Final Project

Hotel Booking Cancellation Prediction

The Report of Explanatory and Advanced Data Analysis.

Prepared by:

Group 4

Lakmini Thamodya – s16340

Sadini Thiranja – s16263

Nihindu Omal – s16276

Abstract

In an industry where customer satisfaction and operational efficiency are critical, hotel booking cancellations pose significant challenges to revenue management and planning. It is no longer a back-end activity to forecast cancellations—today it has turned into a strategic imperative. In this analysis, we study the Hotel Booking Cancellation Kaggle dataset to uncover the underlying causes of booking cancellations. By applying a blend of machine learning approaches and statistical techniques, such as FAMD, XGBoost, and Random Forest, we filter out pivotal behavioral and transactional features impacting the propensity of cancellation. While enhancing predictive accuracy, our findings also offer useful insights for hoteliers to optimize resource allocation, enhance customer targeting, and reduce revenue loss within a dynamic hospitality landscape.

Table of Contents

List of tables

Introduction

Hotel booking cancellations have become a serious concern in the hospitality industry. Cancellations directly affect hotel revenue, planning, and customer satisfaction. As mentioned in the **research by Aditya Dole (2023)** [1], *"The reservation cancellation rate is one aspect that has a significant impact on the sector."*

This report presents an analysis of hotel booking cancellations and identifies the key factors influencing them. To support this analysis, machine learning models were developed to predict the likelihood of a booking being canceled. These predictive models can help hotels to take proactive measures in managing cancellations.

Description of the problem

Managing hotel bookings has become more difficult due to the rising rate of cancellations. According to **Duetto's 2022 Revenue Management Report** [2], the average hotel cancellation rate is between 25% to 35%, which shows how serious this issue has become. These cancellations often lead to financial losses, especially when rooms remain empty at the last minute. **Research by Mirai (HospitalityNet, 2018)** [3] found that bookings made through third-party platforms like Booking.com and Expedia have significantly higher cancellation rates compared to direct hotel website bookings. This makes it harder for hotels to plan room availability, set the right prices, and manage staff and resources efficiently. Without accurate cancellation predictions, hotels may face challenges in making good pricing decisions, maintaining customer satisfaction, and maximizing profits. Therefore, understanding cancellation patterns and factors such as booking lead time, deposit type, reservation channel, and seasonal trends is important for hotels to manage their operations better and minimize losses.

Description of the dataset

The Hotel Booking Prediction dataset from Kaggle contains 119,390 records and 33 variables, including both numerical and categorical features. The main response variable is **"is_cancelled,"** which indicates whether a booking was canceled or not. Below, we have summarized the key variables used in our analysis.

Numerical Variables	Categorical Variables
lead_time	hotel
arrival_date_week_number	arrival_date_month
arrival_date_day_of_month	meal
stays_in_weekend_nights	country
stays_in_week_nights	market_segment (Shows various market segments that individuals belong to when making reservations)
adults	distribution_channel (Specifies different channels through which bookings were made)
children	is_repeated_guest
babies	deposit_type
previous_cancellations	customer_type
previous_bookings_not_cancelled	
booking_changes	
adr (Represents the average daily rate (price per room) for the booking.)	
days_in_waiting_list	
required_car_parking_spaces	
total_of_special_requests	
room_mismatch	

Table 1: List of the Numerical and Categorical Variables

Data set: https://www.kaggle.com/datasets/thedevastator/hotel-bookings-analysis

Data Preprocessing

- No duplicate values were found.
- The **company** column had 94%-95% missing values, making it unhelpful for predictions. It was dropped.
- The **agent** column had many unique values (high cardinality) and 14% missing values. Keeping it could lead to overfitting, so it was dropped.
- The **reservation_status** column contained the same information as **is_cancelled**. If a guest had "Checked Out," the booking was not canceled. If the status was "No Show" or "Canceled," it indicated cancellation. Since the same information existed in **is_cancelled**, this column was removed.
- The **arrival_date_year** column had only three unique years, which did not provide much insight for predicting cancellations, so it was dropped.
- The **country** column had 488 missing values (0.41%), which were replaced with "Unknown." This column originally had 178 unique values, making it difficult to use directly. Instead, countries were categorized into their respective continents: **Europe, North America, Asia, South America, Oceania, Africa**, and "Unknown."
- The **adr** (average daily rate) column had one negative value, which is not possible for a booking price. Since the distribution was skewed, the negative value was replaced with the **median**.

- The **children** column had four missing values, which were also imputed with the **median** due to its skewed distribution.
- Some records had bookings where only **children or babies were registered without an adult**, which is against hotel policy. These rows were likely data entry errors and accounted for only **0.34%** of the dataset. To avoid adding noise to the data, they were removed.
- **Index** column was used only to identify unique rows and had no predictive power, so it was dropped.
- After processing, the number of columns decreased from **33 to 26**.

Feature Engineering

- The **reserved_room_type** and **assigned_room_type** columns contained values like A, B, C, etc. Their exact meanings were unknown and keeping them in this format would not be useful for predictions. Instead, a new feature, **room_mismatch**, was created. This feature indicates whether the assigned room type differs from the reserved room type.

Results of Explanatory Data Analysis

Descriptive Analysis

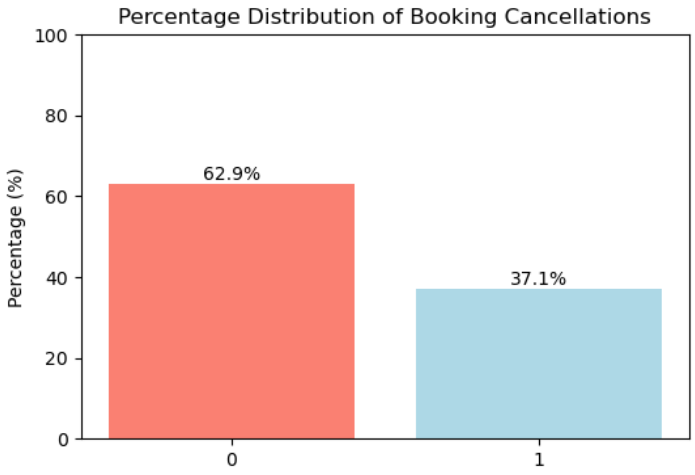


Figure 1: Percentage Distribution of Booking Cancellations

Figure 1 shows that approximately 63% of hotel bookings were not canceled, while 37% were canceled. This indicates that the **majority of bookings are completed as planned**, with nearly **two-thirds** going through successfully. However, the **significant proportion of cancellations** (over one-third) highlights a pattern worth exploring further in our analysis.

Figure 2 indicate that most of the bookings placed from Europe that is 90%. Distribution of Average Daily rate values indicate a right skewed distribution and most of the prices are clusterd at the lower end (likely between 0 and 100), with frequencies sharply declining as rates increase.

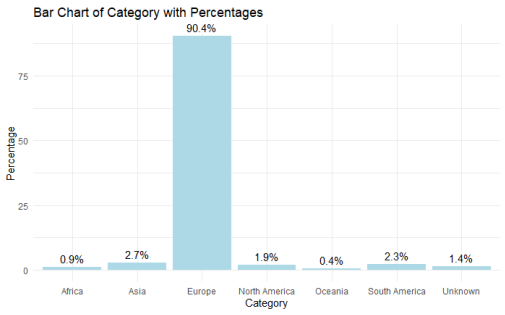


Figure 2: Bar plot of bookings by countries

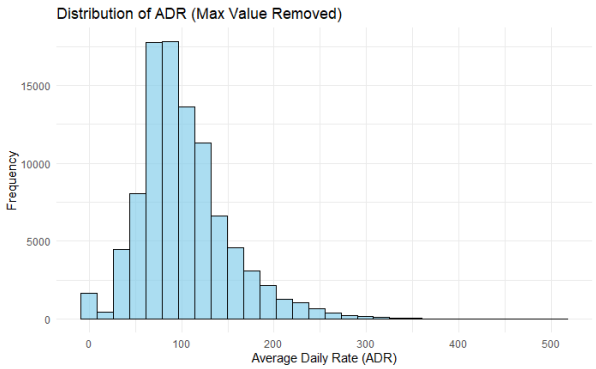


Figure 3: Distribution of adr of per room

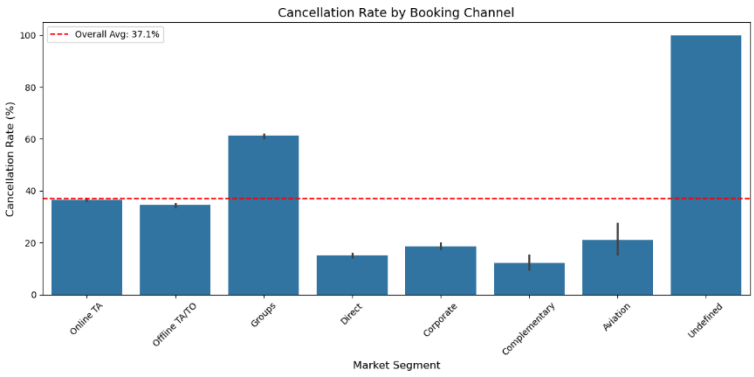


Figure 4: Cancellation rate by market_segment

By figure 3, The cancellation rate varies significantly across booking channels, with the overall average at 37.1%. Online travel agencies (labeled "Online Tri") and offline channels ("Offline TATO") show notably high cancellation rates, while direct bookings ("Direct") and corporate channels exhibit lower rates. This suggests that customers booking through intermediaries may be less committed, while direct or corporate bookings are more stable.

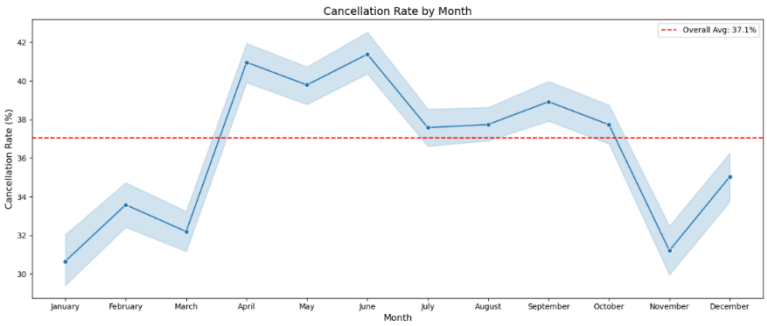


Figure 5: Cancellation Rate by Month

By figure 5, A significant spike is observed in April and June, where cancellation rates surpass 40%. In contrast, November shows the lowest rate, dipping just below 31%. The red dashed line represents the overall average cancellation rate of **37.1%**, indicating that more than half the months experienced rates above this benchmark. These trends may reflect seasonal behaviors or operational challenges.

By fig 6,Transient customers have the highest cancellation rate, exceeding 40%, possibly because they typically book individually and may have more flexible or uncertain travel plans. Contract and Transient-Party customers follow, with moderate cancellation rates around 31% and 25% respectively. Group customers, on the other hand, have the lowest cancellation rate, approximately 10%, indicating higher booking commitment—likely due to coordinated travel and logistical planning involved in group bookings.

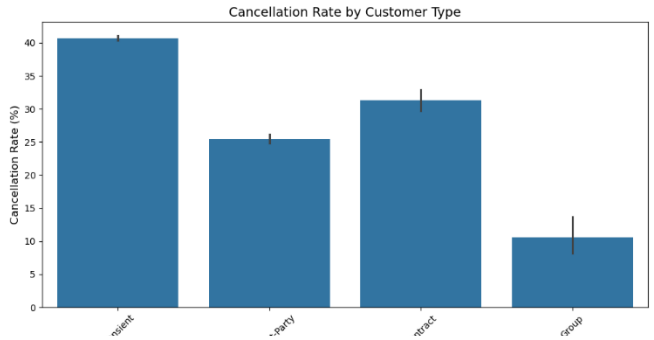


Figure 6 : Cancellation Rate by Customer Type

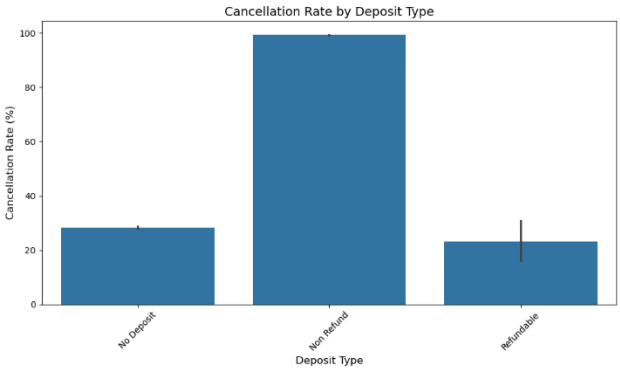


Figure 7: Cancellation Rate By Deposit Type

The bar chart (figure 7) displays the cancellation rate segmented by deposit type. Bookings with a non-refundable deposit type exhibit the highest cancellation rate, nearing 100%, which may initially seem paradoxical. This suggests that guests who select non-refundable options might be more uncertain in their plans or are booking speculatively. In contrast, refundable and no-deposit bookings show substantially lower cancellation rates, around 23%–28%, indicating that guests who choose more flexible booking conditions may be more intentional and committed to their stays.

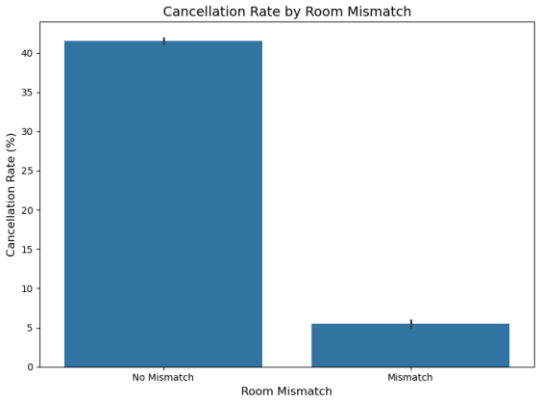


Figure 8: Cancellation Rate by Room mismatch

The bar chart by figure 7 illustrates the cancellation rate based on room mismatch status. Interestingly, the cancellation rate is significantly higher when there is no room mismatch (approximately 42%) compared to when a room mismatch occurs (just over 5%). This counterintuitive trend suggests that guests who experience a mismatch between their reserved and assigned room types are less likely to cancel their bookings. One possible explanation could be that mismatches are more common among last-minute or less flexible bookings, where guests proceed with their stay regardless of the discrepancy. This insight indicates that room mismatch may not be a strong driver of cancellations and could potentially serve as an indirect indicator of booking commitment.

Correlation Between variables

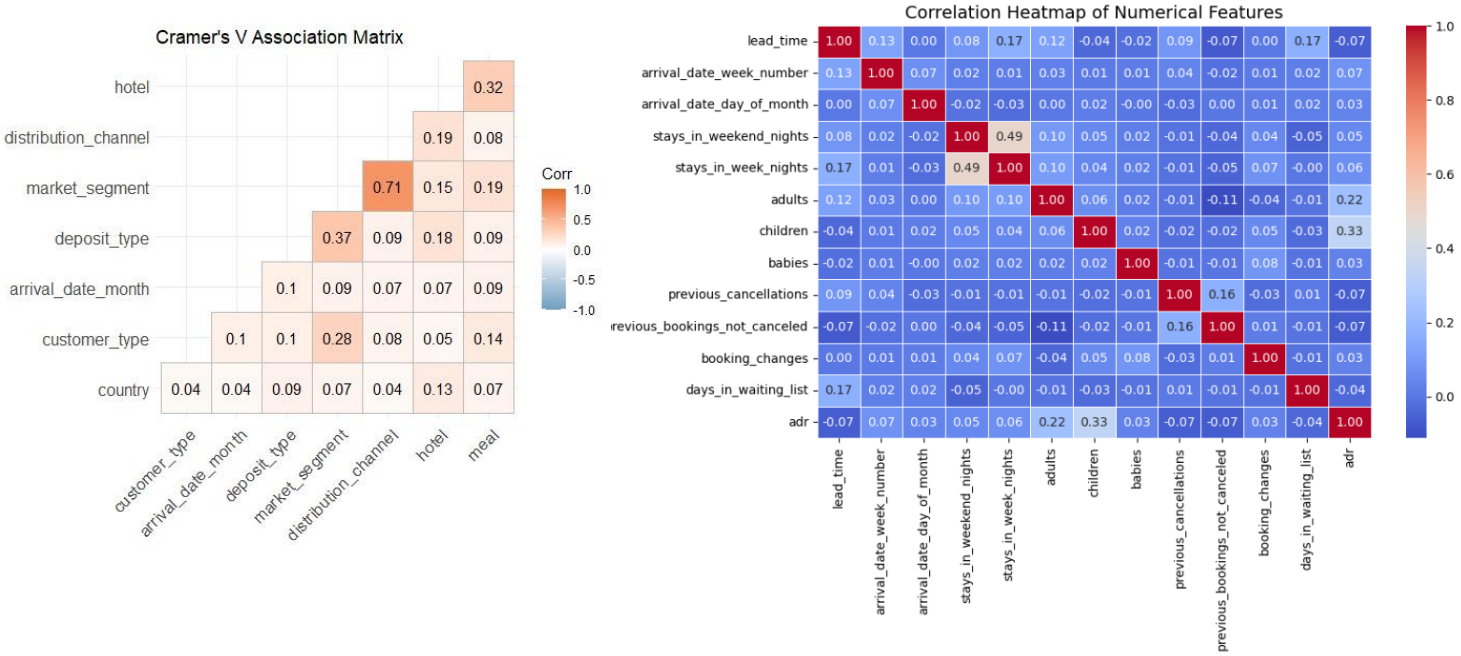
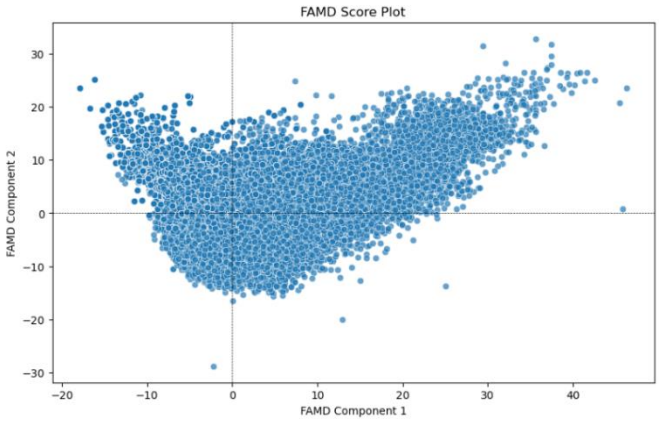


Figure 9: Correlation heatmaps

From Figure 9, the Cramér's V association matrix shows the strength of association between categorical variables. The correlation between the `market_segment` and `distribution_channel` variables is 0.71, indicating a strong relationship. This suggests that the type of market segment a guest belongs to is closely related to the channel they use to book the hotel. In figure correlation heatmap shows the relationship between numerical features in a hotel booking dataset. It highlights that most features have weak correlations except for a moderate positive correlation between **stays_in_week_nights** and **stays_in_weekend_nights (0.49).**

Cluster analysis



The overall crescent-shaped distribution of the points suggests the presence of potential natural groupings within the data even though these clusters are not distinctly separated in the visualization.

Figure 10: Score Plot

Outliers

Isolation Forest is used to identifying outliers because it is well suited with large datasets and does not assume any assumptions of a distribution.

```
Detected 4760 outliers out of 95189 samples.
```

It detects 5% of outliers in the dataset.

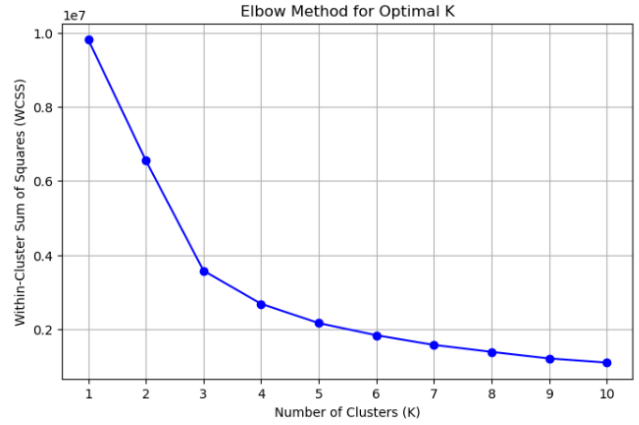
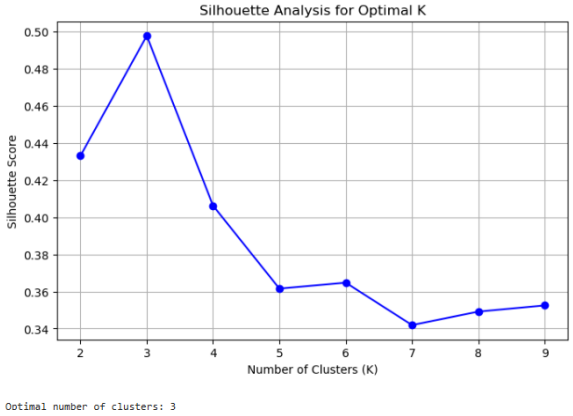


Figure 11: Cluster analysis - Elbow method for Optimal K




```
Optimal number of clusters: 3
```

Figure 12: Cluster analysis - Silhouette Analysis for optimal K

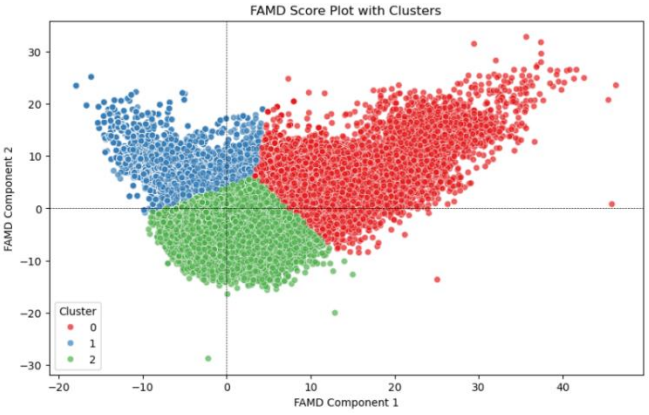


Figure 13: FAMD score plot with clusters

The optimal number of clusters was determined as 3 using both the Elbow Method and Silhouette Analysis. The Elbow Method shows a clear bend at k=3, while the highest silhouette score of 0.49 at k=3 indicates moderate clusters. Therefore, after fitting models, we checked if these clusters affect the model accuracy.

FAMD Analysis

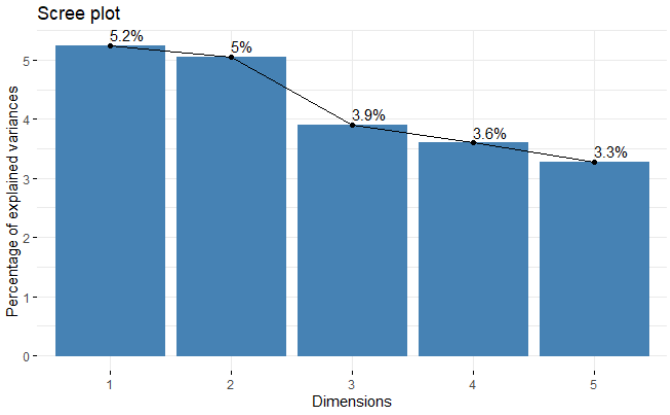
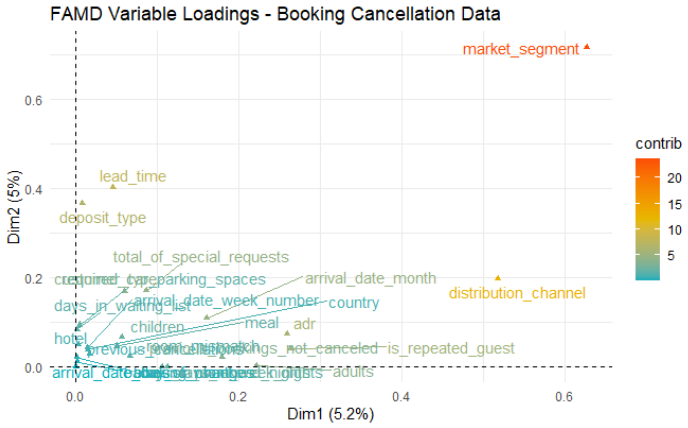


Figure 14: FAMD Analysis – Scree plot

Figure 15: FAMD analysis – Loading Plot



The scree plot from FAMD indicate that only 21% of variation explain by first 5 principal components. So that is not enough to understand full structure with principal components.The FAMD variable loadings plot reveals that market_segment and distribution_channel are the most influential variables, contributing strongly to the first two dimensions and highlighting their importance in explaining variations in booking cancellation behavior. However remainig 23 variables are clustered around origin indicating lower contribution for first 2 principal components.Additionally, correlation checks revealed no strong evidence of multicollinearity among the variables. Therefore, we decided not to perform further multicollinearity treatment at this stage.

Results of Advanced Analysis

Logistic Regression

Due to multicollinearity between the variables 'market segment' and 'distribution channel,' the logistic regression model violated key assumptions. To address this, we removed one of the correlated variables and reevaluated the model's accuracy. The results of this approach are presented in the table below.

	Original Model with multicollinearity	Reduced model 1 (- market segment)	Reduced Model 2 (- Distributional Channel)
Training Accuracy	81.45%	80.58%	81.47%
Testing Accuracy	81.62%	80.66%	81.65%

Table 2: Performance of Logistic Regression

The results exhibit that applying reduced model 2 that remove distributional channel variable gives the highest accuracy.

KNN

The tuned KNN classifier (k=6, weights='distance', metric='euclidean') processed numerical features via standardization and categorical variables through one-hot encoding, delivering 82.5% accuracy with 0.74 F1-score for cancellations.

	Accuracy	Precision		Recall		F1-score	
Train	86%	0	0.85	0	0.95	0	0.90

		1	0.9	1	0.72	1	0.80
Test	82%	0	0.82	0	0.92	0	0.87
		1	0.84	1	0.66	1	0.74

Table 3: Performance of KNN classifier model

The KNN model effectively identifies non-cancellations (92% recall) but shows moderate performance in predicting cancellations (66% recall). The model's balanced performance across classes (F1=0.74-0.87) makes it suitable for initial deployment, with future enhancements planned through comparative modeling approaches.

Classficication Tree

We performed classification tree using "GridSearchCV". Best Parameters: {'criterion': 'gini', 'max_depth': 20, 'min_samples_leaf': 5, 'min_samples_split': 20}

	Accuracy	Precision		Recall		F1-score	
Train	87%	0	0.88	0	0.93	0	0.90
		1	0.87	1	0.78	1	0.82
Test	84%	0	0.85	0	0.90	0	0.87
		1	0.81	1	0.73	1	0.77

Table 4: Performance of Classification Tree

XGBoost

GridSearchCV is used to optimize the hyperparameters of XGBoost model, which resulted in these optimal values: {'colsample_bytree': 0.9, 'learning_rate': 0.1, 'max_depth': 8, 'n_estimators': 350, 'scale_pos_weight': 1.6971834976765272, 'subsample': 0.8}. scale_pos_weight parameter, is used to address class imbalance since using SMOTE caused a significant drop in recall.

	Accuracy	Precision		Recall		F1-score	
Train	90%	0	0.93	0	0.91	0	0.92
		1	0.86	1	0.89	1	0.87
Test	85%	0	0.89	0	0.87	0	0.88
		1	0.79	1	0.82	1	0.81

Table 5: Performance of XGBoost

Random Forest

GridSearchCV was initially used for parameter optimization but recall for class 1 was significantly poor. Manual hyperparameter tuning was performed due to the high computational cost of large grid searches. The optimal hyperparameters found were: {n_estimators: 250, 'criterion':gini, 'max_depth': 24, 'max_depth': 24, 'min_sample_split': 4, 'min_sample_leaf': 2, 'random_state': 0, 'n_jobs': -1}. The criterion was set to 'gini' to address the imbalance because SMOTE caused a significant drop in recall.

	Accuracy	Precision		Recall		F1-score	
Train	90%	0	0.94	0	0.90	0	0.92
		1	0.85	1	0.90	1	0.87
Test	86%	0	0.89	0	0.87	0	0.88
		1	0.80	1	0.82	1	0.81

Table 6: Performance of Random Forest model

Best Model

After evaluating multiple machine learning models; both Random Forest and XGBoost performed similarly achieving the highest accuracy, precision, recall, and F1-score among all the models tested. However, **Random Forest was selected as the best model because it showed slightly higher test accuracy(86%)** compared to XGBoost and has better generalization without overfitting. Several techniques were applied to further improve the Random Forest model's performance, including feature selection, creating new features from existing ones, and adding cluster labels (derived from unsupervised clustering) as an additional feature and retrained the model. However, these efforts did not improve the model performance, confirming that the Random Forest model is already well-optimized for this task

Discussion and Conclusion

Hotel booking cancellations pose a major challenge for the hospitality industry, leading to significant revenue losses and operational inefficiencies. As highlighted in the problem description, cancellation rates average 25-35%, with third-party bookings showing even higher rates. This unpredictability renders it challenging for hotels to set prices, personnel, and resources for optimal efficiency. Our analysis explored the underlying causes of cancellations, revealing that factors such as booking channels (higher cancellations for online travel agencies), deposit types (paradoxically, non-refundable bookings had the highest cancellation rates), and seasonal trends (peaking in April and June) play critical roles.

To address this challenge, we developed predictive models using machine learning. Among these, the Random Forest model delivered the best performance, achieving 86% accuracy in identifying potential cancellations. This model enables hotels to take proactive measures; such as targeted promotions, flexible deposit policies, or dynamic pricing adjustments to mitigate losses.

However, real-world implementation should also consider customer behavior trends and market conditions for on-going improvement. Future enhancements could include integrating real-time data or exploring advanced AI techniques to improve prediction accuracy. Ultimately, this research provides hotel managers with actionable insights to reduce cancellations, optimize revenue management, and enhance operational efficiency, ensuring better stability in an otherwise unpredictable industry.

References

- Dataset and Code : https://drive.google.com/drive/folders/16zMiI4A5GX77APDdOQovaOsugtTzU0ff?usp=drive_link
- https://www.researchgate.net/publication/375272367_An_Analysis_of_Hotel_Booking_Cancellations_and_Factors_Affecting_Revenue_Generation.
- https://www.duettocloud.com/special-reports/rebooting-revenue-2022.
- https://www.hospitalitynet.org/opinion/4110788.html
- https://ujangriswanto08.medium.com/hands-on-with-lasso-and-ridge-logistic-regression-in-r-5e25a83dfd07.
- https://www.lexjansen.com/scsug/2018/Shreiber-Gregory-SCSUG2018-Assumption-Violations.pdf