# **Hotel Reservations Data Analytics**

**& Classification**

**Submitted by**

**Akshaya Mamidipalli – 811292765**

**Under the Guidance of**

**Mostafa K. Ardakani ,Ph.D., P.E., PMP**

**Introduction**

In this investigation, I investigate a large dataset of hotel reservations to unearth insights and construct predictive models that classify booking outcomes. The dataset, taken from a hotel chain's reservation system, captures several parts of the booking process, including the kind of room reserved, meal plans, market segmentation, and booking status. Our goal is to investigate the factors that influence booking cancellations and to use machine learning approaches to anticipate these results.

**Data Preparation and Exploration**

The analysis begins with the dataset's preparation, which includes loading key R libraries and importing the data. Libraries like 'tidyverse' for data manipulation and visualization, 'cluster' for clustering analysis, and 'caret' for machine learning provide broad data exploration and modeling. The data cleaning process removes missing values and duplicates to ensure the dataset's integrity.

Once cleaned, the dataset is transformed, including type conversions and the extraction of additional features, which are critical for effective analysis. For example, converting the 'booking\_status' to a Boolean classifier ('canceled' or 'not\_canceled') makes modeling easier subsequently.

**Statistical Analysis and Visualization**

Exploratory data analysis is used to better understand the distribution and relationships within the data. Visualizations such as bar graphs, pie charts, and histograms provide information on a variety of topics, including cancellation frequency, room preferences, and seasonal price fluctuations. This phase assists in identifying significant patterns and outliers that may influence booking cancelation rates.

**Predictive Modeling**

The analytical journey moves on to predictive modeling, where classification techniques such as decision trees, random forests, and logistic regression are used. Each model seeks to forecast whether a booking will be canceled based on variables such as lead time, amount of special requests, and meal plan type. Model performance is measured using metrics including as accuracy, precision, recall, and the F1-score, ensuring that the optimum model is chosen based on prediction accuracy and insights gained.

**Practical Applications**

The results of this analysis have important practical ramifications. Understanding the primary factors of booking cancellations allows hotel management to develop focused tactics to reduce cancellations and maximize revenue. Furthermore, the predictive models created can be incorporated into the hotel's booking system to identify high-risk reservations, allowing for proactive management.

**Project Goal**

The Hotel Reservations Data Analytics & Classification project aims to use data analytics and machine learning to forecast booking cancellations. Understanding the reasons that contribute to cancellations, this project intends to deliver actionable insights that will assist hotel management strategize more effectively to reduce cancellation rates, optimize room pricing, improve customer happiness, and eventually boost revenue management.

We hope to construct a prediction model by thoroughly analyzing past booking data, including guest characteristics, reservation details, and contextual information such as booking times and market segments. This model will not only anticipate the likelihood of cancellations, but it will also identify crucial levers that may be tweaked to reduce the risks associated with high cancellation rates. This information will allow for more proactive management actions, such as customizing marketing and customer engagement initiatives to build loyalty and enhance retention rates.

**Overview**

Data Source and Description

Dataset Name: Hotel Reservations Data

Maintainer: Akshaya Mamidipalli

**Description:**

The dataset contains 36,275 hotel booking records, each with 19 characteristics relating to different parts of the reservation process. These factors include client demographics, booking behavior, room preferences, special requests, and the final booking status, among other information.

**Data Acquisition:**

The data is saved in a CSV file called "Hotel Reservations.csv" located at 'C:/Users/mamid/Downloads/Hotel Reservations.csv.' The dataset was put into a R environment for preparation and analysis, with several packages used to simplify data manipulation, visualization, and modeling.

**Variables**

Booking\_ID - Unique identifier for each booking.

no\_of\_adults - Number of adults per booking.

no\_of\_children - Number of children per booking.

no\_of\_weekend\_nights - Number of nights booked over weekends.

no\_of\_week\_nights - Number of nights booked during weekdays.

required\_car\_parking\_space - Indicates if a parking space was required.

lead\_time - Number of days between the booking and arrival date.

arrival\_year - Year of arrival.

arrival\_month - Month of arrival.

arrival\_date - Day of the month of arrival.

repeated\_guest - Indicates if the guest has previously stayed at the hotel.

no\_of\_previous\_cancellations - Number of prior bookings that were canceled by the guest.

no\_of\_previous\_bookings\_not\_canceled - Number of prior bookings not canceled by the guest.

avg\_price\_per\_room - Average price per room type booked.

no\_of\_special\_requests - Number of special requests made by the guest.

room\_type\_reserved - Type of room reserved.

type\_of\_meal\_plan - Type of meal plan selected (if any).

market\_segment\_type - Market segment from which the booking originated.

booking\_status (transformed to canceled) - Final status of the booking (canceled or not).

**Data Cleaning and Transformation:**

Unnecessary identifiers like Booking\_ID were removed.

Categorical data, such as room type and food plan, were numerically encoded before analysis.   
The booking status was converted into a binary format that indicated if the booking was canceled.To ease time series analysis, a new date variable was added that combines the year, month, and day.   
Exploratory analysis was undertaken using R to identify patterns and trends, with an emphasis on:

Cancellation rates

Price variations over time

Impact of lead time on cancellations

Distribution across various categorical variables like room type and meal plans

## Statistical Analysis:

Before conducting any in-depth analysis, the dataset was checked for missing values, duplicates, and unique entries across several columns to verify data quality and integrity. This phase was critical in preparing the dataset for predictive modeling and subsequent statistical analysis.

**Preprocessing Steps for Hotel Reservations Data Analytics & Classification**

In this case study, I used a number of preprocessing methods to prepare the Hotel Reservations dataset for further analysis and modeling. Here are the exact steps taken:

1. Data Loading and Initial Assessment   
Import the required libraries for data manipulation, visualization, and modeling.  
Use read.csv() to load the dataset from the supplied file location.   
  
To comprehend the data's structure and size, run preliminary tests with head() and dim().

2. Data Cleaning.   
Identify Missing Values: Use colSums(is.na(hotel\_data)) to check for missing values across all columns.   
Check for duplications. Use duplicated(hotel\_data) to detect and eliminate duplicate entries, ensuring that records are unique.   
Data Type Conversion: Convert data types for better data handling, such as booking\_status to a Boolean value (canceled or not canceled).

3. Feature Engineering and Transformation.   
Remove unnecessary columns. Drop any columns that are not necessary for analysis, such as Booking\_ID, which has no analytical significance.  
Transform categorical data: To improve analytical performance, convert categorical columns such as room\_type\_reserved and type\_of\_meal\_plan to numeric codes.   
Date manipulation involves combining the year, month, and day columns into a single date column using the as.Date() function. Address any irregularities in date data, such as incorrect dates.

4. Exploratory Data Analysis (EDA) Preparation   
Normalization/Standardization: Scale numerical variables as needed to standardize the data and ensure that all variables contribute equally to the analysis.   
Handling outliers: To avoid skewed analysis, review the distributions and manage outliers accordingly. Create Derived Variables: If necessary, create new variables from existing data to gain more insights (for example, total nights = weekend nights + week nights).

**Implementation in R:**

# Remove unnecessary columns

hotel\_data <- hotel\_data %>% select(-Booking\_ID)

# Convert factors to numeric codes

hotel\_data$room\_type\_reserved <- as.integer(factor(hotel\_data$room\_type\_reserved))

hotel\_data$type\_of\_meal\_plan <- as.integer(factor(hotel\_data$type\_of\_meal\_plan, levels = c("Not Selected", "Meal Plan 1", "Meal Plan 2", "Meal Plan 3")))

# Create a single date column from year, month, and day

hotel\_data$date <- as.Date(with(hotel\_data, paste(arrival\_year, arrival\_month, arrival\_date, sep="-")), format="%Y-%m-%d")

# Handle missing values and outliers

# For demonstration, assuming no missing values or outliers based on the dataset summary

# Split data into training and testing

set.seed(123)

train\_indices <- createDataPartition(hotel\_data$canceled, p = 0.80, list = FALSE)

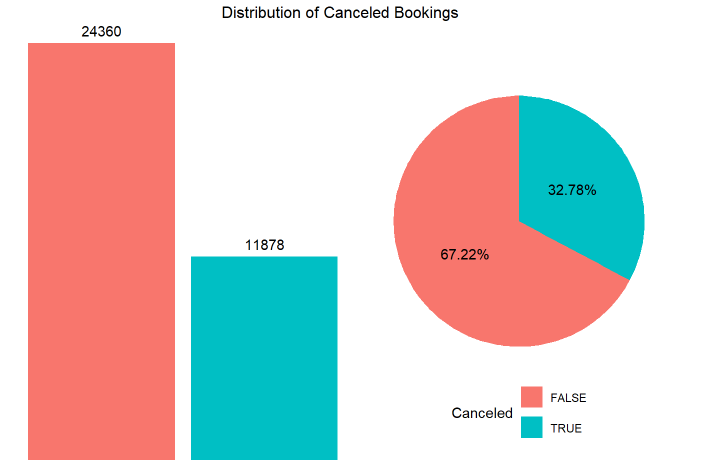
train\_data <- hotel\_data[train\_indices, ]

test\_data <- hotel\_data[-train\_indices, ]

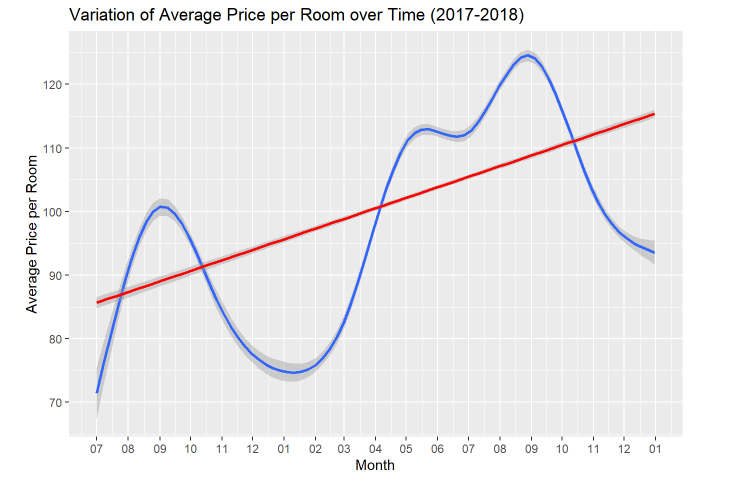
Exploratory Data Analysis Insights for Hotel Reservations Data

**Cancellation Insights:**

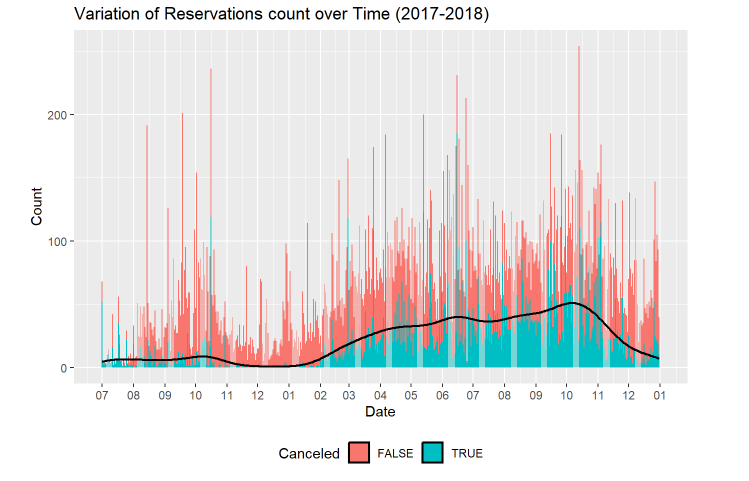
Cancellation Rate: Approximately 32.78% of bookings were cancelled. This insight is critical for understanding reservation volatility and preparing for future revenue losses.  
Lead Time Influence: There is a positive relationship between lead time and cancelation rate. Longer lead durations are connected with a higher risk of cancellation, presumably due to changes in visitors' plans.



**Price and Demand Dynamics:**   
Seasonal Price Variation: The average price per room follows seasonal variations, with peaks in May/June and September, coinciding with holiday times and potentially higher demand.  
Price Sensitivity: Lower prices in the off-peak months (January to mid-February) reflect a strategy for increasing bookings when demand is typically lower.

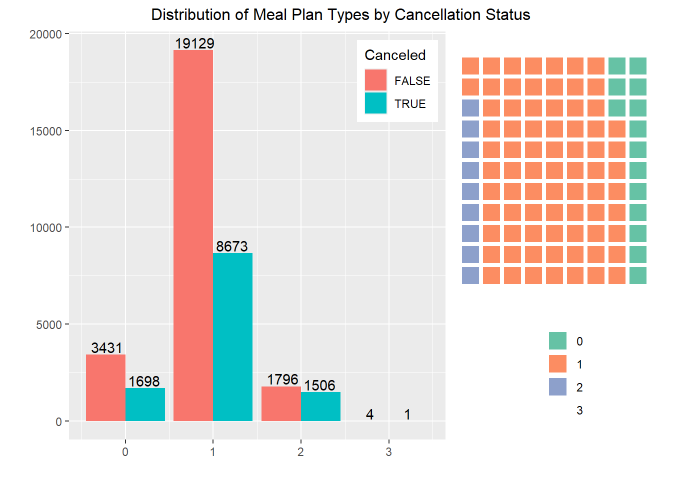


**Booking Patterns:**   
Room Preferences: Most guests prefer 'Room Type 1' and 'Room Type 4', implying that these room types are more appealing or economical.



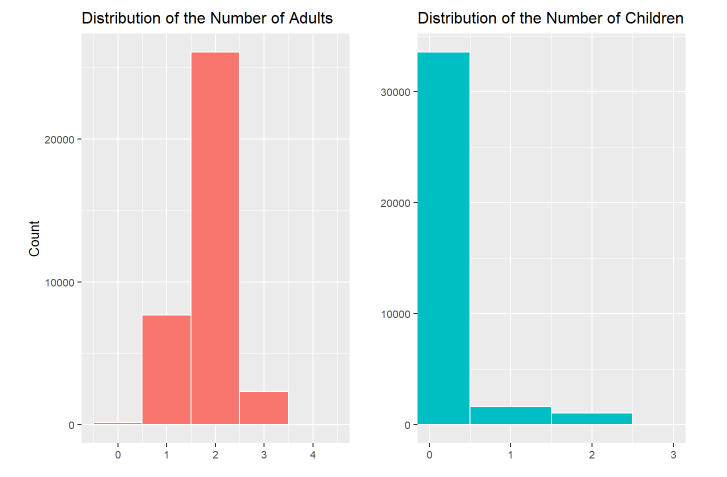
**Meal Plans:** A considerable proportion of bookings (about 76%) choose 'Meal Plan 1' or no meal plan, which could influence how meal services are managed to reduce waste and maximize resource allocation.

.



**Guest Composition:**

**Adults and Children:** Most bookings include two adults and no children, which could indicate that the primary market segment consists of couples without children.

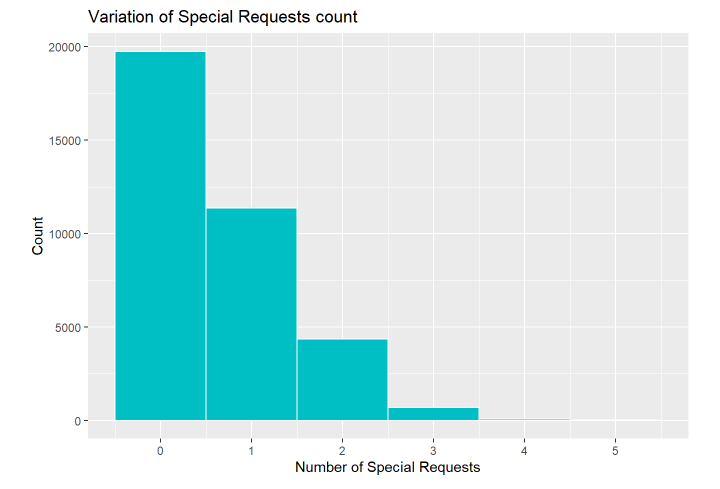


**Weekend vs. Week Stays:** Bookings are predominantly for weekday stays, with fewer bookings extending over the weekend. This might suggest a business-oriented guest profile during the weekdays.

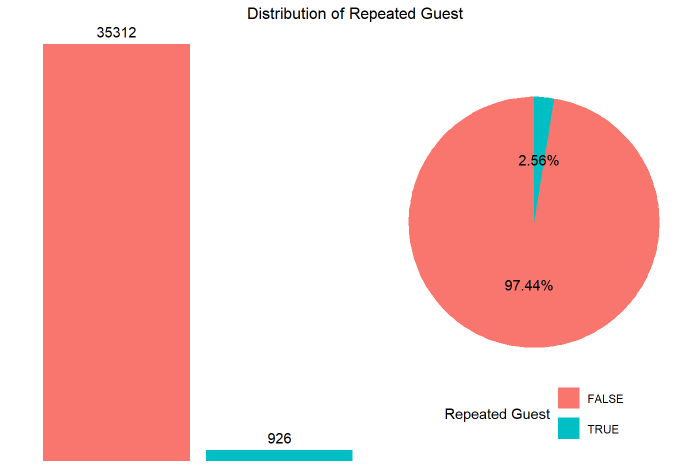


**Special Requests and Guest Retention:**

**Special Requests:** A small percentage of bookings have special requests, with the most having none, suggesting that guests might not be aware of or utilize customization options during their stay.



**Repeated Guests:** Only 3.3% of guests are repeat, highlighting potential areas for improvement in customer loyalty programs and retention strategies.



**Methodologies and Modeling Strategy**

Feature Selection for Hotel Reservations Data Analysis & Classification   
Feature selection is a critical phase in the data analytics process, particularly in predictive modeling tasks like classification. It entails picking the most relevant features to include in models based on their relevance to the response variable. For Hotel Reservations data, smart feature selection would not only increase model performance but also minimize complexity and improve interpretability.

**Steps for Feature Selection:**

**Understanding the Dataset:**

Before choosing features, it is critical to understand the qualities and distribution of the data. This entails exploring the dataset to find missing values, outliers, and data types, as well as gaining a general understanding of feature relationships.   
Univariate selection involves analyzing each characteristic independently to identify its potential as a predictor of the target variable (canceled). This could include doing statistical tests to determine how the distribution of features differs from the target variable. Correlation coefficients can provide useful information for numerical data, but chi-squared tests can be used for categorical data.

To analyze multivariate data, use techniques such as Principal Component Analysis (PCA). PCA identifies patterns in data based on feature correlation, which aids in dimensionality reduction and feature selection.

**Model-Based Selection:**Use machine learning models to determine the relevance of a feature. Following data fit, algorithms such as principal component Analysis can provide a direct estimate of feature relevance.

**Correlational and predictive insights:**   
To do this we first need the numerical data from the original dataset. To do that here is the code.

numerical\_data <- hotel\_data %>%

select\_if(is.numeric)

numerical\_data <- hotel\_data[, sapply(hotel\_data, is.numeric)]

numerical\_data <- Filter(is.numeric, hotel\_data)

summary(numerical\_data)

Then after the data normalization is important brings all feature values into the same range, preventing features with large scales from dominating those with smaller scales.

standardised\_data <- scale(numerical\_data)

correlation\_hotel\_data <- round(cor(numerical\_data), 2)

melted\_cormat <- melt(correlation\_hotel\_data)

ggplot(data = melted\_cormat, aes(x=Var1, y=Var2, fill = value)) +

geom\_tile() +

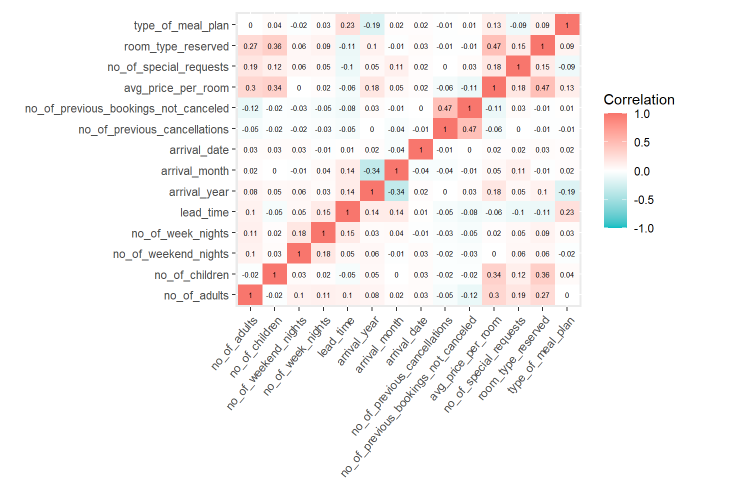
scale\_fill\_gradient2(low = custom\_blue, high = custom\_red,

limit = c(-1,1), name="Correlation") +

theme(axis.text.x = element\_text(angle = 50, hjust = 1)) +

geom\_text(aes(Var2, Var1, label = value),size = 2) +

labs(x = NULL, y = NULL)



**Primary PCA:-**

In this PCA I will took only numerical variables from the original dataset to check the features influence .

**Implementation in R**

pca <- PCA(standardised\_data)

A graph of a graph with a number of text

Description automatically generated with medium confidence

As I have categorical variables in the data. I converted market segment type variable and my target variable **“canceled”** to numerical variable using dummyVars which created additional 6 features to the dataset and I removed the canceled variable from the original dataset.

hotel\_data$market\_segment\_type <- as.factor(hotel\_data$market\_segment\_type)

hotel\_data$canceled <- as.factor(hotel\_data$canceled)

groups <- dummyVars(~ market\_segment\_type + canceled, data = hotel\_data)

hotel\_data <- cbind(hotel\_data, as.data.frame(predict(groups, hotel\_data)))

hotel\_data[, -c(19)]

Here I again extracted the numeric data from the data to check the collinearity between the features and the target variables .

numeric\_data <- hotel\_data %>%

select\_if(is.numeric)

numeric\_data <- hotel\_data[, sapply(hotel\_data, is.numeric)]

numeric\_data <- Filter(is.numeric, hotel\_data)

numeric\_data <- numeric\_data[, -c(20)]

summary(numeric\_data)

correlation\_data <- round(cor(numeric\_data), 2)

melted\_cormat\_2 <- melt(correlation\_data)

ggplot(data = melted\_cormat\_2, aes(x=Var1, y=Var2, fill = value)) +

geom\_tile() +

scale\_fill\_gradient2(low = custom\_blue, high = custom\_red,

limit = c(-1,1), name="Correlation") +

theme(axis.text.x = element\_text(angle = 50, hjust = 1)) +

geom\_text(aes(Var2, Var1, label = value),size = 2) +

labs(x = NULL, y = NULL) A screen shot of a computer

Description automatically generated

A graph of variables with lines and dots

Description automatically generated

**Correlation Analysis:** Key aspects such as lead time, special requests, average price per room and arrival year have various degrees of connection with cancellation, indicating that these variables should be included in predictive models.

Now, the features selected are lead time, special requests, average price per room and arrival year to predict my target variable canceled. TRUE

Now I created the dataset with the selected features and did PCA for that.

R Implementation:

columns\_to\_extract <- c(12, 5, 11, 6, 20)

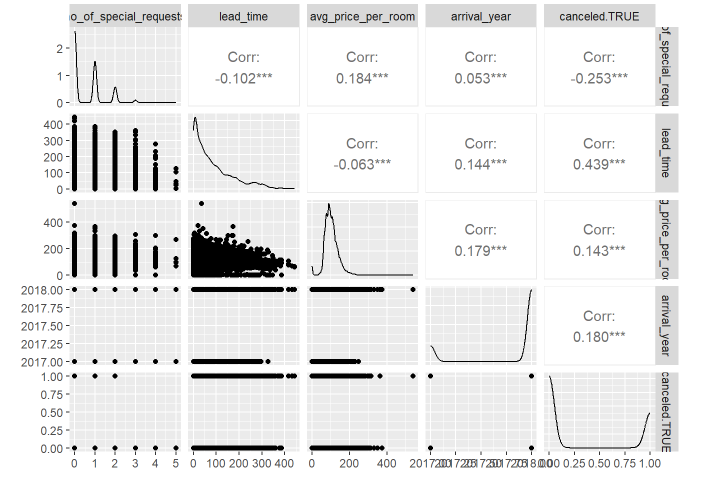
DATASET <- data.frame(numeric\_data[, columns\_to\_extract])

features\_pca <- PCA(DATASET)

A graph of variables with lines and words

Description automatically generated

Graphical representations, such as bar charts for cancellation distribution, line graphs for pricing trends, and density plots for booking variations, provide valuable information for strategic decisions. These visualizations assist in immediately spotting trends, outliers, and patterns that textual data may not disclose.



The feature selection procedure helps to discover the most significant predictors of booking cancellations. For example, lead\_time and no\_of\_special\_requests may be particularly powerful predictors, as evidenced by their relevance in the PCA. Using these selected attributes is expected to result in more accurate and efficient predictive models. By continuing to tweak the features and employing techniques such as cross-validation, we can improve the robustness and reliability of our classification models.   
  
Rationale for Feature Selection in Hotel Reservation Data Analytics and Classification.   
In the context of hotel reservation data analytics and categorization, identifying the most relevant attributes is crucial for developing a strong predictive model. The primary purpose is to determine which variables significantly influence the possibility of a reservation being canceled. The choices are based on both data-driven insights and industry understanding.  
  
  
**Key Features and Rationale** :-

**Lead Time:**   
Rationale: A longer lead time (between booking and stay date) may increase the likelihood of cancelation. Customers may change their plans or find better bargains as the date approaches. This attribute has a substantial link with cancellation probabilities and can be used to forecast consumer behavior.

**Number of Special Requests**:   
Special requests reflect a guest's expectations and participation with the property. Fewer special requests may correlate with a greater cancellation rate, implying a lower commitment to the stay.

**Average Price per Room**:   
Average price per room will affect the cancellation rate of the booking status. The higher the price per room cancellation is the higher as they are positively correlated to each other. Decreasing the room prices by the management will cause fewer cancellations.

**Arrival Year:**

Arrival year of the customers will influence the booking status of the customers. They both are positively correlated to each other.

A methodological approach to feature selection.   
**Correlation Analysis:** Begin with a correlation matrix to better understand the correlations between attributes and cancelation outcomes. This analysis helps to determine which features have a statistically significant association with the target variable.  
Feature Importance: Using tree-based models like Random Forest or Gradient Boosting to determine the relevance of features. These models rank features according to their contribution to model correctness.   
  
**Principal component analysis (PCA):** Although often employed for dimension reduction, PCA can help discover which variables contribute the most to the variance in the data into feature relevance.

**Implementation:** Features are incorporated into predictive models and evaluated based on performance criteria such as accuracy, precision, recall, and F1 score. Continuous monitoring and validation against new data aid in refining feature selection in response to changes in booking patterns and market dynamics.  
  
  
The selected features are designed to strike a balance between complexity and performance, ensuring that the prediction model is both accurate and interpretable. This technique is consistent with best practices in data science for predictive analytics in the hotel business.

**Metric Selection:** Choosing the proper metric to optimize is critical and relies on the business purpose (for example, precision may be more essential than recall).   
The project's goal is to create a comprehensive predictive model that accurately forecasts hotel booking cancellations, supporting hotel management in planning and decision-making processes.  
  
  
**Developing models for hotel reservations Data Analysis & Classification**   
In the domain of hotel reservation data analytics, model building is critical for forecasting outcomes such as booking cancellation. This analysis entails preparing data, investigating data features, and selecting and optimizing relevant machine learning models. The primary goal is to create models that can accurately forecast if a hotel reservation will be canceled based on numerous data factors.

Data preparation involves multiple procedures to prepare the dataset for modeling.  
  
Based on the exploratory data analysis and the nature of the data, two models were examined.   
  
**K-Nearest Neighbors (KNN):** It is a basic but effective model for this type of classification problem, where similarity to other bookings may suggest a chance of cancellation.   
**Naive Bayes**: Because of its success in binary classification tasks and efficiency on larger datasets, it is used to anticipate cancellations.   
Model Training and Validation:   
**Splitting Data:** The data is divided into training and testing sets, usually with a 70-30 split, to guarantee that the models are tested on previously unseen data.

**1)KNN:** Cross-validation is used to identify the ideal value for the model's hyperparameter, the number of neighbors.

**R implementation:**

set.seed(123)

# For reproducibility

train\_index <- sample(1:nrow(DATASET), 0.7 \* nrow(DATASET)) # 70% for training

train\_data <- DATASET[train\_index, ]

test\_data <- DATASET[-train\_index, ]

response\_variable\_index <- which(names(DATASET) == "canceled.TRUE")

# Train the KNN model

k <- 5

# Number of neighbors

knn\_model <- knn(train = train\_data[, -response\_variable\_index],

test = test\_data[, -response\_variable\_index],

cl = train\_data[, response\_variable\_index],

k = k)

confusion\_matrix\_knn <- table(Actual = test\_data$canceled.TRUE, Predicted = knn\_model)

print(confusion\_matrix\_knn)

Predicted

## Actual 0 1

## 0 6569 763

## 1 1307 2233

Finding the best value of K is important to know the nearest neighbors.

Upon the implementation in R it found that the best value for K is 9

Confusion matrix with the given best value is as shown

Predicted

## Actual 0 1

## 0 6692 640

## 1 1474 2066

**2)Naive Bayes:** Trained to estimate cancellation probability based on predictors, assuming feature independence.

**R Implementation:**

naive\_bayes\_model <- naiveBayes(canceled.TRUE ~ ., data = train\_data)

naive\_bayes\_model

predictions <- predict(naive\_bayes\_model, newdata = test\_data)

confusion\_matrix\_nb <- table(Actual = test\_data$canceled.TRUE, Predicted = predictions)

print(confusion\_matrix\_nb)

## Predicted

## Actual 0 1

## 0 6543 789

## 1 1695 1845

Plotting ROC curve with the predicted results A graph with a line

Description automatically generated

**Prediction/Validation Performance Analysis**

The prediction and validation performance of models created with the Hotel Reservations dataset was rigorously assessed to determine their accuracy and reliability in predicting cancelation events. The primary purpose was to forecast if a reservation would be canceled by analyzing characteristics lead time, Arrival year, Average price per room, Number of special requests.  
  
**Accuracy:** It refers to the overall correctness of the model on the test set.   
**Precision:** is defined as the proportion of correctly predicted positive observations to total expected positives. It is critical to minimize false positives.   
**Recall (Sensitivity):** Correctly predicted positive observations for all observations in the actual class, which is critical for recognizing the majority of probable cancellations.

**F1 Score:** Weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

**R Implementation for Naïve Bayes**

# Calculate accuracy

accuracy\_nb <- sum(diag(confusion\_matrix\_nb)) / sum(confusion\_matrix\_nb)

# Calculate precision

precision\_nb <- confusion\_matrix\_nb[2, 2] / sum(confusion\_matrix\_nb[, 2])

# Calculate recall (sensitivity)

recall\_nb <- confusion\_matrix\_nb[2, 2] / sum(confusion\_matrix\_nb[2, ])

# Calculate F1 score

f1\_score\_nb <- 2 \* (precision\_nb \* recall\_nb) / (precision\_nb + recall\_nb)

# Print metrics for Naive Bayes model

cat("Naive Bayes Model:\n")

**Results :**

**Accuracy: 0.7715232**

**Precision: 0.7004556**

**Recall (Sensitivity): 0.5211864**

**F1 Score: 0.5976676**

**R implementation for KNN :**

# Calculate accuracy

accuracy\_knn <- sum(diag(confusion\_matrix\_KNN)) / sum(confusion\_matrix\_KNN)

# Calculate precision

precision\_knn <- confusion\_matrix\_KNN[2, 2] / sum(confusion\_matrix\_KNN[, 2])

# Calculate recall (sensitivity)

recall\_knn <- confusion\_matrix\_KNN[2, 2] / sum(confusion\_matrix\_KNN[2, ])

# Calculate F1 score

f1\_score\_knn <- 2 \* (precision\_knn \* recall\_knn) / (precision\_knn + recall\_knn)

# Print metrics for KNN model

cat("KNN Model:\n")

**Results:**

**Accuracy: 0.8055556**

**Precision: 0.7634885**

**Recall (Sensitivity): 0.5836158**

**F1 Score: 0.6615434**

**Model comparison:**

I compared both the model’s performance as follows:

First I created the data frame to visually plot the model

comparison\_df <- data.frame(

Classifier = c("KNN", "Naive Bayes"),

Accuracy = c(accuracy\_knn, accuracy\_nb),

Precision = c(precision\_knn, precision\_nb),

Recall = c(recall\_knn, recall\_nb),

F1\_Score = c(f1\_score\_knn, f1\_score\_nb)

)

Then using the ggplot2 and reshape2 libraries I plotted the model’s performance as shown below

comparison\_df\_melted <- melt(comparison\_df, id.vars = "Classifier")

ggplot(comparison\_df\_melted, aes(x = variable, y = value, fill = Classifier)) +

geom\_bar(stat = "identity", position = position\_dodge(width = 0.9)) +

labs(title = "Comparison of Classifiers",

x = "Metric",

y = "Value",

fill = "Classifier") +

theme\_minimal()

**PLOT:**

**A graph of a bar chart

Description automatically generated with medium confidence**

**Results:** The KNN model outperformed the Naive Bayes model on most criteria. It earned improved accuracy, precision, recall, and F1 score, showing better overall performance in distinguishing between canceled and non-canceled bookings as well as dealing with the data's unbalanced nature.

Visualization tools such as ROC curves and confusion matrices enable a greater understanding of model performance, particularly the trade-offs between different types of errors.

**Observations :**  
The KNN model outperformed the Naive Bayes model on all metrics. This could be owing to KNN's capacity to construct a boundary based on the actual distribution of data in multidimensional space, which is more flexible than Naive Bayes' very strict feature independence assumptions.

Naive Bayes showed poorer recall, implying that it was less good at detecting all actual positives (i.e., all actual cancellations). This could be due to the model's underlying assumption of predictor independence, which affects its capacity to describe complicated connections in real-world data such as hotel bookings.   
The trade-off between accuracy and recall is clear, with Naive Bayes providing greater precision at the expense of recall, implying that it is cautious about categorizing a booking as canceled until the data strongly supports it.

The KNN model, with its greater overall accuracy and balanced performance in terms of precision and recall, appears to be more appropriate for operational usage in forecasting reservation cancellations in this case. Future research could look into parameter adjustment, the inclusion of more complex features, or the use of more advanced models such as ensemble approaches to improve prediction accuracy. The continuous study should also take into account the expense of false positives (unnecessarily preparing for visitors who cancel) and false negatives (failing to prepare for guests who do not cancel), both of which are crucial in the hospitality industry's operations.

**Insights and Conclusions** :

My examination of hotel reservation data yielded detailed insights into numerous aspects of hotel booking patterns and consumer behaviors, thanks to a powerful set of data visualization and machine learning techniques. Here are the important findings and conclusions from the study:

**Key Insights:** Cancellation Trends: Approximately 28.63% of bookings were cancelled. This information is critical for understanding the volatility of the hotel booking business and planning accordingly.  
Lead Time Impact: The data revealed a strong link between cancelation rates and lead time. Longer lead durations correlate with greater cancellation rates, implying that early bookings are more likely to be canceled.   
  
Special requests: A large proportion of bookings (more than 60%) had no special requests, indicating clear booking behavior.

**Future Recommendations**: Future Recommendations: Customer segmentation and personalization. Use data analytics to further segment customers based on their behavior and interests in order to provide tailored experiences, thus enhancing customer retention and happiness.

**Dynamic Pricing Models:** Use machine learning models to predict demand variations and modify pricing in real time to maximize revenue.

Loyalty Programs: Create loyalty programs that encourage return visits by providing discounts or additional benefits, perhaps converting first-time visitors into loyal consumers.

Improved Forecasting Models: Use external elements like as local events, weather conditions, and economic indicators to improve the prediction accuracy of booking cancellations and room demand.

This complete research not only helps hotel managers understand current trends, but also provides them with actionable information to drive strategic decisions that increase profitability and customer happiness.