Tilburg University

PREDICTING HOTEL BOOKING DEMAND AND CANCELLATIONS USING MACHINE LEARNING AND COMPARISON OF FEATURE IMPORTANCE

RAFAIL CHATZILADAS

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2112855

COMMITTEE

dr. Sharon Ong
Bosong Ding

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 24th, 2024

WORD COUNT

7391

ACKNOWLEDGMENTS

I am deeply thankful to my family, and especially my mother, whose unwavering faith in me has given me motivation and strength throughout my academic journey. I would also like to express my gratitude to my partner who played a pivotal role in my journey to the Netherlands, providing invaluable guidance on navigating the path to finding my passion and achieving my goals. Her support was instrumental in helping me discover my academic pursuits, and her presence and encouragement has enriched every moment of this academic journey. I extend my heartfelt appreciation to Dr. Sharon Ong, my supervisor, for the exceptional guidance, insightful feedback, and understanding throughout my thesis writing process. Additionally, I am grateful to Bosong Ding for kindly serving as the second reader for this thesis.

PREDICTING HOTEL BOOKING DEMAND AND CANCELLATIONS USING MACHINE LEARNING AND COMPARISON OF FEATURE IMPORTANCE

RAFAIL CHATZILADAS

Abstract

This thesis attempts to classify to what extent machine learning can be used to predict whether a reservation in a hotel will be canceled and compares the feature importance of each machine learning model with eXplainable Artificial Intelligence (XAI) methods. For this purpose, the thesis utilized the hotel booking demand dataset, which is publicly available on Kaggle. The database combines reservation data from two distinct hotels, a resort and a city hotel, in Portugal. The data consists of information such as the time, the status, and the country of the bookings. The results indicate that among Logistic regression, Random Forest, and Extreme Gradient Boost, the most accurate model is RF (accuracy, 07844; precision, 0.8012; recall, 0.9341; F1-score, 0.8626). In terms of the feature importance comparison, SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and permutation feature importance test were conducted for each ML model and revealed that lead_time is the most influential among the features. Furthermore, the thesis employed Jaccard similarity and Spearman correlation to evaluate the XAI methods. Results show that SHAP and permutation tests have the higher similarity and correlation when ranking the essential features. This study shows that machine learning can be used to predict hotel booking cancellations and XAI techniques can be used to evaluate feature importance. Additionally, it aims to encourage the usage of machine learning models and artificial intelligence (AI) systems in the hospitality industry

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

1.1 *Source/Code/Ethics/Technology Statement*

Data Source: The dataset that the thesis utilized was collected by Antonio, de Almeida, and Nunes (2019b). The dataset is publicly available and can be obtained freely from Kaggle, and the data were collected through the servers of the hotels' PMS databases (Antonio, de Almeida, & Nunes, 2019b). All the images and figures contained in this paper are created by the paper's author. This study does not involve any data collection from human participants or animals. The dataset is publicly available for use on Kaggle. Any data elements that could expose identification information about hotels or customers have been removed. The current thesis contains visualizations and presentations of the data obtained from cited sources. All the libraries and frameworks that the thesis employed are presented in a paragraph, along with their version numbers. For paraphrasing, correct spelling, and grammar checking, this study applied Grammarly and Thesaurus. No other typesetting tools or services were used.

2 INTRODUCTION

2.1 *Project Definition*

This thesis uses machine learning models to predict hotel booking cancellation from a city and a resort hotel based in Portugal (Antonio, de Almeida, & Nunes, 2019b). Predicting cancellation rates in the hospitality industry can be beneficial for hotel businesses in order to maximize their revenue. The current thesis makes use of different Machine Learning models, such as Logistic Regression, Random Forests, and Extreme Gradient Boosting, to accurately predict future cancellation rates. Additionally, the thesis compares different post-hoc eXplainable Artificial Intelligence (XAI) techniques, including SHapley Additive exPlanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME), and permutation feature importance. Although related work in predicting hotel room cancellation exists, none of the existing works compares the features evaluated by different post-hoc XAI techniques for this application. Therefore, this thesis compares SHAP, LIME, and permutation importance and evaluates the similarity of these XAI techniques and the feature importance, by using Jaccard similarity and Spearman correlation techniques.

2.2 *Societal Motivation*

The tourism industry has been developing rapidly over the last decades, resulting in a growth in the hotel industry (Claveria et al., 2015), and uncovering its weaknesses (Zheng et al., 2020). In the hospitality industry many difficulties exist, which causes weaker businesses to disappear. Daily competition between them is one of their main challenge (Tang et al., 2015), but the businesses are also vulnerable to external events, such as economic crises (Webb et al., 2020) and unexpected weather conditions (Qiu et al., 2021). The spread of COVID-19 and the subsequent lockdowns during that period are further examples of unexpected and challenging situations from which the industry has not yet recovered from (Hao et al., 2020).

These difficulties force hotel managers to seek strategies to keep their businesses competitive and maximize their profits (Veiga et al., 2020). Accurately predicting future booking demand is crucial, as it can provide a significant advantage in the marketplace (Archer et al., 1987; Huang & Zheng, 2021; Wandner & Erden, 1981). Moreover, hotels, in their pursuit to attract more customers, apply more competitive pricing and cancellation policies, such as flexible cancellation plans , which often leads to empty hotel rooms and therefore loss of income (C.-C. Chen et al., 2011). To avoid that, hotels design cancellation policies that are often strict in order of timing with cancellation deadlines, allowing them a short window of time to resell canceled rooms. Additionally, penalties are imposed on customers who cancel their reservations after these deadlines. (M. Yoo et al., 2024). Therefore, hotel booking cancellations play a vital role in a hotel's performance, as the cancellation rates could lead to unsold rooms and thus, to a loss of revenue (Sierag et al., 2015; M. Yoo et al., 2024). The use of machine learning models and XAI methods could increase the accuracy of hotel booking demand and cancellation prediction.

2.3 *Scientific Motivation*

The vast tourism industry has come to significant attention, as it contributes an incredible amount of income for many countries (Tsang & Benoit, 2020). Hotel occupancy, on the other hand, is comparatively under-studied and has only recently gained broader interest (Huang et al., 2022). This could be attributed due to the complexity of the available data (Huang & Zheng, 2021). Additionally, prediction of hotel booking demand has primarily been conducted based on time series (Zhu et al., 2021), which struggle when they are used to handle nonlinear and complex data (Kırtıl & Aşkun, 2021) resulting in a decrease in prediction's accuracy (Zheng et al., 2020). ML models can handle complex data and enormous datasets, and capture complicated relationships between various factors, and thus, have a greater impact on prediction accuracy (Zhu et al., 2021). Moreover, results deriving from machine learning models can be difficult to be explained, due to the lack of providing clear insights on their process (Gilpin et al., 2018). EXplainable Artificial Intelligence (XAI) methods can evaluate features' contribution to the ML models' results (Elkhawaga et al., 2023). These techniques do not treat the models as a whole, but instead they handle the model's specific outputs (Leslie, 2019; Selbst & Barocas, 2018). In the research that have been conducted over the years, there are not enough studies that used XAI models to assess the explainability of their models, especially in the hotel industry.

2.4 *Research Strategy*

2.4.1 *Description*

There are not many applications of machine learning on data from the hospitality industry and therefore, it would be beneficial to employ these methods. This study aims to explore a large dataset of hotel booking data to identify the most accurate ML model for forecasting cancellations and which features have the largest impact on the results. To derive reliable insights and predictions from a complex dataset, selecting the most accurate models is essential. Comparing various machine learning algorithms can help us identify the best classification model, especially when dealing with large and complex datasets. Therefore, the main question of this project is: *To what extent can Machine Learning models (Logistic regression, Random Forest, Extreme Gradient Boost) accurately predict hotel room cancellations, and what are the main cancellation drivers?*

Sub Questions: The thesis aimed to identify the most important features that could increase the performance of each model using univariate feature selection. Moreover, to understand the models predictive results, the thesis applied three eXplainable Artificial Intelligence methods, which is a novel approach in the field of hotel industry. Specifically, SHAP, LIME, and permutation feature importance were applied on the most accurate model. Therefore the first sub question is:

What are the predictive features used by the ML models, and can different eXplainable Artificial Intelligence techniques provide explanations for these features? Furthermore, the thesis investigated potential links and similarities between the eXplainable Artificial Intelligence models using Jaccard similarity and Spearman correlation. Through this extensive analysis, the project aimed to identify any notable effects caused by individual features. Thus, the second sub question is:

How do the findings of diverse eXplainable Artificial Intelligence techniques contrast?

2.4.2 *Research Findings*

The current thesis developed three ML models to predict hotel booking cancellations. Logistic Regression, Random Forest, Extreme Gradient Boost were optimized using their top hyperparameters and evaluated based on accuracy, precision, recall, and F1 metrics. The findings indicate that the RF model is the most accurate with an accuracy score of 0.784. The XGB model's accuracy score of 0.781 also shows a strong predictive performance. To assess the ML models' performance three eXplainable Artificial Intelligence methods were utilized. Their results were compared using Jaccard similarity and Spearman correlation. The results of the XAI methods highlighted that lead_time is the most influential feature across all models. Additionally, the comparison of the XAI methods revealed that SHAP and permutation importance have higher similarity and correlation in each model.

3 RELATED WORK

3.1 *Machine learning models for hotel cancellation*

Demand in the hospitality industry has to be managed strategically due to the industry's uncertainty, where various internal and external factors, such as time of booking, season, brand reputation, weather, and natural disasters, can impact cancellation rates (M. Yoo et al., 2024). Moreover, not renting rooms on particular dates equals a loss in revenue for the hotels (M. M. Yoo & Yang, 2021). Thus, efficiently predicting future demand is essential for hotel managers to maintain profitability (M. M. Yoo & Yang, 2021). In recent studies that attempted to predict cancellation rates, and applied either classification or regression algorithms, different results have been achieved. While Morales and Wang (2010) suggested treating the prediction of hotel booking cancellation as a regression problem, later studies showed the effectiveness of classification algorithms, achieving cancellation rates greater than 0.90 (Antonio, De Almeida, & Nunes, 2019; Antonio, de Almeida, & Nunes, 2019a; Antonio et al., 2017; Falk & Vieru, 2018; Sánchez-Medina, Eleazar, et al., 2020). Furthermore, the existing literature suggests that, in order to achieve a more substantial prediction power, a predictive model must contain many variables (Antonio, De Almeida, & Nunes, 2019; Antonio, de Almeida, & Nunes, 2019a; Antonio et al., 2017; C.-C. Chen et al., 2011; Sánchez-Medina, Eleazar, et al., 2020). Furthermore, currently, neural networks are the most well-known machine learning model for cancellation predictions (Y. Chen et al., 2022). This models, though, have many limitations, as they are computationally expensive, especially compared to other traditional models (Tu, 1996). Thus, as Y. Chen et al. (2022) noted, more literature on hotel booking cancellation prediction is needed.

Other studies have also treated hotel booking cancellation as a classification problem. Different methods have been used, and various results have been obtained. Andriawan et al. (2020) employed four tree-based models and found that Random Forest had the highest accuracy score of 87%. That study concluded that Machine Learning models could significantly reduce hotel income loss (Andriawan et al., 2020). Another significant research study is the one by Antonio, de Almeida, and Nunes (2019a), where the project built an automated model to make predictions. Based on the XGBoost algorithm, the automated model resulted in 84% correct classifications from the reduction in cancellations that the algorithm caused. This reductions in cancellations led to annual savings of 39 millions euros for the hotels implementing the model (Antonio, de Almeida, & Nunes, 2019a).

3.2 *Important features in hotel room cancellation prediction*

The results of the different studies, even if they give a scope of how the models normally perform, can only be accepted for the specific dataset each study uses (Antonio, de Almeida, & Nunes, 2019a). For example, a model created for a specific hotel and made its predictions based on a specific dataset can have different prediction results and be less or more accurate in different circumstances (Gartvall & Skånhagen, 2022).

An important step to be able to explain a model's result is to find which features are more helpful in predicting a target variable (Lundberg & Lee, 2017). For hotel booking cancellation prediction, the existing literature has used a tool named feature importance. Feature importance is a widely used tool that helps in understanding each variable's contribution (Adler & Painsky, 2022). It calculates the importance of each variable the model uses and assigns a score based on its usefulness for the prediction (Zien et al., 2009). Furthermore, feature importance can specify the importance and the affection, negative or positive, each variable has on the performance of each model (Gartvall & Skånhagen, 2022). Various studies have employed feature importance on different datasets, and discovered interesting results. For instance, Andriawan et al. (2020), whose Random Forest model had an accuracy of 87%, and Gartvall and Skånhagen (2022), whose Random Forest had an accuracy of 78,6% indicated the most influential feature called lead_time, which is the deviation between the day of the booking and the arrival_day. Antonio et al. (2017), where tree-based models had an accuracy of around 0.95, the feature land_of_origin was the most influential.

3.3 *Comparing eXplainable Artificial Intelligence on different tasks*

Prediction models must be precise, and even minor adjustments can affect the results (Biecek et al., 2021). Therefore, individuals must understand the datasets they are working with in depth. XAI techniques would allow for a better understanding of a complex model's result (Gilpin et al., 2018) by detecting the most essential features on which the model bases its performance and obtaining insights into the decision process (Biecek, 2018).

Even though they are generally new, XAI techniques are used in different fields, primarily within medical, healthcare, and finance (Bhargava & Gupta, 2022). While some eXplainable Artificial Intelligence techniques can significantly explain certain data types, they perform in a different way based on the context (ElShawi et al., 2021). For instance, Guleria et al. (2022) used the XAI techniques to assess the results of cardiovascular disease prediction. That study calculated the mean of absolute SHAP values between all the features and found that the most crucial feature was sex, followed by age (Guleria et al., 2022). Furthermore, Singh et al. (2022) employed the SHAP technique to explain the results from each ML model they used to estimate the N statues of wheat from hyperspectral data. That study stated that according to SHAP, the N status of each plant increased from lower wavelength value (Singh et al., 2022). Several studies have reviewed the possibility of comparing different explainable AI methods. For instance, Rebane et al. (2021) employed the top-k Jaccard index to compare the similarity of the different XAI techniques used to predict drug events. Another example is Krishna et al. (2022), where Spearman correlation is used to analyze the agreement of the feature ranking that the XAI methods provided.

3.4 *Critical Gap in the Literature*

Forecasting future cancellation has come into play in recent years, and several studies exist aiming to increase the prediction's accuracy. However, there is a lack, in the hospitality industry, of studies utilizing ML models for hotel booking cancellation predictions. Furthermore, the application of eXplainable Artificial Intelligence techniques, essential for interpreting model decisions, remains largely unexplored in the hospitality industry. The XAI methods have proven effective in domains like healthcare (Dan et al., 2020; Duell et al., 2021; Peres et al., 2020) and finance (Biecek et al., 2021) and could provide valuable insights into the factors driving cancellation prediction.

4 METHOD

This thesis applies several machine learning techniques to predict hotel room cancellation. Logistic Regression serves as the baseline model against which the thesis compares Random Forest and XGBoost. In addition, the thesis investigates the importance of each feature in influencing the performance of these models by employing the SHAP, LIME, and permutation feature importance tests. Furthermore, the thesis attempts to identify the similarities and the correlations between the different XAI techniques using Jaccard similarity and Spearman correlation metrics. Figure 1 shows the process overview.

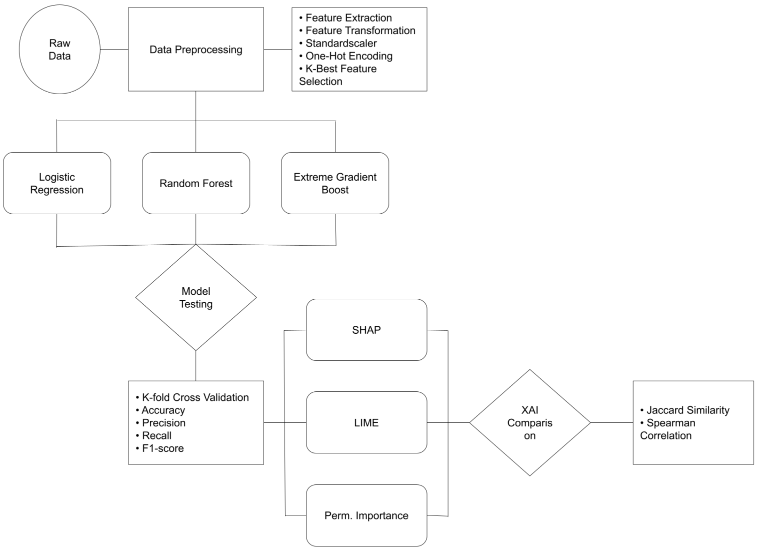

Figure 1: Process overview.

4.1 *Machine Learning Models*

4.1.1 *Logistic regression*

Logistic regression, introduced by Cox (1958), is a fundamental classification algorithm widely used due to its interpretability and efficiency (Hosmer Jr et al., 2013). It models the relationship between the dependent and independent variables by estimating probabilities bounded between

0 and 1 (LaValley, 2008). This simplicity allows for clear interpretation of how each feature influences the likelihood of an outcome, making logistic regression an ideal baseline model for classification tasks (Hosmer Jr et al., 2013).

4.1.2 *Random Forest*

Random Forest is a classification and regression (CART) model introduced by Breiman (2001). Through randomization, it creates a large number of decision trees, and the output is aggregated using voting into a single output, which is then trained through bagging and boosting methods. The bagging method produces distinct training sets sampled from the original data with a replacement, and the boosting method converts the aggregation of weak learners into strong learners (Rigatti, 2017). The thesis employs the Random Forest model as it is known for its robustness, high accuracy, and reduction of over-fitting (Barreñada et al., 2024).

4.1.3 *Extreme Gradient Boosting*

Extreme Gradient Boosting (XGBoost), an implementation of gradient boosting, was introduced by T. Chen and Guestrin (2016). Known for its versatility in handling regression and classification tasks, XGBoost has become widely adopted in machine learning due to itss robustness and computation efficiency (T. Chen et al., 2015). The decision to utilize XGB in this thesis stems from its ability to mitigate overfitting while maintaining high predictive performance, a feature highlighted in recent studies (Wade & Glynn, 2020). By employing XGBoost, the thesis aims to enhance the accuracy and interpretability of its predictive models, particularly in the context of hotel room cancellation prediction.

4.2 *Explainable Artificial Intelligence*

This study will employ three explainable artificial intelligence (XAI) methods to assess the feature importance of the different machine learning models: Sharpley additive explanation (SHAP), local interpretable model-agnostic explanations (LIME), and permutation feature importance. The XAI methods will calculate the explanation score for the best-performing model, and their results will be compared.

4.2.1 *Shapley Additive Explanations*

SHAP (Shapley Additive Explanations) is a technique capable of providing both local and global explanations in machine learning models (Duell et al., 2021). First introduced by Lundberg and Lee (2017), SHAP assigns an importance value to each feature, enabling a deeper understanding of how each feature influences model predictions. SHAP is effective in enhancing model interpretability and decision-making processes by clarifying feature contributions in a more comprehensive manner (Lundberg et al., 2020).

4.2.2 *Local Interpretable Model-Agnostic Explanations*

The LIME (Local Interpretable Model-Agnostic Explanations) technique was introduced by Ribeiro et al. (2016). This method can identify an interpretable model that remains locally faithful to the classifier while utilizing the interpretable representation (Ribeiro et al., 2016). LIME explains the instance of interest by fitting an interpretable model to a perturbed sample surrounding the input instance of interest (ElShawi et al., 2021).

4.2.3 *Permutation feature importance*

Permutation feature importance is a method that first permutes the feature and then calculates any increase in the model's prediction error. Shuffling increases the error of 'important' features and does not affect the values of the 'unimportant' ones (Molnar, 2020). This method was introduced by Breiman (2001).

4.3 *Hyperparameter Tuning Exploration*

To maximize each model's predictive performance, this thesis used the additional step of hyperparameter tuning based on the Grid Search method. During this optimization process, the thesis used the 10-fold cross validation technique to solve the models' bias. Hyperparameter tuning searches for the optimal model's parameters and applies them to each algorithm (Alibrahim & Ludwig, 2021).

Table 1: Hyperparameter Exploration for Logistic Regression.

Hyperparameter	**Search Range**	**Best Results**
solver	'newton-cg', 'lbfgs', 'sag' 'liblinear', 'saga'	'liblinear'
penalty	'l1', 'l2', 'elasticnet'	'l1'
C	uniform(0,001, 10)	0.97
max_iter	randint(100, 500)	387

Table 2: Hyperparameter Exploration for Random Forest.

Hyperparameter	Search Range	Best Results
bootstrap	'True', 'False'	'True'
criterion	'gini', 'entropy'	'gini'
max_depth	'None' or randint(10, 50)	35
max_features	'sqrt', 'log2'	'sqrt'
min_samples_leaf	randint(1, 20)	1
min_samples_split	randint(2, 20)	19
n_estimators	randint(100, 500)	323

Table 3: Hyperparameter Exploration for XGBoost.

Hyperparameter	Search Range	Best Results
colsample_bytree	uniform (0.6, 0.4)	0.65
gamma	uniform(0, 0.5)	0.19
learning_rate	uniform(0.001, 0.1)	0.08
max_depth	randint(3, 10)	7
min_child_weight	randint(1, 6)	5
n_estimators	randint(100, 1000)	146
reg_alpha	uniform(0, 1)	0.41
reg_lambda	uniform (0, 1)	0.77
scale_pos_weight	uniform (1, 10)	1.12
subsample	uniform (0.6, 0.4)	0.84

4.4 *Evaluation process*

4.4.1 *Classification Models Evaluation*

The Machine Learning models used to forecast hotel booking cancellations were evaluated using Accuracy, Prediction, Recall, and F1-score metrics. The accuracy metric measures how correct the results of the ML models are.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

The precision metric measures the proportion of the true positive predictions out of all positive predictions made by the models.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

The recall metric measures the proportion of the actual positive cases that each model correctly identifies.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

The F1-score metric measures the harmonic mean between precision and recall and provides a balance between these two metrics.

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{4}$$

4.4.2 *XAI Methods Evaluation*

The thesis evaluated the different scores obtained from XAI methods by taking the mean absolute importance score for SHAP and LIME and the original importance score of the permutation test. Furthermore, Jaccard similarity and Spearman correlation were used to compare the scores of the different XAI methods. The Jaccard similarity coefficient measured the similarity between the top essential features from each XAI method. Spearman correlation then measured how well the different explanations agreed on feature importance.

5 EXPERIMENTAL SETUP

5.1 *Dataset Description*

This thesis used the dataset collected by Antonio, de Almeida, and Nunes (2019b). This dataset contains information about two hotels located in Portugal. One is a city hotel in Lisbon, and the other is a resort hotel in the Algarve resort region. The dataset includes 31 variables describing more than 100,000 observations for the two hotels. These observations are bookings due to arrive between July 1, 2015, and August 31, 2017, and they represent both bookings that actually arrived and canceled bookings. All the data in this dataset was obtained from the servers of the hotels' PMS databases using a TSQL query on SQL Server Studio Manager (Sirkin & Hughes, 2017). To protect the privacy of the hotels and their customers, any data elements that could expose hotel or customer identification information have been deleted (Abbott, 2014). The current thesis uses variables from various tables within the Property Management System (PMS) database (Antonio, de Almeida, & Nunes, 2019b). The different variables have been either extracted through the bookings, change log database tables or engineered from different variables across the tables

(Antonio, de Almeida, & Nunes, 2019b). Appendix A provides the list of the dataset's features and their description.

5.2 *Data Preparation*

In the feature engineering process, the thesis selected 29 out of 31 features for processing. The features agent and company were dropped because they contained mostly NaN values. The company column had missing values accounting for 94.3% of the total values, making its removal necessary. As for the agent column, the missing values accounting for 13.6% of the total were also dropped as they were deemed not quite relevant. Furthermore, two more features with missing values were identified. The children feature contained 4 NaN values, filled with 0. The country feature contained 488 NaN values, filled with the most frequent value of the country column, PRT. Additionally, the thesis dropped duplicate columns to ensure the integrity of the data. Appendix B gives a list of the features that contained missing values.

After the initial preprocessing, additional adjustments were made. The meal feature contained the values SC and Undefined, which both mean no_meal_package (Antonio, de Almeida, & Nunes, 2019b). These were combined under the SC feature. Moreover, the thesis removed irrational data with values 9 and 10 in the feature babies. Additionally, the thesis filtered out rows where both stays_in_weekend_nights and stays_in_week_nights were zero. A new dataframe, stays, was created to contain instances where weekend and weekday stays were zero. 651 instances with zero values were found and dropped as irrelevant. Furthermore, when combining the children and babies features under the kids features, the thesis checked for rows with no adults but with kids. It found 219 such rows and dropped them as unrealistic.

5.3 *Exploratory Data Analysis*

The Exploratory Data Analysis (EDA) provides a thorough dataset overview and highlights specific details such as outliers, correlations, and trends. A numerical correlation matrix reveals correlations lower than 0.5, suggesting no potential multicollinearity. The highest correlation is 0.41 and is detected between adr and total_people, which is a shortening of variables adults, children, and babies for computational conveniences. The correlation matrix can be found in Figure C2 in Appendix C. Furthermore, the thesis addresses a multicollinearity check. The Variance Inflation Factor (VIF), was measured, and it was found that VIF values are relatively low, mostly below 1.5. The results are provided by Table C1 in Appendix C.

The current thesis uses 74,737 bookings, of which 22,364 are canceled reservations. The city hotel has the highest cancellation rate among the two hotels, with 31.9% canceled reservations, instead of 26.7% from the Resort Hotel. The thesis measured the correlation between the features reserved_room_type, assigned_room_type, and is_canceled to examine any possibility for a customer to have canceled their booking in a case where the customer reserved a specific type of room, and the hotel provided them a different one. The higher correlation is detected between Room Type F and Room Type D as it can be seen in Figure C3 in Appendix C. Furthermore, the study examines whether days_in_waiting_list or lead_time increase the possibility for a customer to cancel the reservation. The thesis found that days_in_waiting_list increases the possibility for a booking to get canceled more than lead_time, as half of the bookings that stay around fifty days without being confirmed by the hotels get canceled. Figure D1 in Appendix D provides visualization. To better understand the data set, the thesis conducted computations between adr and arrival_date_month for each hotel to find the most valuable months based on their income. Thus, it will be easier to understand in which months each hotel would have difficulty dealing with late cancellations. For the resort hotel, the higher adr is in August and July, and for the city hotel, it is in August and May. A visualization can be found in Figure D2 in Appendix D.

5.3.1 *t-SNE analysis*

This thesis uses the t-SNE (t-distributed stochastic neighbor embedding) on the test set for the machine learning models to better understand the data and gain information about the distribution of different classes within the dataset used for the hotel booking predictions (Van der Maaten & Hinton, 2008). This technique was chosen because it can maintain the local structure of the data points in low-dimensional space. Thus, it is really useful as it visualizes complex datasets. The visualization can be found in Figure 3,4, and 5. The thesis performed t-SNE with different perplexities to visualize the distribution on the true labels of the test set as accurately as possible. The different perplexity values used were 10, 50, and 100. Furthermore, each perplexity visualization used different color coding for the two classes (class 0 and class 1).

Figure A shows that the perplexity = 10 did not accurately separate the data. However, increasing the perplexity value made the separation more distinct. Figures b and c, with perplexities of 50 and 100, respectively, show that our models were able to distinguish the different classes in the dataset. The separation between class 0 and class 1 is particularly noticeable between the values of 60 and 80 on the y-axis and between -40 and -20 on the x-axis.

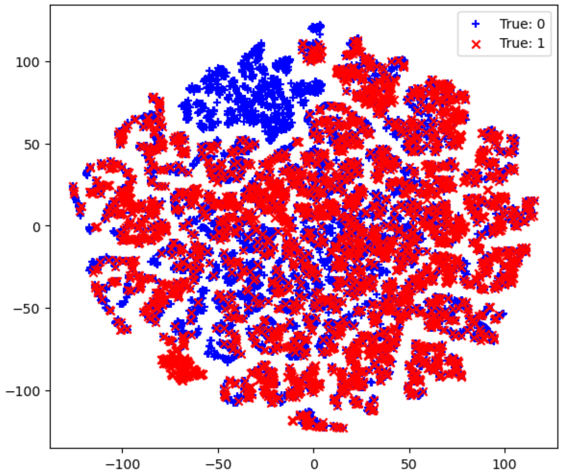

Figure 2: t-SNE analysis with perplexity = 10.

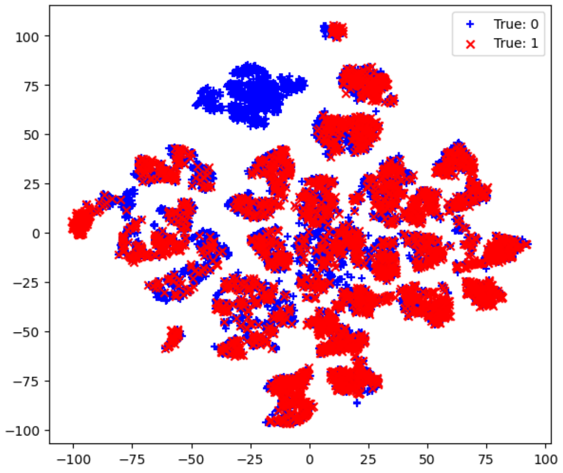

Figure 3: t-SNE analysis with perplexity = 50.

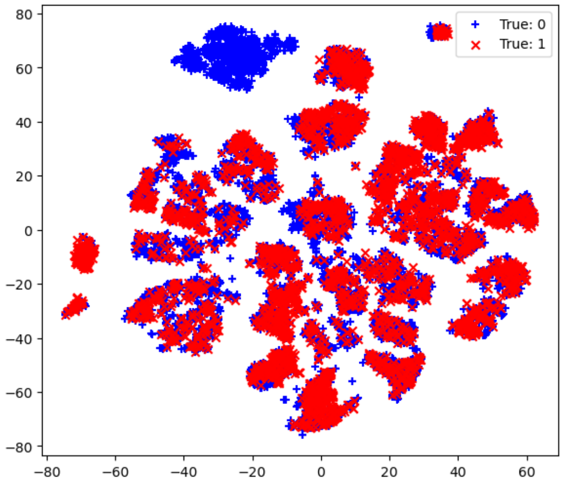

Figure 4: t-SNE analysis with perplexity = 100.

5.4 *Experimental process*

The data splitting process was used as a preliminary stage to separate the dataset into input components enclosing static features and output elements corresponding to the target variable. Furthermore, 80% of the data was allocated to the training set, while 20% was used for the testing set that evaluated their performance. Additionally, the thesis used feature selection to find the features that improve the model's performance and discard the irrelevant ones, as well as cross-validation employed for the hyperparameter tuning. Finally, the three XAI techniques, which assess and explain the significance of the features, are employed and evaluated.

5.5 *Algorithms and software*

For this experiment, the thesis used the programming language of Python 3.12.3 (Van Rossum & Drake, 2009). The coding activities in this study were conducted in Anaconda Navigator ("Anaconda Software Distribution", 2020). The packages Pandas (pandas development team, 2024), and Numpy (Harris et al., 2020) were used to manipulate the dataset. Furthermore, the thesis utilized the scikit-learn library for data transformation and implementation of machine learning algorithms, including permutation feature importance (Pedregosa et al., 2011) and XGBoost (T. Chen & Guestrin, 2016). As for Shapley additive explanations and Local Interpretable Model-Agnostic Explanations, the SHAP library (Lundberg & Lee, 2017) and the LIME library (Ribeiro et al., 2016) were used. Moreover, the thesis used Matplotlib (Hunter, 2007) and Seaborn (Waskom, 2021) for visualization.

6 RESULTS

6.1 *Models Comparison*

In the classification results section of this thesis, we assessed three ML models. Logistic Regression served as the baseline model, while Random Forest and XGBoost were analyzed for their perofrmance improvements over the baseline.

Initially, Logistic Regression exhibited moderate performance metrics: an accuracy score of 0.6585, precision of 0.8546, recall of 0.6370, and F1 score of 0.7299. These metrics indicated a higher precision due to fewer false positive predictions but revealed lesser effectiveness across other metrics compared to the subsequent models. Random Forest demonstrated superior performance with an accuracy score of 0.7666, precision of 0.8068, recall of 0.8913, and F1 score of 0.8469. XGBoost further enhanced these results with an accuracy score of 0.7834, precision of 0.8063, recall of 0.9226, and F1 score of 0.8605. These models showed competitive performance, particularly XGBoost, which excelled in accuracy and recall while maintaining a balanced precision and F1 score. Table 4 provides a list with the results

Table 4: Basic Evaluation Results

Models	**Accuracy**	**Precision**	**Recall**	**F1-score**
Logistic regression	0.6585	0.8546	0.6370	0.7299
Random Forest	0.7666	0.8063	0.8913	0.8469
XGB	0.7834	0.8063	0.9226	0.8605

After the first evaluation of the three models, the thesis performed a hyperparameter tuning in each model. Then, the thesis retrained the models based on each model's best hyperparameters and re-evaluated them. The new evaluation results were quite similar for Random Forest and Extreme Gradient Boost but significantly better for Logistic Regression. Specifically, this time, Random Forest emerged as the most accurate model with an improved accuracy score of 0.7844, precision of 0.8012, recall of 0.9341, and F1 score of 0.8626. XGBoost followed closely with an accuracy score of 0.7811, and precision, recall, and F1 score of 0.8078, 0.9156, and 0.8583, respectively. Surprisingly, Logistic Regression, after hyperaparameter tuning, showed significant improvement with an accuracy score of 0.7510, precision of 0.7623, recall of 0.9534, and F1 score of 0.8473. Although still less accurate than the ensemble models, its recall score notably increased, making it a more competitive option post-tuning. The results can be found in Table 5.

Table 5: Final Basic Evaluation Results

Models	Accuracy	Precision	Recall	F1-score
Logistic regression	0.7510	0.7623	0.9534	0.8473
Random Forest	0.7844	0.8012	0.9341	0.8626
XGB	0.7811	0.8078	0.9156	0.8583

In conclusion, while LR model initially achieved high precision, the RF and the XGB models prove to be superior overall after both initial evaluation and hyperparameter tuning. Random Forest excelled in accuracy and recall, while XGBoost maintained strong performance across all metrics.

6.2 *Feature Importance Score Results*

The feature-important results play a significant role in this thesis, as they widely explain each feature's role in the overall performance of machine learning models. Explainable artificial intelligence (XAI) techniques including SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), and permutation feature importance were emplloyed across all machine learning models used for classification. The thesis printed the five most influential predictors from each technique for the three models. Additionally, the thesis employed Jaccard similarity and Spearman correlation techniques to better understand the consistency and agreement of the XAI methods.

6.2.1 *Logistic Regression*

Table 7 presents the results of the SHAP analysis for the Logistic Regression model. The analysis identified required_car_parking_spaces as the most influential feature with a SHAP score of 3.2094, followed by total_of_special_requests with a score of 0.3175. Additionally, features such as lead_time (0.2985), adr (0.2496), and previous_bookings_not_canceled (0.1834) also demonstrated notable influence.

Table 6: Top 5 features in SHAP for LR model

Feature	Importance
required_car_parking_spaces	3.2094
total_of_special_requests	0.3175
lead_time	0.2985
adr	0.2496
previous_bookings_not_canceled	0.1834

Table 7 presents the results of the LIME analysis, which produced less extreme scores compared to SHAP. The most influential feature according to LIME was previous_cancellations with a score of 0.5450, followed by required_car_parking_spaces (0.1782), and previous_bookings_not_canceled (score = 0.1415). The rest of the features had scores below 0.1.

Table 7: Top 5 features in LIME for LR model

Feature	Importance
previous_cancellations	0.5450
required_car_parking_spaces	0.1782
previous_bookings_not_canceled	0.1415
lead_time	0.0581
booking_changes	0.0543

Permutation importance revealed even lower importance results compared to SHAP and LIME. Six of the ten most important features had scores more than 0.0. Notably, total_of_special_requests emerged as the most inluential feature with a score of 0.0317, followed by required_car_parking_spaces (0.0215), and lead_time (0.0213). Table 8 provide a comprehensive list of the top features ranked by permutation importance.

Table 8: Top 5 features in permutation importance for LR model

Feature	Importance
total_of_special_requests	0.0317
required_car_parking_spaces	0.0215
lead_time	0.0213
adr	0.0203
booking_changes	0.0113

To assess consistency across XAI techniques, Jaccard similarity and Spearman correlation were computed. The results indicated that SHAP and permutation importance exhibited the highest similarity with a Jaccard score of 0.82 and Spearman correlation of 0.94. In contrast, the relationships between SHAP and LIME achieved lower agreement with scores of 0.46 in Jaccard and 0.45 in Spearman. Similarly, LIME and permutation feature importance showed moderate agreement with scores of 0.46 in Jaccard similarity and 0.34 in Spearman correlation. Table 10 and 11 present the detailed results.

Table 9: Jaccard similarity for LR model

Similarity	**Score**
SHAP and LIME	0.46
SHAP and perm. importance	0.82
LIME and perm. importance	0.46

Table 10: Spearman correlation for LR model

Similarity	**Score**
SHAP and LIME	0.45
SHAP and perm. importance	0.94
LIME and perm. importance	0.34

The feature importance analysis utilizing SHAP, LIME, and permutation importance techniques provide valuable insights into the LR model's predictive mechanisms. The feature required_car_parking_spaces consistently emerged as highly influential across all methods, highlighting its critical role predicting outcomes. The high agreement between SHAP and permutation feature importance underscores their robustness in identifying feature importance, while LIME interpretations showed some separation.

6.2.2 *Random Forest*

The feature importance results of the Random Forest model, which achieved the best overall evaluation results, are slightly different. The top five features identified by SHAP analysis were total_of_special_requests, lead_time, adr, required_car_parking_spaces, and booking_changes. The most influential one being total_of_special_requests with a score of 0.0778. Table 11 presents the complete list.

Table 11: Top 5 features in SHAP for RF model

Feature	Importance
total_of_special_requests	0.0778
lead_time	0.0767
adr	0.0558
required_car_parking_spaces	0.0528
booking_changes	0.0333

Table 12 presents the LIME analysis results, which revealed minimal differences among the features. The most influential feature was previous_bookings_not_canceled with a score of 0.0688. The other significant features were lead_time (0.0604), previous_cancellations (0.0592), booking_changes (0.0394), and total_of_special_requests (0.0377).

Table 12: Top 5 features in LIME for RF model

Feature	Importance
previous_bookings_not_canceled	0.0688
lead_time	0.0604
previous_cancellations	0.0592
booking_changes	0.0394
total_of_special_requests	0.0377

Table 13 presents the permutation feature importance results for the Random Forest model. Similar to Logistic Regression, total_of_special_requests was the most influential feature with a score of 0.0457. The other influential variables were lead_time (score = 0.0435), adr (score = 0.036), arrival_date_year (score = 0.0183), and previous_cancellations (score = 0.0168).

The agreement between the three methods was generally high, as revealed by Jaccard similarity and Spearman correlation. SHAP and permutation importance had the highest scores, with 0.81 in Jaccard and 0.80 in Spearman. The similarity between SHAP and LIME, as well as between

Table 13: Top 5 features in permutation importance for RF model

Feature	Importance
total_of_special_requests	0.0457
lead_time	0.0435
adr	0.0366
arrival_date_year	0.0183
previous_cancellations	0.0168

LIME and permutation importance, 0.66 Jaccard. However the Spearman correlations were low, with SHAP and LIME at 0.13 and LIME with permutation importance at 0.07. The results are presented in tables 14 and 15.

Table 14: Jaccard similarity for RF model

Similarity	Score
SHAP and LIME	0.67
SHAP and perm. importance	0.8
LIME and perm. importance	0.67

Table 15: Spearman correlation for RF model

Similarity	Score
SHAP and LIME	0.14
SHAP and perm. importance	0.81
LIME and perm. importance	0.08

The feature importance analysis of SHAP, LIME, and permutation feature importance for the RF model, revealed some crucial insights into the model's decision-making process. The consistent identification of feature total_of_special_requests as the most influential feature across all methods underscores its pivotal role in the model's predictive capabilities. The high level of agreement between SHAP and permutation importance techniques demonstrates their reliability in feature ranking. Meanwhile, LIME's different perspective highlights the values of using multiple XAI techniques to capture a more comprehensive picture of feature significance.

6.2.3 *Extreme Gradient Boost*

The Extreme Gradient Boosting method, the second best-performing model for the evaluation metrics, provided interesting results on feature importance. In SHAP analysis, similarly to Logistic Regression, required_car_parking_spaces emerged as the most influential feature, significantly outsourcing other features with a score of 1.4485. In comparison, lead_time scored 0.4796, total_of_special_requests scored 0.3836, adr scored 0.3080, and booking_changes had a score of 0.2237 (Table 16).

Table 16: Top 5 features in SHAP for XGB model

Feature	**Importance**
required_car_parking_spaces	1.4485
lead_time	0.4796
total_of_special_requests	0.3836
adr	0.3080
booking_changes	0.2237

The LIME analysis shown in Table 17, indicated that previous_cancellations was the most influential feature with a score of 0.4777. This was followed by required_car_parking_spaces (0.2783) and previous_bookings_not_canceled (0.2007). Other notable features where lead_time (0.1986) and days_in_waiting_list (0.1755).

Table 17: Top 5 features in LIME for XGB model

Feature	**Importance**
previous_cancellations	0.4777
required_car_parking_spaces	0.2783
previous_bookings_not_canceled	0.2007
lead_time	0.1986
days_in_waiting_list	0.1755

Similar to Logistic Regression and Random Forest, permutation importance analysis highlighted total_of_special requests as the feature most impacting the model's results, with a score of 0.0484. This score was closely followed by lead_time (0.0423) and adr (0.041). Detailed results can be found in Table 18.

Table 18: Top 5 features in permutation importance for XGB model

Feature	**Importance**
total_of_special_requests	0.0484
lead_time	0.0423
adr	0.0411
previous_cancellations	0.0163
booking_changes	0.0127

High agreement was observed among the three techniques based on Jaccard similarity analysis. SHAP and permutation importance showed perfect agreement with a score of 1, indicating identical influential features. Both methods had a Jaccard score of 0.5 with LIME. In terms of Spearman correlation, SHAP and permutation importance exhibited a strong correlation of 0.85, while LIME and permutation importance had a correlation of 0.12. The correlation between SHAP and LIME was 0.29. Detailed results are provided in Tables 29 and 20.

Table 19: Jaccard similarity for XGB model

Similarity	**Score**
SHAP and LIME	0.5
SHAP and perm. importance	1
LIME and perm. importance	0.5

Table 20: Spearman correlation for XGB model

Similarity	**Score**
SHAP and LIME	0.29
SHAP and perm. importance	0.85
LIME and perm. importance	0.13

The feature importance analysis for the XGB model highlighted feature required_car_parking_spaces as a consistently influential feature, particularly in the SHAP and LIME analysis. Additionally, the high agreement between the SHAP and permutation importance and the distinct perspective of the LIME analysis, similarly to the analysis of the other two models, emphasize the necessity of employing multiple XAI techniques to achieve a comprehensive understanding of model behaviour.

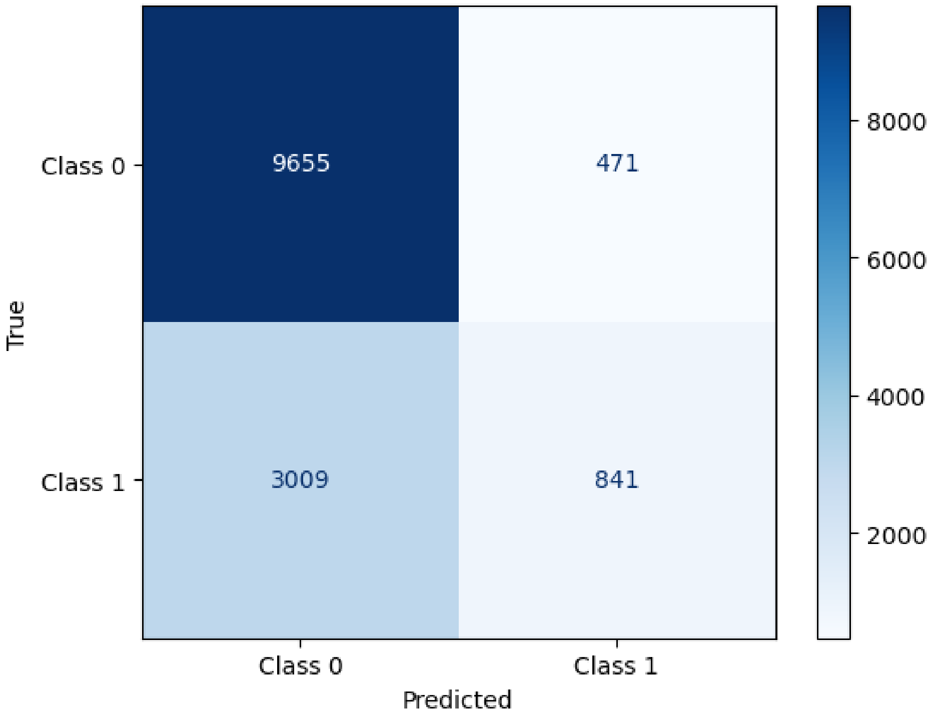

Figure 5: Confusion Matrix for Logistic Regression.

6.3 *Error Analysis*

Error analysis was conducted for the three machine learning models to gain a deeper understanding of their performance and identify reasons for incorrect predictions. By analyzing the confusion matrices and examining specific instances where miss-classifications occurred, valuable insights were gathered about each model's strengths and weaknesses.

The confusion matrix of the LR model revealed that out of 13,976 instances, the model correctly classified 9,026 instances as class 0 (True Negatives) and 1,689 instances as class 1 (True Positives). However, the model misclassified 3,480 instances. This resulted in 471 False Positives, where the model predicted class 1 but the label was class 0, and 3,009 False Negatives, where the model predicted class 0 but the true label was class 1. By examining the misclassified instances, notable deviations were observed in certain features. For example, in instance 61086, the features lead_time, stays_in_weekend_nights, booking_changes, and adr deviated by 1.13, -0.97, -0.37, and 0.27, respectively, from the mean values.

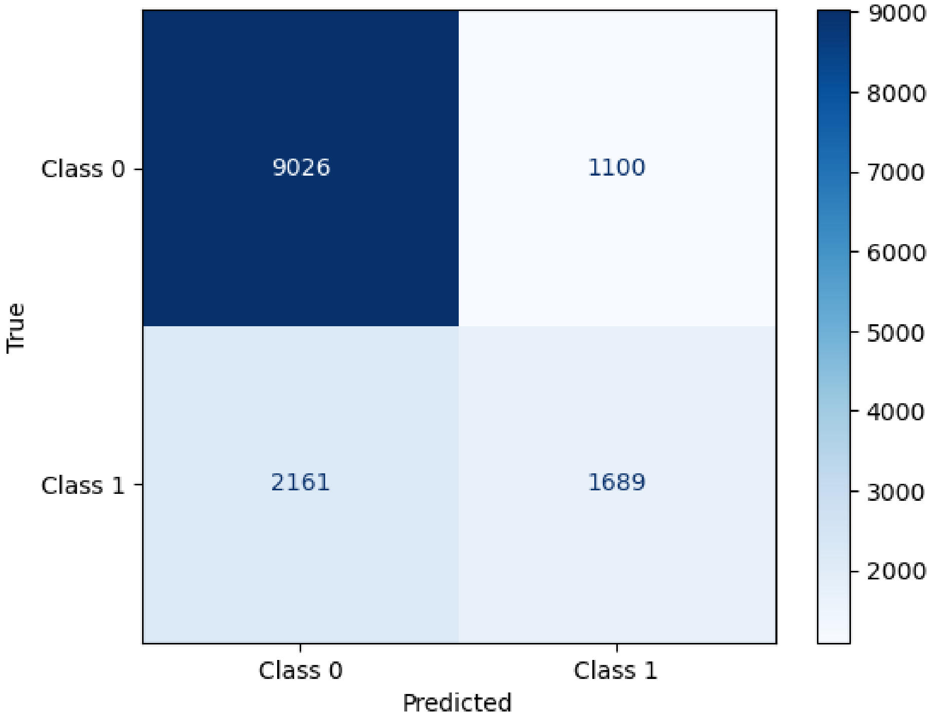

Figure 6: Confusion Matrix for Random Forest.

Figure 6 provides notable findings in error analysis for Random Forest. The confusion matrix showed 9,026 instances correctly classified as class 0 (TN) and 1,689 as class 1 (TP). However, it misclassified 1,110 instances as class 1 when they were class 0 (FP), and 2,165 instances as class 0 when they were class 1 (FN). Further analysis highlighted significant deviations in certain instances. For example, in instances 67292 and 36342, the feature lead_time showed deviations of -0.638103 and 2.702550, respectively.

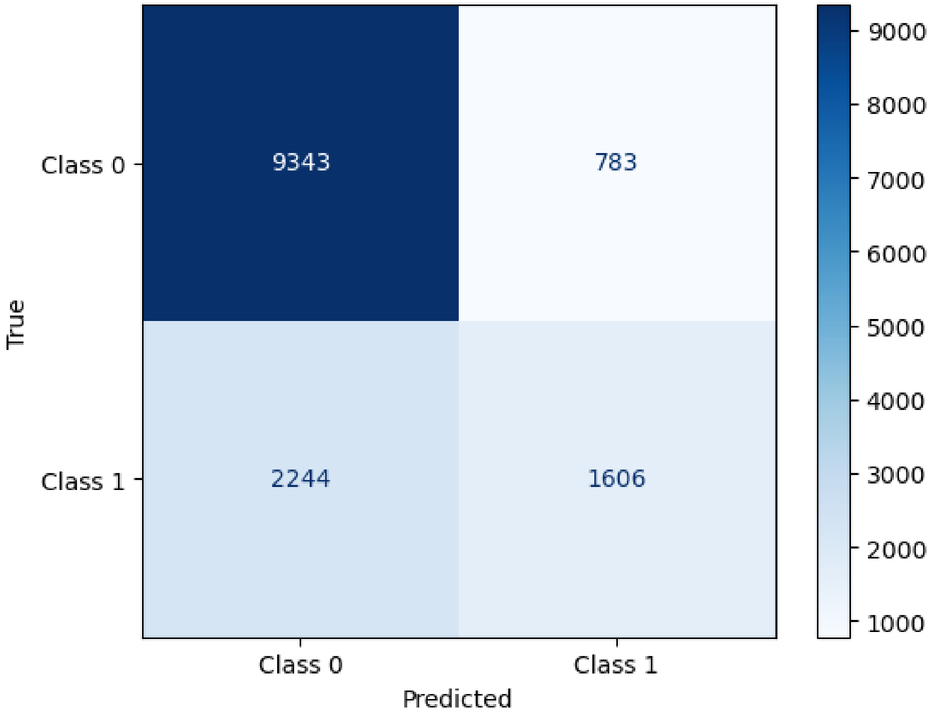

Figure 7: Confusion Matrix for XGBoost.

Figure 7 presents error analysis results for the XGB model. The model correctly predicted 10,949 instances, including 9,343 TN and 1,606 TP. It misclassified 783 instances as class 1 when they were actually class 0 (FP) and 2.244 instances as classo where they were actually class 1 (FN) indicating better performance compared to the other models. Significant deviations were observed in misclassified samples, notably in instance 50432 where arrival_date_year had a deviation of 1..152019 and stays_in_weekend_nights had a deviation of -973501.

The error analysis across the three models revealed both similarities and differences in their predictive performance, Logistic Regression showed a higher number of False Negatives, indicating challenges in identifying class 1 instances correctly. Random Forest, while improving in some aspects, still exhibited considerable misclassifications with notable deviations. Additionally, the XGB model achieved the best overall performance, with fewer misclassifications and slightly better handling of feature deviations.

7 DISCUSSION

This thesis section is about the results and responses to the research question. It includes an in-depth evaluation of the findings, the limitations of the research, and some proposals for additional future research in the field.

7.1 *Thesis goal*

The main goal of the current thesis is to find additional solutions to cancellation complications for the hotel industry by examining three different Machine Learning models and evaluating their performances. Additionally, the project examines and measures the influence of various features on each model, and through that, it attempts to justify their performances. The thesis introduces the eXplainable Artificial Intelligence methods in the hotel industry to measure the importance of features.

7.2 *Analysis of study findings*

The primary research question in the current thesis is: How accurate are the ML models in predicting hotel room cancellations, and what are the main cancellation drivers? The thesis constructed and evaluated three different ML models to answer the main question. Logistic regression was the baseline model, while Random Forest and Extreme Gradient Boost were also evaluated. The thesis divided the dataset into a training and a testing set and evaluated the three models to obtain their predicting results. The first evaluation found XGB as the best predictive model with an accuracy score of 0.7834. Afterwards, the thesis conducted a hyperparameter tuning process for each model to optimize their performances. Post-tuning, Random Forest showed a slight improvement over XGB emerging as the most accurate model overall.

The thesis further delves into the main research question through the two additional sub-questions. The first sub-question is: What are the predictive features used by the ML models, and can different explainable artificial intelligence techniques provide explanations for these features? To address, the thesis employed three eXplainable AI techniques. SHAP, LIME, and permutation importance were applied across all models, and the results were compared. lead_time was found to be the most influential among the dataset variables, as it appears as an essential factor in every eXplainable AI method in all three models. Additionally, total_of_special_requests was consistently highlighted as an important feature by all three XAI methods in the best performing model. Subsequently, the thesis attempts to answer the final sub-question: How do the findings of diverse explainable artificial intelligence techniques contrast? To answer this, the thesis employed Jaccard similarity and Spearman correlation to find the agreement of the XAI techniques. Jaccard similarity showed high agreement between SHAP and permutation, with scores of 0.81 for both the LR and RF models, and a perfect score of 1 for the XGB models. Similarly, the Spearman correlation test revealed the highest correlation between the SHAP and permutation test, with a correlation score of 0.94 for the LR model, 0.80 for the RF model, and 0.85 for the XGB model. In contrast, LIME's similarity and correlation with SHAP and permutation importance analysis were significantly lower.

Overall, the analysis in this thesis underscores the robustness in predicting hotel room cancellations of the Random Forest model. Furthermore, the analysis shows the critical role of lead_time and total_of_special_requests in driving these predictions. The comparison of XAI techniques also provides valuable insights into their relative effectiveness in features explanation, with SHAP analysis and permutation feature importance showing better agreement and reliability.

7.3 *Comparison with Existing Literature*

This study's results show that advanced models like Random Forest and XGBoost outperform traditional models like Logistic Regression in predicting hotel booking cancellations. These results, for example, achieved accuracy of RF and XGB, 0.7844 and 0.7811 respectively, in line with the existing literature and the findings by Andriawan et al. (2020) and Antonio, de Almeida, and Nunes (2019a), who reported high accuracy and effectiveness of tree-based models in cancellation prediction (RF accuracy score of 87% and XGB accuracy score of 84%, respectively). Additionally, the results show that treating a hotel booking cancellation as a classification problem provides effective outcomes, and supports the transition from regression to classification approaches as pointed out by studies like Antonio, de Almeida, and Nunes (2019a) and Antonio et al. (2017) and others. The results of the eXplainable Artificial Intelligence techniques revealed the importance of lead_time across various models in this study and reflect the findings of the existing literature, where this feature was highlighted as important predictor of cancellation by Andriawan et al. (2020) and Gartvall and Skånhagen (2022). Furthermore, the discovery of required_car_parking_spaces as an important feature in both Random Forest and XGBoost add new insights to the existing literature, suggesting additional variables that might be considered in future studies.

The use of the XAI techniques in this thesis provided deeper understanding of model behavior and it was aligned with the need for model explainability in complex prediction tasks, highlighted by Gilpin et al.

(2018). Furthermore, employing agreement metrics of Jaccard similarity and Spearman correlation to compare the consistency of the different XAI methods demonstrated strong validation approaches. This application provides new insights to the literature, as these metrics have mainly been used in other fields, as Rebane et al. (2021) and Krishna et al. (2022) highlighted. The use and comparison of different XAI methods in this study addresses the identified gap in the literature regarding the use of XAI techniques in the hospitality industry for hotel booking cancellation prediction. Additionally, this approach contributes to the broader field by indicating the applicability and effectiveness of the XAI techniques beyond healthcare and finance, as suggested by Bhargava and Gupta (2022).

7.4 *Limitations and Future Work*

The current thesis addresses the possibility of predicting hotel booking cancellation by employing machine learning models and the importance of individual features on each model's performance. The thesis encountered limitations that should be considered when interpreting the results. A considerable missing aspect from the current dataset was weather data. Variables such as precipitation and temperature could potentially improve model performance as weather conditions often influence travel decisions. Future work could incorporate weather data, focusing on forecasting booking cancellations based on weather patterns close to the booking date.

Even though the thesis used a dataset that contained time-related information for each reservation and cancellation, it did not fully explore these aspects. Treating the data from a time series perspective could provide valuable insights. Future research could evaluate the performance of ML models in predicting cancellations over different time periods, such as cancellations occurring shortly after booking versus those closer to arrival date. Additionally, exploring how dynamic pricing changes over time might influence cancellation rates would be beneficial, given that price sensitivity is a known factor in decision making by customers. Moreover, future works could investigate the booking sources, such as booking through hotel's websites or through third-party websites, and find if they could potentially influence the cancellation rates.

Moreoever, future works can explore alternative ML models such as Support Vector Machines and CatBoost to different datasets, comparing their performance and analyzing their feature importance. While the thesis evaluates the prediction of the machine learning models using metrics such as accuracy, precision, recall, and F1-score, future work could include additional metrics like G-Mean (Geometric Mean), Macro-Average, and AUC-score (Area Under the Curve) for a more comprehensive evaluation.

Additionally, instead of solely using Jaccard similarity and Spearman correlation to assess the decision-making process of the XAI models, future studies could employ alternative metrics to evaluate and better understand the XAI methods.

8 CONCLUSION

The current thesis built three Machine Learning methods to predict hotel booking cancellations. The thesis constructed logistic regression, Random Forest, and Extreme Gradient Boost based on their top hyperparameter and evaluated using accuracy, precision, recall, and F1 metrics. Furthermore, the thesis employed three different XAI methods. SHAP, LIME, and permutation importance were computed to assess each model's performance. Additionally, the thesis compared the results of these methods by employing Jaccard similarity and Spearman correlation.

The results found that the RF model is slightly more accurate than the XGB model, with an accuracy score of 0.784, compared to XGB's 0.781. The Logistic regression, even though it had a higher recall score of 0.95, had a lower accuracy score of 0.751. Hence, it is less accurate. Furthermore, the employment of the eXplainable AI methods revealed that lead_time is the most influential feature in the performance of each model. Additionally, using Jaccard's and Spearman's tests, the thesis found that the relationship between SHAP and permutation tests is the most similar and correlated in each model.

Even though the thesis attempted to find the most accurate ML model between LR, RF, and XGB, it only evaluated their performance based on accuracy, precision, recall, and F1 metrics. Future works could enhance additional metrics for evaluating the performance of the models and the performance of the XAI methods. Additionally, it could be beneficial if future studies attempt to predict cancellations based on weather and time data.

REFERENCES

Abbott, D. (2014). *Applied predictive analytics: Principles and techniques for the professional data analyst*. John Wiley & Sons.

Adler, A. I., & Painsky, A. (2022). Feature importance in gradient boosting trees with cross-validation feature selection. *Entropy*, *24*(5), 687.

Alibrahim, H., & Ludwig, S. A. (2021). Hyperparameter optimization: Comparing genetic algorithm against grid search and bayesian optimization. *2021 IEEE Congress on Evolutionary Computation (CEC)*, 1551–1559.

Anaconda software distribution. (2020). https://docs.anaconda.com/

Andriawan, Z. A., Purnama, S. R., Darmawan, A. S., Wibowo, A., Sugiharto, A., Wijayanto, F., et al. (2020). Prediction of hotel booking cancellation using crisp-dm. *2020 4th International Conference on Informatics and Computational Sciences (ICICoS)*, 1–6.

Antonio, N., De Almeida, A., & Nunes, L. (2017). Predicting hotel booking cancellations to decrease uncertainty and increase revenue. *Tourism & Management Studies*, *13*(2), 25–39.

Antonio, N., De Almeida, A., & Nunes, L. (2019). Big data in hotel revenue management: Exploring cancellation drivers to gain insights into booking cancellation behavior. *Cornell Hospitality Quarterly*, *60*(4), 298–319.

Antonio, N., de Almeida, A., & Nunes, L. (2019a). An automated machine learning based decision support system to predict hotel booking cancellations. *Data Science Journal*, *18*, 32–32.

Antonio, N., de Almeida, A., & Nunes, L. (2019b). Hotel booking demand datasets. *Data in brief*, *22*, 41–49.

Archer, B., et al. (1987). Demand forecasting and estimation. *Demand forecasting and estimation.*, 77–85.

Barreñada, L., Dhiman, P., Timmerman, D., Boulesteix, A.-L., & Van Calster, B. (2024). Understanding random forests and overfitting: A visualization and simulation study. *arXiv preprint arXiv:2402.18612*.

Bhargava, D., & Gupta, L. K. (2022). Explainable ai in neural networks using shapley values. In *Biomedical data analysis and processing using explainable (xai) and responsive artificial intelligence (rai)* (pp. 59–72). Springer.

Biecek, P. (2018). Dalex: Explainers for complex predictive models in r. *Journal of Machine Learning Research*, *19*(84), 1–5.

Biecek, P., Chlebus, M., Gajda, J., Gosiewska, A., Kozak, A., Ogonowski, D., Sztachelski, J., & Wojewnik, P. (2021). Enabling machine learning algorithms for credit scoring–explainable artificial intelligence (xai) methods for clear understanding complex predictive models. *arXiv preprint arXiv:2104.06735*.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Chen, C.-C., Schwartz, Z., & Vargas, P. (2011). The search for the best deal: How hotel cancellation policies affect the search and booking decisions of deal-seeking customers. *International Journal of Hospitality Management*, *30*(1), 129–135.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., et al. (2015). Xgboost: Extreme gradient boosting. *R package version 0.4-2*, *1*(4), 1–4.

Chen, Y., Ding, C., Ye, H., & Zhou, Y. (2022). Comparison and analysis of machine learning models to predict hotel booking cancellation. *2022 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022)*, 1363–1370.

Claveria, O., Monte, E., & Torra, S. (2015). A new forecasting approach for the hospitality industry. *International Journal of Contemporary Hospitality Management*, *27*(7), 1520–1538.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *20*(2), 215–232.

Dan, T., Li, Y., Zhu, Z., Chen, X., Quan, W., Hu, Y., Tao, G., Zhu, L., Zhu, J., Jin, Y., et al. (2020). Machine learning to predict icu admission, icu mortality and survivors' length of stay among covid-19 patients: Toward optimal allocation of icu resources. *2020 IEEE international conference on bioinformatics and biomedicine (BIBM)*, 555–561.

Duell, J., Seisenberger, M., Aarts, G., Zhou, S., & Fan, X. (2021). Towards a shapley value graph framework for medical peer-influence. *arXiv preprint arXiv:2112.14624*.

Elkhawaga, G., Elzeki, O., Abuelkheir, M., & Reichert, M. (2023). Evaluating explainable artificial intelligence methods based on feature elimination: A functionality-grounded approach. *Electronics*, *12*(7), 1670.

ElShawi, R., Sherif, Y., Al-Mallah, M., & Sakr, S. (2021). Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Computational Intelligence*, *37*(4), 1633–1650.

Falk, M., & Vieru, M. (2018). Modelling the cancellation behaviour of hotel guests. *International Journal of Contemporary Hospitality Management*, *30*(10), 3100–3116.

Gartvall, E., & Skånhagen, O. (2022). Predicting hotel cancellations using machine learning.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, 80–89.

Guleria, P., Naga Srinivasu, P., Ahmed, S., Almusallam, N., & Alarfaj, F. K. (2022). Xai framework for cardiovascular disease prediction using classification techniques. *Electronics*, *11*(24), 4086.

Hao, F., Xiao, Q., & Chon, K. (2020). Covid-19 and china's hotel industry: Impacts, a disaster management framework, and post-pandemic agenda. *International journal of hospitality management*, *90*, 102636.

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., et al. (2020). Array programming with numpy. *Nature*, *585*(7825), 357–362.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

Huang, L., Li, C., & Zheng, W. (2022). Daily hotel demand forecasting with spatiotemporal features. *International Journal of Contemporary Hospitality Management*, *35*(1), 26–45.

Huang, L., & Zheng, W. (2021). Novel deep learning approach for forecasting daily hotel demand with agglomeration effect. *International Journal of Hospitality Management*, *98*, 103038.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in science & engineering*, *9*(03), 90–95.

Kırtıl, İ. G., & Aşkun, V. (2021). Artificial intelligence in tourism: A review and bibliometrics research. *Advances in Hospitality and Tourism Research (AHTR)*, *9*(1), 205–233.

Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., & Lakkaraju, H. (2022). The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*.

LaValley, M. P. (2008). Logistic regression. *Circulation*, *117*(18), 2395–2399.

Leslie, D. (2019). Understanding artificial intelligence ethics and safety. *arXiv preprint arXiv:1906.05684*.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, *2*(1), 56–67.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Morales, D. R., & Wang, J. (2010). Forecasting cancellation rates for services booking revenue management using data mining. *European Journal of Operational Research*, *202*(2), 554–562.

pandas development team, T. (2024, April). *Pandas-dev/pandas: Pandas (v2.2.2)* (Version 2.2.2). Zenodo. https://doi.org/10.5281/zenodo.10957263

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, *12*, 2825–2830.

Peres, I. T., Hamacher, S., Oliveira, F. L. C., Thomé, A. M. T., & Bozza, F. A. (2020). What factors predict length of stay in the intensive care unit? systematic review and meta-analysis. *Journal of Critical Care*, *60*, 183–194.

Qiu, R. T., Liu, A., Stienmetz, J. L., & Yu, Y. (2021). Timing matters: Crisis severity and occupancy rate forecasts in social unrest periods. *International journal of contemporary hospitality management*, *33*(6), 2044–2064.

Rebane, J., Samsten, I., Pantelidis, P., & Papapetrou, P. (2021). Assessing the clinical validity of attention-based and shap temporal explanations for adverse drug event predictions. *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, 235–240.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Rigatti, S. J. (2017). Random forest. *Journal of Insurance Medicine*, *47*(1), 31–39.

Sánchez-Medina, A. J., Eleazar, C., et al. (2020). Using machine learning and big data for efficient forecasting of hotel booking cancellations. *International Journal of Hospitality Management*, *89*, 102546.

Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham L. Rev.*, *87*, 1085.

Sierag, D. D., Koole, G., van der Mei, R. D., Van der Rest, J., & Zwart, B. (2015). Revenue management under customer choice behaviour with cancellations and overbooking. *European journal of operational research*, *246*(1), 170–185.

Singh, H., Roy, A., Setia, R., & Pateriya, B. (2022). Estimation of nitrogen content in wheat from proximal hyperspectral data using machine learning and explainable artificial intelligence (xai) approach. *Modeling Earth Systems and Environment*, *8*(2), 2505–2511.

Sirkin, J., & Hughes, A. (2017). Microsoft sql server management studio (ssms). *TechTarget, febrero*.

Tang, C. M. F., King, B. E., & Kulendran, N. (2015). Estimating future room occupancy fluctuations to optimize hotel revenues. *Journal of Travel & Tourism Marketing*, *32*(7), 870–885.

Tsang, W. K., & Benoit, D. F. (2020). Gaussian processes for daily demand prediction in tourism planning. *Journal of Forecasting*, *39*(3), 551–568.

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of clinical epidemiology*, *49*(11), 1225–1231.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(11).

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.

Veiga, P. M., Ambrósio, F., & Ferreira, R. R. (2020). Competitiveness of the hotel industry: A knowledge management approach. In *Multilevel approach to competitiveness in the global tourism industry* (pp. 9–25). IGI Global.

Wade, C., & Glynn, K. (2020). *Hands-on gradient boosting with xgboost and scikit-learn: Perform accessible machine learning and extreme gradient boosting with python*. Packt Publishing Ltd.

Wandner, S. A., & Erden, J. v. (1981). Estimating the demand for international tourism using time series analysis.

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021.

Webb, T., Schwartz, Z., Xiang, Z., & Singal, M. (2020). Revenue management forecasting: The resiliency of advanced booking methods given dynamic booking windows. *International Journal of Hospitality Management*, *89*, 102590.

Yoo, M. M., & Yang, S. (2021). Forecasting demand. In *Operations management in the hospitality industry* (pp. 71–94). Emerald Publishing Limited.

Yoo, M., Singh, A. K., & Loewy, N. (2024). Predicting hotel booking cancelation with machine learning techniques. *Journal of Hospitality and Tourism Technology*, *15*(1), 54–69.

Zheng, T., Liu, S., Chen, Z., Qiao, Y., & Law, R. (2020). Forecasting daily room rates on the basis of an lstm model in difficult times of hong kong: Evidence from online distribution channels on the hotel industry. *Sustainability*, *12*(18), 7334.

Zhu, M., Wu, J., & Wang, Y.-G. (2021). Multi-horizon accommodation demand forecasting: A new zealand case study. *International Journal of Tourism Research*, *23*(3), 442–453.

Zien, A., Krämer, N., Sonnenburg, S., & Rätsch, G. (2009). The feature importance ranking measure. *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part II 20*, 694–709.

APPENDIX A

Dataset Feature Description

Feature	Description
adr	Average Daily Rates
adults	Number of adults
agent	Travel agency ID
arrival_date_day_of_month	Arrival's date day of the month
arrival_date_month	Arrival's date month
arrival_date_week_number	Arrival's date week number
arrival_date_year	Arrival's date year
assigned_room_type	Code of the type of the room assigned to each customer
babies	Number of babies
booking_changes	Number of changes made to each booking until checkout or cancellation
children	Number of children
company	ID of the company from where the reservation was made
country	Country of origin
customer_type	Type of booking
days_in_waiting_list	Number of days the booking was in waiting list until it was confirmed
deposit_type	Indication on if the customer provided a deposit
distribution_channel	Booking distribution channel
is_canceled	Value (0 or 1) of if the booking was canceled or not
is_repeated_guest	Value (0 or 1) of if the customer was a repeated guest
lead_time	Number of days elapsed between the entering of the booking into the PMS and the arrival date
market_segment	Market segment designation
meal	Type of meal booked
previous_bookings_not_canceled	Number of previous bookings not canceled by the same customer
previous_cancellations	Number of previous bookings canceled by the same customer
required_car_parking_spaces	Number of car parking spaces required by the customer
reservation_status	Last status of the reservation
reservation_status_date	Date at which the last status was set
reserved_room_type	Code of the type of the room reserved by each customer
stays_in_weekend_nights	Number of weekend nights a guest booked to stay at the hotel
stays_in_week_nights	Number of week nights a guest booked to stay at the hotel
total_of_special_requests	Number of special requests made by the customer

APPENDIX B

Missing Values in the Dataset

Feature	**Missing Values**
children	4
country	488
agent	16340
company	112593

APPENDIX C

Table C1: Top VIF Scores

Feature	VIF
stays_in_week_nights	1.539210
stays_in_weekend_nights	1.469007
previous_bookings_not_canceled	1.253772
adr	1.242359
previous_cancellations	1.227099
lead_time	1.196476

Figure C2: Correlation Matrix

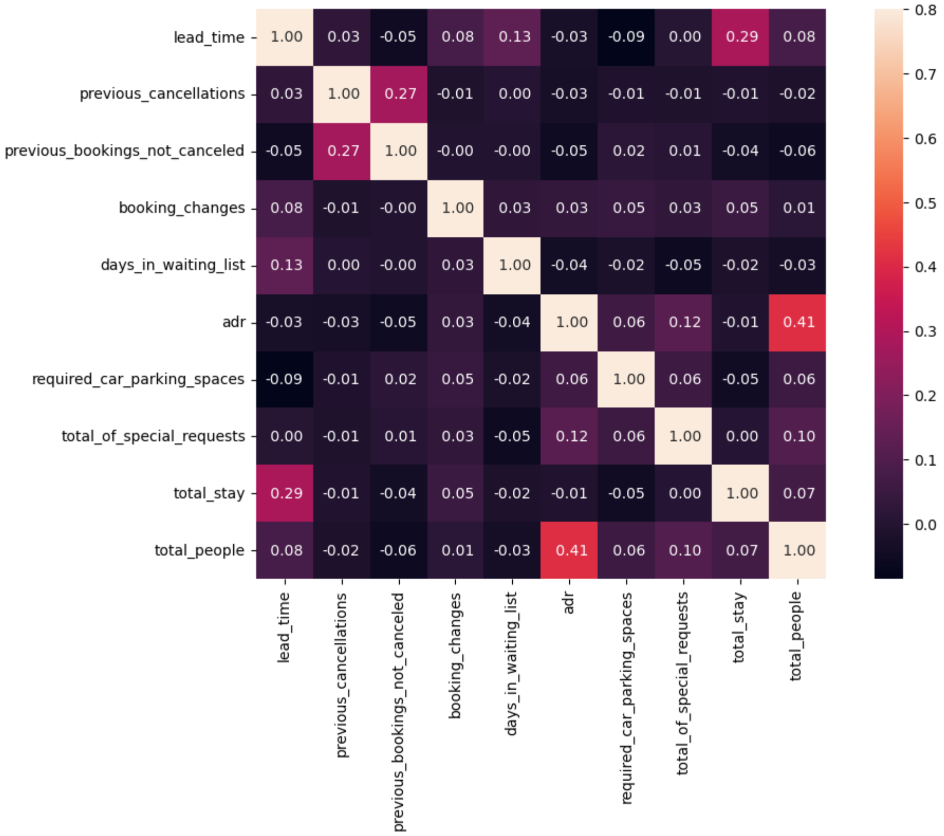

Figure 8: Correlation Matrix.

Figure C3: Correlation Between ADR, Reserved and Assigned Room Type

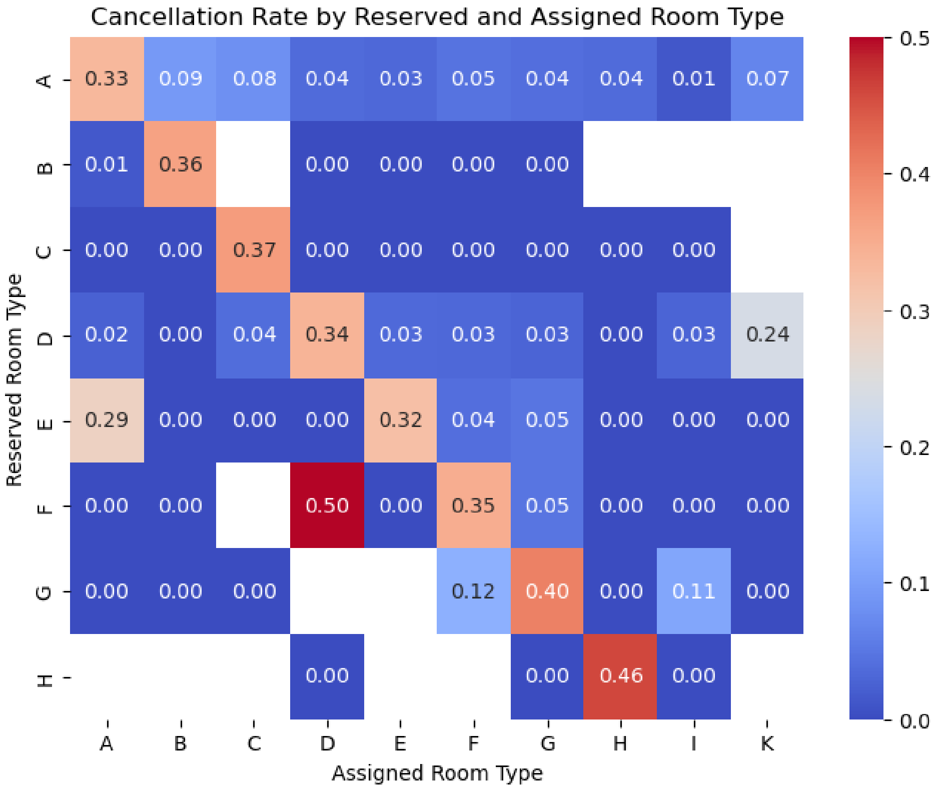

Figure 9: Correlation Between ADR, Reserved and Assigned Room Type.

APPENDIX D

Figure D1: Days in Waiting List and Lead Time Cancellation Measure

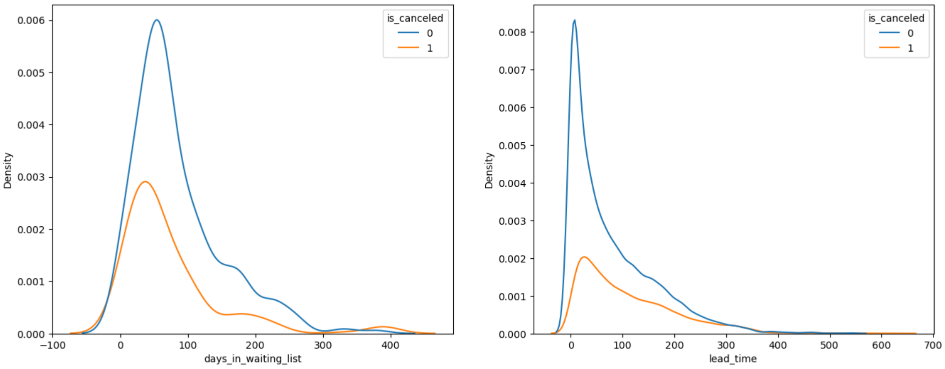

Figure 10: Days in Waiting List and Lead Time Cancellation Measure.

Figure D2: ADR Distribution per Month for Each Hotel

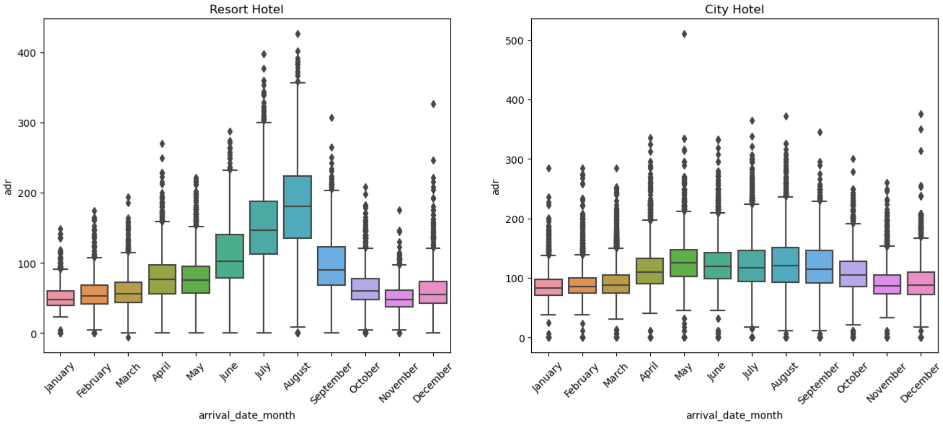

Figure 11: ADR Distribution per Month for Each Hotel.