UNIVERSITY OF SURREY

THE FACULTY OF ARTS AND SOCIAL SCIENCES

Surrey Business School

Data Driven Approach to Predict Hotel Reservation Cancellations – Business Report

16th August 2023

Prepared By:

Group 9

1. Bhanwar Sachdeva – 6780526
2. Sagar Saxena - 6781347
3. Thensih Kaithavalappil Abu - 6782614
4. Nattabhat Sirisoonthorn – 6780137

Total Pages: 30

Word Count: 3918

Table of Contents

[1. Executive Summary 4](#_Toc142569799)

[2. Background 4](#_Toc142569800)

[3. Problem Statement & Objectives 4](#_Toc142569801)

[4. Data Sources and Pre-preprocessing 5](#_Toc142569802)

[5. Data Exploration and Visualization 7](#_Toc142569803)

[6. Modelling and Analysis 13](#_Toc142569804)

[6.1. K-Nearest Neighbor (KNN) 13](#_Toc142569805)

[6.2. Support Vector Machine (SVM) 14](#_Toc142569806)

[6.3. Random Forest 16](#_Toc142569807)

[6.4. Logistic Regression 18](#_Toc142569808)

[7. Model Evaluation 19](#_Toc142569809)

[8. Return on Investment (ROI) 20](#_Toc142569810)

[9. Conclusions 21](#_Toc142569811)

[10. Recommendations 21](#_Toc142569812)

[11. References 23](#_Toc142569813)

[Appendix A – Generative AI Flowchart 25](#_Toc142569814)

[Appendix B - Algorithms 26](#_Toc142569815)

[Appendix C – Logistic Regression Models 27](#_Toc142569816)

[Appendix D - BACCM 28](#_Toc142569817)

[Appendix E – Project Plan 29](#_Toc142569818)

[Appendix F – R Scripts 30](#_Toc142569819)

Tables

[Table 1: Data Description (19 Fields) 5](#_Toc142312669)

[Table 2: Performance Metrics (KNN) 14](#_Toc142312670)

[Table 3: Performance Metrics post Kernel Trick (SVM) 15](#_Toc142312671)

[Table 4: Performance Metrics post Cost Update (SVM) 16](#_Toc142312672)

[Table 5: Performance Metrics with different terminal nodes (RF) 17](#_Toc142312673)

[Table 6: Performance Metrics (LR) 19](#_Toc142312674)

[Table 7: Model Evaluation Tabe along with Key Metrices 19](#_Toc142312675)

[Table 8: Confusion Metrics of Random Forest Algorithm 20](#_Toc142312676)

Figures

[Figure 1: Distribution of Target Field 6](#_Toc142312677)

[Figure 2: Relationship of Average Room Price with Other Variables 8](#_Toc142312678)

[Figure 3: Distribution of Room Type Reserved with Booking Status 8](#_Toc142312679)

[Figure 4: Distribution of Market Segment by Booking Status 9](#_Toc142312680)

[Figure 5: Variation of Reservation Count over Time 10](#_Toc142312681)

[Figure 6: Variation of Average Room Price over Time 10](#_Toc142312682)

[Figure 7: Impact of Lead Time over Booking Status 11](#_Toc142312683)

[Figure 8: Distribution of Repeated Guests 11](#_Toc142312684)

[Figure 9: Correlation Coefficients 12](#_Toc142312685)

[Figure 10: Elbow Method to determine K Value 14](#_Toc142312686)

# Executive Summary

In the era marked by dynamic booking behaviors and technological advancements, the hospitality industry grapples with the persistent challenge of hotel reservations cancellations and their associated revenue losses. To confront this issue head-on, our study introduces an innovative Hotel Cancellation Prediction Model, leveraging cutting-edge machine learning algorithms and comprehensive dataset. The dataset was cleaned to address any duplicate and missing values, reformat certain features and engineer new ones. Exploratory analysis of the dataset including analysis of the target variable (Booking Status) and its relationship with other variables by utilizing data visualization techniques to identify trends and gather insights from the analysis. A set of supervised algorithms like K-Nearest Neighbor, Logistic Regression, Random Forest and Support Vector Machine (SVM) were utilized to generate models and their profiles were analyzed. The model produced using Random Forest algorithm was selected as the predictive model because of its high Accuracy (89.5%) and Kappa (75.8%) in predicting the cancellations successfully. Return on Investment of approx. 65% was achieved using the results from Random Forest Algorithm. Subsequently, conclusions were drawn and recommendations were formulated to establish a harmonious synergy between customer expectations and management objectives.

# Background

A substantial number of hotel reservations are lost due to cancellations or no-shows. The most common reasons include unexpected sickness, accidents, schedule difficulties, unexpected obligations, and natural disasters. Although the ability to cancel a hotel reservation (ideally at a low cost) is advantageous to prospective hotel guests, it is less desired and revenue-diminishing element for hotel managers to cope with. Such losses are particularly high on last-minute cancellations. However, new technologies involving online booking channels have dramatically changed customers’ booking possibilities and behavior. This increases the problem of how hotels manage cancellations, which no longer restricted to traditional booking and passenger characteristics, resulting in missed opportunities for revenue optimization and operational efficiencies. (Falk & Vieru, 2018).

# Problem Statement & Objectives

Lack of accurate cancellation prediction hinders effective decision-making regarding pricing, inventory management, staffing and personalized guests’ services. To address these challenges, our aim is to leverage machine learning algorithms and historical Hotel Reservation Dataset to build a reliable predictive model for hotel reservation cancellations. The model will make use of relevant data and trends to assess the likelihood of cancellation for each reservation, providing real-time insights and recommendations to revenue managers, front-desk workers, and decision-makers.

*The study's objective is to minimize monetary losses, identification of high-risk reservations, enhance customer retention initiatives and flexible booking opportunities to provide superior customer experience which will be a win-win situation for both customers and hotel management.*

# Data Sources and Pre-preprocessing

The quality of the data is fundamental to the success of any data-driven analysis. In this section, the data source utilized in the study is presented, along with a detailed account of pre-processing procedures employed to ensure the accuracy and consistency of the data.

The secondary dataset used in this study is a hotel reservations data obtained from Kaggle (Raza, 2023). With 36,275 observations and 19 features (including the target variable), the dataset appears to be well suited for effectively training and testing a machine learning algorithm to predict the hotel cancellations. The description of these fields has been provided in the table below.

Table 1: Data Description (19 Fields)

|  |  |  |
| --- | --- | --- |
| S. No. | Features | Description |
| 1 | Booking\_ID | Unique identifier of each booking |
| 2 | no\_of\_adults | Number of adults |
| 3 | no\_of\_children | Number of children |
| 4 | no\_of\_weekend\_nights | Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel |
| 5 | no\_of\_week\_nights | Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel |
| 6 | type\_of\_meal\_plan | Type of meal plan booked by the customer |
| 7 | required\_car\_parking\_space | Does the customer require a car parking space? (0 - No, 1- Yes) |
| 8 | room\_type\_reserved | Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels. |
| 9 | lead\_time | Number of days between the date of booking and the arrival date |
| 10 | arrival\_year | Year of arrival date |
| 11 | arrival\_month | Month of arrival date |
| 12 | arrival\_date | Date of arrival |
| 13 | market\_segment\_type | Market segment designation |
| 14 | repeated\_guest | Is the customer a repeated guest? (0 - No, 1- Yes) |
| 15 | no\_of\_previous\_cancellations | Number of previous bookings that were cancelled by the guest before the current booking |
| 16 | no\_of\_previous\_bookings\_not\_canceled | Number of previous bookings not cancelled by the guest before the current booking |
| 17 | avg\_price\_per\_room | Average price per day of the reservation; prices of the rooms are dynamic. (In euros) |
| 18 | no\_of\_special\_requests | Number of special requests from the guest (e.g., high floor, view from the room, etc.) |
| 19 | booking\_status | Binary Target Field - Flag indicating if the booking was cancelled or not |

A pie chart with numbers and a few percentages

Description automatically generated

Figure 1: Distribution of Target Field

The following pre-processing steps were undertaken during the study to ensure that the dataset is ready for wide range of data-driven tasks, from exploratory analysis to machine learning algorithms.

* Missing and Duplicate Values: No duplicates or missing values were detected in the dataset.
* Feature Engineering: An updated Arrival Date column was created using the existing day, month and year columns. The engineered column was used later in the algorithms and the original columns were removed from the study as retaining them would lead to unnecessary duplication and risk of multicollinearity.
* Erroneous data points detection: In order to ensure the accuracy and validity of the analysis, it is imperative that the erroneous entries were identified and removed from the dataset. 32 observations with 29th February 2018 date were detected, which is an invalid date as it does not exist, were excluded from the study. Since the number is very small as compared to total observations, excluding them seems reasonable as it would not impact the study.
* Outlier Detection: Several outliers were detected in lead time and average prices per room by plotting the box plots. However, due to the lack of background about the data and the cause of such outliers, a conscious decision was taken not to remove them to avoid introduction of any bias in the dataset (Goyal, 2021).
* Data Balancing: The target variable has 70-30 distribution with cancelled bookings as minority class (can be seen in fig. 1), which reflects a closer to real-world scenario then the 50-50 balanced distribution. Since the aim of the study is to predict the cancellations, balancing it could lead to its underemphasis, impacting the practical utility of model. Hence, it was decided to use the unbalanced data going forward in the algorithms (Verzino, 2021).
* Data Encoding: In order to convert the text to numerical format, ASCII (American Standard Code for Information Exchange) encoding standards were used which assign a unique numerical code to each character such that algorithms could understand the textual information (Chen, 2022).

# Data Exploration and Visualization

The Exploratory Data Analysis (EDA) serves as a foundational aspect of the data-driven approach, enabling understanding of the variables, their interrelationships, and the overall distribution of the data points.

In this section, a comprehensive EDA was undertaken to thoroughly investigate the dataset at hand and the relationships have been presented using various visualization techniques such as bar graphs, pie charts, histograms etc.

The average price of a room charged to customers while making a reservation is majorly dependent on 3 factors i.e., market segment, room type and the meal plan chosen by the customer. It was observed that the prices were highest if rooms were booked online and lowest if it was a complimentary stay. A variety of rooms were available for bookings where type 1 and type 2 rooms are the cheapest and type 7 is the most expensive room with more than 100% increase in room fare. Lastly, more the number of meals included in the package, higher would be the room price.

A graph of a number of rooms

Description automatically generatedA graph of a bar

Description automatically generated

A graph of a meal plan

Description automatically generated

Figure 2: Relationship of Average Room Price with Other Variables

Room type 1 is the most preferred room with over 75% of reservations opting for it followed by room type 4 with 16% reservations. Though, 25% reservations of room type 1 and 16% of room type 4 chose to cancel their reservations at later dates. Room type 1 is one of the cheapest room available (as seen in fig. 2) which could be the possible reason for high demand.

A graph of different colored bars

Description automatically generated

Figure 3: Distribution of Room Type Reserved with Booking Status

Online bookings constitute the primary mode of reservations, followed by offline booking. Hence, we can observe the highest cancellation rate in the online reservations as well. It is also noteworthy that customers coming through complimentary segment have no cancellations at all.

A graph of sales and sales

Description automatically generated with medium confidence

Figure 4: Distribution of Market Segment by Booking Status

Fig. 5 shows the increasing trend in number of reservation count for both cancelled and non-cancelled bookings from July 2017 to December 2018. A rise in non-cancellations was observed in February, reaching its initial peak in April and then slightly decreasing in June-July, followed by another much higher peak in in September-October. The number of cancellations tend to move in the same direction as the number of confirmed bookings. However, the cancellations decreased drastically toward the end of the year indicating lower cancellation rate during the holiday period.

A blue and pink line graph

Description automatically generated

Figure 5: Variation of Reservation Count over Time

A consistent upward trend in the average prices of the rooms were observed over time with peaks occurring around September 2017, May-June 2018 and September 2018 and considerably lower prices during Dec-2017 and Dec-2018. This trend seems to be cyclic in nature and is in sync with the demand of the hotel rooms. As seen in fig. 5 (above), the demand of hotel rooms is lowest in the months of December and January, hence the rooms are available at cheaper prices whereas the reservation count is highest in the summer months from May to September leading to expensive rooms.

A graph with a blue line and a red line

Description automatically generated

Figure 6: Variation of Average Room Price over Time

There is a clear trend indicating that as lead time increases, likelihood of cancelling the reservations also increases (refer to fig. 7). There are higher chances of confirmed bookings if the lead time is lower.

A graph of a line graph

Description automatically generated

Figure 7: Impact of Lead Time over Booking Status

As seen in fig. 8, it is evident that majority of the guests (approx. 97%) have made reservation in the hotel for the first time while only 2.6% are the repeated guests who have stayed in the hotel before. This shows that there is an unexplored opportunity to increase the revenue by reaching out to our previous guests through marketing means and offering various discounts for their staycations. The cost of acquiring new customers is much more in case of new customers then the repeated guests (Wertz, 2018).

A pie chart with numbers and a number of values

Description automatically generated

Figure 8: Distribution of Repeated Guests

Lastly, determining the relationships between input variables and the target variable holds significant importance in machine learning models. Including features that have a strong correlation with the target variable can enhance predictive power and ultimately improve classification performance, resulting in more accurate predictions (builtin.com).

The correlation graph (refer fig. 9) reveals that booking status had strong positive correlation with “no\_of\_special\_requests” (25%), “market\_segment\_type” (15%), and “repeated\_guest” (11%) when compared with other fields. On the other hand, the most negatively correlated features were “arrival\_year” (-18%), “Arrival\_date” (-17%), “lead\_time” (-44%) and “avg\_price\_per\_room” (-14%). However, it is important to note that the other features have minimal or negligible impact on the customer's choice to cancel their booking.

A close-up of a graph

Description automatically generated

Figure 9: Correlation Coefficients

# Modelling and Analysis

The aim of the study is to predict if the customer would cancel the hotel reservation. With the booking status as our binary target field, our primary objective is to construct predictive models that discern the factors influencing cancellations and enable us to anticipate these outcomes. For this purpose, various supervised learning algorithms such as Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Logistic Regression and Random Forest were considered. Details related to these have been provided in the below subsections.

The 6 fields identified in the above section except “arrival\_year” (to avoid multicollinearity) to have high positive and negative correlation with the target field were only considered for the study.

Performance metrics like Accuracy, Sensitivity, Specificity, Precision and Kappa were used to evaluate the model. These metrics provide valuable insights into various aspects of a model's behavior and effectiveness, helping to gauge its quality, suitability, and potential for real-world applications. However, high weightage was given to Kappa, since it excludes the agreement by chance from Accuracy (Pykes, 2022).

## K-Nearest Neighbor (KNN)

The K-Nearest Neighbors (KNN) algorithm was used as one of the classification methods for prediction due to its wide usage and intuitive nature in supervised machine learning. It falls under the instance-based learning category, where data points are categorized based on their closest neighbors, relying on the majority class among them. KNN is categorized as non-parametric since it does not assume any specific data distribution, making it adaptable and suitable for various classification tasks (Lantz, Brett, 2019).

The performance of the model was enhanced through the following techniques:

* Feature scaling: KNN is sensitive to the scale of features, so normalizing or standardizing the data can improve the model's performance. The standardization technique (z-score normalization) was employed for feature scaling.
* K parameter selection: The Elbow method was utilized to determine the optimal value of k. Selecting the right k value is critical for balancing bias and variance in the model. The best result was achieved when k=9.

A graph with a line and a point

Description automatically generated

Figure 10: Elbow Method to determine K Value

* The cross-validation technique ensures generalization by testing various subsets of data. The elbow method was employed, and the best result was obtained with a cross-validation value of 10.
* Two distance metric was tried, and the best result was obtained with the Euclidean distance.

1. Euclidean distance, which represents the measurement one would obtain by using a ruler to connect two points (Lantz, Brett, 2019).
2. Manhattan distance calculates the distance based on the paths a pedestrian would take by walking city blocks (Lantz, Brett, 2019).

* Different training and test data proportions were experimented with, the model performed well with 70% training data and 30% test data.

When comparing model performance on test data before feature scaling and after feature scaling, the model gave the best results with standardized data (Table); Accuracy - **86.43%** and Kappa - **68.62%**.

Table 2: Performance Metrics (KNN)

|  |  |  |
| --- | --- | --- |
| **Metric** | **Without Feature Scaling** | **With Feature Scaling** |
| **Accuracy** | 80.42% | 86.43% |
| **Kappa** | 54.16% | 68.62% |
| **Sensitivity** | 64.16% | 75.61% |
| **Specificity** | 88.34% | 91.71% |
| **Precision** | 72.85% | 81.64% |

## Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful and widely used supervised machine learning algorithm. The main goal of (SVM) algorithms is to find an optimal boundary called the hyperplane to create fairly homogeneous partitions on either side and is most easily understood when used for Binary Classification. In a binary classification setting, the hyperplane acts as a decision boundary that maximizes the margin between the closest data points of the two classes, also known as support vectors. These support vectors are the critical data points that influence the position and orientation of the separating hyperplane (Lantz, Brett, 2019). Hence, it was considered in the project for binary classification of the target field.

Multiple iterations with varied training and test data proportions were run and the best results were observed with 60% training and 40% testing data. In order to improve the model performance, the following changes were carried out -

* Kernel Trick

The key feature of SVM is their ability to map the problem into higher dimensional space using the process called Kernel Trick in order to make a non-linear relationship appear to be quite linear (Lantz, Brett, 2019). Hence, easier to find the hyperplane.

The base SVM model used the simple linear kernel function i.e., “vanilladot” kernel. The model was updated with more complex kernel functions like “rbfdot”, “tanhdot” and “polydot” to potentially obtain the better model fit. By changing the kernel to “rbfdot”, there was significant increase in kappa and slight improvement in accuracy as well.

Table 3: Performance Metrics post Kernel Trick (SVM)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metrics** | **vanilladot** | **rbfdot** | **polydot** | **tanhdot** |
| **Accuracy** | 79.60% | 83.19% | 79.39% | 61.64% |
| **Kappa** | 50.59% | 60.45% | 50.25% | 12.84% |
| **Sensitivity** | 56.56% | 67.46% | 56.89% | 41.25% |
| **Specificity** | 90.84% | 90.86% | 90.35% | 71.57% |
| **Precision** | 75.06% | 78.25% | 74.20% | 41.44% |
| **Recall** | 56.56% | 67.46% | 56.89% | 41.25% |

* Cost Function

Another method to improve the model performance is by applying a penalty to all the misclassified samples. Higher the cost/penalty, less will be the misclassified samples and higher the computation time. However, this could lead to an increase the risk of overfitting on noisy data (Lantz, Brett, 2019). Hence, it is imperative to find an appropriate cost value to avoid the aforementioned problems.

In order to do so, different cost values were tried on the rbfdot kernel model (from the above step) and the best results were observed with C = 5. The impact of different cost function on the metrics can be seen in the below table.

Table 4: Performance Metrics post Cost Update (SVM)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Metrics** | **1** | **3** | **5** | **7** | **9** |
| **Accuracy** | 83.19% | 83.57% | 84.24% | 83.64% | 83.89% |
| **Kappa** | 60.45% | 61.53% | 62.65% | 61.57% | 61.77% |
| **Sensitivity** | 67.46% | 68.97% | 67.69% | 68.47% | 66.87% |
| **Specificity** | 90.86% | 90.68% | 92.30% | 91.03% | 92.19% |
| **Precision** | 78.25% | 78.30% | 81.09% | 78.82% | 80.68% |
| **Recall** | 67.46% | 68.97% | 67.69% | 68.47% | 66.87% |

Therefore, SVM model with 60-40 data partition, “rbfdot” kernel and C = 5 gave the best results on the test data.

## Random Forest

Random Forest is an ensemble supervised machine learning algorithm used for both classification and regression problem which combines many decision trees that eventually converts it into forest with the objective of overcoming the overfitting problem of the individual decision trees. ‘Random’ in Random Forest refers to mainly two processes-

1. Random observations to grow each tree
2. Random variables selected for splitting each node (Goyal, 2022)

As mentioned, the analysis was started by applying C5.0 Decision Tree algorithm on the different combination of training and test data set and model accuracy achieved as ***68.19%.*** Post that couple of performance improvement techniques of decision tree was executed:

* Controlling No. of Boosting Iterations – By performing normalization and controlling the no. of boosting iterations for creating C5.0 decision trees with trials(n)=1, 3, 5, 7, 10 and the best result received with n=7 at 80:20 (Training and Test Proportion) with Accuracy =**68.66%** and extremely low *Kappa=****7.75%***
* Cost Matrix – Implementation of cost matrix in the algorithm to specify the costs associated with various types of errors for creating C5.0 decision trees. The best results on the test data received at **80:20** proportion with *Accuracy=****85.46%*** *& Kappa=****66.75%***

Post implementation of performance improvement techniques on Decision Tree, it was important to implement Random Forest which is an ensemble of decision trees method called Bootstrap Aggregation or bagging. It uses sampling with replacement technique, in which new training datasets pick sample of data points with replacement (known as bootstrap samples) from the original datasets. The model is generated by combining different bootstrap samples and for prediction of a categorical variable, majority voting is used in which means the output which was generated maximum times would be the outcome of the model.

Random Forest algorithm was applied, and the best result received at the training and test proportion of **70:30** with *Accuracy=****87.49%*** *& Kappa=****71.06%.*** The major challenge with Decision Tree was overfitting as it has unlimited flexibility, meaning it keeps growing unless for every observation there is presence of only one leaf node which leads to high variance. To overcome this challenge, another performance improvement technique of ***“Setting the minimum number of terminal nodes (leaves) in each decision tree”*** and the results are in the table below:

Table 5: Performance Metrics with different terminal nodes (RF)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Metrics** | **5** | **10** | **20** | **30** | **45** | **60** |
| **Accuracy** | 86.46% | 86.99% | 86.22% | 87.27% | 86.62% | 86.77% |
| **Kappa** | 69.93% | 69.64% | 70.10% | 70.45% | 69.74% | 69.08% |
| **Sensitivity** | 69.72% | 75.20% | 75.97% | 76.17% | 74.96% | 75.16% |
| **Specificity** | 89.07% | 90.68% | 90.65% | 92.68% | 90.24% | 90.36% |
| **Precision** | 76.66% | 80.36% | 80.19% | 83.54% | 79.33% | 79.60% |
| **Recall** | 69.72% | 75.20% | 75.97% | 76.17% | 74.96% | 75.16% |

Therefore, the best result received at setting the minimum number of terminal nodes as **30**, i.e., *Accuracy=****87.27%*** & *Kappa=****70.45%***

## Logistic Regression

According to Lantz (2019), regression is identifying a relationship between a dependent and an independent variable. The assumption of the simplest form of regression is the relationship follows a straight line. There are numerous models in the regression algorithm, for instance, simple regression, multiple regression, Poisson regression, etc., in this project, we decided to use logistic regression because our target field is in a binary form. The dependent variable is the booking status column, and the independent variables are the remaining columns.

In the process, trials with varied proportions of test and training data were performed and multiple techniques were applied to the logistic regression model to achieve the best result. Firstly, all fields were included as independent variables in the regression model to predict the dependent variable or booking status. Secondly, only 6 fields that have a high impact on the target field were included as the independent variables. After that, the performance improvement techniques were used:

* Adding non-linear relationships – the relationship between independent and dependent variables in the logistic regression is assumed to be linear, however, this is not always true. For example, the number of guests who cancelled their booking or not may not be constant over the increasing lead time for booking.
* Transformation – for instance, at a certain price the room may affect the guest’s decision-making, whether to finalize the booking or not. There might be another cheaper room price for the same standard to be the guest’s option. Given that, turning the average price per room to binary form with the specified threshold could improve the model.
* Adding Interactions – an individual feature might have a high impact on the dependent variable, however, the impact from the combination effect of two or more features could be more.
* Putting all together – is the combination of adding nonlinear relationships, transformation, and adding interactions to the regression model.

Finally, the best performance came from adding non-linear relationships (power to 2) to all fields because relationships between 6 significant fields are not linear to the target field with the 70:30 proportion of training to test data.

Table 6: Performance Metrics (LR)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Metrics** | **lr\_model1** | **lr\_model3** | **lr\_model6** | **lr\_model11** | **lr\_model17** |
| **Accuracy** | 80.22% | 79.92% | 80.72% | 80.71% | 79.64% |
| **Kappa** | 52.73% | 51.58% | 52.44% | 53.85% | 50.75% |
| **Sensitivity** | 59.95% | 57.87% | 61.94% | 60.51% | 56.83% |
| **Specificity** | 90.11% | 90.67% | 89.87% | 90.56% | 90.76% |
| **Precision** | 74.71% | 75.15% | 74.89% | 75.76% | 75.00% |
| **Recall** | 59.95% | 57.87% | 61.94% | 60.51% | 56.83% |

*Note: The details of each logistic regression model can be seen in Appendix C.*

# Model Evaluation

Model evaluation is the process of quantifying a Machine Learning model’s performance to find the best performing one that fits the given problem. Evaluating models are crucial to ensure that the model works optimally and correctly when deployed in production or put to test in real life unseen datasets. There are different evaluation metrics and for the existing business problem Kappa was selected. The best practice for model evaluation is to test a model using the metrics to understand the model’s suitability (Comet, 2022)

Kappa is a statistical measure used in machine learning that compares the accuracy of predictions made by a model to the accuracy that would be expected by random chance. It assesses the model's capacity to learn from data and make predictions that hold up when applied to new, unseen data (Joseph, 2022). Based on thorough assessment of four different models’ performance, best ***Cohen’s Kappa of*** ***75.80%*** obtained for model prepared ***by Random Forest Algorithm*** which indicates substantial agreement between its predictors and the actual outcomes, beyond what would be expected by random chance, which will ultimately conclude that the model is making predictions that align well with true outcomes (McHugh, 2012).

Table 7: Model Evaluation along with Key Metrices

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metrics** | **Support Vector Machine** | **Logistics Regression** | **K-Nearest Neighbour** | **Random Forest** |
| **Accuracy** | 84.24% | 80.71% | 86.43% | 87.27% |
| **Kappa** | 62.65% | 54.22% | 68.62% | 70.45% |
| **Sensitivity** | 67.69% | 62.03% | 75.61% | 76.17% |
| **Specificity** | 92.30% | 89.82% | 91.71% | 92.68% |
| **Precision** | 81.09% | 74.81% | 81.64% | 83.54% |

# Return on Investment (ROI)

Return on Investment (ROI) is the performance measure used to evaluate the profitability or efficiency of an investment in any business (Fernando, 2023).

With successful identification of high-risk reservations using the **Random Forest Algorithm**, the hotels can proactively reach out to the customers to try to retain them by offering them discounts or other marketing schemes like giving the option to change the dates. However, the hotels would have to bear the marketing cost and in return the revenue would be generated by the customers who decide to ‘not cancel’ their reservation post the discounts or offers. Hence, it is imperative to calculate the hotels’ return on their investment (marketing cost).

Table 8: Confusion Metrics of Random Forest Algorithm

|  |  |  |  |
| --- | --- | --- | --- |
| Random Forest Algorithm | | Predicted Results | |
| Cancelled | Not Cancelled |
| Actual Results | Cancelled | TP = 3619 | FN = 1132 |
| Not Cancelled | FP = 713 | TN = 9031 |

*\* ”Cancelled” is considered as the positive entity while generating the confusion metric as the aim is to identify cancellations.*

As per the provided confusion metric, the random Forest Algorithm was able to identify that 3619 + 713 = 4332 customers would cancel their reservation. Therefore, the cost would be incurred to reach out to these customers. As a result of marketing efforts, 713 customers would not actually cancel their reservation. Therefore, Revenue would be generated from these customers.

Formula to calculate ROI (Teplu, 2023):



Revenue – Revenue from not cancelled Guests i.e., Average cost of room per guest

Expenses – Marketing Cost, which is approx. 10% of the average price of room (Turner, 2022)

Therefore,

ROI = \* 100% = **64.59%**

# Conclusions

The Random Forest classifier outperformed the other models and achieved an impressive accuracy rate of 89.5% with a Kappa value of 75.8% in predicting booking status. This implies that it is highly effective at differentiating between customers who are likely to cancel bookings and those who are not. The model's accuracy makes it a dependable tool for data-driven decision-making in marketing.

Based on the model's output, effective marketing activities on customers who are predicted to cancel bookings could result in a substantial 65% return on investment (ROI) on marketing expenditure, demonstrating the effectiveness of marketing strategies in retaining a significant portion of potentially lost business.

Although the model's accuracy and return on investment are impressive, it is crucial to consistently monitor its performance and adapt marketing strategies accordingly. Keeping the model up to date with new data will ensure its continued relevance and effectiveness. By incorporating these insights into marketing strategies, hotels can optimize its marketing efforts, improve customer retention rates, and ultimately increase overall profitability.

# Recommendations

Building upon the findings and conclusions, the subsequent section presents actionable recommendations tailored to the hospitality industry's business point of view.

* The hotel’s marketing team should spend most of the marketing budget on the online platform because, from the dataset, approximately 64% of the guests came through online booking, followed by offline booking at 29%. Additionally, the advent of websites and social media has made hospitality businesses more competitive than ever before. There are benefits that the hotel can get from digital marketing. Firstly, it is the most cost-effective way to get attention from the audience. Secondly, it helps the hotel to globally target a specific demographic of potential guests. Finally, it delivers valuable and usable analytics for further service improvement (Chechi, 2022).
* By knowing the customers’ booking patterns, the hotel can know in advance whether they will cancel their booking or not (Lee, 2018). From the analysis, the guests who book room type 1 via online booking in advance for many months are more likely to have a higher cancellation rate than the others. Given that, the hotel can increase the overbooking limits to get more actual guests or implement a less forgiving rate policy such as 7-day free cancellation after the booking, etc. By using these strategies for the identified periods and specific customer demographic, the hotel can remain and increase its profit.
* The peak season in Europe starts from late May until July, and sometimes it continues and ends in September or October (Firebird Tours, 2023). Along with the historical reservation data, the hotel could heavily throw its marketing budget during this period to attract guests as much as possible and to gain a higher return on investment.
* From the analysis, most booking cancellations come from online booking. Additionally, Lee (2018) claimed that customers who book their rooms via online platforms tend to have a higher cancellation rate than those who book directly with the hotel. The explanation for this is that when potential customers find a hotel within their price range, they will book it right away and cancel it later if they find a better option. To retain the guests, the hotel should try to offer them promotional discounts or exclusive offers. This strategy could make the hotel to be the first choice for the guests.
* In the hospitality business, the value of repeated guests cannot be overlooked. The amount of money needed to be spent to gain a new customer is 5 to 8 times more than the cost of retaining the existing one (Social Tables, 2021). To retain the guests, the hotel has to focus on its Customer Relationship Management (CRM) system to gain insight into its customers’ behaviors and experiences. After knowing these values, the hotel should try to craft guest experiences that create relationships and a sense of home, which eventually will embrace brand loyalty to the guests.

# References

Lantz, Brett. *Machine Learning with R - Second Edition : Discover How to Build Machine Learning Algorithms, Prepare Data, and Dig Deep into Data Prediction Techniques with R*, Packt Publishing, Limited, 2015

Goyal, C. (2022) “Bagging- 25 questions to test your skills on Random Forest Algorithm,” Analytics Vidhya [Preprint]. Available at: <https://www.analyticsvidhya.com/blog/2021/05/bagging-25-questions-to-test-your-skills-on-random-forest-algorithm/>.

Raza, A. (2023) *Hotel Reservations dataset*, *Kaggle*. Available at: <https://www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset>  
(Accessed: 03 August 2023).

Fernando, J. (2023) *Return on investment (ROI): How to calculate it and what it means*, *Investopedia*. Available at: <https://www.investopedia.com/terms/r/returnoninvestment.asp>

(Accessed: 03 August 2023).

Wertz, J. (2018) “Don’t Spend 5 Times More Attracting New Customers, Nurture The Existing Ones,” Forbes, 12 September. Available at: <https://www.forbes.com/sites/jiawertz/2018/09/12/dont-spend-5-times-more-attracting-new-customers-nurture-the-existing-ones/>.

Goyal, C. (2021) “Why You Shouldn’t Just Delete Outliers,” Analytics Vidhya [Preprint]. Available at: <https://www.analyticsvidhya.com/blog/2021/05/why-you-shouldnt-just-delete-outliers/>.

Verzino, G. (2021) *Why balancing classes is over-hyped*, *Medium*. Available at: <https://towardsdatascience.com/why-balancing-classes-is-over-hyped-e382a8a410f7>

(Accessed: 05 August 2023).

Comet, T. (2022) Why is Model Evaluation Important in Machine Learning? Available at: <https://www.comet.com/site/blog/why-is-model-evaluation-important-in-machine-learning/>.

Chen, J. (2022) *American code for information interchange (ASCII)*, *Investopedia*. Available at: <https://www.investopedia.com/terms/a/american-code-for-information-interchange.asp>

(Accessed: 05 August 2023).

Joseph (2022) “What is Kappa in Machine Learning? - reason.town,” reason.town, 16 August. Available at: <https://reason.town/kappa-in-machine-learning/>.

McHugh, M.L. (2012) Interrater reliability: the kappa statistic. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/#:~:text=Cohen%20suggested%20the%20Kappa%20result%20be%20interpreted%20as,as%20substantial%2C%20and%200.81%E2%80%931.00%20as%20almost%20perfect%20agreement>.

Chechi, H. (2022) *Hospitality Digital Marketing*, *Les Roches*. Available at: <https://lesroches.edu/blog/hospitality-digital-marketing/#:~:text=Highly%20targeted%3A%20It%20can%20be,your%20guests’%20needs%20and%20wants>.

(Accessed: 04 August 2023).

Firebird Tours ® (2023) When to Go to Europe: Peak Season vs Off-Season. Available at: <https://www.firebirdtours.com/blog/when-go-europe-peak-season-vs-season>.

(Accessed: 04 August 2023).

Zajac, J. (2023) *Hotel Budget Planning: Your 2023-2024 hotel marketing budget*, *Five Star Content*. Available at: <https://fivestarcontent.co/blog/hotel-budget-guide>

(Accessed: 04 August 2023).

Lee, J.E. (2018) *Three strategies to tackle hotel cancellations*, *Pegasus*. Available at: <https://www.pegs.com/blog/three-strategies-to-tackle-hotel-cancellations/>

(Accessed: 05 August 2023).

Social Tables (2021) *8 effective customer retention strategies for Hotels*, *Social Tables*. Available at: <https://www.socialtables.com/blog/hotel-sales/customer-retention-strategy/>

(Accessed: 05 August 2023).

Pykes, K. (2022) *Cohen’s Kappa explained*, *Built In*. Available at: <https://builtin.com/data-science/cohens-kappa>

(Accessed: 05 August 2023).

Turner, J. (2022). *What Marketing Budgets Look Like in 2022*. [online] Gartner. Available at: <https://www.gartner.com/en/articles/what-marketing-budgets-look-like-in-2022>.

‌Teplu, E. (2023) *The ROI formula: How to calculate it and why your marketing needs it*, *Dashly blog*. Available at: <https://www.dashly.io/blog/roi-formula/>

(Accessed: 05 August 2023).

# Appendix A – Generative AI Flowchart



# Appendix B - Algorithms

The Excel file provided below contains a comprehensive collection of outcomes resulting from the application of various methodologies to individual models. This Excel spreadsheet gives a valuable overview of the effectiveness and importance of the techniques used, driving our efforts to optimize the performance of each machine model algorithm. Take a closer look at the Excel sheet to explore the specific results that have guided our progress towards improved model efficiency.



# Appendix C – Logistic Regression Models

* *lr\_model1 <- train(booking\_status ~ ., data = train.data, method = "glm", family = "binomial")*
* *lr\_model3 <- train(booking\_status ~ lead\_time + Arrival\_date + market\_segment\_type + avg\_price\_per\_room + repeated\_guest + no\_of\_special\_requests, data = train.data, method = "glm", family = "binomial")*
* *lr\_model6 <- train(booking\_status ~ lead\_time + Arrival\_date + market\_segment\_type + avg\_price\_per\_room + repeated\_guest + no\_of\_special\_requests + lead\_time2 + Arrival\_date2 + market\_segment\_type2 + avg\_price\_per\_room2 + repeated\_guest2 + no\_of\_special\_requests2 , data = train.data\_nl, method = "glm", family = "binomial")*
* *lr\_model11 <- train(booking\_status ~ market\_segment\_type\*avg\_price\_per\_room\*no\_of\_special\_requests\*lead\_time\*Arrival\_date\*repeated\_guest, data = train.data, method = "glm", family = "binomial")*
* *lr\_model17 <- train(booking\_status ~ lead\_time\*repeated\_guest + Arrival\_date + market\_segment\_type + avg\_price\_per\_room\_BINARIZED + repeated\_guest + no\_of\_special\_requests + Arrival\_date2 + market\_segment\_type2 + no\_of\_special\_requests2 , data = train.data\_nl, method = "glm", family = "binomial")*

# Appendix D - BACCM



# Appendix E – Project Plan



# Appendix F – R Scripts









