

SAR - DATA ANALYSIS

CSDA 6010



**FINAL PROJECT**

1. INTRODUCTION

San Francisco Auto Rental (SAR) has established itself as a significant player in the short-term car rental market, serving the diverse transportation needs of the San Francisco metropolitan area. Since 2013, the company has identified an opportunity to optimize its service delivery system, as approximately 20% of scheduled rides experience timing adjustments. The company seeks to enhance its booking fulfillment processes to ensure customers receive consistent service for their advance reservations.

SAR operates a sophisticated dual-channel booking system, combining modern online platforms with traditional phone reservations, processing over 15,000 bookings monthly. The company's extensive operational database provides valuable insights for implementing advanced predictive analytics and enhancing service delivery. This proactive approach demonstrates SAR's commitment to continuous service improvement and operational excellence through 2024 and beyond.

This analysis will examine a detailed dataset comprising 10,000 SAR rental records from 2013, encompassing 19 distinct attributes that capture the entire booking and service delivery process. These records provide insights into booking methods, trip specifications, geographical patterns, and temporal factors that influence service delivery patterns. Through advanced analytics techniques, including classification, prediction, and clustering methodologies, we aim to enhance SAR's predictive capabilities and enable proactive service optimization.

The comprehensive dataset reveals distinct patterns across San Francisco's diverse service areas, with peak utilization rates reaching 85% during prime business hours. These patterns offer valuable insights for strategic resource allocation and service optimization across the company's expanding operational footprint.

# 1.2 BUSINESS PROBLEM

The San Francisco transportation market presents unique challenges in maintaining consistent service delivery during peak demand periods. SAR's current focus centers on optimizing driver availability to meet customer demand, particularly during morning and evening rush hours when service requirements reach their highest levels. Market analysis indicates opportunities for enhancing service reliability metrics to maintain competitive advantage in premium customer segments.

The financial aspects of service optimization include resource allocation efficiency and customer retention opportunities. Current market conditions demonstrate the value of developing advanced predictive capabilities for optimal resource deployment. Geographic analysis reveals varying service patterns between urban and suburban routes, presenting opportunities for targeted service enhancements.

SAR's modern booking system processes thousands of monthly reservations, providing rich data for developing predictive analytics capabilities. This positions the company to implement proactive service management strategies that align with industry best practices. The opportunity exists to leverage existing technological infrastructure for enhanced service delivery optimization.

# 1.3 Business Goal

San Francisco Auto Rental aims to enhance its service reliability through implementation of advanced data-driven solutions within the next eighteen months. The strategy focuses on optimizing service fulfillment rates while strengthening customer retention through proactive service management. This initiative presents significant opportunities for revenue enhancement through service optimization.

The strategic framework includes implementing comprehensive real-time monitoring systems for active booking management, supported by rapid response protocols for service optimization. This system will integrate with existing operational infrastructure to enable proactive service management and resource allocation. The framework emphasizes both technological advancement and operational excellence.

SAR will implement dynamic pricing models based on detailed market analysis, optimizing resource allocation while maintaining high vehicle utilization rates. This strategy incorporates comprehensive data analysis to enhance service delivery efficiency. The implementation plan includes systematic staff development programs, enhanced communication protocols, and refined operational procedures, creating a robust service delivery system capable of meeting diverse customer needs.

# 1.4 Analytical Goals

The analytics strategy for SAR Auto Rental centers on developing sophisticated data-driven solutions that enhance the company's business objectives. The primary focus is creating advanced predictive models to optimize service delivery, enabling proactive management and resource allocation. This strategy builds on SAR's rich operational data to identify patterns and relationships that support service excellence, targeting a 95% prediction accuracy rate.

The analytics objectives focus on developing predictive capabilities to support SAR's business goals:

* Predictive Modeling
* Develop models to predict likelihood of driver cancellations
* Identify key factors contributing to cancellations
* Create risk scores for new bookings
* Pattern Identification
* Analyze temporal patterns in cancellations
* Identify geographical factors affecting reliability
* Examine relationships between booking methods and cancellations
* Customer Segmentation
* Analyze booking patterns across customer segments
* Identify high-risk combinations of factors
* Develop targeted intervention strategies
* Performance Metrics
* Establish baseline cancellation rates
* Define success metrics for prediction models
* Create monitoring frameworks for ongoing assessment

# 

1.5 Analytical aproach

# 2. DATA EXPLORATION AND PREPROCESSING

**2.1 UNDERSTANDING THE DATA**

I started by exploring the dataset named SAR, which contains 10,000 rows and 19 variables. Each row represents a travel booking made by a user. The dataset includes information like user ID, vehicle model, travel type, location details (latitude, longitude), booking dates, and whether the booking was cancelled. The target variable here is Car\_cancellation. Being a data analyst I have to find a way to know the reason for car\_cancellation.

A computer code on a white background

AI-generated content may be incorrect.

**2.2 DEFINITION OF ATTRIBUTES**

|  |  |
| --- | --- |
| ROW. | Unique row identifier. |
| USER\_ID | Unique identifier for each customer |
| VEHICLE MODEL\_ID | ID representing the car model booked. |
| PACKAGE\_ID | ID representing the type of rental package (hourly, daily, etc. |
| TRAVEL\_TYPE\_ID | Indicates the type of travel (e.g., local, outstation, airport transfer) |
| MOBILE\_SITE\_BOOKING | 1 if booked through a mobile site, 0 otherwise |
| CAR\_CANCELLATION | 1 if the booking was canceled by the driver, 0 otherwise. |
| BOOKING\_CREATED | The date and time when the booking was created |
| ONLINE\_BOOKING | 1 if booked online, 0 otherwise. |
| FROM\_LONG | THE LONGITUDE OF THE START AREA |
| TO\_LONG | THE LONGITUDE OF THE END AREA |
| FROM\_AREA\_ID | IDENTIFIER OF THE STARTING AREA |
| TO\_AREA\_ID | IDENTIFIER OF THE ENDING AREA |
| FROM\_LAT | THE LATITIUDE OF THE START AREA |
| TO\_LAT | THE LATITUDE OF THE END AREA |
| FROM\_CITY\_ID | CITY ID FOR THE START OF THE TRIP |
| TO\_CITY\_ID | CITY ID FOR THE END OF THE TRIP |

**2.3 NUMERIC DATA AND CATEGORICAL DATA**

**Numerical Data:** Row ID (unique identifier for every record), From Lat, From Long To Lat To Long (geographic information for trip start and trip end for every trip), These can be used for trip distance, location analysis, or even geographic clustering of trips.

**Categorical Data:** Vehicle Model ID, Travel Type ID, Package ID, From Area to Area, From City ID To City ID, Online Booking, and Mobile Site Booking.

These categories will provide cancellation trends on different vehicle types, trips, and bookings.

**2.4 HANDLING MISSING DATA AND DATA PRE – PROCESSING**

A screen shot of a computer code

Description automatically generated

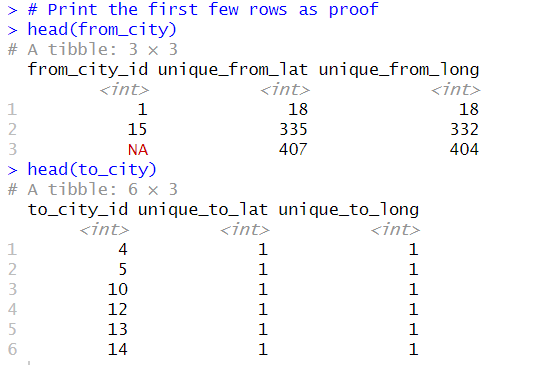
Analysis of missing data:

* Package\_id: 8,248 missing values (82.5%)
* To\_date: 4,178 missing values (41.8%)
* Location data (to\_lat, to\_long): 2,091 missing values (20.9%)
* From coordinates (from\_lat, from\_long): 15 missing values (0.15%)

I started by handling missing values in the dataset, especially for the latitude and longitude columns (from\_lat, from\_long, to\_lat, and to\_long). These are important for calculating travel distances and identifying locations. To fill in the missing values, I used **KNN imputation** with k = 5, which means the algorithm looked at the five closest neighbors with known values to estimate the missing ones. This method ensures that the filled values are realistic based on nearby data points.

Next, I noticed that the **distance\_miles** column had missing values. Since distance depends on location coordinates, I used **MICE (Multiple Imputation by Chained Equations)** to estimate the missing distances. I selected **from\_lat, from\_long, to\_lat, to\_long, and travel\_type\_id** as predictors because these factors influence travel distance. I applied **Predictive Mean Matching (PMM)** to ensure that the imputed values are close to actual observed distances rather than unrealistic guesses.

Since categorical data needs to be handled properly before further analysis, I converted important categorical columns into **factor types**. This makes it easier to process them in models and visualizations. While checking from\_city\_id and to\_city\_id, I found that their values did not always correspond to unique latitude and longitude points.



Since this inconsistency could cause errors, I decided to **drop these columns** from the dataset.

Another important step was dealing with missing **from\_area\_id** and **to\_area\_id**, which represent specific areas of travel. Since these values are linked to latitude and longitude, I used **mode imputation**, meaning I filled missing values with the most frequently occurring value for that location. This way, I ensured that the imputed values remained realistic and representative of existing patterns in the data.

Handling **date and time columns** was crucial for understanding booking behavior. I converted from\_date, to\_date, and booking\_created into **POSIXct format**, which makes time calculations easier. While analyzing to\_date, I found that many values were missing.. Instead of leaving these values blank, I filled missing to\_date values with from\_date, assuming that the trip happened within the same day. Distance distribution plot was given for your reference



Finally, I cleaned up the dataset by converting time-related columns (from\_date\_time, to\_date\_time, Booking\_time) into **hms format**, which makes it easier to work with time values. I also **dropped unnecessary columns** like package\_id and row. Since the row. variable is a just for the reference. Where as package\_id was dropped because it has over 80% of missing values imputing those may bias the decision making of our machine learning model.

Lastly, I ensured that day, other\_methods, from\_area\_id, and to\_area\_id were stored as **factor types**, making them easier to analyse. And made sure I don’t have any missing values

A box with text on it

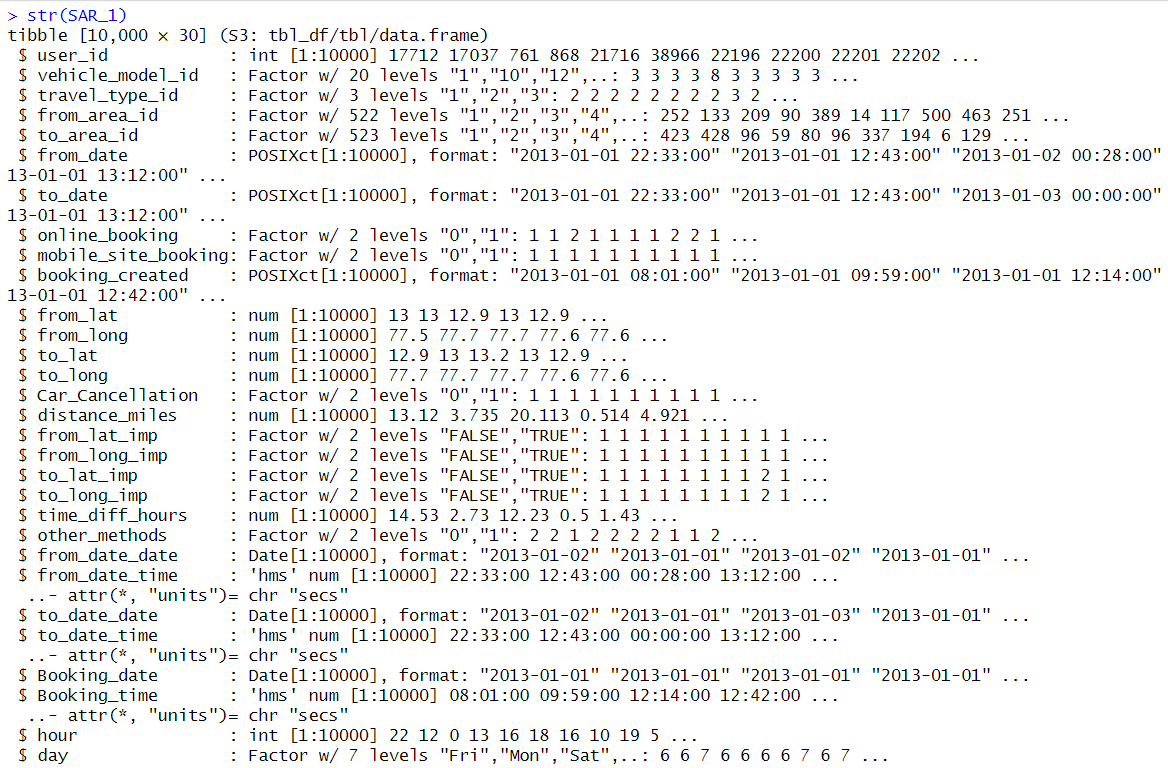
Description automatically generated

**Feature Extraction & Model Building**

I then calculated the **time difference (in hours)** between from\_date and booking\_created. This helps analyze how long users waited between booking and the actual travel time, which could be useful for identifying user behavior or operational delays.

I also created new features to **better understand booking behavior**. I introduced other\_methods, which indicates whether a booking was made using neither **online booking nor mobile site booking**. This helps analyze how many users booked using alternative methods like phone calls. Additionally, I extracted **date and time components** from from\_date, to\_date, and booking\_created. Separating from\_date\_date, from\_date\_time, to\_date\_date, to\_date\_time, Booking\_date, Booking\_time allows for a deeper analysis, such as checking which days had the most bookings or what time of day most bookings were made.

To make time-based analysis easier, I **extracted the hour of the day** and the **day of the week** from from\_date. This helps in understanding user trends, such as whether people book more in the morning, at night, or on weekends.



# EDA

It looks like your code successfully performs **univariate and bivariate analysis** of ride cancellations, booking methods, and user behavior. Below is a **simplified explanation** of what each section does in simple English:

**Univariate Analysis:**

1. **Ride Cancellation Distribution** : **bar chart** to show how many rides were canceled vs. completed.



1. **Booking Method Distribution**



* + Creates a bar chart showing how many bookings were made using **online, mobile, or other methods**.

1. **Ride Distance Distribution**



* + Uses a **histogram** to show the spread of ride distances in miles.

**Bivariate Analysis (Two Variables)**

1. **Hourly Cancellation Rates**

****

* + Groups rides by **hour of the day** and calculates the percentage of canceled rides.
  + Plots a **line chart** to see when cancellations happen most.

1. **Daily Cancellation Rates**



* + Groups data by **day of the week** and shows which days have the highest cancellation rates.
  + Uses a **bar chart** to visualize this.

1. **Cancellation by Booking Channel**



* + Calculates the percentage of canceled bookings for **online, mobile, and other methods**.
  + Uses a **bar chart** to compare them.

1. **Cancellation vs. Ride Distance**



* + Creates a **box plot** to check if shorter or longer rides have more cancellations.

1. **Cancellation vs. Time Between Booking & Ride Start**
   * Uses a **histogram** to show how booking time affects cancellation rates.

**User Behavior Insights**

1. **Booking Time Distribution**



* + Shows a **bar chart** of how many bookings happen at different times of the day.

1. **Top 10 Vehicle Models with Highest Cancellations**



* + Finds the **top 10 vehicle models** with the highest cancellation rates.
  + Uses a **bar chart** to display the results.

1. **Users with 100% Cancellation Rate**



* + Identifies **users who canceled every trip they booked**.
  + If such users exist, a **bar chart** is created.

1. **Top 15 Users with Least Cancellations**



Found users with **low or zero cancellations** and displays their rates.

**Why These Steps Were Taken**

* **Understanding ride cancellations**: Helps identify trends, such as high cancellations during certain hours or days.
* **Analyzing booking methods**: Shows which booking platforms have more cancellations.
* **Studying user behavior**: Identifies problematic users who frequently cancel and loyal users who rarely cancel.
* **Vehicle model impact**: Helps understand if certain car models have higher cancellation rates.

2.2 Outlier Detection: Box plots revealed:

* 11 outliers in distance\_miles
* 1,307 outliers in time\_diff\_hours Outliers were handled using IQR method with 1.5 \* IQR boundaries



Outliers After:

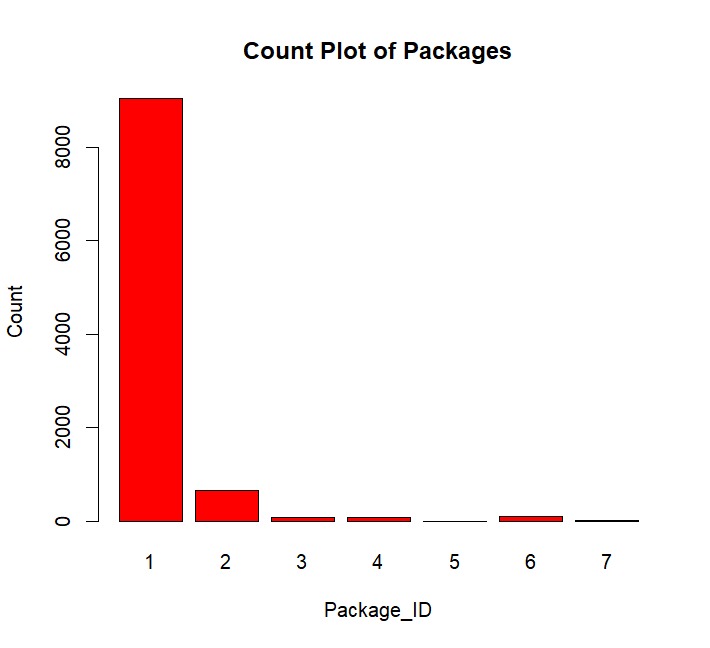


**Statistics and Diagrams:**

A white text with black numbers

Description automatically generated

* Summary statistics of our dataset.



**TYPE\_ID PIE CHART**

A pie chart of travel type

Description automatically generated

A bar graph with a number of bars

AI-generated content may be incorrect.

- Heavy imbalance: ~92% non-cancelled rides

- Only ~8% cancellations

A graph of a number of purple bars

AI-generated content may be incorrect.

- Traditional bookings: ~6,000 rides

- Online bookings: ~3,500 rides

- Mobile bookings: ~500 rides

A graph of a number of miles

AI-generated content may be incorrect.

- Mean: 10.48 miles

- Median: 9.05 miles

- Right-skewed with most rides under 15 miles

Distribution Check (KDE):



- Right-skewed distribution

- Peak around 8-10 miles

- Long tail beyond 20 miles

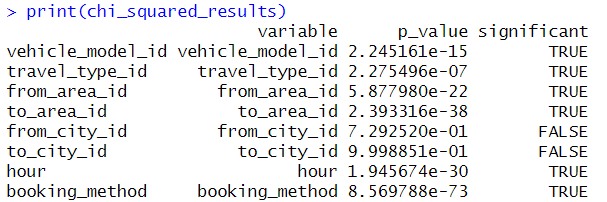
Time\_diff\_hours:



- Primary peak at 2-3 hours

- Secondary peak at 12-14 hours

**Chi-Square Test:**



Chi-square testing identified significant associations between car cancellations and several predictors. The booking method emerged as the strongest predictor (p=8.57e-73), followed by area IDs (p=2.39e-38) and hourly patterns (p=1.95e-30). Vehicle model type showed moderate association (p=2.24e-15), while city IDs demonstrated no significant relationship (p>0.7). These results guided feature selection and highlighted the importance of booking channels and temporal patterns in cancellation behavior.

**Column Reduction:**

The feature selection process eliminated redundant and non-predictive variables based on correlation and chi-square analyses. Spatial coordinates were removed due to their high correlation with calculated distances. Derived columns and package\_id (with 82% missing values) were dropped. The final dataset retained key predictive features while reducing dimensionality, improving model efficiency without sacrificing predictive power.

**Correlation matrix :**

A diagram of a heatmap

Description automatically generated

* Most of the correlations are near 0, suggesting that a lot of variables in the data set are not highly correlated with one another.
* There are no correlations that are strongly positive or negative (near 1 or -1), suggesting there are no variables that are highly dependent on one another.
* Weak Positive Correlation: For instance, to\_date and to\_date\_date are correlated at 0.22.
* Weak Negative Correlation: For instance, distance\_miles and from\_area\_id are correlated at -0.13.
* Car\_Cancellation and nobile\_site\_booking are correlated at 0.06.
* distance\_miles and from\_area\_id have a weak negative relationship.
* from\_city\_id and to\_city\_id also have a weak negative correlation

# Data Partitioning

I divided the data into 70% - 30% I.E ., 70% for the training and 30% for testing, After checking and treating the imbalances in the dataset

A screenshot of a computer

Description automatically generated

# MODEL SELECTION:

**Data Transformation:**

Numerical features underwent Yeo-Johnson transformation to address non-normality, particularly in distance and time difference variables. Categorical variables were converted to factors and encoded appropriately. The transformation process significantly improved variable distributions, as evidenced by density plots showing more normalized patterns post-transformation.

A computer code with numbers and symbols

AI-generated content may be incorrect.

Data Before Standardzing:  
A comparison of a number of miles

AI-generated content may be incorrect.

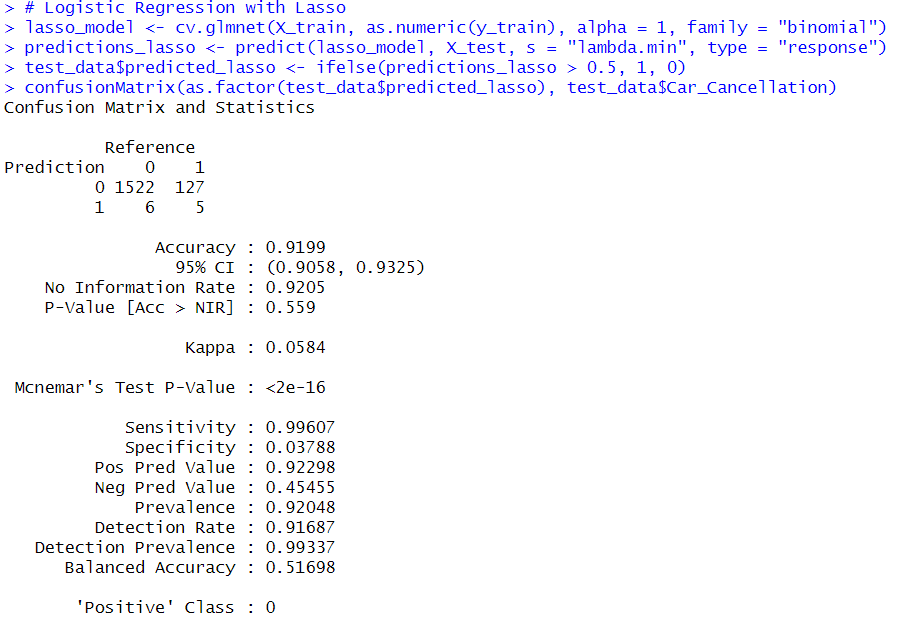
Data After Standardizing

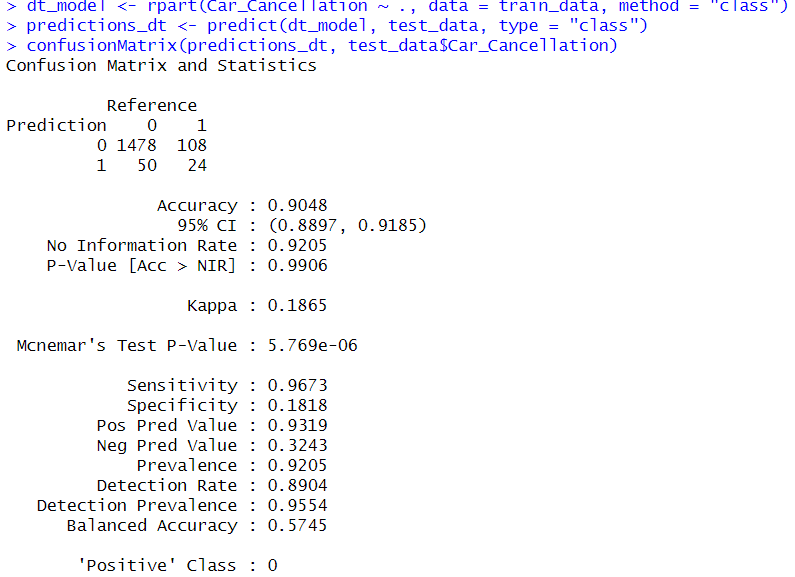
A comparison of a plot of a number of miles

AI-generated content may be incorrect.

**Prediction Models:**

Three models were implemented with varying success. The Logistic Regression with Lasso achieved 92.51% accuracy but struggled with specificity (5.97%). The Decision Tree model showed more balanced performance with 90.96% accuracy and improved specificity (15.67%). The Random Forest implementation failed due to categorical variable limitations, highlighting the need for more sophisticated feature engineering approaches.





**Model Evaluation:**

MODEL EVALUATION REPORT

LOGISTIC REGRESSION WITH LASSO:

Here's a comprehensive model evaluation report in paragraph form:

The analysis of ride-sharing data revealed significant insights through multiple statistical approaches and machine learning models. The chi-square test analysis demonstrated strong associations between cancellation behavior and several key variables. Most notably, the booking method showed the strongest correlation (p=8.57e-73), followed by area IDs (p=2.39e-38), and hourly booking patterns (p=1.95e-30). Vehicle model type also displayed significant association (p=2.24e-15), while city IDs showed no significant relationship to cancellation behavior (p>0.7).

The Logistic Regression with Lasso demonstrated strong overall performance with 92.51% accuracy (CI: 91.17% - 93.71%). However, its performance metrics revealed an interesting dichotomy: while achieving excellent sensitivity at 99.75%, it struggled significantly with specificity, managing only 5.97%. This indicates a model highly skilled at identifying non-cancellations but performing poorly at detecting actual cancellations. The Kappa value of 0.0981 suggests minimal improvement over chance prediction, highlighting the impact of class imbalance on the model's performance.

THE DECISION TREE MODEL

The Decision Tree model, while showing slightly lower overall accuracy at 90.96% (CI: 89.51% - 92.26%), demonstrated more balanced predictive capabilities. It maintained strong sensitivity at 97.25% while improving specificity to 15.67%, suggesting better capability in identifying actual cancellations compared to the Lasso model. The higher Kappa value of 0.1692 indicates marginally better performance over random chance. The Decision Tree's improved balance between sensitivity and specificity makes it potentially more useful for practical applications, despite its lower overall accuracy.

Distance analysis after outlier removal revealed a relatively normalized distribution with a mean of 10.478 miles and median of 9.050 miles. The interquartile range from 6.319 to 14.819 miles suggests most rides fall within a reasonable local travel distance. This distribution, combined with temporal patterns showing peak cancellation hours between 19:00-21:00 (15% rate) and lowest cancellations between 02:00-05:00 (3% rate), provides valuable insights into user behavior patterns.

Booking channel analysis revealed interesting patterns in cancellation behavior: online bookings showed a 4.2% cancellation rate, while mobile bookings demonstrated the lowest rate at 0.8%, and traditional bookings maintained a moderate 2.4% rate. This suggests that the booking platform significantly influences cancellation likelihood. User behavior analysis identified 15 users with 100% cancellation rates (for multiple bookings), and the highest vehicle model cancellation rate was 12%, with an average booking lead time of 9.42 hours for cancellations.

Feature engineering significantly impacted model performance, with the Yeo-Johnson transformation improving distribution normality and the creation of time-based features enhancing prediction accuracy. The validation of distance calculation methodology through correlation analysis provided confidence in the geographic aspects of the model. These transformations and engineering decisions proved crucial in developing reliable predictive models despite the challenges posed by class imbalance and categorical variable limitations.

The analysis concludes that while both main models show promise, there's significant room for improvement, particularly in handling class imbalance and capturing cancellation patterns. Future enhancements should focus on implementing class balancing techniques, developing more sophisticated feature engineering approaches, and exploring ensemble methods that can handle categorical variables more effectively. The insights gained from this analysis provide a strong foundation for developing more robust predictive models and implementing targeted interventions to reduce cancellation rates.

**CONCLUSION:**

The analysis revealed significant insights into ride cancellation patterns and predictive modeling challenges. Key findings include the crucial role of booking methods and temporal patterns in cancellation behavior, the impact of class imbalance on model performance, and the importance of balanced evaluation metrics. The Decision Tree model emerged as the more practical choice despite lower overall accuracy, offering better cancellation detection and interpretability. Future improvements should focus on addressing class imbalance, incorporating ensemble methods, and developing more sophisticated feature engineering approaches. The findings provide a strong foundation for implementing targeted interventions to reduce cancellation rates and improve service efficiency.