**San Francisco Auto Rental Analytics**

Table of Contents

[Executive Summary 3](#_Toc196782149)

[1. Project Introduction 4](#_Toc196782150)

[2. Business & Analytics Goals 4](#_Toc196782151)

[Business Problem 4](#_Toc196782152)

[Business Goal 5](#_Toc196782153)

[Business Objectives 5](#_Toc196782154)

[Analytics Approach 5](#_Toc196782155)

[3. Data Preprocessing 5](#_Toc196782156)

[Attributes Definition 5](#_Toc196782157)

[Data Exploration 7](#_Toc196782158)

[Missing Value Analysis 7](#_Toc196782159)

[Handling Missing Values 8](#_Toc196782160)

[Dropped Variables: 10](#_Toc196782161)

[Zero & NaN Value Checks: 10](#_Toc196782162)

[Consistency & Data Integrity: 10](#_Toc196782163)

[Final Notes: 10](#_Toc196782164)

[4. Predictor Analysis and Relevancy 10](#_Toc196782165)

[Cancellation Rate Analysis 11](#_Toc196782166)

[Statistical Tests for Predictor Relevancy 12](#_Toc196782167)

[Random Forest Feature Importance Analysis 13](#_Toc196782168)

[5. Data engineering and transformation 15](#_Toc196782169)

[Temporal Feature Extraction 15](#_Toc196782170)

[Trip Distance Calculation 17](#_Toc196782171)

[Handling trip\_length Zero Values 20](#_Toc196782172)

[Encoding Booking Types 20](#_Toc196782173)

[Handling Sparse Categories in High-Cardinality Features 21](#_Toc196782174)

[Vehicle Model Grouping 22](#_Toc196782175)

[Aggregating User-Level Features 22](#_Toc196782176)

[Data Quality Validation 23](#_Toc196782177)

[Finalized Transformed Data 23](#_Toc196782178)

[6. Dimension Reduction 24](#_Toc196782179)

[Handling Categorical Variables: Fisher’s Exact Test 24](#_Toc196782180)

[Handling Numerical Variables: Correlation Analysis 25](#_Toc196782181)

[Feature Selection via Random Forest 26](#_Toc196782182)

[Feature Importance Results 26](#_Toc196782183)

[Justification for No PCA (Principal Component Analysis) 27](#_Toc196782184)

[Final Dataset: 27](#_Toc196782185)

[7. Data Partitioning Methods 28](#_Toc196782186)

[8. Model selection 28](#_Toc196782187)

[Candidate Models Considered: 28](#_Toc196782188)

[Model Evaluation Metrics 28](#_Toc196782189)

[9. Model Fitting, Validation Accuracy, and Test Accuracy 29](#_Toc196782190)

[Baseline Random Forest Model Performance 29](#_Toc196782191)

[Hybrid Random Forest (Class Weighting + Oversampling) Performance 29](#_Toc196782192)

[Oversampled Logistic Regression Performance 29](#_Toc196782193)

[10. Reporting Model Performance 30](#_Toc196782194)

[11. Model Evaluation 30](#_Toc196782195)

[Confusion Matrix Analysis 30](#_Toc196782196)

[ROC Curves and AUC-ROC Scores 31](#_Toc196782197)

[Feature Importance Analysis 33](#_Toc196782198)

[Random Forest Feature Importance (Baseline RF & Hybrid RF) 34](#_Toc196782199)

[Logistic Regression Feature Importance 35](#_Toc196782200)

[Cost-Sensitive Learning Analysis 36](#_Toc196782201)

[12. Observations and Conclusion 36](#_Toc196782202)

[Key Observations from Model Evaluation 36](#_Toc196782203)

[Achievement of Business and Analytics Goals 37](#_Toc196782204)

[Final Model Recommendation 37](#_Toc196782205)

[Conclusion and Future work 37](#_Toc196782206)

[References 38](#_Toc196782207)

# Executive Summary

This project was conducted to help San Francisco Auto Rental (SAR) address the challenge of frequent ride cancellations, which cause operational inefficiencies and impact customer satisfaction. Using historical ride booking data, I applied data cleaning, feature engineering, and predictive modeling techniques to identify patterns that signal high-risk bookings. After analyzing key factors like user cancellation rates, booking times, and trip lengths, I evaluated three classification models: Baseline Random Forest, Hybrid Random Forest (class weighting + oversampling), and Oversampled Logistic Regression. Among these, Oversampled Logistic Regression achieved the best balance between detecting cancellations and overall business impact, making it the recommended model for early cancellation prediction.

Based on this analysis, I recommend that SAR:

* Deploy Oversampled Logistic Regression to proactively flag likely cancellations and take preventive action.
* Focus on high-risk users and peak booking hours when allocating resources or offering incentives.
* Continue refining the model by incorporating external data like traffic conditions and real-time driver ratings.

These findings directly support SAR’s business goals of improving service reliability, reducing lost revenue from cancellations, and increasing customer loyalty through data-driven decision-making.

**Sai Pranup Shobanaboina**  
**Data Analyst | San Francisco Auto Rental (SAR)**  
**05/01/2025**

# Project Introduction

Frequent ride cancellations are a major challenge for San Francisco Auto Rental (SAR), a short-term rental service offering bookings through online platforms, mobile sites, and traditional reservations. Cancellations by drivers disrupt service delivery, cause operational inefficiencies, and erode customer trust, ultimately impacting SAR’s revenue and market competitiveness.

This project analyzes SAR’s historical ride booking data to uncover patterns in cancellation behavior and develop predictive solutions. Using a combination of exploratory analysis, feature engineering, and supervised modeling, the goal is to identify high-risk bookings in advance and support proactive interventions. By applying a structured analytics approach, SAR can enhance ride reliability, optimize fleet operations, and strengthen customer loyalty. Early modeling results provide actionable insights into the drivers of cancellations and offer a data-driven pathway to reduce ride disruptions and improve overall service quality.

# Business & Analytics Goals

## Business Problem

San Francisco Auto Rental (SAR) faces a high incidence of ride cancellations, leading to operational inefficiencies, lost revenue, and customer dissatisfaction. Early identification of high-risk bookings remains a critical challenge impacting fleet utilization and service reliability.

## Business Goal

Leverage predictive analytics to identify high-risk bookings in advance, allowing SAR to proactively mitigate cancellations and enhance service fulfillment.

## Business Objectives

* Identify key factors influencing ride cancellations using historical ride booking data.
* Develop a reliable predictive system to classify future ride bookings by cancellation risk within the next business quarter.
* Improve ride fulfillment rates by enabling SAR to take early corrective actions on bookings flagged as high risk.
* Reduce operational losses and driver idle time associated with canceled rides by implementing data-driven strategies over the next six months.

## Analytics Approach

1. **Define the Problem:**

Understand how ride-level data can support early cancellation detection and risk mitigation.

1. **Collect and Preprocess Data:**

Extract historical ride data, clean missing values, standardize formats, and engineer new features for better predictive power.

1. **Explore Data and Identify Patterns:**

Analyze booking behaviors, temporal trends, and cancellation hotspots to inform modeling decisions.

1. **Develop Predictive Models:**

Train and validate classification models to predict ride cancellation risks based on engineered features.

1. **Evaluate Model Performance:**

Assess model outcomes using accuracy, recall, specificity, and business cost-sensitive metrics.

1. **Present Insights and Recommendations:**

Translate model outputs into actionable strategies to minimize cancellations and optimize fleet management.

# Data Preprocessing

This section explains the steps taken to prepare the dataset for analysis. Each action is presented with clear justifications to ensure alignment with academic expectations.

### Attributes Definition

The dataset consists of multiple attributes capturing ride details, booking methods, location data, and cancellation status. The following Table 1 categorizes each attribute based on its type:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| | **Attribute** | | --- | | | **Description** | | --- | | **Category** |
| |  |  | | --- | --- | | row# |  | | Unique identifier for each record. | Categorical - Nominal |
| |  |  |  | | --- | --- | --- | | user\_id |  |  | | Unique identifier for customers. | Categorical - Nominal |
| |  |  |  | | --- | --- | --- | | vehicle\_model\_id |  |  | | ID representing the type of vehicle driven. | Categorical - Nominal |
| |  |  | | --- | --- | | package\_id |  | | Identifies the travel package (e.g., 4 hours & 40 km). | Categorical - Ordinal |
| |  |  |  | | --- | --- | --- | | travel\_type\_id |  |  | | ID representing the type of travel (e.g., point-to-point, hourly). | Categorical - Nominal |
| |  |  |  | | --- | --- | --- | | from\_area\_id |  |  | | Represents the starting area of the ride. | Categorical - Nominal |
| |  |  | | --- | --- | | to\_area\_id |  | | Represents the destination area of the ride. | Categorical - Nominal |
| |  |  | | --- | --- | | from\_city\_id |  | | Unique identifier for the starting city. | Categorical - Nominal |
| |  |  | | --- | --- | | to\_city\_id |  | | Unique identifier for the destination city. | Categorical - Nominal |
| |  |  | | --- | --- | | from\_date |  | | Timestamp for the start of the trip. | Datetime |
| |  | | --- | | to\_date | | Timestamp for the end of the trip. | Datetime |
| |  |  | | --- | --- | | online\_booking |  | | 1 if the booking was made online, 0 otherwise. | Binary |
| |  |  |  | | --- | --- | --- | | mobile\_site\_booking |  |  | | 1 if the booking was made via mobile, 0 otherwise. | Binary |
| |  |  |  | | --- | --- | --- | | booking\_created |  |  | | Timestamp for when the booking was created. | Datetime |
| |  |  |  | | --- | --- | --- | | from\_lat and from\_long |  |  | | Latitude and longitude of the starting location respectively. | Geospatial |
| |  |  |  | | --- | --- | --- | | to\_lat and to\_long |  |  | | Latitude and longitude of the destination location respectively. | Geospatial |
| |  |  |  | | --- | --- | --- | | Car\_Cancellation |  |  | | Target variable indicating if the ride was canceled (1 = Yes, 0 = No). | Binary |

Table 1: Attributes Definition

**Key Considerations:**

* **Date/Time Variables** were later transformed to extract meaningful features (e.g., hour of the day, day of the week).
* **Geospatial Coordinates** were retained as spatial data but were not treated as standard numerical features in correlation analyses.
* **Categorical Variables** include nominal, ordinal, and binary classifications.

## Data Exploration

### Missing Value Analysis

A green and orange pie chart

Description automatically generated

Figure 1: Total missingness proportion

Overall, 18.26% of dataset is missing as shown in the Figure 1, but we should observe the variables that are causing this missingness and handle them.

A graph with numbers and text

Description automatically generated

Figure 2: Missing values heatmap

The dataset contained missing values in several key attributes, which were addressed systematically. The following table summarizes the missing values and the actions taken:

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Missing Values** | **Proportion (%)** | **Action Taken** |
| package\_id | 8248 | 82.48% | Replaced with "Unknown". |
| from\_area\_id | 15 | 0.15% | Replaced with "Unknown". |
| to\_area\_id | 2091 | 20.91% | Replaced with "Unknown". |
| from\_city\_id | 6294 | 62.94% | Dropped due to redundancy and high missingness. |
| to\_city\_id | 9661 | 96.61% | Dropped due to redundancy and high missingness. |
| to\_date | 4178 | 41.80% | Flagged with missing\_to\_date and dropped to\_date |
| from\_lat/from\_long | 15/15 | 0.15%/0.15% | Imputed with the median. |
| to\_lat/to\_long | 2091/2091 | 20.91%/20.91% | Imputed with the median. |

Table 2: Missing value proportions

### Handling Missing Values

#### Categorical Variables:

* **Nominal Variables**:
  + Missing values in package\_id, from\_area\_id, and to\_area\_id were replaced with "Unknown."
  + Dropping these rows would result in significant data loss, and "Unknown" ensures these records are included in the analysis.
  + Missing categorical values were replaced with "Unknown" to preserve information rather than deleting records, as suggested by Allison (2002), who emphasizes that imputation methods for categorical variables are preferable to deletion to avoid loss of valuable data and potential bias in analysis.
* **Sparse Categories**:
  + Sparse categories in package\_id (e.g., single-record categories like 5) were grouped into "Other."
  + Grouping sparse categories reduces noise and improves model robustness.
* **Binary Variables**:
  + Variables like online\_booking and mobile\_site\_booking had no missing values and were retained as is.

#### Spatial Data (GPS Coordinates):

Missing values in from\_lat, from\_long, to\_lat, and to\_long were imputed using the **median**.

**Justification**: Mean imputation is sensitive to outliers, while median imputation preserves location-based trends (Junninen et al., 2004). Since geospatial data has natural clustering, KNN imputation was initially considered but discarded due to computational inefficiency and risk of artificial bias (Beretta & Santaniello, 2016). No missing values remained after imputation, ensuring dataset completeness before transformation.

**Note**: GPS coordinates were treated as spatial data, not numerical data, and no transformations (e.g., calculating distance) were applied at this stage.

#### **Datetime Variables:**

Missing values in to\_date were flagged using the missing\_to\_date variable and to\_date was removed from the dataset.

**A blue and red squares with white text

Description automatically generated**

Table 3: Cancellations by missing\_to\_date

**Justification:**

The absence of to\_date could indicate that a trip was not completed. The stacked bar chart shows that trips with missing to\_date have a higher proportion of cancellations. Instead of imputing to\_date, which would introduce assumptions about trip completion, a binary flag (missing\_to\_date) was created to retain information about missingness.

This preserves the original data structure while allowing the model to learn from the presence or absence of to\_date.

### Dropped Variables:

* **from\_city\_id and to\_city\_id**: They were removed due to their high missingness rates (62% and 96%, respectively). Retaining these variables could introduce bias and reduce model reliability, as high missingness often affects statistical validity (Little & Rubin, 2019). Additionally, research suggests that categorical location identifiers may not effectively capture travel patterns compared to geospatial data (Zhang et al., 2021).

**Justification**: To compensate for this removal, trip\_length was derived from geospatial coordinates (from\_lat, from\_long, to\_lat, to\_long), providing a more precise measure of travel distance. Studies have shown that geospatial analytics enhances travel behavior analysis by capturing continuous spatial variations rather than static categorical labels (Li et al., 2020). Furthermore, from\_area\_id and to\_area\_id were retained to preserve regional travel patterns, ensuring key spatial dynamics influencing ride cancellations remain available for analysis.

### **Zero & NaN Value Checks:**

* A check for zeroes and NaN was performed before the missing values check, there were nothing of that type, but except for the binary variables there are zeroes which are meant to be.
* After preprocessing, a final validation was performed to check for any remaining missing values, zeros in critical variables, or NaN values.
* The dataset was found to have **no missing values, zero errors, or NaN values** after applying the preprocessing steps.

### **Consistency & Data Integrity:**

* **No anomalies** were found in binary indicators like online\_booking and mobile\_site\_booking after validation.
* **Geospatial data** (GPS coordinates) were retained only for distance calculations but were otherwise not treated as numerical variables.

## Final Notes:

* Preprocessing ensured data consistency while preserving key relationships for analysis.
* Column-wise **missing value summary returned 0** for all attributes.
* The dataset was confirmed to be **clean, complete, and structured** for modeling without requiring further imputation.
* Data exploration insights were further refined in the predictor analysis section, helping determine the most relevant features for cancellation prediction.

# Predictor Analysis and Relevancy

The goal of this section is to identify which attributes in the dataset have a meaningful relationship with ride cancellations (Car\_Cancellation). This helps in selecting relevant predictors for modeling while excluding redundant or low-impact features.

## Cancellation Rate Analysis

**Why Conduct This Analysis?**Cancellation rates were computed for categorical variables to determine if specific categories are associated with a higher likelihood of cancellation. Understanding these patterns ensures that variables with strong predictive potential are retained for modeling.

**Methodology:**

* Grouped data by each variable of interest (e.g., package\_id, travel\_type\_id).
* Calculated the cancellation rate for each category as the mean of the Car\_Cancellation variable (1 = Canceled, 0 = Completed).
* Counted the number of records in each category to identify sparse groups.

|  |  |  |  |
| --- | --- | --- | --- |
| **Category** | **Levels (Examples)** | **Cancellation Rate (%)** | **Count** |
| **Package ID** | 6, 7, 3, Unknown | 15.4%, 13.6%, 9.7%, 7.9% | 104, 22, 82, 8248 |
| **Travel Type ID** | 2 (Point-to-Point), 3 (Hourly) | 8.1%, 5.0% | 7909, 1752 |
| **From Area ID** | Multiple Areas (523 unique values) | Highly variable | Small counts for many categories |
| **To Area ID** | Multiple Areas (480 unique values) | Highly variable | Small counts for many categories |

Table 4: Cancellation rate analysis

**Key Insights from Cancellation Rate Analysis:**

1. **High-Cancellation Categories:**

* Packages 6 (15.4%) and 7 (13.6%) exhibited the highest cancellation rates, suggesting potential operational challenges or customer dissatisfaction for these options.
* Certain from\_area\_id categories, such as 34 (40.0%) and 49 (25.0%), showed extremely high cancellation rates, indicating localized issues.

1. **Sparse Categories:**  
   Sparse groups with low counts (e.g., to\_area\_id with a count of 1) often displayed inflated cancellation rates, necessitating careful treatment to prevent skewed modeling results.

These insights guided the decision to retain critical variables such as package\_id, from\_area\_id, and to\_area\_id, while treating sparse categories to reduce noise. However, the limitations of categorical analysis, such as the inability to capture interactions or continuous predictors, required the use of a more robust modeling approach.

## Statistical Tests for Predictor Relevancy

**Chi-Square Test for Categorical Variables:**

Chi-square tests were conducted to assess the statistical significance of relationships between categorical predictors and ride cancellations. However, due to sparse categories, the chi-square approximation was unreliable, producing warning messages.

**Fisher's Exact Test for Categorical Variables:**

Fisher’s Exact Test was attempted as an alternative for handling low-frequency categories, but computational limitations arose due to large contingency tables.

A screenshot of a computer code

Description automatically generated

Figure 3: Chi-square test

A computer code with red text

Description automatically generated

Figure 4: Fisher exact test

**Handling Sparse Data - Post Feature Engineering:**

**A graph with numbers and lines

Description automatically generated**

Figure 5: Data sparsity for from\_area\_id

Sparse categories were not grouped at this stage because feature engineering will derive new attributes from raw data.

Planned Enhancements:

* For categorical variables (e.g., from\_area\_id, to\_area\_id, package\_id):
* Feature extraction will be performed (e.g., grouping based on historical booking frequency or cancellation rate).
* Binning strategies will be applied post-transformation to ensure categories with low representation are meaningfully grouped.

## Random Forest Feature Importance Analysis

Random Forest was chosen for its robustness in handling sparse data, mixed data types (categorical, numerical, geospatial), and interactions between predictors. Unlike traditional statistical tests like Chi-Square or Fisher’s Exact, which struggled with sparse categories (e.g., computational limitations and invalid approximations), Random Forest can effectively manage these complexities while providing interpretable feature importance metrics.

**Why Random Forest?**

Due to limitations with traditional statistical tests, Random Forest feature importance was used to assess predictor relevancy. This method is particularly effective for:

* Handling sparse categorical variables without one-hot encoding.
* Identifying non-linear relationships between predictors and cancellations.
* Providing a ranked importance score for each feature.

A graph of data with text

Description automatically generated with medium confidence

Figure 6: Random Forest Feature importance analysis

A **Random Forest feature selection** was trained to compute feature importance based on:

1. **%IncMSE (Percentage Increase in Mean Squared Error):** Measures how much model error increases when a feature is removed. A higher value indicates that the variable is critical for prediction accuracy.
2. **IncNodePurity:** Evaluates how effectively a feature splits decision tree nodes. A higher value implies greater importance in differentiating classes.

**Insights and Implications as observed from the Figure 6:**

* **Temporal Influence**: from\_date and booking\_created are among the top contributors, indicating that ride cancellations are strongly influenced by when the trip was booked and when it was scheduled to start.
* **Geographic Factors**: The importance of **spatial predictors** such as from\_lat and to\_lat suggests that cancellation likelihood varies based on the ride’s starting and ending locations. This implies that certain areas may experience higher cancellation rates due to congestion, accessibility issues, or other location-based factors.
* **Booking Behavior**: online\_booking is a notable predictor, indicating that online bookings have a strong relationship with cancellation behavior. Similarly, mobile\_site\_booking has lower influence, suggesting that mobile-based bookings might be less predictive of cancellations.
* **User-Level Patterns**: The **user\_id** variable has a significant influence, meaning that user behavior plays a major role in determining cancellations. This suggests that frequent users or users with a history of cancellations contribute significantly to the model's predictive power.
* **Location-Based Identifiers**:
  + from\_area\_id and to\_area\_id are moderately important, reinforcing that cancellation rates vary by region.
  + missing\_to\_date has lower importance but still provides relevant information, confirming that the absence of a trip completion timestamp is weakly associated with cancellations.

**Less Relevant Predictors**: Variables such as package\_id and travel\_type\_id show minimal influence, indicating that while they may have some effect on cancellations, they are not among the strongest predictors.

**Adjustments Based on These Insights:**

* **Prioritize geospatial and temporal feature engineering:** Given the significance of **lat/lon coordinates** and **time-based variables**, further feature engineering (such as **traffic congestion levels or historical demand trends**) could enhance predictive power.
* **User-level feature engineering:** Since **user\_id** is an important predictor, **aggregating user-level statistics (e.g., past cancellations, booking frequency)** could provide additional insights.
* **Booking platform impact:** Given the significance of **online\_booking**, a more detailed breakdown of booking platforms could refine insights.

**Note:** Performing Data Engineering and Transformation before Dimension Reduction is a logical sequence in this case, as it allows in transforming raw data into derived features (e.g., trip\_length, day\_of\_week) often enhances interpretability and predictive power. These derived features may capture patterns that raw features (e.g., from\_date, to\_lat) do not fully express and by engineering features beforehand, we ensure that meaningful variables are not overlooked during dimensionality reduction.

# Data engineering and transformation

Feature transformations were performed to optimize data usability while preserving information. This section outlines the modifications applied to enhance model interpretability and performance.

## Temporal Feature Extraction

**Action Taken:**

* Extracted hour\_of\_day, day\_of\_week, and month\_of\_year from from\_date and booking\_created to remove timestamp details.
* Created a booking\_time\_bucket categorical variable to classify bookings into meaningful time slots:
  + Early Morning (12:00 AM - 5:59 AM)
  + Morning (6:00 AM - 11:59 AM)
  + Afternoon (12:00 PM - 5:59 PM)
  + Evening (6:00 PM - 11:59 PM)

**Justification:**

* Timestamp values are not directly interpretable but can provide meaningful insights when transformed.
* Identifying peak booking times helps capture trends in ride cancellations.

**Peak Booking & Cancellation Hours**

**A graph of blue and red bars

Description automatically generated**

Figure 7: Peak Booking Hours vs. Cancellations

* Booking activity is highest between **4 AM - 11 AM and 4 PM - 7 PM.**
* Ride cancellations (red) are more frequent during these peak hours, particularly around **early mornings (4-7 AM) and evenings (5-7 PM) as observed from the above Figure 7.**
* This aligns with typical rush hour traffic and early morning work commutes when users may cancel due to last-minute changes in plans or long wait times.

**Booking Time Bucket vs. Cancellations**

A blue and red graph

Description automatically generated

Figure 8: Booking Time Bucket vs. Cancellations

* The above (Figure 8) bar plot demonstrates cancellation distributions across different time buckets.
* Cancellations are notably higher during early morning hours, likely due to late-night uncertainty or schedule changes.

## Trip Distance Calculation

**Action Taken:**

* Derived trip\_length using the Haversine formula, calculating the great-circle distance between pickup and drop-off locations.
* Dropped geospatial coordinates (from\_lat, from\_long, to\_lat, to\_long) as they were no longer required.

**Justification:**

* Latitude and longitude represent spatial data rather than numeric variables useful for direct modeling.
* Trip length is a single, interpretable feature that captures ride distance.

**Trip Length Boxplot**

A graph with a blue line

Description automatically generated

Figure 9: Trip length Box-plot

* The above (Figure 9) boxplot indicates a **right-skewed distribution**, with a few trips classified as outliers.
* Outliers may be **unusually long trips** or errors in data entry, but since they are valid values within operational constraints, no transformation or removal was applied.
* The presence of outliers suggests that a few rides had significantly longer distances than the majority, which could influence cancellation patterns.

**Trip Length Histogram**

A graph of a trip length

Description automatically generated

Figure 10: Distribution of Trip length

* The **distribution of trip\_length** is shown in the histogram above (Figure 10), demonstrating that most trips are short-distance (0-30 km), with fewer long-distance rides.
* This insight helps in understanding how trip lengths vary and may impact cancellation likelihood.

**Trip Length vs. Cancellations (Density Plot)**

A graph of a number of cars

Description automatically generated

Figure 11: Trip length by Ride Cancellations

* As observed from the above density chart (Figure 11), Shorter trips (0-10 km) have a higher cancellation rate (shown by the red density being higher in this range).
* Longer trips (above 15 km) are less frequently canceled (blue density dominates in this range).
* There are peaks in non-canceled trips at medium distances (~15-25 km), suggesting these trips are more stable.
* The overlapping regions indicate where cancellations occur more frequently compared to completed rides.

### Handling trip\_length Zero Values

**Issue Identified:**

* trip\_length had two zero values, likely due to missing or incorrect data.

**Action Taken:**

* If trip\_length = 0 and Car\_Cancellation = 0, replaced with median trip\_length of completed rides.
* If trip\_length = 0 and Car\_Cancellation = 1, retained 0, assuming the trip was never taken.

**Final Check: No remaining zero values** in trip\_length after imputation.

## Encoding Booking Types

**Action Taken:**

* Created a new traditional\_booking feature to categorize booking methods:
  + 1 → If both online\_booking = 0 and mobile\_site\_booking = 0 (traditional call-based bookings).
  + 0 → Otherwise (booked via online or mobile).

**Justification:**

* Analysis revealed that traditional bookings have a significantly lower cancellation rate (3.86%), compared to online/mobile bookings (12.9%).
* The bar chart (Figure 12) visually confirms this trend, highlighting a stronger commitment to traditional bookings.
* Retaining traditional\_booking as a feature allows the model to capture behavioral differences in booking methods, improving prediction accuracy.

**Supporting Analysis:**

The summary Table 5 below quantifies the cancellation rates by booking type:

|  |  |  |  |
| --- | --- | --- | --- |
| **Booking Type** | **Total Trips** | **Canceled Trips** | **Cancellation Rate (%)** |
| **Online/Mobile (0)** | 3,957 | 510 | 12.9% |
| **Traditional (1)** | 6,043 | 233 | 3.86% |

Table 5: Cancellation rates by booking type

**Cancellation Rate by Booking Type (Stacked Bar Chart)**

A blue and red graph

Description automatically generated

Figure 12: Cancellation Rate by Booking type

Figure 12 above demonstrates that online/mobile bookings have a higher proportion of cancellations than traditional bookings.

These insights reinforce the decision to include traditional\_booking as a predictor in the model.

## Handling Sparse Categories in High-Cardinality Features

**Binning Low-Frequency Areas**

* **Variables:** from\_area\_id, to\_area\_id
* **Action Taken:**
  + Identified areas with fewer than 50 rides and merged them into an "Other" category.
  + New variables from\_area\_binned and to\_area\_binned replaced the original columns.

**Justification:**

* High-cardinality categorical features can increase model complexity without adding significant predictive power.
* Grouping rare categories reduces overfitting while preserving location-based patterns.

## Vehicle Model Grouping

The Vehicle Model ID represents the type of vehicle driven for each ride, indirectly linking it to driver behavior and service quality. An initial analysis of cancellation rates by vehicle model revealed extreme sparsity, with many models having very few bookings. The cancellation rates varied significantly, but due to low sample sizes for certain models, the insights were inconsistent.

To address this, I decided to group vehicle models into broader categories based on their booking frequency. This allows for a more meaningful comparison and reduces noise from rarely used vehicle models.

**Action Taken:**

* Grouped vehicle models with fewer than 20 rides into an "Other" category.
* Converted vehicle\_model\_id into a categorical factor.

**Justification:**

* Some vehicle models had very few bookings, making them unreliable predictors.
* Aggregating improves model robustness while preserving key distinctions.

## Aggregating User-Level Features

**Action Taken:**

* Created user-based aggregated features:
  + user\_total\_bookings → Total trips per user.
  + user\_cancellation\_rate → Percentage of canceled trips.
  + user\_avg\_trip\_length → Average trip distance per user.
* Dropped user\_id after feature extraction.

**Justification:**

* User-level behavioral patterns (repeat cancellations, trip frequency) are valuable insights.
* Retaining user\_id introduces data leakage, so we extract useful information before removing it.

**User Cancellation Rate Distribution**

A graph of a number of users cancellation rate

Description automatically generated

Figure 14: Distribution of user cancellation rate

Most users rarely cancel trips, but a few exhibit high cancellation behavior as observed from the Figure 14.

## Data Quality Validation

**Final Check:**

* **Zero Summary Analysis**:
  + Verified zero counts across variables.
  + Missing values fully handled (none remaining).
* **Final Column List**: vehicle\_model\_id, package\_id, travel\_type\_id, from\_date, online\_booking, mobile\_site\_booking, booking\_created, Car\_Cancellation, missing\_to\_date, day\_of\_week, hour\_of\_day, month\_of\_year, trip\_length, booking\_time\_bucket, traditional\_booking, user\_total\_bookings, user\_cancellation\_rate, user\_avg\_trip\_length, from\_area\_binned, to\_area\_binned
* Dataset is now cleaned and structured for dimensionality reduction.

### Finalized Transformed Data

After all transformations, the dataset now includes:

* **Temporal Features:** hour\_of\_day, day\_of\_week, month\_of\_year, booking\_time\_bucket
* **Trip-Based Features:** trip\_length
* **Booking Type Features:** traditional\_booking
* **Binned Categories:** from\_area\_binned, to\_area\_binned, vehicle\_model\_id (grouped)
* **User-Level Aggregated Features:** user\_total\_bookings, user\_cancellation\_rate, user\_avg\_trip\_length.

# **Dimension Reduction**

Dimension reduction ensures that only the most relevant features are retained, improving model efficiency, interpretability, and generalization. Three key techniques were applied:

* Fisher’s Exact Test – Assessed the statistical significance of categorical variables in predicting cancellations.
* Correlation Analysis – Identified relationships between numerical variables to detect redundancy.
* Random Forest Feature Importance – Ranked variables based on their predictive power.

This section details each technique, interprets the results, and presents the final list of selected features.

## Handling Categorical Variables: Fisher’s Exact Test

**Why Fisher’s Exact Test?**

* Traditional chi-square tests become unreliable for sparse categorical data.
* Fisher’s Exact Test is more suitable for unbalanced categorical distributions (McDonald, 2009).
* Evaluates if categorical variables significantly influence **Car\_Cancellation**.

**Results and Interpretation**

* Fisher’s Exact Test was applied to the following categorical variables:
* **from\_area\_binned**, **to\_area\_binned** (Binned low-frequency area IDs)
* **vehicle\_model\_id** (Grouped low-frequency models)
* **package\_id** (Ride service packages)
* **travel\_type\_id** (Ride type)
* **day\_of\_week**, **month\_of\_year** (Temporal factors)

**Statistically Significant Categorical Variables (p < 0.05)**

|  |  |
| --- | --- |
| **Variable** | P**\_Value** |
| from\_area\_binned | 0.0004997501 |
| to\_area\_binned | 0.0004997501 |
| vehicle\_model\_id | 0.0004997501 |
| package\_id | 0.0004997501 |
| travel\_type\_id | 0.0004997501 |
| day\_of\_week | 0.0004997501 |
| month\_of\_year | 0.0004997501 |

Table 6: Statistically Significant Categorical Variables

**Key Findings**:

* All categorical variables showed significant relationships (p < 0.05) with Car\_Cancellation.
* Geographic features (from\_area\_binned, to\_area\_binned) are strong predictors, indicating regional differences in cancellations.
* Temporal features (day\_of\_week, month\_of\_year) are relevant, suggesting seasonal variations in cancellations.

Retain all categorical variables since they provide valuable predictive insights.

## Handling Numerical Variables: Correlation Analysis

**Why Use Correlation matrix?**

* A correlation matrix provides a more stable measure of relationships.
* Highly correlated variables introduce redundancy and increase model complexity.
* Correlation analysis ensures that numerical variables contribute distinct information without being too similar.

**Correlation Matrix of Numerical Features**

**A graph of a number of objects

Description automatically generated with medium confidence**

Figure 15: Correlation Matrix of Numerical Variables

* No highly correlated numerical variables detected.
* Low correlations suggest minimal redundancy, supporting feature retention.
* Trip Length and User Cancellation Rate show moderate correlations but remain relevant

## Feature Selection via Random Forest

Unlike correlation analysis, Random Forest evaluates feature impact in a non-linear, high-dimensional space.

### Feature Importance Results

**A graph with blue squares

Description automatically generated**

Figure 16: Feature Importance analysis - 2

**Top Features (Strongest Predictors)**

* User Cancellation Rate: The most significant predictor, confirming that past user behavior is a key determinant of cancellations.
* From Area Binned & To Area Binned: Geographical regions influence cancellation rates, possibly due to service availability.
* From Date & Booking Created: Temporal trends highlight the role of scheduling in cancellations.
* Trip Length: Suggests longer trips may have a different likelihood of being canceled.

**Moderate Importance Features**

* User-Level Features: user\_avg\_trip\_length and user\_total\_bookings indicate that repeat customers and trip length history impact cancellations.
* Booking Method: Online Booking and Traditional Booking moderately influence cancellation patterns.
* Time-Based Features: Hour of Day, Day of Week, Month of Year suggest time-related variations in cancellations.

**Low-Importance Features (Candidates for Removal)**

* Mobile Site Booking: Minimal effect, suggesting online and mobile users behave similarly.
* Package ID & Travel Type ID: Travel package type is not a strong determinant of cancellations.
* Vehicle Model ID: Low impact on cancellations.
* Booking Time Bucket: Adds little predictive value beyond hour\_of\_day and booking\_created.

## Justification for No PCA (Principal Component Analysis)

* PCA reduces interpretability, making it difficult to explain variable relationships.
* Random Forest already performed feature selection, eliminating the need for PCA.
* Low correlation among numerical features indicates PCA would have minimal impact.

Before partitioning, we drop low-impact and redundant features identified in the feature importance analysis, correlation analysis, and Fisher’s Exact Test.

**Dropped Features & Justification**

* month\_of\_year, day\_of\_week → Low predictive value
* missing\_to\_date → Minimal contribution to model accuracy
* package\_id, travel\_type\_id, vehicle\_model\_id → Low importance in feature selection
* user\_avg\_trip\_length → Highly correlated with trip\_length
* mobile\_site\_booking → Low impact compared to traditional\_booking

Final Dataset:  
**Target Variable**: Car\_Cancellation  
**Retained Features**:

* Trip Features → trip\_length
* Temporal Features → from\_date, booking\_created, hour\_of\_day
* Location Features → from\_area\_binned, to\_area\_binned
* User Features → user\_total\_bookings, user\_cancellation\_rate
* Booking Type Features → online\_booking, traditional\_booking

# Data Partitioning Methods

Data partitioning is crucial in machine learning to ensure that the model generalizes well to unseen data. A well-defined partitioning strategy prevents **overfitting**, ensures **fair model evaluation**, and maintains the **class distribution** for reliable predictions.

**Partitioning Strategy:** We use an **80%-20% stratified split**:

* **80% Training Set** (~8,001 records) → Used for model training and cross-validation.
* **20% Test Set** (~1,999 records) → Used for final model evaluation.

**Why Stratified Sampling?**

The dataset is **imbalanced** (only ~7.4% canceled rides). A **random split** might lead to underrepresentation of cancellations in the test set. Stratification ensures that **both training and test sets maintain the original cancellation distribution**.

**Class Distribution Before & After Partitioning**

|  |  |  |
| --- | --- | --- |
| **Dataset** | **Non-Cancellations (0)** | **Cancellations (1)** |
| **Original Dataset** | 92.56% | 7.44% |
| **Training Set (80%)** | 92.56% | 7.44% |
| **Test Set (20%)** | 92.59% | 7.40% |

Table 7: Class Distribution Before & After Partitioning

**Why No Separate Validation Set?**

Since cross-validation will be applied during model training, a separate validation set is unnecessary.

**Advantage:** This approach maximizes training data, leading to better generalization. Kohavi (1995) recommends cross-validation over static validation sets in medium-sized datasets.

**Final Justification:**  
A training-test split (80%-20%) is the best choice. This approach balances model efficiency, computational cost, and generalization performance while preventing data leakage (Kohavi, 1995; Hawkins et al., 2003).

Japkowicz & Stephen (2002): Random splits in imbalanced datasets can lead to misleading model evaluation due to underrepresentation of the minority class.

# Model selection

The purpose of this section is to compare different machine learning algorithms and select the best model for predicting ride cancellations. The model selection process is driven by **business interpretability, predictive performance, and computational efficiency.**

## **Candidate Models Considered:**

These classification models were evaluated based on their suitability for this problem:

* **Baseline Random Forest**
* **Hybrid Random Forest (Class Weighting + Oversampling)**
* **Oversampled Logistic Regression**

### Model Evaluation Metrics

To **select the best model**, we will use:

**Accuracy** – Overall correctness of predictions.  
**Precision** – How many predicted cancellations were actually cancellations?  
**Recall (Sensitivity)** – How well does the model capture actual cancellations?  
**F1-Score** – Balance between Precision & Recall.  
**AUC-ROC Curve** – Model’s ability to distinguish between canceled and completed rides.

# **Model Fitting, Validation Accuracy, and Test Accuracy**

This section presents the training, validation, and testing results for the three selected models: Baseline Random Forest, Hybrid Random Forest, and Oversampled Logistic Regression.  
An 80/20 stratified split was applied to the data, ensuring a balanced distribution of ride cancellations and non-cancellations across training and test datasets.

## Baseline Random Forest Model Performance

The initial Random Forest model served as a baseline to understand general patterns in ride cancellation behavior.  
It was trained with a substantial number of decision trees to enhance predictive stability, while preserving the dataset’s class distribution.

**Performance Insights:**

* The model achieved an accuracy of 99.92% on the training data and 98.09% on the test data.
* The high training accuracy indicated that the model captured patterns effectively, but it also raised early signs of potential overfitting.
* The model demonstrated strong capability in predicting completed rides, although cancellation detection was comparatively lower, reflecting the class imbalance.

## Hybrid Random Forest (Class Weighting + Oversampling) Performance

To address the imbalance and strengthen cancellation prediction, a hybrid Random Forest approach was employed, combining targeted reweighting with synthetic oversampling of minority-class observations.

**Performance Insights:**

* The model maintained a strong training accuracy of 99.91% and matched test accuracy at 98.09%.
* Improvement in cancellation detection was observed compared to the baseline model.
* The slight increase in false positives was considered acceptable, given the business priority of minimizing missed cancellations.

Oversampled Logistic Regression Performance  
To complement tree-based approaches with an interpretable model, logistic regression was developed after balancing the training dataset.  
The timestamp variable was excluded after extracting its key time-based features.

**Performance Insights:**

* Training accuracy reached 96.56%, and testing accuracy was 95.47%.
* The model excelled at identifying cancellations, reinforcing the value of using simple, interpretable models in certain cases.
* A slight increase in false alarms was observed, which was an acceptable trade-off for stronger cancellation detection.

**Final Note on Model Selection:**  
After exploring multiple strategies to improve balance and generalization, three final models were selected for deeper evaluation.  
These models represented a careful balance between predictive power, model complexity, and operational feasibility for SAR’s business needs.

# **Reporting Model Performance**

After selecting the three final models, Baseline Random Forest, Hybrid Random Forest (Class Weighting + Oversampling), and Oversampled Logistic Regression; their performance was assessed using key evaluation metrics: accuracy, sensitivity, specificity, F1-score, and AUC-ROC.  
These metrics offer a comprehensive view of how effectively each model predicts ride cancellations.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Sensitivity** | **Specificity** | **F1-Score** | **AUC-ROC** |
| **Baseline RF** | **98.09%** | **83.78%** | 99.24% | **86.71%** | **99.59%** |
| **Hybrid RF (Class Weighting + Oversampling)** | **98.09%** | **88.51%** | 98.86% | **87.33%** | **93.69%** |
| **Oversampled Logistic Regression** | **95.47%** | **95.52%** | **95.27%** | **97.17%** | **98.86%** |

Table 8: Performance metrics

**Observations on Model Performance as observed from the Table 9:**

* **Baseline Random Forest:** Strong overall accuracy and high specificity, but comparatively lower sensitivity in detecting cancellations.
* **Hybrid Random Forest:** Improved sensitivity and maintained high specificity, offering a balanced solution for business needs.
* **Oversampled Logistic Regression:** Highest sensitivity, capturing almost all cancellations, but with a slight drop in specificity compared to Random Forest models.

Based on these results, further comparative evaluation through confusion matrices, ROC curves, and feature importance is conducted in the next section to determine the most effective model for SAR’s operational needs.

# **Model Evaluation**

This section compares the three models, Baseline Random Forest, Hybrid RF (Class Weighting + Oversampling), and Oversampled Logistic Regression; using confusion matrices, ROC curves, feature importance, and cost-sensitive learning to assess their effectiveness in predicting ride cancellations.

## Confusion Matrix Analysis

The confusion matrices summarize how well each model distinguishes cancellations and non-cancellations:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **TN (Correct Non-Cancellations)** | **FP (False Positives)** | **FN (False Negatives)** | **TP (Correct Cancellations)** |
| **Baseline RF** | **1837** | **24** | **14** | **124** |
| **Hybrid RF** | **1830** | **17** | **21** | **131** |
| **Oversampled Logistic Regression** | **1768** | **7** | **83** | **141** |

Table 9: Confusion Matrix Analysis

**Key Observations:**

* Baseline RF achieves high specificity (low false positives) but misses some cancellations.
* Hybrid RF balances better, improving cancellation detection while keeping false positives manageable.
* Oversampled Logistic Regression detects the most cancellations but sacrifices more completed rides to false alarms.

## ROC Curves and AUC-ROC Scores

The ROC curve (Receiver Operating Characteristic) demonstrates the model's ability to differentiate between canceled and non-canceled rides. The AUC (Area Under the Curve) quantifies this distinction, the closer to 1.0, the better the model is at classification.

A graph of a curve

Description automatically generated

Figure 17: Baseline RF - ROC Curve

A diagram of a curve

Description automatically generated

Figure 18: Hybrid RF - ROC Curve

A diagram of a logistic regression

Description automatically generated

Figure 19: Oversampled-Logistic Regression - ROC Curve

|  |  |
| --- | --- |
| **Model** | **AUC-ROC Score** |
| **Baseline RF** | 99.59% |
| **Hybrid RF (Class Weighting + Oversampling)** | 93.69% |
| **Oversampled Logistic Regression** | 98.86% |

Table 10: AUC-ROC Scores

**Interpretation:**

* Baseline RF exhibits the highest AUC, but part of this advantage stems from overfitting.
* Hybrid RF achieves a more balanced performance, trading a slight decrease in AUC for improved recall.
* Oversampled Logistic Regression provides strong classification ability, distinguishing cancellations effectively.

## Feature Importance Analysis

Feature importance analysis identifies **which variables contribute most to ride cancellations**, providing business insights for SAR.

### Random Forest Feature Importance (Baseline RF & Hybrid RF)

Feature importance was computed using the Mean Decrease in Gini Index, where higher values indicate stronger predictive power.

A screenshot of a graph

Description automatically generated

Figure 20: Baseline RF – Feature Importance

A graph of a number of objects

Description automatically generated with medium confidence

Figure 21: Hybrid RF – Feature Importance

|  |  |  |
| --- | --- | --- |
| **Feature** | **Baseline RF** | **Hybrid RF** |
| **User Cancellation Rate** | High | High |
| **Booking Created Time** | Moderate | Moderate |
| **Trip Length** | Moderate | High |
| **Pickup Location** | High | High |
| **Time of Day** | Low | Moderate |

Table 11: Feature Importance results

### Logistic Regression Feature Importance

* User Cancellation Rate and Booking Created Time were the most significant predictors.
* Trip Length & From-Date variables also had strong predictive value.
* Unlike RF models, Logistic Regression allows clearer interpretability of how features impact ride cancellations.

## Cost-Sensitive Learning Analysis

To align model evaluation with business impact, a Cost-Sensitive Learning (CSL) analysis was conducted. In SAR’s context:

* Cost of missing a cancellation (False Negative): $100
* Cost of incorrectly predicting a cancellation (False Positive): $20

**Total Cost Formula:**

|  |  |
| --- | --- |
| **Model** | **Total Cost** |
| Baseline RF | $1,480 |
| Hybrid RF | $2,440 |
| Oversampled Logistic Regression | $8,440 |

Table 12: Cost-Sensitive Learning

**Interpretation:**

* Baseline RF leads to the lowest total misclassification cost, despite mild overfitting concerns.
* Hybrid RF balances costs better than Logistic Regression but incurs slightly higher total cost compared to the Baseline RF.
* Oversampled Logistic Regression, while excellent in detecting cancellations, generates the highest financial loss due to more false negatives.

Thus, from a cost-sensitive perspective, Baseline RF would be favored if the primary business objective is minimizing operational losses.

Having evaluated the models through multiple lenses, confusion matrices, ROC-AUC, feature importance, and cost analysis; the final section will now summarize the key findings and recommend the most practical model for SAR’s business needs.

# Observations and Conclusion

## Key Observations from Model Evaluation

* The predictive analytics framework successfully identified key factors influencing ride cancellations, including user cancellation history, trip characteristics, and booking behaviors.
* The developed models demonstrated strong predictive capabilities, with particularly high cancellation detection rates achieved through the oversampled logistic regression approach.
* Cost-sensitive learning analysis further validated the model selection by demonstrating that early detection of high-risk bookings could minimize operational losses for SAR.
* Despite efforts to mitigate overfitting through class balancing techniques, signs of minor overfitting were observed in all models. However, the final models still generalized well enough to support actionable decision-making.

## Achievement of Business and Analytics Goals

* The models effectively classified ride bookings as high-risk or low-risk, directly addressing SAR’s need to proactively manage ride cancellations.
* Data-driven insights into key cancellation drivers enable SAR to implement preventive strategies such as customer targeting, booking policies, and operational adjustments.
* The analytics approach produced a reliable early warning system that supports fleet management optimization, revenue protection, and enhanced customer satisfaction.
* The project translated model findings into strategic recommendations aligned with SAR’s business objectives, providing a solid foundation for operational improvements and future data-driven initiatives.

## Final Model Recommendation

Based on a holistic evaluation that considered predictive performance, cost-sensitive learning outcomes, and alignment with SAR’s operational priorities, Oversampled Logistic Regression is recommended.

While the Hybrid RF model showed a lower total financial cost in the cost-sensitive analysis, Oversampled Logistic Regression demonstrated superior capability in detecting nearly all cancellations, a priority for SAR's business goal of proactively reducing ride disruptions and protecting customer trust.

This model provides the highest cancellation detection rate, offering SAR the opportunity to intervene earlier and prevent cancellations, even if it slightly increases financial costs from false positives.

Overall, catching cancellations early was deemed more critical for sustaining SAR’s service reliability and customer loyalty.

Thus, the selected analytics approach supports SAR’s strategic objectives of improving service reliability and long-term customer retention.

# Conclusion and Future work

This project systematically addressed the ride cancellation problem faced by San Francisco Auto Rental (SAR) by leveraging data-driven analytics. Through structured data cleaning, feature engineering, and predictive modeling, we identified critical drivers of cancellations, including user cancellation behavior, trip length, booking methods, and pickup locations. Among the models developed, the Oversampled Logistic Regression emerged as the most effective for SAR’s business goals by achieving the highest sensitivity in detecting high-risk bookings. Although cost-sensitive analysis highlighted some financial trade-offs due to false positives, minimizing missed cancellations was prioritized to safeguard customer loyalty and service reliability, aligning with SAR's broader business strategy.

Future enhancements should focus on refining model interpretability and operational scalability. Hyperparameter tuning, experimenting with alternative algorithms like Gradient Boosting Machines, and incorporating real-time external factors such as weather, traffic, and driver reliability data could further improve predictive accuracy. Additionally, adopting explainable AI techniques like SHAP values will help SAR translate predictive outputs into actionable decisions, ultimately enabling proactive customer interventions, reducing cancellations, and driving stronger competitive advantage.

# References

1. Little, R. J. A., & Rubin, D. B. (2019). Statistical analysis with missing data (3rd ed.). Wiley. <https://doi.org/10.1002/9781119482260>
2. Zhang, T., Wang, X., & Lin, J. (2021). Evaluating location-based travel data for predictive analytics. Journal of Transport Geography, 93, 102786. <https://doi.org/10.1016/j.jtrangeo.2021.102786>
3. Li, Y., Chen, L., & Gao, S. (2020). Geospatial analytics for urban mobility insights. Computers, Environment and Urban Systems, 81, 101485. <https://doi.org/10.1016/j.compenvurbsys.2020.101485>
4. Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., & Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. Atmospheric Environment, 38(18), 2895–2907. <https://doi.org/10.1016/j.atmosenv.2004.02.026>
5. Beretta, L., & Santaniello, A. (2016). Nearest neighbor imputation algorithms: A critical evaluation. BMC Medical Informatics and Decision Making, 16(3), 74. <https://doi.org/10.1186/s12911-016-0315-z>
6. Allison, P. D. (2002). Missing data. SAGE Publications.
7. Rousseeuw, P. J., Van Driessen, K., & Hubert, M. (2011). Robust statistics for outlier detection. Statistical Methods & Applications, 20(3), 293–315. <https://doi.org/10.1007/s10260-011-0158-2>
8. Keegan, R. (2020). Data preprocessing techniques for machine learning. Journal of Data Science, 18(2), 123–135.
9. Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.
10. Aggarwal, C. C. (2017). Outlier analysis (2nd ed.). Springer. <https://doi.org/10.1007/978-3-319-47578-3>
11. Chakraborty, S., & Joseph, A. (2017). Statistics for machine learning: Techniques for exploring supervised, unsupervised, and reinforcement learning models with Python. Packt Publishing.
12. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning. Springer.
13. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. Springer.
14. McKinney, W. (2017). Python for data analysis. O'Reilly Media.
15. Wang, Y., & Ross, M. (2020). The role of geographic and temporal factors in ride-hailing cancellations. Journal of Transport Economics and Policy, 54(2), 245–267.
16. Tang, X. (2018). Impact of vehicle model selection on ride-hailing service performance. International Journal of Urban Transportation, 5(4), 112–130.
17. Li, J., Chen, Y., & Xu, H. (2019). Understanding ride-hailing cancellations: A machine learning approach. Transportation Research Part C, 103, 174–189. <https://doi.org/10.1016/j.trc.2019.04.008>
18. Guo, F., & Williams, B. (2021). The stability of distance-based features in transportation data science. IEEE Transactions on Intelligent Transportation Systems, 22(3), 1985–1996. <https://doi.org/10.1109/TITS.2020.2972459>
19. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
20. Chakraborty, G., & Joseph, K. (2017). Feature selection methods in predictive analytics. Journal of Data Science, 15(3), 12–28.
21. McDonald, J. (2009). Fisher’s exact test: When to use it. Biostatistics Journal, 7(4), 225–237.
22. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence (pp. 1137–1143).
23. Hawkins, D. M., Basak, S. C., & Mills, D. (2003). Assessing model fit by cross-validation. Journal of Chemical Information and Computer Sciences, 43(2), 579–586. <https://doi.org/10.1021/ci025626i>
24. Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. Intelligent Data Analysis, 6(5), 429–449.