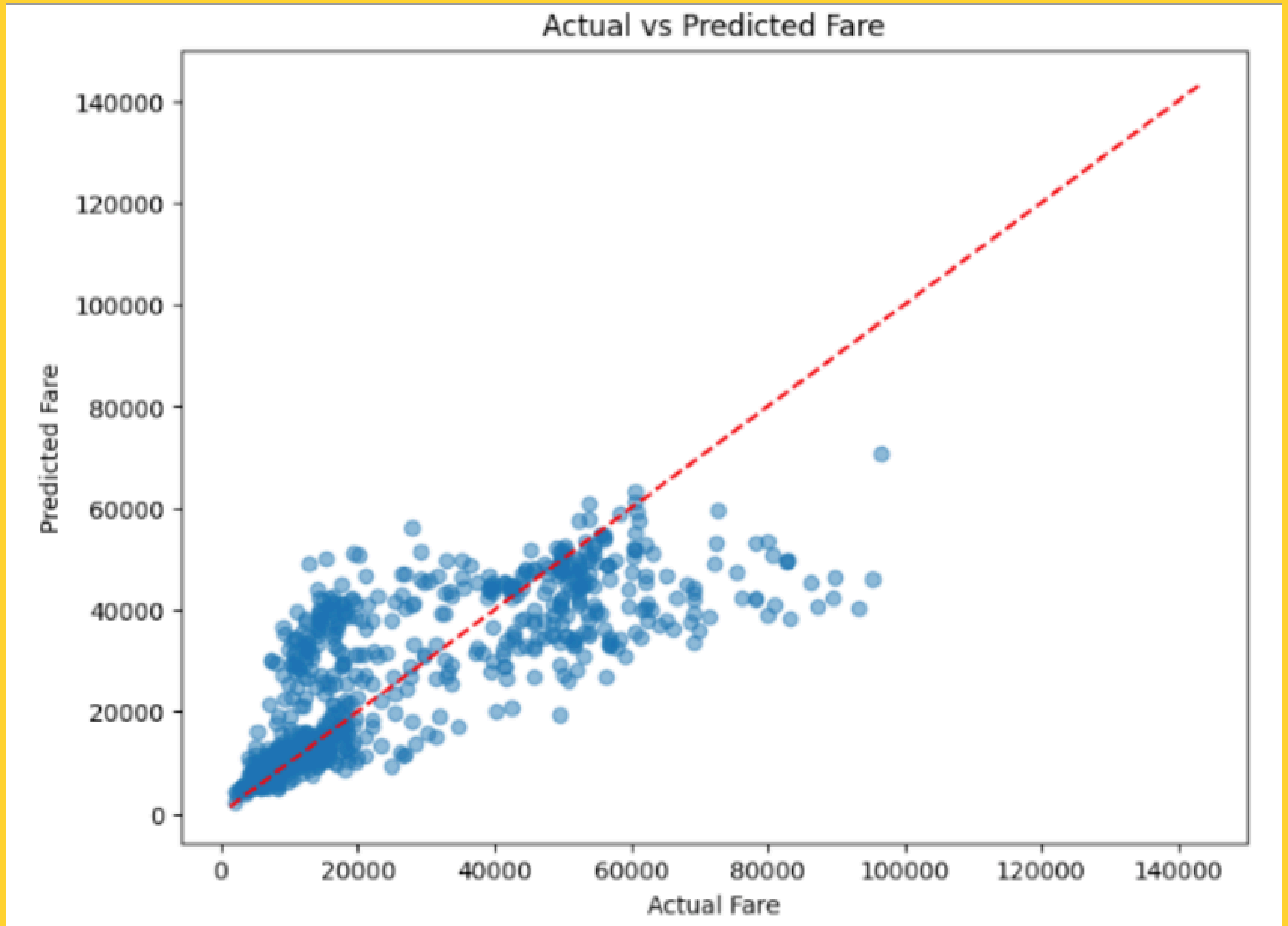

Figure 8

Conclusion

- **Business Impact:**
  - **Helps optimize pricing strategies and target premium customers.**
  - **Empower millions of travelers and enhance travel platforms with accurate airfare prediction tools.**
- **Areas for Improvement:**
  - **Advanced feature engineering (e.g., OrdinalEncoder).**
  - **XGBoost, LightGBM**
  - **Log transforms to improve predictions for higher fare ranges.**
  - **Focusing strictly on economy**

Conclusion:

HGBR provides a solid foundation for fare prediction but could benefit from further enhancements to increase accuracy and customer trust.