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.

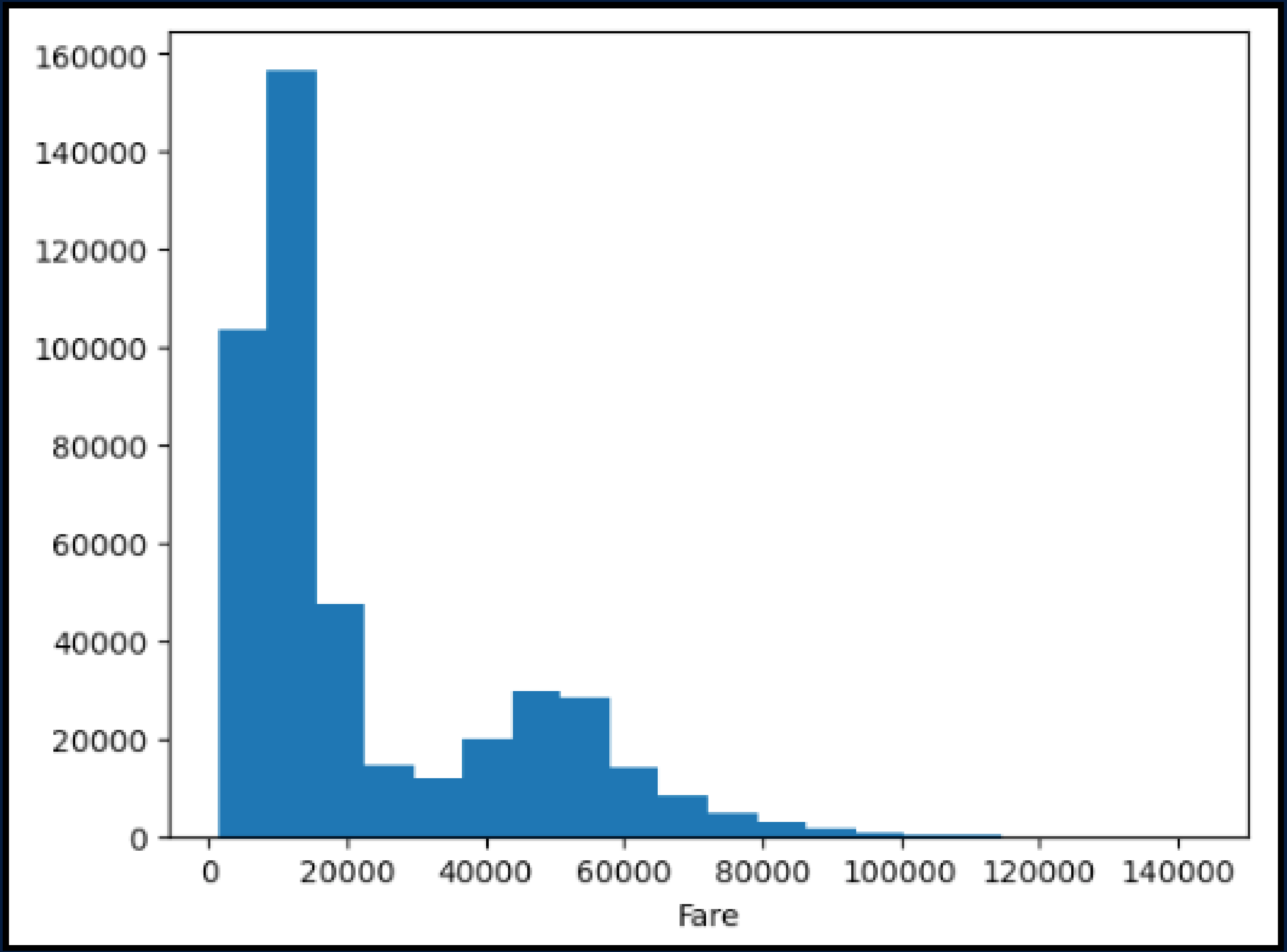

Figure 1

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).

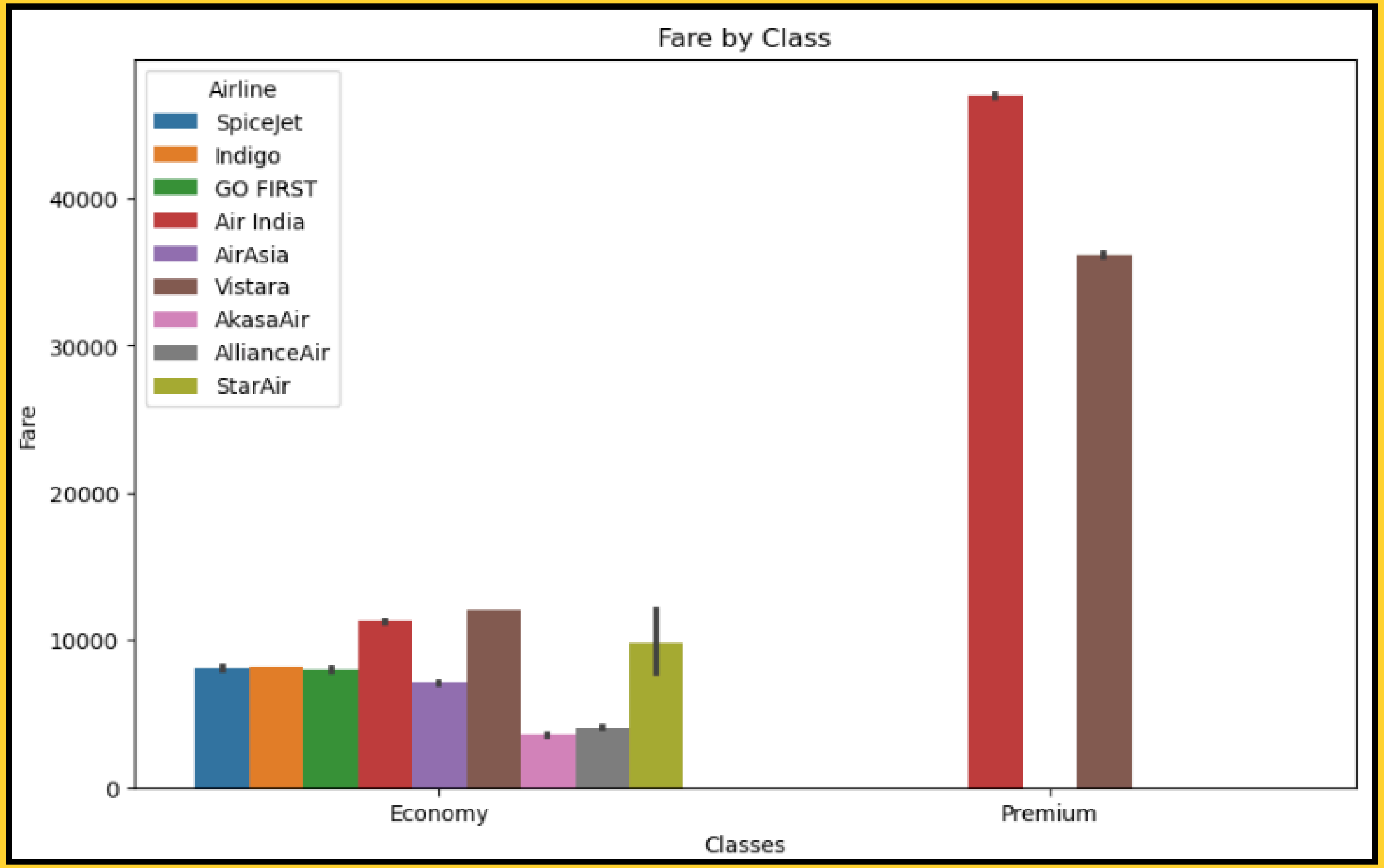

Figure 2

Frequency Patterns from EDA

- Common Travel Day: Mondays for most airlines
- Frequent Class: Economy for most airlines; Vistara's frequent class is Premium.
- Popular Flight Times:
  - 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.
- Fare Insights by Route:
  - Ahmedabad to Mumbai - Most expensive, ₹18,712.
  - Bangalore to Delhi - Cheapest ₹10,338, frequently operated by budget airline AirAsia.

Fare Trends

- **Fare vs. Booking Category:**
  - **Last-minute bookings - Most expensive**
  - **Planned bookings - Cheapest**
  - **Short-notice bookings - Most frequent, driven by business travel or emergencies**

Key Takeaway:

- **Median analysis highlighted critical factors like route, booking category, travel time, and flight type, setting the stage for further exploration through visualizations.**

- **Other Trends:**
  - **Travel Days: Sunday (most expensive), Thursday (cheapest).**
  - **Departure Times: Before 6 AM (cheapest), 6 AM–Noon (most expensive).**
  - **Stops Impact: More stops = higher fares.**
  - **End of Month Flights: Higher demand leads to higher fares.**

Visualized Patterns

Flight Class vs. Fare (Figure 3):

- Premium fares exhibit a higher range and median compared to Economy fares.
- Economy fares are concentrated at lower price points, reflecting affordability and demand.

Airlines and Median Fare (Figure 4):

- Highest fares - Vistara and Air India
- Lowest fares - Akasa Air, AirAsia, Alliance Air, and Star Air

- Significant features driving fares: Class, Route, Airline, Journey_day, Departure/Arrival times, and Total_stops.

- Last-minute bookings: Highest fares.

- Newly created features (e.g., Route, Part_of_the_month) and existing ones hold strong predictive potential.

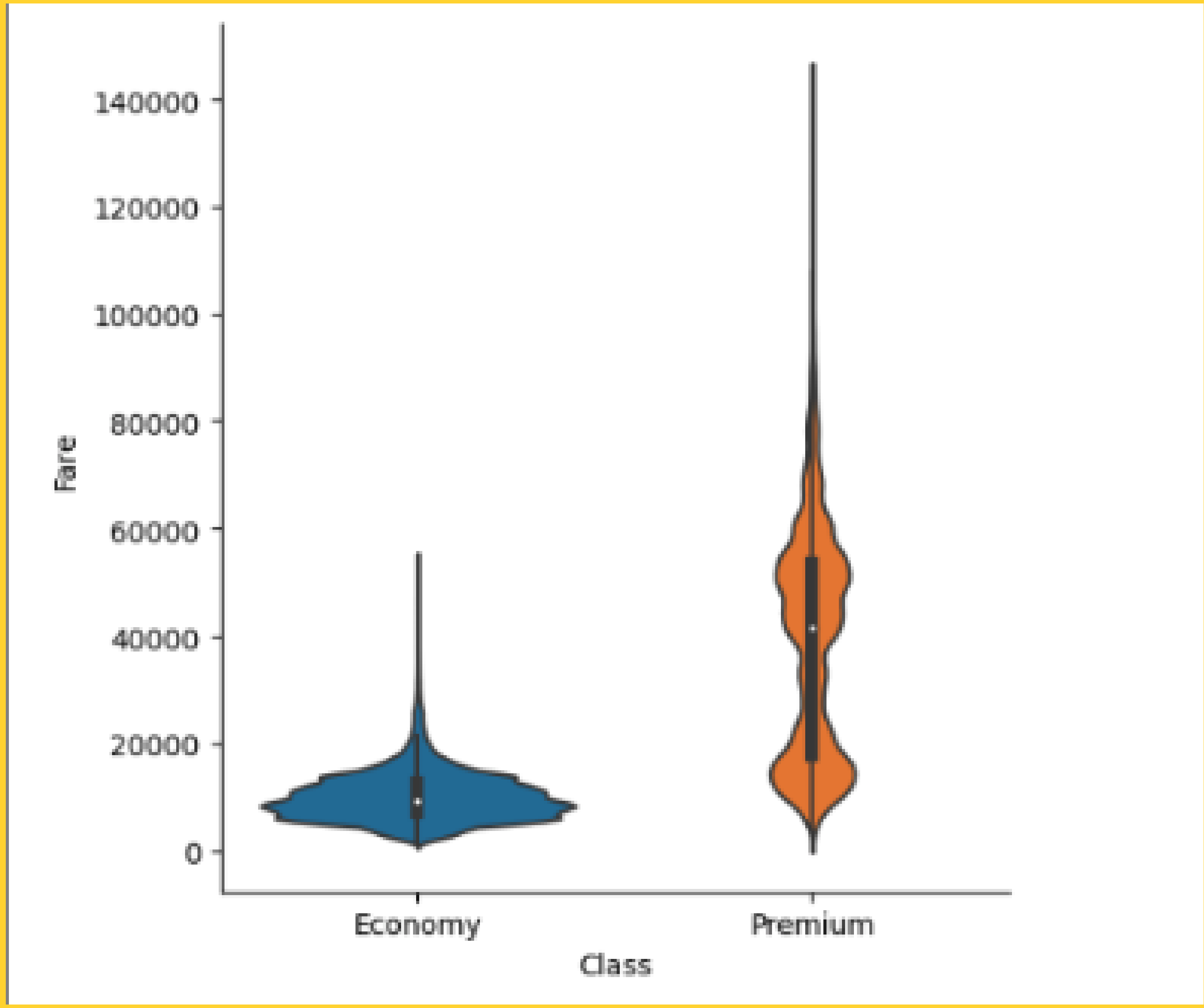

Figure 3

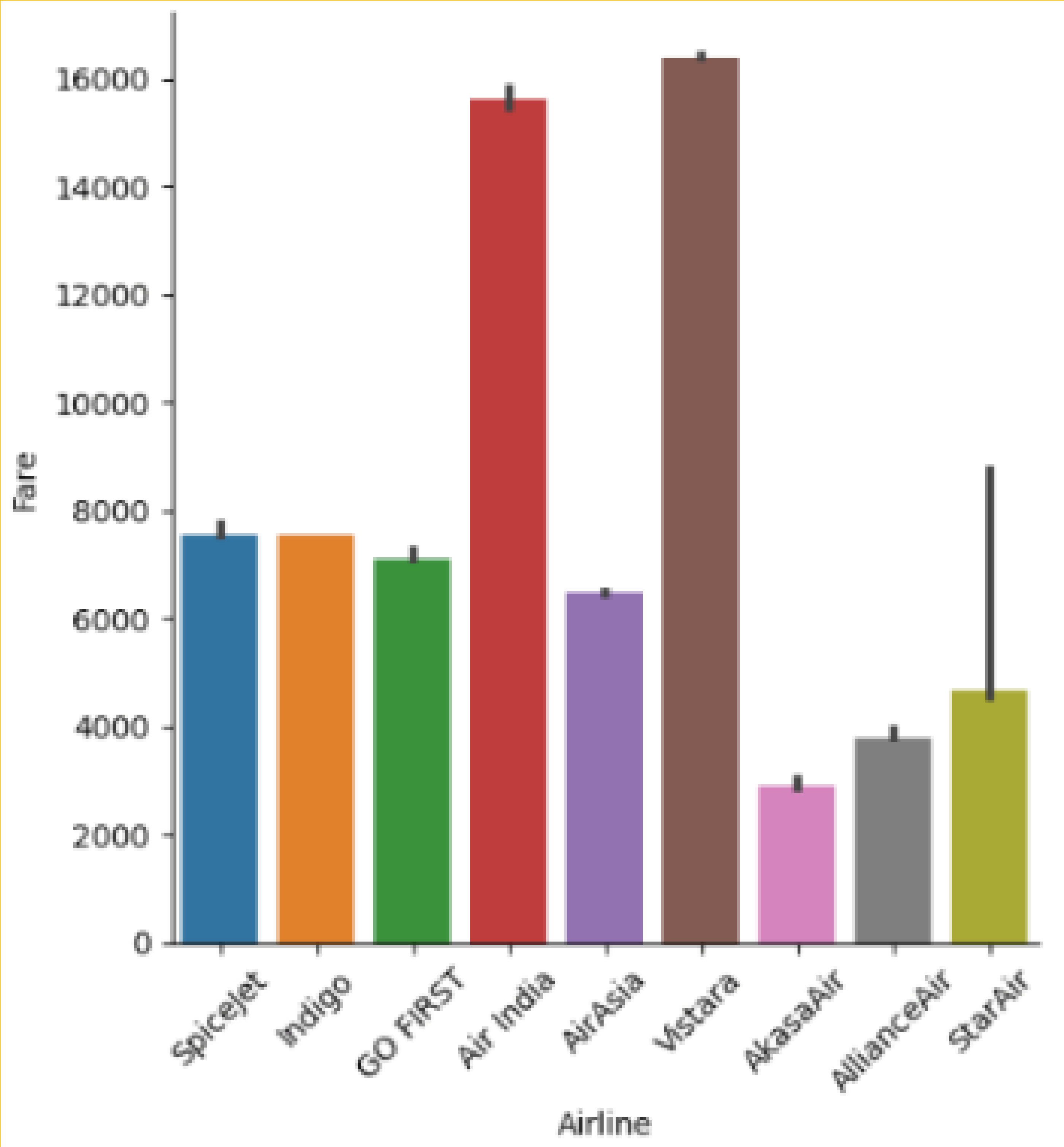

Figure 4

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.
- **Model Performance:**
  - Negative R-squared value (-0.2195, -0.2172) indicate poor performance
- **Evaluation Metrics:**
  - MAE: Average error of ~14,289 rupees.
  - MSE: Extremely high due to outliers (506460343, 508088827)
  - RMSE: (~22,505, ~22,541) rupees.

- **Key Insight:** Similar metrics for train and test data suggest good generalization, with the model learning underlying patterns.

- **Conclusion:** The dummy regressor highlights the data’s complexity, suggesting the need for a more advanced model with additional features and hyperparameter tuning for improved performance.

Linear Regression Model

- **Performance Metrics:**
  - **R-squared: ~56% for train and test sets.**
  - **Train MAE: 9424 | Test MAE: 9431.**
- **GridSearch CV identified the best K as 68 (all features), still explaining 56% of the variation.**

- **Key Features Identified:**
  - **Premium Class: Strong pos. impact**
  - **Routes from Kolkata: Pos.**
  - **Non-Stop Flights: Neg. impact**

Conclusion: While the model provides decent performance, more complex models with advanced hyperparameter tuning are better suited for this dataset.

Gradient Boosting Regressor Model

- **Performance Metrics:**
  - R-squared: 0.60, improved to 0.66 after hyperparameter tuning.
  - Outperforming the linear regression model.
  - Train MAE: 8151 | Test MAE: 8176.
- **Advantages:**
  - Handles complex datasets and robust to outliers.

- Random Search CV used for efficient hyperparameter tuning due to large search space.
- **Key Features Identified:**
  - Class_Premium: Strong pos. impact on airfare.
  - Total_stops_non-stop: Neg. airfare.
  - Airline_Vistara: Pos. feature
  - Booking_Category_Planned: Neg. feature

Conclusion: Gradient Boosting Regressor shows strong potential and better performance, with opportunities for further improvement through fine-tuning.

Histogram Gradient Boosting Regressor Model

- **Performance Metrics:**
  - Basic HGBR model (no tuning): R^2 = 0.64 (better than the simple GBR model).
  - Tuned HGBR model (Random Search CV): R^2 = 0.65.
  - Train MAE: 7514 | Test MAE: 7558.
- **Advantages:**
  - Chosen for its efficiency with high-dimensional categorical data and reduced computation time.

- **Key Features Identified:**
  - Class_Premium (inc. price).
  - Total_stops_Non-stop (dec. price).
  - Route_Delhi_to_Kolkata and Total_stops_2+-stop.

Conclusion: Outperformed the GBR model; selected as the final model.

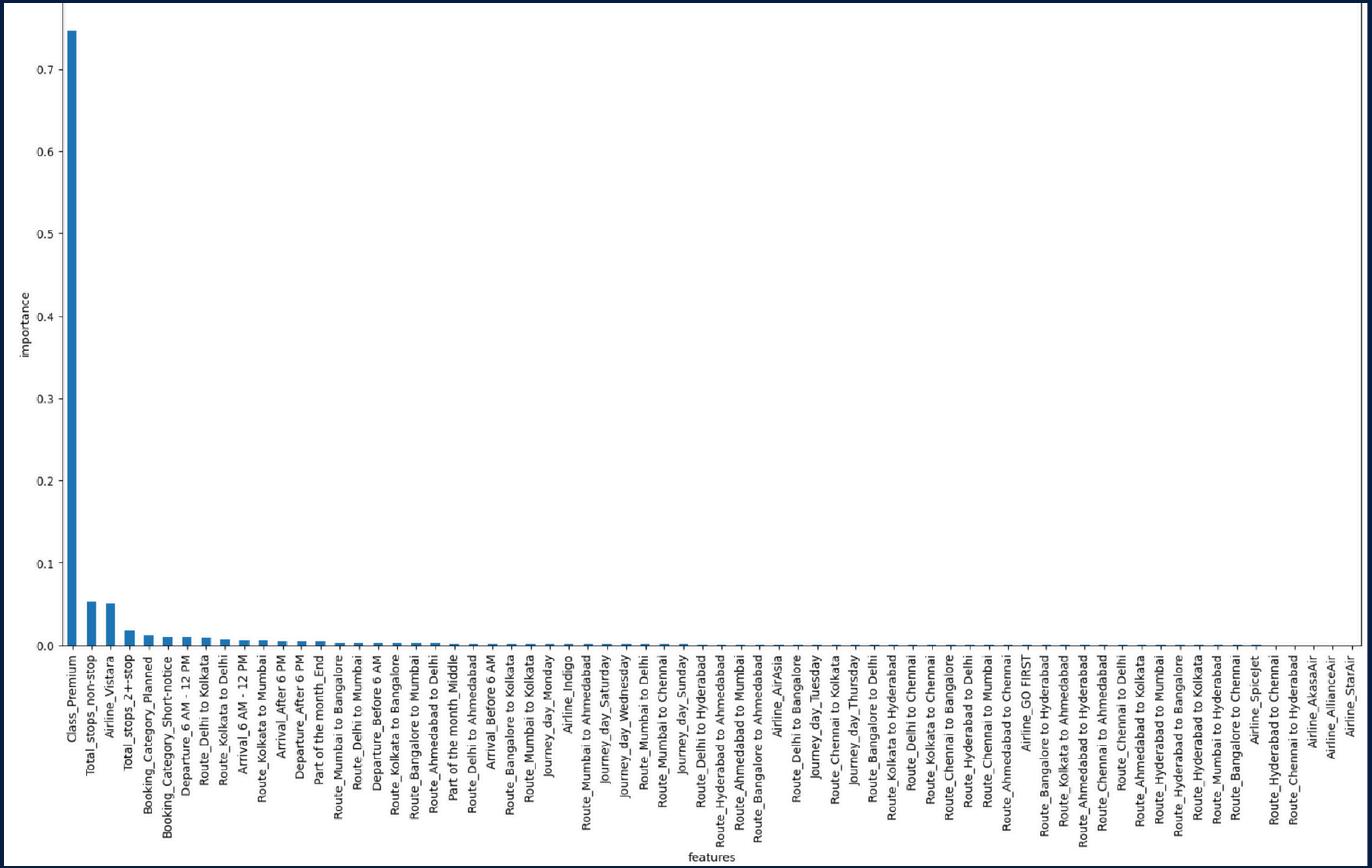

Figure 5

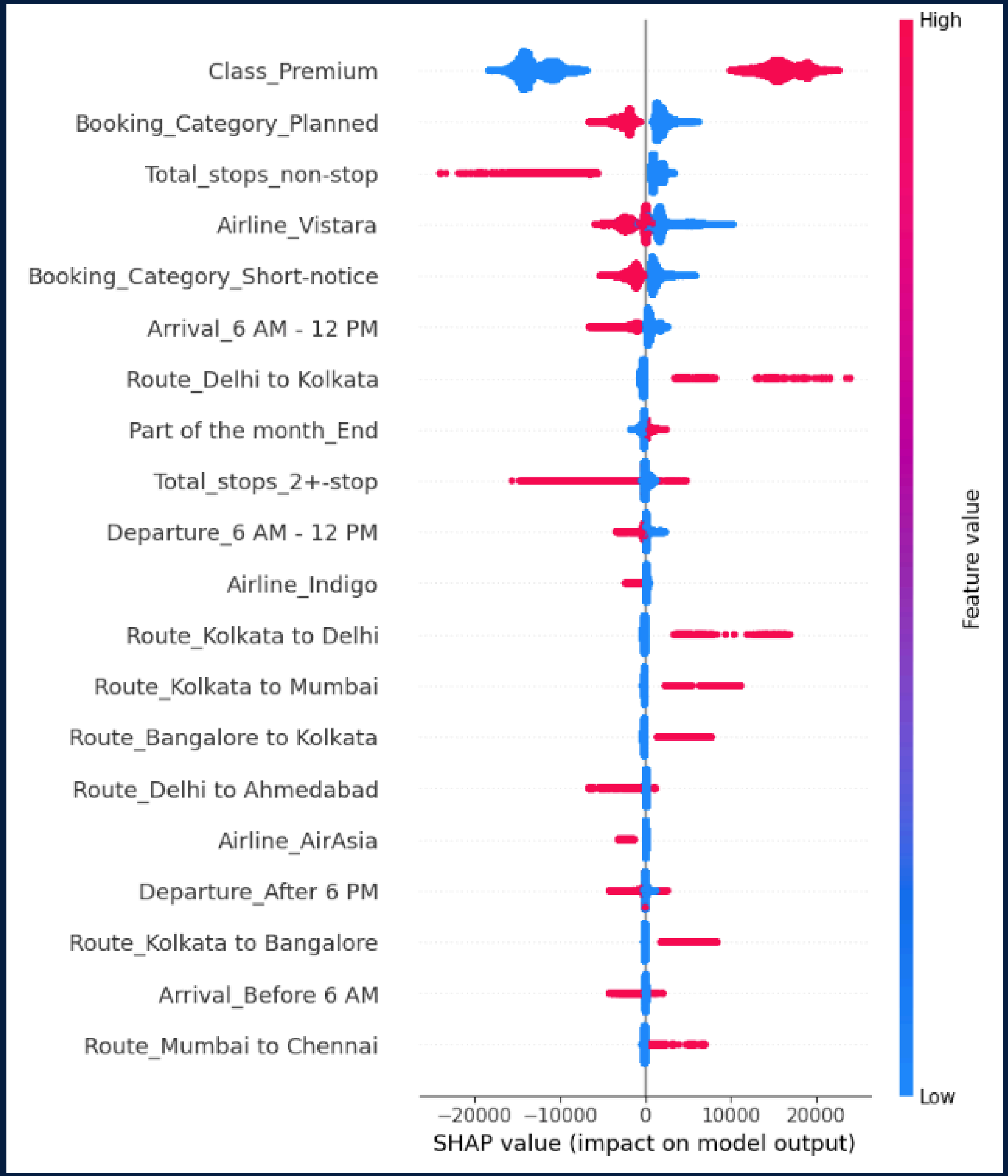

Figure 6

Model Selection

- **Selected Model: Histogram Gradient Boosting Regressor (HGBR)**
- **Best predictive power for airfare.**
- **Good correlation between actual and predicted values.**
- **Scatter increases as fare values rise, indicating higher prediction errors for expensive fares.**
- **Underprediction for fares above 80,000 rupees.**
- **Noticeable outliers due to data noise and model limitations.**

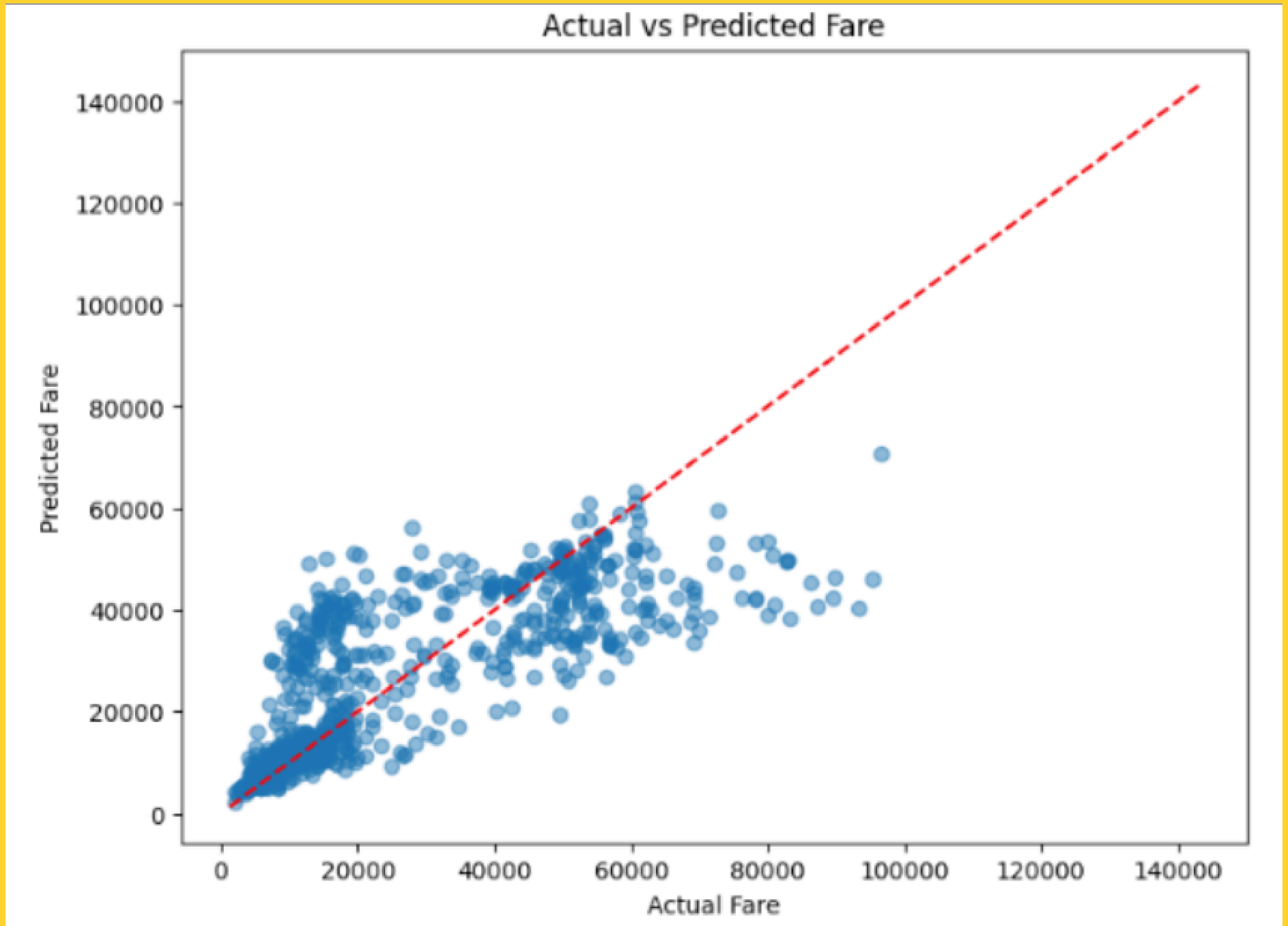

Figure 7

Conclusion

- **Business Impact:**
  - Helps optimize pricing strategies and target premium customers.
  - Consistent train/test metrics reduce the risk of pricing errors during high-demand periods.
- **Areas for Improvement:**
  - Advanced feature engineering (e.g., OrdinalEncoder).
  - Experiment with alternative models (XGBoost, LightGBM).
  - Explore log transformations 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.