Drashti Khatra - 11053803

Sahil Patel - 11045739

Yashkumar Kalariya - 11045980

**HOTEL DEMAND**

1. **Introduction and Objectives**

The hotel industry operates in a highly competitive and dynamic market where making data-driven decisions is crucial for success. With new data mining and predictive analytics techniques, hotels can use their customer and operational data to gain valuable insights and make informed decisions. By using data, hotels can improve revenue management, marketing strategies, customer satisfaction, and gain a competitive advantage.

The primary objective of this project is to utilize data mining and predictive analytics methods to analyze hotel booking demand data and provide actionable insights that support strategic decision-making. The project aims to address two key business objectives:

1. Identifying the main factors influencing hotel booking cancellations and developing predictive models to forecast the likelihood of a booking being canceled.
2. Segmenting customers based on their lifetime value using techniques such as RFM (Recency, Frequency, Monetary) analysis to enable targeted marketing efforts and maximize customer profitability.

To achieve these objectives, the project will follow a structured approach, utilizing the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. This methodology provides a systematic framework for conducting data mining projects, ensuring a comprehensive and iterative process that aligns with business goals.

The project will focus on two research questions that align with these business objectives:

1. What are the key factors that significantly impact hotel booking cancellations, and how can machine learning algorithms be employed to develop a predictive model that identifies bookings with a high likelihood of cancellation?
2. How can customer segmentation techniques, such as RFM (Recency, Frequency, Monetary) analysis, be applied to the hotel booking data to identify distinct customer groups based on their lifetime value, enabling the development of targeted marketing strategies to maximize customer profitability?

By addressing these research questions and using data mining and predictive analytics, this project aims to provide valuable insights and recommendations that can directly support hotel management in making informed decisions. The findings will contribute to the development of strategies for revenue optimization, customer segmentation, and targeted marketing initiatives, ultimately leading to increased profitability and customer satisfaction in the highly competitive hotel industry.

1. **Literature Review**

The hotel industry is using data analysis and prediction techniques more and more to improve different parts of their business. This includes forecasting demand, grouping customers, and managing revenue. This literature review summarizes the important research in these areas and highlights the key findings and how they relate to the current project.

One of the most important aspects of managing a hotel is accurately predicting demand, as this directly impacts revenue management and operational planning. Rajopadhye et al. (2001) developed a neural network model to predict hotel room demand. It performed better than traditional time series methods. The study showed the potential of machine learning techniques to find complex patterns and improve the accuracy of demand predictions.

Building on this, Haensel and Koole (2011) used a combination of time series forecasting and machine learning techniques to predict hotel booking cancellations. Their research emphasized how important it is to consider cancellation behavior when forecasting demand, as it can have a big impact on hotel operations and revenue management strategies.

Customer segmentation is another important area where data mining techniques have been applied to improve hotel management. By understanding the distinct characteristics and behavior patterns of different customer groups, hotels can develop targeted marketing strategies and offer personalized services. Furthermore, customer lifetime value (CLV) has emerged as an important metric for customer segmentation and targeted marketing in the hotel industry. Reinartz and Kumar (2003) emphasized the importance of CLV in identifying high-value customer segments and developing strategies to maximize their profitability. Khajvand et al. (2011) applied RFM (Recency, Frequency, Monetary) analysis to identify high-value customer segments in the hospitality sector. Their research showed how RFM analysis can provide insights into customer behavior and value, enabling hotels to focus their marketing efforts on the most profitable segments.

The integration of data mining and predictive analytics in hotel revenue management has shown promising results. El Gayar et al. (2011) developed a decision support system that incorporates data mining techniques to optimize hotel room pricing. The system used association rule mining and decision tree algorithms to identify patterns in customer booking behavior and dynamically adjust prices accordingly. This research demonstrates how data mining techniques can be used to make data-driven pricing decisions, optimizing revenue and occupancy rates. The insights gained from such systems can help hotels remain competitive in the dynamic and ever-changing market conditions.

In conclusion, the literature review shows that data mining and predictive analytics are being used more and more in the hotel industry. Previous studies have used various techniques, such as neural networks, time series forecasting, and machine learning, to solve problems in predicting hotel demand and booking cancellations. They have also used customer segmentation techniques, like RFM analysis and CLV, to find high-value customer groups and create targeted marketing plans. Using data mining in hotel revenue management systems has also shown good results in optimizing room prices and increasing revenue and occupancy rates.

This project aims to build on this existing knowledge by focusing on two main areas: predicting hotel booking cancellations and grouping customers based on their lifetime value. By using machine learning and RFM analysis, the project will develop models that can identify bookings that are likely to be cancelled and divide customers into groups based on their profitability. These insights will help hotels make decisions to reduce revenue loss from cancellations and create targeted marketing strategies to increase customer profitability.

1. **Data, Data Sources, and Data Characteristics**

The dataset used for this project contains 119,390 observations and 32 variables(Table 1), including both categorical and numerical variables. The data covers hotel bookings from July 1st, 2015, to August 31st, 2017, providing a comprehensive view of the hotel's booking activities over a period of more than two years.

One of the key characteristics of this dataset is the presence of a binary target variable, 'is\_canceled', which indicates whether a booking was canceled (1) or not (0). This variable is crucial for addressing the first research question of identifying the main factors influencing hotel booking cancellations and developing predictive models to forecast the likelihood of a booking being canceled.

Another important aspect of the data is the availability of variables related to customer segmentation, such as 'customer\_type', 'distribution\_channel', 'market\_segment', and 'is\_repeated\_guest'. These variables provide valuable information for addressing the second research question of segmenting customers based on their lifetime value using techniques like RFM (Recency, Frequency, Monetary) analysis.

The dataset also includes variables related to the booking details, such as 'lead\_time', 'arrival\_date\_year', 'arrival\_date\_month', 'stays\_in\_weekend\_nights', 'stays\_in\_week\_nights', 'adults', 'children', 'babies', 'meal', 'reserved\_room\_type', 'assigned\_room\_type', 'deposit\_type', and 'adr' (Average Daily Rate). These variables can be used to gain insights into the booking patterns and preferences of the hotel's customers.

Additionally, the dataset contains variables related to the customers' previous booking behavior, such as 'previous\_cancellations', 'previous\_bookings\_not\_canceled', and 'booking\_changes'. These variables can be useful for understanding the customers' loyalty and predicting their future behavior.

Overall, the dataset provides a rich and diverse set of variables that can be utilized to address the research questions and gain valuable insights into hotel booking demand and customer behavior.

3.1 Data overview

The data dictionary (Table 1) provides a comprehensive overview of the 32 variables included in the dataset, their data types, and descriptions. This information is crucial for understanding the nature of the data and selecting appropriate variables for analysis and modeling.  
  
3.2 Clean and Preprocess the Data

To clean and preprocess the hotel demand dataset, several steps were taken. First, missing data for each column was summarized to identify any columns with a significant number of missing values. The 'agent' and 'company' columns were removed due to NULL values.

Next, the 'adr' (Average Daily Rate) variable was examined. A histogram and boxplot revealed outliers, including a maximum value of 5400(Figure 1). Observations with missing or zero 'adr' were filtered out, and the single observation with 'adr' equal to 5400 was removed as an outlier (Figure 2).

The 'arrival\_date\_week\_number', 'arrival\_date\_day\_of\_month', ‘country’ and 'arrival\_date\_month' columns were removed as they were deemed not useful for the analysis. The 'meal' variable was also investigated. The 'Undefined' category was replaced with 'SC' (no meal package) based on the data dictionary(Figure 3). Bar plots were created to visualize the distribution of meal types before and after this change(Figure 4).

Similar steps were taken for the 'market\_segment' and 'distribution\_channel' variables. Observations with 'Undefined' categories were filtered out, and bar plots were used to compare the distributions before (Figure 5,7) and after the removal(Figure 6,8).

The 'hotel' variable was converted into a dummy variable, with 'Resort Hotel' coded as 1 and 'City Hotel' as 0. All character columns were then converted to factors, and integer columns with low cardinality were also converted to factors.

Correlation analysis was performed on the numeric variables before adjecting the outliers (Figure 9). A correlation matrix heatmap was created to visualize the correlations between variables.

To handle outliers, a function was created to find the bounds for outliers using the interquartile range (IQR) method. This function was applied to several variables, including 'stays\_in\_week\_nights', 'adults', 'children', 'babies', 'previous\_cancellations', 'previous\_bookings\_not\_canceled', and 'booking\_changes'. Observations falling outside the acceptable bounds were removed.

Correlation analysis was performed on the numeric variables after adjecting the outliers (Figure 10). A correlation matrix heatmap was created to visualize the correlations between variables. The 'lead\_time' variable was further analyzed using a violin plot and an average lead time bar plot to understand its relationship with the cancellation status.

Interesting subsets of the data, such as canceled bookings (Figure 11) and high-value customers (Figure 12), were selected for further investigation. Histograms were used to visualize the distribution of lead time for canceled bookings and the distribution of average daily rate for high-value customers.

Finally, principal component analysis (PCA) was applied to the numeric variables. The dataset was transformed to ensure all variables were numeric, and any constant columns were removed. The PCA results showed that the first principal component explained 15.01% of the variance, while the first 13 components together explained 87.19% of the variance.

These preprocessing steps aimed to clean the data, handle missing values, remove outliers, and explore relationships between variables. The resulting dataset was then ready for further analysis and modeling, but we didn’t do anything using PCA.

3.3 Reduce the Data Dimension

In the process of preparing the hotel demand dataset for analysis, several techniques were employed to reduce the number of variables and simplify the dataset. This step is crucial as it helps to focus on the most relevant variables, reduces computational complexity, and improves the interpretability of the results.

One approach used for dimension reduction was the removal of variables based on their relevance to the analysis. For example, the 'arrival\_date\_week\_number', 'arrival\_date\_day\_of\_month', and 'arrival\_date\_month' columns were removed as they were considered not useful for the purpose of the study. Similarly, the 'agent' and 'company' columns were removed due to a significant amount of missing data.

Another technique applied was the conversion of categorical variables into numeric forms. The 'hotel' variable was converted into a dummy variable, where 'Resort Hotel' was coded as 1 and 'City Hotel' as 0. This transformation allows for easier integration of the variable into mathematical models. Other categorical variables, such as 'meal', 'market\_segment', 'distribution\_channel', 'reserved\_room\_type', 'assigned\_room\_type', 'deposit\_type', 'customer\_type', 'reservation\_status', and 'reservation\_status\_date', were converted to numeric factors, which assigns a unique numeric value to each category.

Additionally, principal component analysis (PCA) was applied as a dimension reduction technique. PCA is a method that transforms the original variables into a new set of variables called principal components. These components are linear combinations of the original variables and are designed to capture the maximum amount of variance in the data. By selecting a subset of the principal components that explain a significant portion of the variance, the dimensionality of the dataset can be reduced while retaining most of the important information.

Before applying PCA, the dataset was preprocessed to ensure that all variables were numeric, and any constant columns (columns with zero variance) were removed. The PCA results indicated that the first principal component explained 15.01% of the variance, and the first 13 components together accounted for 87.19% of the variance in the data.

The decision to select a subset of principal components depends on the trade-off between dimensionality reduction and the amount of information retained. In this case, selecting the first 13 components would reduce the dimensionality of the dataset while still capturing a significant amount of the underlying variability.

3.4 Determine the Data Mining Task and Its Importance in Marketing Analytics

The main data mining task in the hotel demand project is to predict hotel booking cancellations. This is evident from the exploratory data analysis and modeling steps performed. The target variable 'is\_canceled' is a binary variable indicating whether a booking was canceled (1) or not (0). The goal is to develop predictive models that can accurately classify bookings as either canceled or not canceled based on various booking attributes and customer characteristics.

Predicting hotel booking cancellations is crucial for effective marketing analytics in the hotel industry for several reasons:

1. Revenue Optimization: Accurate cancellation predictions enable hotels to optimize their revenue management strategies. By anticipating cancellations, hotels can adjust their overbooking policies, pricing, and inventory allocation to minimize the impact of cancellations on revenue. This helps maximize occupancy rates and revenue per available room (RevPAR).

2. Resource Planning: Cancellation predictions assist hotels in efficiently planning their resources, such as staff, rooms, and amenities. By having a reliable estimate of the expected number of guests, hotels can allocate resources appropriately, reducing waste and improving operational efficiency.

3. Targeted Marketing Interventions: The predictive models can identify key factors and customer segments associated with higher cancellation probabilities. This information enables hotels to develop targeted marketing campaigns and interventions to reduce cancellations. For example, hotels can offer incentives or personalized communication to customers who are more likely to cancel, encouraging them to maintain their bookings.

4. Enhanced Customer Experience: By proactively identifying potential cancellations, hotels can reach out to customers to address any concerns or issues they may have. This proactive approach demonstrates care and attention towards customers, enhancing their overall experience and reducing the likelihood of cancellations.

Several data mining techniques and models were employed to predict hotel booking cancellations. These include:

1. Logistic Regression: Logistic regression models were trained to predict the probability of a booking being canceled. The models were evaluated using metrics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC).

2. Decision Trees: Decision tree models were constructed to identify the most important predictors of cancellations and to provide interpretable rules for classification. The models were visualized using tree diagrams to understand the decision-making process.

3. Random Forest: Random Forest models were trained to predict cancellations, leveraging an ensemble of decision trees. The variable importance measures from the random forest models helped identify the most influential predictors of cancellations.

4. XGBoost: XGBoost, an optimized gradient boosting algorithm, was applied to predict cancellations. The XGBoost models demonstrated high predictive performance and provided insights into the key factors driving cancellations.

3.5 Partition the Data

In hotel demand dataset was partitioned into training and testing sets. This partitioning was performed to facilitate the development and evaluation of predictive models. The purpose of splitting the data into training and testing sets is to assess how well the models generalize to unseen data and to prevent overfitting.

The data partitioning process was carried out using a 70-30 split, where 70% of the observations were allocated to the training set, and the remaining 30% were assigned to the testing set. This split ratio is commonly used in data mining projects as it provides a balance between having sufficient data for model training and reserving an adequate portion for model evaluation.

The partitioning process involved several steps. First, dummy variables were created for the categorical predictors using the dummyVars function from the caret package. This step converted the categorical variables into binary indicator variables, making them suitable for use in predictive models.

Next, the indices of the dataset were shuffled randomly using the sample function. This step ensured that the observations were randomly distributed between the training and testing sets, avoiding any bias or systematic patterns that could affect the model's performance.

Finally, using the shuffled indices, the dataset was split into training and testing sets. The train\_index and test\_index variables were created to store the indices of the observations belonging to the training and testing sets, respectively. The actual observations corresponding to these indices were then assigned to the train\_data and test\_data dataframes.

1. **Methodology**

Several data mining techniques were employed to predict hotel booking cancellations. The choice of these techniques was based on their suitability for the binary classification task at hand, as well as their ability to provide interpretable results and handle the dataset's characteristics. The main techniques used in the project were logistic regression, decision trees, random forests, XGBoost, and neural networks.

The dataset used in the project contained a total of 32 variables, including the dependent variable 'is\_canceled' and 31 independent variables. However, during the data preprocessing and feature selection stages, some variables were removed due to missing values or low relevance. The final set of independent variables based on step-wise model used in the predictive models included 19 variables:

1. hotel

2. lead\_time

3. stays\_in\_weekend\_nights

4. stays\_in\_week\_nights

5. adults

6. meal

7. market\_segment

8. distribution\_channel

9. is\_repeated\_guest

10. previous\_cancellations

11. previous\_bookings\_not\_canceled

12. reserved\_room\_type

13. assigned\_room\_type

14. deposit\_type

15. days\_in\_waiting\_list

16. customer\_type

17. adr

18. required\_car\_parking\_spaces

19. total\_of\_special\_requests

These variables were selected based on their potential influence on booking cancellations and their relevance to the problem at hand. They encompass various aspects of the booking process, customer characteristics, and hotel attributes.

Each model utilized a specific set of independent variables to predict the dependent variable 'is\_canceled'. Below, we will discuss the variables used in each model and provide an explanation of the methodology.

Logistic Regression (Model 1):

Independent Variables:

1. hotel

2. lead\_time

3. arrival\_date\_year

4. stays\_in\_weekend\_nights

5. adults

6. market\_segment

7. distribution\_channel

8. is\_repeated\_guest

9. reserved\_room\_type

10. assigned\_room\_type

11. deposit\_type

12. days\_in\_waiting\_list

13. customer\_type

14. adr

15. required\_car\_parking\_spaces

16. total\_of\_special\_requests

17. reservation\_status

18. reservation\_status\_date

Logistic regression is a statistical method used for binary classification problems. It models the probability of an event occurring (in this case, a booking cancellation) based on a set of independent variables. The coefficients in the logistic regression model represent the change in the log odds of the dependent variable for a one-unit change in the independent variable, holding other variables constant. The model's performance is evaluated using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC.

Logistic Regression (Model 2):

Independent Variables:

1. hotel

2. lead\_time

3. stays\_in\_weekend\_nights

4. meal

5. reserved\_room\_type

6. assigned\_room\_type

7. deposit\_type

8. days\_in\_waiting\_list

9. customer\_type

10. adr

11. required\_car\_parking\_spaces

12. total\_of\_special\_requests

This logistic regression model uses a different set of independent variables compared to Model 1. The model's performance is evaluated using the same metrics as Model 1.

Decision Trees and Random Forests:

Independent Variables (Random Forest):

1. hotel

2. lead\_time

3. stays\_in\_weekend\_nights

4. meal

5. reserved\_room\_type

6. assigned\_room\_type

7. deposit\_type

8. days\_in\_waiting\_list

9. customer\_type

10. adr

11. required\_car\_parking\_spaces

12. total\_of\_special\_requests

Decision trees are a non-parametric method that recursively partitions the data based on the independent variables to create a tree-like model. They are useful for both classification and regression tasks. In the hotel demand project, decision trees were used to predict booking cancellations. The dependent variable is the 'is\_canceled' variable, and the independent variables are the same as those used in the logistic regression model.

Random forests, an ensemble method that combines multiple decision trees, were also employed to improve the predictive performance and reduce overfitting. Random forests create many decision trees on random subsets of the data and variables and then aggregate their predictions.

XGBoost:

Independent Variables:

1. hotel

2. lead\_time

3. stays\_in\_weekend\_nights

4. meal

5. reserved\_room\_type

6. assigned\_room\_type

7. deposit\_type

8. days\_in\_waiting\_list

9. customer\_type

10. adr

11. required\_car\_parking\_spaces

12. total\_of\_special\_requests

XGBoost (Extreme Gradient Boosting) is an optimized implementation of the gradient boosting algorithm. It combines multiple weak learners (decision trees) to create a strong predictive model. The model iteratively trains decision trees, with each tree attempting to correct the errors made by the previous trees. XGBoost is known for its high predictive accuracy, scalability, and ability to handle complex datasets.

Neural Networks:

Independent Variables:

1. hotel

2. lead\_time

3. stays\_in\_weekend\_nights

4. meal

5. reserved\_room\_type

6. assigned\_room\_type

7. deposit\_type

8. days\_in\_waiting\_list

9. customer\_type

10. adr

11. required\_car\_parking\_spaces

12. total\_of\_special\_requests

The neural network model uses the same set of independent variables as the random forest and XGBoost models. Neural networks are a class of machine learning models inspired by the structure and function of the human brain. They consist of interconnected nodes (neurons) organized in layers. The model learns complex, non-linear relationships between the independent variables and the dependent variable through an iterative training process.

The choice of these data mining techniques was based on their strengths and their ability to provide valuable insights into the factors influencing hotel booking cancellations. Logistic regression offers interpretability and estimates the impact of each independent variable on the cancellation probability. Decision trees and random forests provide a clear understanding of the decision-making process and can handle both numerical and categorical variables. XGBoost is known for its high predictive accuracy and efficiency, making it suitable for large datasets. Neural networks have the capacity to capture complex, non-linear relationships and can automatically learn representations from the data.

By employing multiple techniques and utilizing a comprehensive set of independent variables, the hotel demand project aims to obtain a robust understanding of the factors driving cancellations and to develop accurate predictive models. The combination of interpretable models like logistic regression and decision trees with powerful ensemble methods like random forests and XGBoost, along with the flexibility of neural networks, allows for a thorough analysis of the problem and the derivation of actionable insights.

Furthermore, the project utilizes appropriate evaluation metrics such as accuracy, precision, recall, F1-score, and AUC-ROC to assess the performance of each model. These metrics provide a quantitative measure of how well the models can predict cancellations and help in comparing their effectiveness.

In summary, the application of various data mining techniques, including logistic regression, decision trees, random forests, XGBoost, and neural networks, to predict hotel booking cancellations. The project utilizes a comprehensive set of 19 independent variables and 12 independent variables, selected based on their relevance and potential impact on cancellations. Each model employs a specific subset of these variables to predict the dependent variable 'is\_canceled'. The choice of these techniques is based on their suitability for the binary classification task, interpretability, and ability to handle the dataset's characteristics. By employing multiple techniques, utilizing different sets of variables for each model, and evaluating their performance using appropriate metrics, the project aims to develop robust predictive models and gain insights into the factors influencing cancellations, ultimately supporting data-driven decision-making in hotel revenue management and customer satisfaction.

1. **Empirical Results**

In the empirical investigation of the hotel demand dataset, several data mining techniques were applied, including logistic regression, decision trees, random forests, XGBoost, and neural networks. The process involved data preprocessing, exploratory data analysis, model training, and evaluation. This section will discuss the challenges encountered and the iterations performed during the analysis.

Data Preprocessing Challenges:

One of the initial challenges was handling missing data and outliers in the dataset. The 'agent' and 'company' columns were removed due to a significant number of NULL values. The 'adr' (Average Daily Rate) variable contained outliers, with a maximum value of 5400. Observations with missing or zero 'adr' were filtered out, and the single observation with 'adr' equal to 5400 was removed as an outlier.

Categorical variables such as 'meal', 'market\_segment', and 'distribution\_channel' contained 'Undefined' categories, which were replaced or filtered out based on the data dictionary. The 'hotel' variable was converted into a dummy variable, and all character columns were converted to factors.

Outliers were addressed using the interquartile range (IQR) method for several variables, including 'stays\_in\_week\_nights', 'adults', 'children', 'babies', 'previous\_cancellations', 'previous\_bookings\_not\_canceled', and 'booking\_changes'. Observations falling outside the acceptable bounds were removed.

Iterations and Model Development:

Multiple iterations of logistic regression models were run with different sets of independent variables. The first model included a comprehensive set of variables, while the second model focused on a smaller subset. The step-wise method was also employed to identify the most significant variables contributing to the prediction of hotel booking cancellations.

The logistic regression models were trained using cross-validation with 10 folds. The models were evaluated using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. The first logistic regression model achieved an AUC value of 0.975, indicating strong predictive performance.

Decision tree and random forest models were also developed to predict hotel booking cancellations. The decision tree model was trained using the rpart package, and the resulting tree was visualized to understand the decision-making process. The random forest model was trained using the randomForest package, and variable importance measures were obtained to identify the most influential predictors.

The XGBoost model was implemented using the xgboost package. The model was trained with cross-validation and hyperparameter tuning to optimize its performance. The XGBoost model demonstrated high predictive accuracy and provided insights into the key factors driving cancellations.

Neural network models were also explored for predicting hotel booking cancellations. The models were trained using the nnet package, and hyperparameter tuning was performed to find the optimal architecture and settings. The neural network models required careful consideration of the number of hidden layers, nodes, and activation functions to achieve satisfactory performance.

Model Evaluation and Comparison:

The performance of each model was evaluated using appropriate metrics and cross-validation techniques. The logistic regression models achieved an accuracy of around 0.76, indicating strong predictive power. The decision tree model had an accuracy of 0.77, with a sensitivity of 0.80 and specificity of 0.75. The random forest model outperformed the decision tree, with an accuracy of 0.84.

The XGBoost model demonstrated comparable performance to the random forest, with an accuracy of 0.81. The neural network model, after hyperparameter tuning, achieved an accuracy of 0.77, with a sensitivity of 0.68 and specificity of 0.83.

Comparing the models, the random forest model stands out with the highest accuracy, followed closely by XGBoost. The decision tree and neural network models also provide valuable insights, each with its own strengths in sensitivity and specificity.

In summary, the empirical investigation encompassed multiple iterations and encountered challenges in data preprocessing and model development. Among the models assessed, random forest exhibited the highest accuracy, closely followed by XGBoost. The decision tree and neural network models, although slightly trailing in accuracy, provided valuable insights with their strengths in sensitivity and specificity. Through iterative processes and thorough model comparisons, the study offered a comprehensive evaluation of diverse data mining techniques, illuminating their suitability for predicting hotel booking cancellations.

1. **Conclusions and Recommendations**

The empirical investigation of the hotel demand dataset using various data mining techniques has yielded valuable insights into the factors influencing hotel booking cancellations and customer segmentation based on lifetime value. The results provide a foundation for making informed decisions in hotel revenue management and marketing strategies.

Several models, including logistic regression, random forest, XGBoost, and neural networks, were trained and evaluated using different algorithms. Table 2 presents a comparison of the performance metrics for each model, along with their significance levels.

The random forest model achieved the highest accuracy (0.8283) and kappa value (0.6304), indicating its strong predictive power in identifying bookings likely to be canceled. The XGBoost model also demonstrated good performance, with an accuracy of 0.8022 and a kappa value of 0.5638. The logistic regression and neural network models, while still effective, had slightly lower performance metrics compared to the random forest and XGBoost models.

From a marketing analytics perspective, these results offer several recommendations for hotel revenue management and customer retention strategies. Firstly, hotels should focus on the key variables identified by the models as significant predictors of cancellations, such as lead time, deposit type, average daily rate (ADR), and the total number of special requests. By monitoring and analyzing these variables, hotels can proactively identify bookings with a higher risk of cancellation and take appropriate measures to mitigate that risk.

For example, hotels can offer targeted incentives or personalized communication to customers with longer lead times or those who have made special requests to encourage them to maintain their bookings. Additionally, implementing stricter deposit policies or offering non-refundable rates for certain booking segments can help reduce cancellations and protect revenue.

Secondly, the significance of 'total\_of\_special\_requests' suggests that hotels should pay attention to customer preferences and strive to accommodate special requests whenever possible. By providing a tailored and satisfactory experience, hotels can increase customer satisfaction and loyalty, leading to reduced cancellations and increased repeat business.

The RFM (Recency, Frequency, Monetary) analysis conducted in the study provides valuable insights into customer segmentation based on their lifetime value. Table 3 presents a summary of the RFM scores and corresponding customer segments.

The RFM analysis enables hotels to identify high-value customers and develop targeted marketing strategies to retain and nurture these profitable segments. By focusing on customers with high RFM scores, hotels can optimize their marketing efforts and resource allocation to maximize customer lifetime value and overall profitability.

Moreover, the study emphasizes the significance of data-driven decision-making in the hotel industry. By utilizing advanced analytics techniques and machine learning models, hotels can gain a deeper understanding of customer behavior, preferences, and cancellation patterns. These insights can inform various aspects of hotel operations, such as pricing strategies, inventory management, and targeted marketing campaigns, allowing hotels to optimize their revenue and minimize the impact of cancellations.

However, it is crucial to acknowledge the limitations of the study and consider them when interpreting the results. The specific dataset used in this analysis may not be representative of all hotel contexts, and the models' performance may vary when applied to different datasets or hotel properties. Therefore, hotels should validate and fine-tune the models using their own data to ensure the best fit for their specific circumstances.

Additionally, while the models provide valuable insights, they should be used in conjunction with human expertise and domain knowledge. Hotel managers and marketers should combine the data-driven insights with their understanding of the industry, market trends, and customer preferences to make well-informed decisions.

In conclusion, this study provides a comprehensive analysis of hotel booking cancellations using various data mining techniques. The empirical results and recommendations offer actionable insights for hotel revenue managers and marketers to optimize their strategies and drive business growth. By leveraging the power of predictive modeling and customer segmentation, hotels can make data-driven decisions to reduce cancellations, enhance customer satisfaction, and maximize profitability.

Future research could explore the integration of additional variables, such as competitor pricing, weather conditions, or economic factors, to further improve the predictive power of the models. Moreover, conducting similar analyses across different hotel segments, geographical locations, or time periods could provide valuable comparative insights and validate the generalizability of the findings.

Overall, this study contributes to the growing body of knowledge in hotel revenue management and marketing analytics. The insights gained from this research can serve as a catalyst for further innovation and data-driven decision-making in the hospitality industry, ultimately leading to improved business performance and enhanced customer experiences.

**References:**

Ahani, A., Nilashi, M., Ibrahim, O., Sanzogni, L., & Weaven, S. (2019). Market segmentation and travel choice prediction in Spa hotels through TripAdvisor's online reviews. International Journal of Hospitality Management, 80, 52-77. https://doi.org/10.1016/j.ijhm.2019.01.003

El Gayar, N. F., Saleh, M., Atiya, A., El-Shishiny, H., Zakhary, A. A. Y. F., & Habib, H. A. A. M. (2011). An integrated framework for advanced hotel revenue management. International Journal of Contemporary Hospitality Management, 23(1), 84-98. https://doi.org/10.1108/09596111111101689

Haensel, A., & Koole, G. (2011). Booking horizon forecasting with dynamic updating: A case study of hotel reservation data. International Journal of Forecasting, 27(3), 942-960. https://doi.org/10.1016/j.ijforecast.2010.10.004

Khajvand, M., Zolfaghar, K., Ashoori, S., & Alizadeh, S. (2011). Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. Procedia Computer Science, 3, 57-63. https://doi.org/10.1016/j.procs.2010.12.011

Rajopadhye, M., Ghalia, M. B., Wang, P. P., Baker, T., & Eister, C. V. (2001). Forecasting uncertain hotel room demand. Information Sciences, 132(1-4), 1-11. https://doi.org/10.1016/S0020-0255(00)00082-7

Reinartz, W. J., & Kumar, V. (2003). The impact of customer relationship characteristics on profitable lifetime duration. Journal of Marketing, 67(1), 77-99. https://doi.org/10.1509/jmkg.67.1.77.18589

Salas-Olmedo, M. H., García-Palomares, J. C., Moya-Gómez, B., & Gutiérrez, J. (2018). Tourists' digital footprint in cities: Comparing Big Data sources. Tourism Management, 66, 13-25. https://doi.org/10.1016/j.tourman.2017.11.001

**Appendix**

**Appendix A: Tables**

**A.1 Table 1: Data Dictionary**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Type** | **Description** | **Source/Engineering** |
| ***ADR*** | Numeric | Average Daily Rate | BO, BL and TR / Calculated by dividing the sum of all lodging transactions by the total number of staying nights |
| ***Adults*** | Integer | Number of adults | BO and BL |
| ***Agent*** | Categorical | ID of the travel agency that made the booking | BO and BL |
| ***ArrivalDateDayOfMonth*** | Integer | Day of the month of the arrival date | BO and BL |
| ***ArrivalDateMonth*** | Categorical | Month of arrival date with 12 categories: “January” to “December” | BO and BL |
| ***ArrivalDateWeekNumber*** | Integer | Week number of the arrival date | BO and BL |
| ***ArrivalDateYear*** | Integer | Year of arrival date | BO and BL |
| ***AssignedRoomType*** | Categorical | Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons | BO and BL |
| ***Babies*** | Integer | Number of babies | BO and BL |
| ***BookingChanges*** | Integer | Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation | BO and BL/Calculated by adding the number of unique iterations that change some of the booking attributes, namely: persons, arrival date, nights, reserved room type or meal |
| ***Children*** | Integer | Number of children | BO and BL/Sum of both payable and non-payable children |
| ***Company*** | Categorical | ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons | BO and BL. |
| ***Country*** | Categorical | Country of origin. Categories are represented in the ISO 3155–3:2013 format | BO, BL, and NT |
|  |  |  |  |
| ***CustomerType*** | Categorical | Type of booking, assuming one of four categories: | BO and BL |
| Contract - when the booking has an allotment or other type of contract associated to it; |
| Group – when the booking is associated to a group; |
| Transient – when the booking is not part of a group or contract, and is not associated to other transient booking; |
| Transient-party – when the booking is transient, but is associated to at least other transient booking |
| ***DaysInWaitingList*** | Integer | Number of days the booking was in the waiting list before it was confirmed to the customer | BO/Calculated by subtracting the date the booking was confirmed to the customer from the date the booking entered on the PMS |
|  |  |  |  |
| ***DepositType*** | Categorical | Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: | BO and TR/Value calculated based on the payments identified for the booking in the transaction (TR) table before the booking׳s arrival or cancellation date. |
| No Deposit – no deposit was made; |
| In case no payments were found the value is “No Deposit”. |
| If the payment was equal or exceeded the total cost of stay, the value is set as “Non-Refund”. |
| Non Refund – a deposit was made in the value of the total stay cost; |
| Otherwise, the value is set as “Refundable” |
| Refundable – a deposit was made with a value under the total cost of stay. |
| ***DistributionChannel*** | Categorical | Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators” | BO, BL, and DC |
| ***IsCanceled*** | Categorical | Value indicating if the booking was canceled (1) or not (0) | BO |
| ***IsRepeatedGuest*** | Categorical | Value indicating if the booking name was from a repeated guest (1) or not (0) | BO, BL, and C/ Variable created by verifying if a profile was associated with the booking customer. If so, and if the customer profile creation date was prior to the creation date for the booking on the PMS database it was assumed the booking was from a repeated guest |
| ***LeadTime*** | Integer | Number of days that elapsed between the entering date of the booking into the PMS and the arrival date | BO and BL/ Subtraction of the entering date from the arrival date |
| ***MarketSegment*** | Categorical | Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators” | BO, BL, and MS |
|  |  |  |  |
| ***Meal*** | Categorical | Type of meal booked. Categories are presented in standard hospitality meal packages: | BO, BL and ML |
| Undefined/SC – no meal package; |
| BB – Bed & Breakfast; |
| HB – Half board (breakfast and one other meal – usually dinner); |
| FB – Full board (breakfast, lunch and dinner) |
| ***PreviousBookingsNotCanceled*** | Integer | Number of previous bookings not cancelled by the customer prior to the current booking | BO and BL / In case there was no customer profile associated with the booking, the value is set to 0. Otherwise, the value is the number of bookings with the same customer profile created before the current booking and not canceled. |
| ***PreviousCancellations*** | Integer | Number of previous bookings that were cancelled by the customer prior to the current booking | BO and BL/ In case there was no customer profile associated with the booking, the value is set to 0. Otherwise, the value is the number of bookings with the same customer profile created before the current booking and canceled. |
| ***RequiredCardParkingSpaces*** | Integer | Number of car parking spaces required by the customer | BO and BL |
|  |  |  |  |
| ***ReservationStatus*** | Categorical | Reservation last status, assuming one of three categories: | BO |
| Canceled – booking was canceled by the customer; |
| Check-Out – customer has checked in but already departed; |
| No-Show – customer did not check-in and did inform the hotel of the reason why |
| ***ReservationStatusDate*** | Date | Date at which the last status was set. This variable can be used in conjunction with the *ReservationStatus* to understand when the booking was canceled or when did the customer checked-out of the hotel | BO |
| ***ReservedRoomType*** | Categorical | Code of room type reserved. Code is presented instead of designation for anonymity reasons | BO and BL |
| ***StaysInWeekendNights*** | Integer | Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel | BO and BL/ Calculated by counting the number of weekend nights from the total number of nights |
| ***StaysInWeekNights*** | Integer | Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel | BO and BL/Calculated by counting the number of weeknights from the total number of nights |
| ***TotalOfSpecialRequests*** | Integer | Number of special requests made by the customer (e.g. twin bed or high floor) | BO and BL/Sum of all special requests |

Table 1 1Table 1

**A.2 Table 2: Model Performance Comparison**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Kappa** | **Sensitivity** | **Specificity** | **Significance level** |
| Logistic Regression | 0.7462 | 0.4061 | 0.9498 | 0.4157 | p < 0.001 |
| Random Forest | 0.8283 | 0.6304 | - | - | p < 0.001 |
| XGBoost | 0.8022 | 0.5638 | - | - | p < 0.001 |
| Neural Network | 0.7678 | 0.5166 | 0.7853 | 0.7396 | p < 0.001 |

**A.3 Table 3: RFM Analysis Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Customer Segment** | **Recency Score** | **Frequency Score** | **Monetary Score** | **RFM Score** |
| Churned | -0.783 | -0.618 | -0.601 | -2.00 |
| Potential | 1.31 | -0.686 | -0.660 | -0.0401 |
| High Value | -0.783 | 1.46 | 1.47 | 2.14 |
| Loyal | 0.261 | -0.153 | -0.209 | -0.101 |

**Appendix B: Figures**

**B.1 Figure 1: Histogram of Average Daily Rate(ADR)**A graph of a graph

Description automatically generated

**B.2 Figure 2: Histogram of Lead Time**A graph of a diagram

Description automatically generated

**B.3 Figure 3: Distribution of Meal Types**

A graph of a number of different types of food

Description automatically generated

**B.4 Figure 4: Distribution of Meal Types**

A graph of a number of different types of food

Description automatically generated

**B.5 Figure 5: Distribution of Market Segment**

A graph of a number of blue bars

Description automatically generated

**B.6 Figure 6: Distribution of Market Segment**

A graph of a bar graph

Description automatically generated

**B.7 Figure 7: Distribution of Distribution Channel**

A graph of a distribution channel

Description automatically generated

**B.8 Figure 8: Distribution of Distribution Channel**

A graph with blue squares

Description automatically generated with medium confidence

**B.9 Figure 9: Correlation Matrix Heatmap**

A graph with text and dots

Description automatically generated with medium confidence

**B.10 Figure 10: Correlation Matrix Heatmap**

A graph with dots and text

Description automatically generated with medium confidence

**B.11 Figure 11: Distribution of Lead Time for Canceled Bookings**

A graph of a number of blue bars

Description automatically generated

**B.12 Figure 12: Distribution of Average Daily Rate(ADR) for High-Value Customers**

A graph of a number of red bars

Description automatically generated