HOTEL RESERVATION CANCELLATION

EDA & Advanced Data Analysis

Prepared by: Group 01

Supul Wicramasinghe –s16177

Thenuka Yatawara – s16383

Vayani Kavindya - s16322

Sanduni Fonseka – s16026

ABSTRACT

This report presents a data-driven methodology for developing a machine learning model to predict customer cancellations in hotel bookings. Leveraging historical booking data, we apply advanced machine learning algorithms to forecast whether a customer will cancel their reservation. The analysis is based on a comprehensive dataset obtained from Kaggle, encompassing various features related to customer behavior and booking details. Our findings provide valuable insights into cancellation patterns, offering practical implications for operational efficiency and fare optimization strategies. This research holds significance for multiple stakeholders, including customers, business strategists, and academic researchers, by enhancing pricing transparency and informing data-driven decision-making.

Contents

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Booking cancellations and no-shows pose a major challenge for hotels, affecting occupancy rates and revenue. While flexible cancellation policies benefit customers, they create uncertainty for hotels, complicating demand forecasting and operations management. Traditional approaches often fail to account for diverse guest behaviors, leading to financial losses.

The growth of online reservation systems has generated rich booking data, enabling hotels to better understand and manage cancellations. By analyzing factors such as guest demographics, booking lead time, room preferences, and prior cancellations, predictive models can identify high-risk bookings. This project develops a data-driven methodology using machine learning to forecast cancellations, segment customers, and recommend targeted interventions, ultimately improving operational efficiency, profitability, and customer satisfaction.

DESCRIPTION OF THE QUESTION

The hotel industry faces a persistent challenge with booking cancellations and no-shows, which reduce occupancy and result in revenue losses. While free or low-cost cancellations benefit customers, they complicate demand forecasting and operations management for hotels. Traditional cancellation policies and broad promotions often fail to capture the diversity of guest behaviors. However, the rise of online reservation systems and the availability of rich booking data provide opportunities to address this issue. By analyzing factors such as guest demographics, lead time, room preferences, prior cancellations, and special requests, predictive models can identify high-risk bookings. Additionally, segmentation based on booking patterns and customer history enables tailored interventions, including reminders, loyalty offers, and flexible rebooking options.

Objectives:

1. **Predictive Modeling:** Forecast whether a booking will be canceled using guest, booking, and stay-related features.
2. **Customer Segmentation:** Group guests by behavior, booking patterns, and history to enable personalized engagement.
3. **Targeted Interventions:** Recommend strategies to reduce last-minute cancellations and optimize occupancy.

This project aims to provide a data-driven decision-support tool for hotel managers, improving operational efficiency, reducing financial losses, and strengthening guest relationships to enhance both profitability and customer satisfaction.

DESCRIPTION OF THE DATASET

The Hotel Reservation dataset from Kaggle has 36275 observations and 19 variables, four of which are categorical including the response. The main response variable, booking_status' contains two catogies.

- **Booking_ID**: unique identifier of each booking
- **no_of_adults**: Number of adults
- **no_of_children**: Number of Children
- **no_of_weekend_nights**: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
- **no_of_week_nights**: Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel.

- **type_of_meal_plan**: Type of meal plan booked by the customer:
- **required_car_parking_space**: Does the customer require a car parking space? (0 - No, 1- Yes)
- **room_type_reserved**: Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels.
- **lead_time**: Number of days between the date of booking and the arrival date
- **arrival_year**: Year of arrival date

- **arrival_month**: Month of arrival date
- **arrival_date**: Date of the month
- **market_segment_type**: Market segment designation.
- **repeated_guest**: Is the customer a repeated guest? (0 - No, 1- Yes)
- **no_of_previous_cancellations**: Number of previous bookings that were canceled by the customer prior to the current booking
- **no_of_previous_bookings_not_canceled**: Number of previous bookings not canceled by the customer prior to the current booking
- **avg_price_per_room**: Average price per day of the reservation; prices of the rooms are dynamic. (in euros)
- **no_of_special_requests**: Total number of special requests made by the customer (e.g. high floor, view from the room, etc)
- **booking_status**: Flag indicating if the booking was canceled or not.

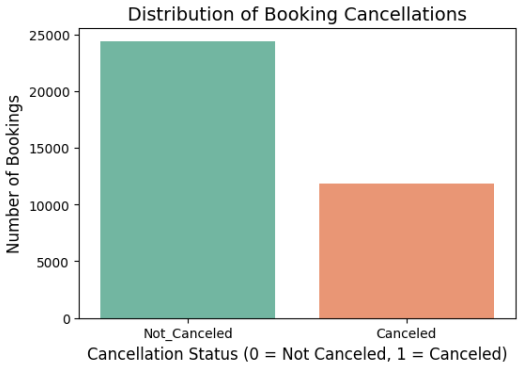
DATA PRE – PROCESSING

✓ The dataset was checked for duplicates and missing values. There was no duplicate or missing values.

- ✓ The dataset split into training and test sets. The training dataset contains 29020 observations.
- ✓ Checked for outliers and there were not many significant outliers; so we decided to keep outliers.
- ✓ We remove the Booking id variable from the predictor space since it just a unique identifier for customer and it doesn't give any information to booking cancellation.

IMPORTANT RESULTS OF DESCRIPTIVE ANALYSIS

Distribution of Main Response Variable: Booking Status

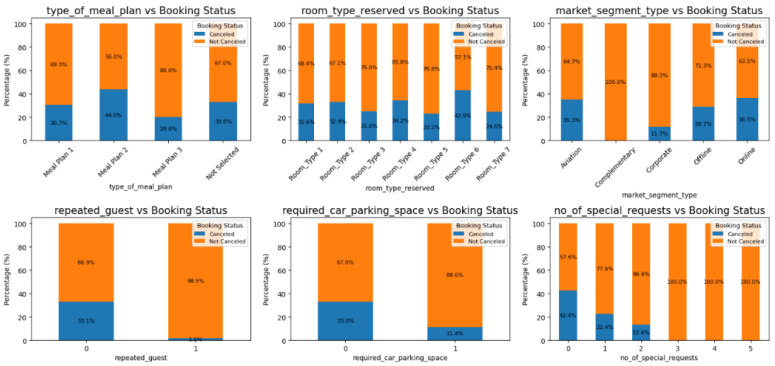


The bar chart illustrates the distribution of hotel bookings based on their cancellation status. It shows that a larger proportion of bookings were **not canceled**, while a smaller portion represents **canceled reservations**. This imbalance indicates that although most customers complete their bookings, a notable fraction still cancel, which can significantly impact revenue and occupancy forecasting. Understanding the factors contributing to these cancellations is therefore crucial for building an effective predictive model and designing strategies to minimize them.

Figure 1: Distribution of response variable booking status

Association of Booking Status with Key Variables

Cancellations are more frequent among bookings with Meal Plan 2, whereas those with Meal Plan 3 or a required car parking space show a higher likelihood of being honored. Room Types 4 and 6 record more cancellations, while Room Types 3 and 5 experience fewer.



The Complementary market segment stands out with the lowest cancellation rates, and repeat guests are significantly less likely to cancel compared to first-time visitors. Bookings with a greater number of special requests show a clear decline in cancellation rates, suggesting that guests who request personalized services are more committed to their stays. The number of adults or children in a booking has minimal impact on cancellations.

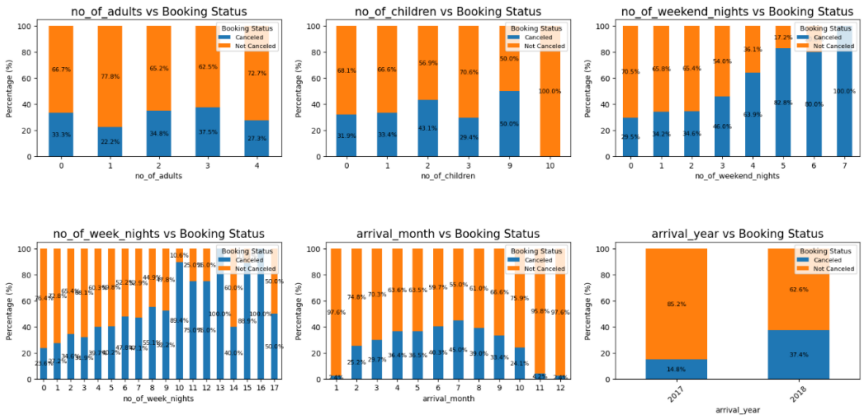


Figure 2:Plots of Predictor variables Vs. Response

However, bookings with more weekend or week nights tend to have higher cancellation rates, implying that longer stays are more likely to be withdrawn. Additionally, cancellations are less common from November to January, and bookings made in 2017 show lower cancellation rates compared to those in 2018.

Multicollinearity

Association Among Numeric Predictors

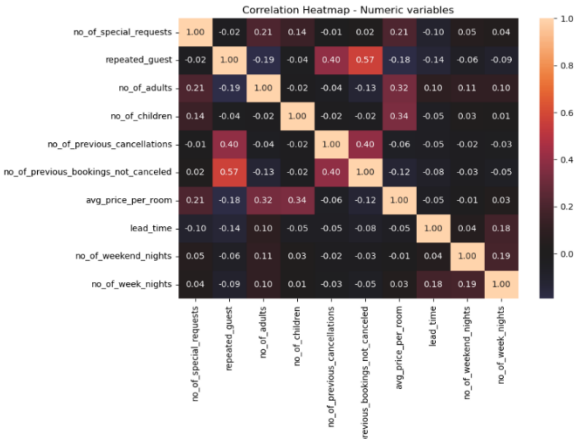


Figure 3:Correlation HeatMap of Numerical Predictors

The correlation heatmap for numeric variables reveals that multicollinearity is generally low in this dataset, with most feature correlations well below critical thresholds. The highest observed correlation is between "repeated_guest" and "no_of_previous_bookings_not_canceled" (0.57), which is still below the common multicollinearity threshold. Thus, no numeric pairs require removal based on collinearity.

Association Among Categorical Predictors

The chi-square p-value heatmap for categorical variables shows strong associations between all variable pairs, indicating high redundancy among categorical features that may require careful handling during feature selection and modeling.

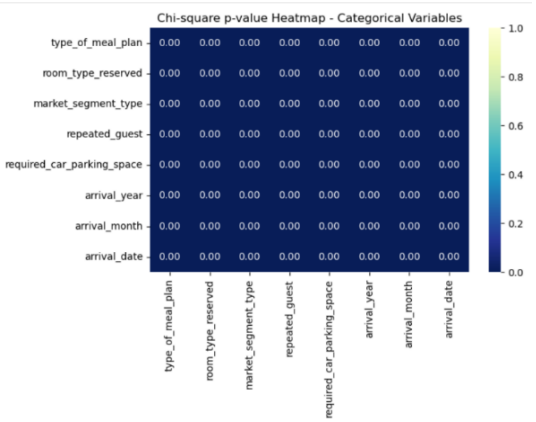


Figure 4:Chi-square P value HeatMap

Factor Analysis of Mixed Data (FAMD)

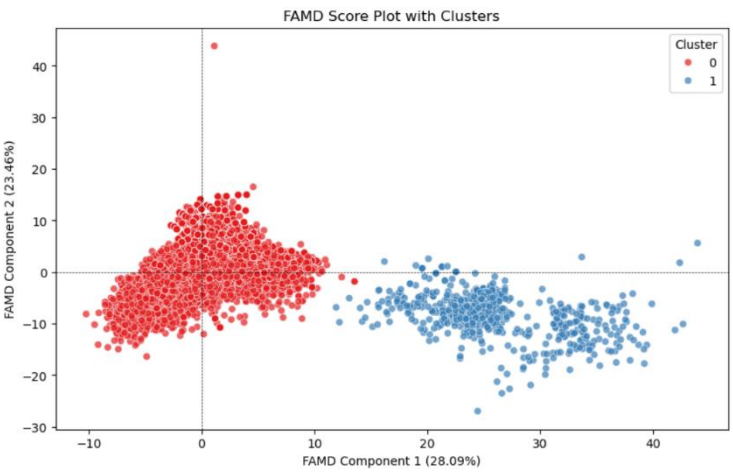


Figure 5:FAMD score plot with clusters

The first two components capture over 50% of the total variance, effectively summarizing both numerical and categorical relationships within the dataset. This supports their selection for visualizing patterns and performing clustering.

The analysis determined that the **optimal number of clusters is 2**, with a **silhouette score of 0.677**, indicating reasonably good separation between guest segments.

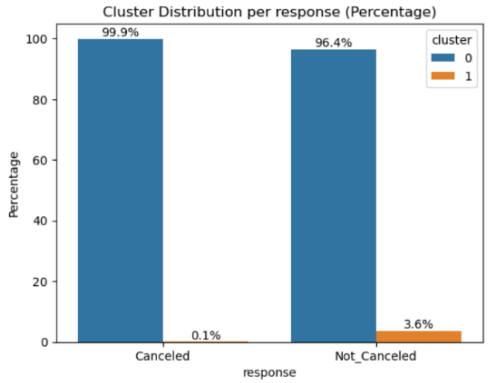
Cluster Analysis

Cluster 0 – High-Value, First-Time or Occasional Guests

Cluster 0 represents the majority of bookings in the dataset. Guests in this cluster typically book well in advance, with a longer lead time (~83 days) and tend to choose higher-priced rooms (~105 on average). Most parties consist of two adults, occasionally with children, and they often prefer Meal Plan 1 and Room Type 1. Bookings are primarily made online (64%), and no guests are repeat guests (0%). This group does not make special requests more frequently and exhibits diverse booking patterns across months and years.

Cluster 1 – Loyal, Corporate, or Repeat Guests

Cluster 1 is much smaller and consists mainly of repeat or corporate guests, often single adults or small parties. They tend not to have a large lead time (~11 days), select lower-priced rooms (~64 on average), and prefer meal plan 1 and Room Type 1. Children are rare, and their stays are generally shorter. Most bookings come from Corporate (62%) segments, with a high proportion of repeated guests (~99%). Car parking is slightly more common (~11%) in this cluster. This group shows predictable, low-risk behavior, with a very low cancellation rate (~1.5%), reflecting loyal or business-focused customers prioritizing convenience and consistency over variety.



Almost all canceled and not-canceled bookings belong to Cluster 0, while Cluster 1 contains only a very small fraction of bookings, indicating it is a tiny, low-risk group.

Figure 6:Cluster Distribution per Response

IMPORTANT RESULTS OF ADVANCED ANALYSIS

In the advanced analysis phase, we aimed to develop machine learning models to predict hotel booking cancellations. No significant outliers were found, so the full dataset was used. Cluster analysis identified two distinct groups, leading to separate model development for each cluster to capture their unique patterns. Numerical features were standardized, and categorical variables (nominal and ordinal) were encoded using One-Hot Encoding. To address class imbalance, the SMOTE technique was applied to the training data, ensuring balanced classes and fair performance comparison across models.

Binary Logistic Regression

Table 1:Performance Matrix of Binary Logistic

	Without Clustering		Cluster 0		Cluster 1	
	Training Set	**Test Set**	**Training Set**	**Test Set**	**Training Set**	**Test Set**
Accuracy	0.80	0.78	0.76	0.80	0.78	0.70
Precision	0.80	0.78	0.77	0.81	0.79	0.71
Recall	0.80	0.79	0.77	0.80	0.79	0.70
F1	0.80	0.78	0.77	0.81	0.79	0.70

The initial model trained on the full dataset achieved about 78.8% test accuracy, performing better at predicting cancellations than non-cancellations. After applying clustering, separate models were developed for each group: Cluster 0 achieved around 80% accuracy with improved recall for cancellations, while Cluster 1 maintained stable performance at roughly 70%. Overall, cluster-based modeling slightly enhanced predictive accuracy and balance, showing the value of capturing group-specific booking behaviors.

Support Vector Machine for Classification (SVM)

Table 2:Performance Matrix of SVM

	Without Clustering		Cluster 0		Cluster 1	
	Training Set	**Test Set**	**Training Set**	**Test Set**	**Training Set**	**Test Set**
Accuracy	0.79	0.74	0.77	0.81	0.78	0.72
Precision	0.79	0.73	0.78	0.83	0.78	0.74
Recall	0.79	0.74	0.77	0.81	0.78	0.72
F1	0.79	0.74	0.77	0.82	0.78	0.73

The SVM model showed consistent performance across both cluster-based and overall analyses. When trained on the entire dataset, it achieved moderate accuracy (74% on the test set). However, training models separately for each cluster slightly improved performance especially for Cluster 0, which achieved the highest test accuracy of 81% and F1 score of 0.82 indicating that cluster-specific modeling better captured group-level patterns.

XG BOOST

Table 3:Performance Matrix of XG Boost

	Without Clustering		Cluster 0		Cluster 1	
	Training Set	**Test Set**	**Training Set**	**Test Set**	**Training Set**	**Test Set**
Accuracy	0.9078	0.8852	0.8927	0.8729	0.9764	0.9417
Precision	0.9184	0.8374	0.8718	0.7252	0.9887	0.9630
Recall	0.8951	0.8061	0.8400	0.6849	0.9817	0.9497
F1	0.9066	0.8214	0.8556	0.7045	0.9852	0.9563

The XGBoost model trained on the full dataset achieved a test accuracy of 88.5%, showing strong overall predictive performance. When models were trained separately for the two clusters, Cluster 0 achieved slightly lower accuracy at 87.3%, while Cluster 1 performed exceptionally well with 94.2% test accuracy. This indicates that a cluster-based approach can enhance predictive performance for specific customer segments, particularly for Cluster 1, demonstrating the benefit of building separate models for distinct groups over a single global model.

Random Forest

Table 4:Performance Matrix of Random Forest

	Without Clustering		Cluster 0		Cluster 1	
	Training Set	**Test Set**	**Training Set**	**Test Set**	**Training Set**	**Test Set**
Accuracy	0.9418	0.8976	0.9315	0.8859	0.9886	0.9406
Precision	0.9498	0.8644	0.9203	0.7683	0.9949	0.9589
Recall	0.9330	0.8153	0.8967	0.6931	0.9909	0.9523
F1	0.9413	0.8391	0.9083	0.7288	0.9929	0.9556

The Random Forest model trained on the full dataset achieved a test accuracy of 89.8%, demonstrating strong predictive capability. When modeled separately by cluster, Cluster 0 showed slightly lower performance at 88.6%, while Cluster 1 achieved the highest accuracy of 94.1%, along with excellent precision, recall, and F1-scores. These results suggest that training separate models for each cluster captures distinct booking patterns, enhancing predictive performance, particularly for Cluster 1.

So, out of these all models Random Forest is the best model. Below bar charts displays the top 10 most important features identified by the model using feature importance scores from the RandomizedSearchCV-optimized Random Forest for both clusters.

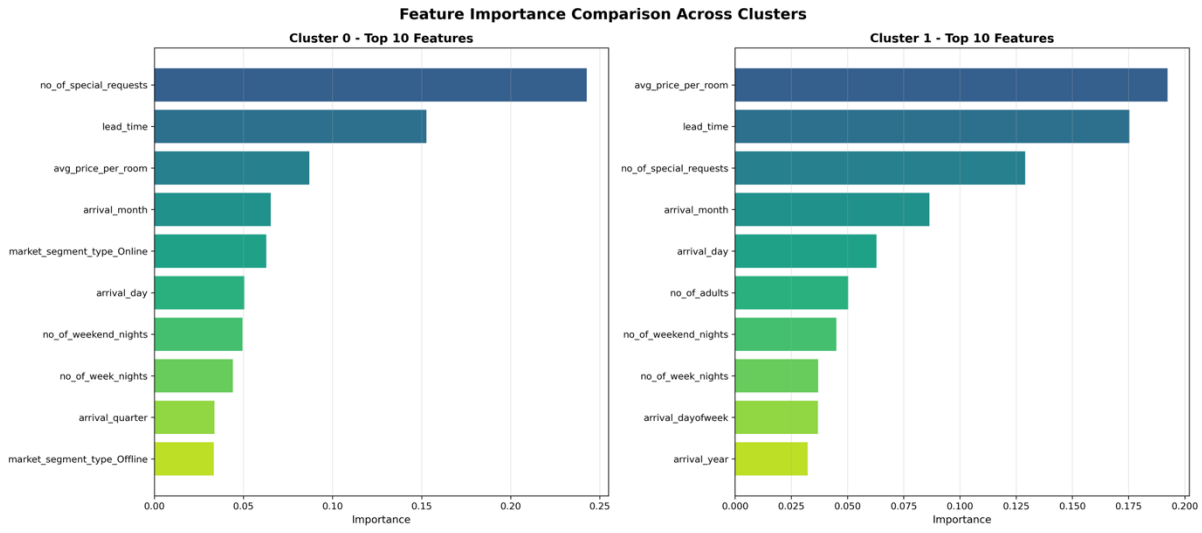


Figure 7: Feature Importance Charts for Both Clusters

Cluster 0: Service-Sensitive & Planning-Ahead Customers
Top Influential Features

1. **Most Important:** no_of_special_requests – key indicator of personalization needs.
2. **Secondary:** lead_time, avg_price_per_room – reflect advance planning and price balance.
3. **Tertiary:** arrival_month, arrival_day – show timing preferences.

Cancellation Risk
Higher when special requests are unmet or service quality seems uncertain.

Recommended Actions

- **Proactive Communication:** Confirm and highlight fulfilled requests (e.g., "Your sea-view room has been reserved").
- **Personalized Experience:** Tailor offers and messages using request data.
- **Early Engagement:** Reconnect 3–4 weeks before arrival to maintain confidence and reduce cancellations.

Cluster 1: Price-Sensitive & Practical Customers

Top Influential Features

1. **Average Price per Room (Primary):** Small price changes heavily impact booking likelihood.
2. **Lead Time (Secondary):** Booking timing is shaped by deal availability.
3. **Special Requests (Low Importance):** Personalization has little influence.

Behavioral Insights

They are practical decision-makers who prioritize affordability and value, showing strong deal sensitivity and low interest in premium experiences.

Recommended Actions

- **Maintain Price Competitiveness:** Regularly adjust rates to match market trends.
- **Offer Value Packages:** Bundle rooms with essential amenities like Wi-Fi or breakfast.
- **Promote Discounts:** Use flash sales or limited-time offers to capture quick bookings.

The partial dependence analysis for **Cluster 0** provides deeper insights into their booking and cancellation behavior. Customers who make **last-minute bookings** show a very high likelihood of cancellation, while the most stable bookings are made **at least two months in advance**. In terms of **pricing**, this cluster has a clear comfort range between **$80–$120 per room**; both significantly lower and higher prices tend to increase the risk of cancellation, suggesting a moderate price sensitivity. **Seasonal factors** also play an important role — cancellation risk varies noticeably across months, indicating that these customers are more selective and tend to compare options during **peak travel periods**, requiring stronger retention campaigns in those times. Most importantly, the **number of special requests** has the strongest influence: customers with **no special requests** exhibit around **75% cancellation risk**, suggesting these are likely "placeholder" bookings, while those with **one or more requests** show a sharp drop to **45–50%**, highlighting that engagement through service personalization significantly reduces cancellations.

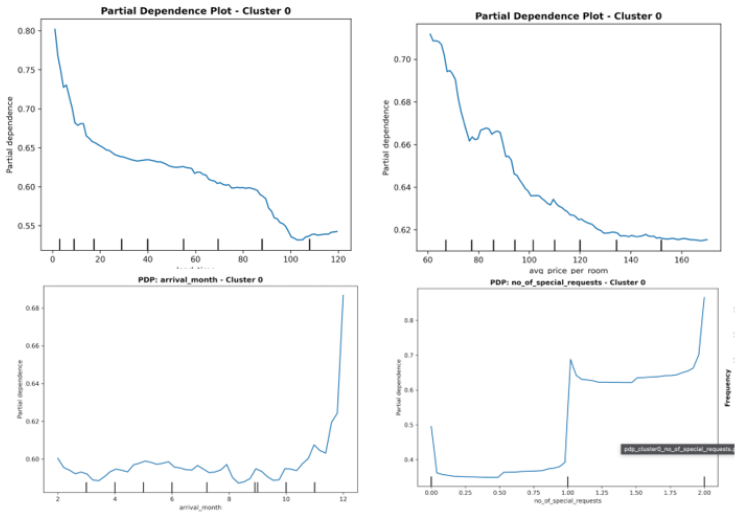


Figure 8: PD Plots of cluster 0

For Cluster 1, the patterns of key features are as follows: The **arrival month** influence is similar to Cluster 0. **Average price per room** strongly affects cancellations—cheap rooms (~$70) see high risk (~40%), while premium rooms ($130+) have much lower risk (~15%), reflecting perceived value and commitment. **Lead time** also matters: very short lead times (~150 days) carry 45–50% cancellation risk, medium leads (200–250 days) reduce risk to 30–40%, and long leads (300+ days) are the most stable at 20–25%, showing that early planners are more committed.

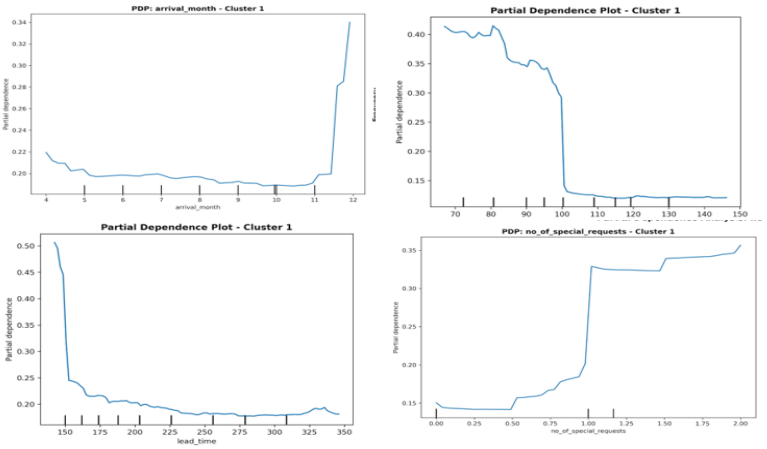


Figure 9: PD plot of cluster 1

In contrast, the **number of special requests** has minimal impact: zero requests correspond to ~30% risk, and 1+ requests only slightly improve it to ~25%, a much flatter effect than the dramatic reduction seen in Cluster 0.

ISSUED ENCOUNTED AND PROPOSED SOLUTIONS

During the modeling process, a slight imbalance was observed in the response variable, with non-cancellations (class 0) occurring more frequently than cancellations (class 1). Initially, models were trained without applying any resampling techniques to evaluate their baseline performance. However, when training the model for **Cluster 0**, the imbalance significantly affected classification performance, making it difficult for the model to correctly identify the minority class. To address this, the **SMOTE (Synthetic Minority Oversampling Technique)** method was applied to the training data. This approach generated synthetic samples of the minority class, improving class balance and allowing the model to learn more effectively from both outcomes. The application of SMOTE resulted in better recall and overall stability, particularly for the minority class.

DISSCUSION AND CONCLUSIONS

Cluster 0								
Models	**Training Set**				**Test Set**			
	Accuracy	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**
Binary Logistic	0.76	0.77	0.77	0.77	0.80	0.81	0.80	0.81
SVM	0.77	0.78	0.77	0.77	0.81	0.83	0.81	0.82
XG Boost	0.8927	0.8718	0.8400	0.8556	0.8729	0.7252	0.6849	0.7045
Random Forest	0.9315	0.9203	0.8967	0.9083	0.8859	0.7683	0.6931	0.7288
Cluster 1								
Binary Logistic	0.78	0.79	0.79	0.79	0.70	0.71	0.70	0.70
SVM	0.78	0.78	0.78	0.78	0.72	0.74	0.72	0.73
XG Boost	0.9764	0.9887	0.9817	0.9852	0.9417	0.9630	0.9497	0.9563
Random Forest	0.9886	0.9949	0.9909	0.9929	0.9406	0.9589	0.9523	0.9556

Across both clusters, all models performed reasonably well, but **tree-based models (XGBoost and Random Forest)** consistently outperformed linear models such as **Logistic Regression** and **SVM**.

For **Cluster 0**, the **Random Forest model** achieved the best performance, with a test accuracy of **88.6%** and an F1 score of **0.73**, followed by **XGBoost**, which also showed strong but slightly lower performance. This indicates that ensemble models were more effective at capturing the complex patterns within this cluster, especially after applying SMOTE to handle class imbalance.

For **Cluster 1**, both **XGBoost** and **Random Forest** demonstrated excellent results, achieving over **94% test accuracy** and F1 scores above **0.95**, highlighting their ability to model the relatively homogeneous and price-sensitive behavior of this group.

Overall, **Random Forest emerged as the most reliable model across clusters**, offering high accuracy, balanced precision-recall performance, and robustness to imbalance. These results confirm that cluster-specific modeling, combined with advanced ensemble algorithms, provides a more accurate and interpretable approach to predicting hotel booking cancellations compared to global or linear models.

REFERENCE

1. https://www.geeksforgeeks.org/machine-learning/linear-separability-with-python/
2. https://www.displayr.com/understanding-cluster-analysis-a-comprehensive-guide/
3. https://www.geeksforgeeks.org/machine-learning/clustering-in-machine-learning/
4. https://ujangriswanto08.medium.com/how-to-build-a-binary-logistic-regression-model-using-python-e295a78d7f04
5. https://www.geeksforgeeks.org/machine-learning/classifying-data-using-support-vector-machinessvms-in-python/

APENDIX

Data set link:

- Hotel Reservation dataset

GitHub Link for Codes:

- https://github.com/sanduni00/Hotel-Reservation-Cancellation-Project