NOVA IMS
Information Management School

MDDM

Master Degree Program in
Data-Driven Marketing

Hospitality AI-Driven Customer Journey Analytics

Predicting touchpoints in hotel Customer Journeys

Duarte Barros Rodrigues

Project Work

presented as partial requirement for obtaining the Master Degree Program in Data-Driven Marketing

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

HOSPITALITY AI-DRIVEN CUSTOMER JOURNEY ANALYTICS

PREDICTING TOUCHPOINTS IN HOTEL CUSTOMER JOURNEYS

By

Duarte Barros Rodrigues

Project Work presented as partial requirement for obtaining the Master's degree in Data-Driven Marketing, with a specialization in Marketing Intelligence

Supervisor: Miguel de Castro Neto

Co-Supervisors*:* Bruno Jardim

NOVEMBER 2023

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

[student signature]

[place, date]

DEDICATION

I would like to dedicate this thesis to my father, José Carlos Barros Rodrigues, for his unwavering support and commitment to my academic life. I hope to one day become the pillar in my children lives that he has been in mine.

ACKNOWLEDGEMENTS

I would like to thank Professor Bruno Jardim and Professor Miguel de Castro Neto for all their support and their endless patience in helping me craft this thesis and my family and friends for their belief in me and my potential.

ABSTRACT

The goal of this project is to develop a Customer Journey Framework that enables the marketing department of two hotels, a resort hotel (H1) and a city hotel (H2), to predict guest's behaviour using classification models throughout the three phases of their journey, Pre-Service, Service and Post-Service, using data obtained from the reservation management system. The marketing department can use the predictions to improve each guest's stay by anticipating the best outcome at each of these phases for each guest and start a conversation, creating opportunities to improve the service provided to guests while also improving the bottom-line results of the hotel at very little cost. It expands on work developed by Andriawan et al. (2020) and Antonio, Almeida, et al. (2017), who successfully developed classification models to predict cancellations using tree-based algorithms, by increasing the prediction scope to a full CJ in a hotel, building and testing separate models, divided per hotel, with each model answering each research question. The first model predicts the cancellation of bookings, scoring a recall above 80% in both hotels, the second model predicts the food and beverage package, scoring a F1 Score of 66% at H1 and 85% at H2, lastly the third model predicts which guests will book another stay, scoring a recall above 90% in both hotels.

KEYWORDS

Hospitality; Customer journey; Touchpoints; Machine Learning; AI; Classification; XGBoost

Sustainable Development Goals (SGD):

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
CJ	Customer Journey
CJF	Customer Journey Framework
CRISP-DM	Cross Industry Standard Process for Data Mining
DT	Decision Tree
RF	Random Forest
XGBoost	Extreme Gradient Boost
NN	Neural Networks
AUC	Area Under the Curve
ROC	Receiver Operating Characteristic
F&B	Food and Beverage
H1	Resort Hotel
H2	City Hotel
ADR	Average Daily Rate

1. INTRODUCTION

1.1. Context

The use of Artificial Intelligence (AI) to improve customer experience is becoming an increasingly important factor for decision-makers in various industries (Daqar & Smoudy, 2019).Particularly, in the Hospitality industry, we see a positive impact from increased usage of different AI tools, such as chatbots and recommender systems (Wilma & Schrotenboer, 2019) in improving the Customer Journey (CJ), and consequently a better relationship between brands and their customers (Ameen et al., 2021), which enhances the sales process.

In a world where successful Hospitality Marketing Strategies are increasing its reliance on AI (Parvez, 2020) , it's very clear that integrating Machine Learning (ML) into the three natural Stages of a CJ: Pre-purchase, Purchase and Post-purchase (Trapani et al., 2019) is a mandatory step to take. The aim of this project is to develop an experimental application of a CJ model based on instances of interaction between a customer and a hotel (touchpoints) using predictive techniques based on ML. Our proposed experimental application will be based on classification algorithms to estimate booking cancellations, Food and Beverage packages and if guests will book again.

It's our goal to conceptualize an optimized Customer Journey Framework (CJF) that improves the customer experience and increases the efficiency and efficacy of a hotel's booking process. By introducing an AI element into the CJ, the Marketing team will be empowered to provide a more effective service, optimized to improve personal and hotel-wide results.

1.2. Research Gap

Integrating AI in Marketing processes is viewed as having a positive impact by professional marketers (Tanveer et al., 2021), but to this day there's little research proposing effective frameworks to improve CJ that successfully combines both the analytical capabilities of AI with the decision-making of a professional. This constitutes a clear gap, with an enormous potential to deploy upgraded CJF that leverage the power of ML into the Sales process of a company (Arco et al., 2019). It is the purpose of this project to produce a CJF model (Halvorsrud et al., 2016), by adding predictive analytics to the CJF to augment the decision-making capabilities of Hospitality professionals, shown in figure 1.

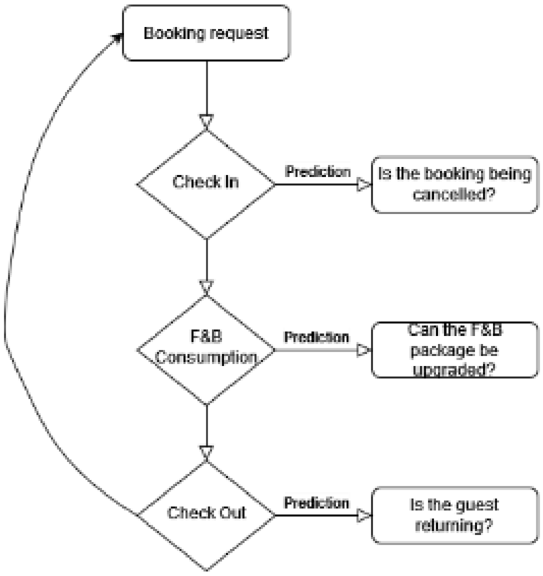


Figure 1 – Customer Journey diagram with prediction models

With the experimental application of these models, this project aims to answer the following research questions:

- Is it possible to accurately predict the cancellation of a booking?
- Is it possible to accurately predict the purchase of a F&B package?
- Is it possible to accurately predict if the guest is going to book again?

1.3. PROJECT DEVELOPMENT

The objective of this project is to deliver a benchmark of the integration of ML in the CJ in a hotel context to show how it can improve the management of guest's journeys, by exceeding their expectations of service and communication while limiting the cost of any marketing actions. What follows is a detailed account of the steps taken to build this project: Background presents key concepts used in the development of this project; Related Work analysis relevant literature published about CJ and their use in Marketing, the broader uses of AI in hospitality and ML models predicting cancellations in hotels; Methodology describes the CRISP-DM methodology applied to our hotel's dataset, where we contextualise the Business Understanding and Data meaning of our predictive models, the steps taken to Prepare the data for modelling, using different algorithms to achieve different results and creating the best Model, and Evaluate and analysing the performance of the models; Results and Discussion which reports discusses the results achieved by each algorithm and each model; Conclusions, Limitations and Future Works which summarizes the key takeaways from the project, while commenting on the limitations found throughout the development of the project as well as recommendations of ideas to develop in the future.

2. BACKGROUND

Most of the research in a Hospitality context regarding ML has started to appear in 2019, with a clear tendency to grow as time moves on (He & Zhang, 2022). The study of how this technology is applied has a certain trend to estimate the impact that Robots will have in the entire value-proposition of hospitality business over the medium to long-term, particularly how advanced robots will replace humans across several supporting positions. Very little research has been done to understand how ML can be applied in the short term, improving not just the efficiency of processes but also how to enhance professionals with analytical information that will translate into more value provided to all stake holders (save for Chatbots and Service Robots, which are already common techniques across different industries).

We start by presenting important concepts within ML that are fundamental for the development of this project, from the broadest concepts of Supervised and Unsupervised Learning to the relevant types of algorithms used in Supervised Learning

2.1. Supervised and Unsupervised Learning in Machine Learning

Unsupervised learning involves unlabelled data, necessitating the model to discern inherent patterns and structures without specific guidance on the output. The primary objective of unsupervised learning is to extract meaningful insights, identify hidden patterns, or reduce the dimensionality of the data. Common techniques in unsupervised learning include clustering, association, and dimensionality reduction (Kotsiantis et al., 2006).

Contrastingly, supervised learning is where models are trained on labelled datasets. The training data consists of input-output pairs, allowing the model to learn the mapping between input features and corresponding labels. This process enables the model to generalize and predict outputs accurately for unseen data. Supervised learning encompasses regression, predicting continuous values, and classification, predicting discrete labels (Kotsiantis et al., 2006).

2.1.1. Types of Supervised Classification Algorithms

These supervised classification algorithms constitute the fundamental methodologies in supervised learning, providing a comprehensive toolkit for handling diverse classification tasks. Decision Trees are a versatile machine learning algorithm used for both classification and regression tasks. They are a type of supervised learning model that creates a tree-like structure to make decisions. Each internal node of the tree represents a decision or a test on a feature, and each leaf node represents a class label (in classification) or a predicted value (in regression). Decision Trees are interpretable and can handle both categorical and numerical data. The algorithm makes decisions by recursively splitting the data into subsets based on the values of features. The choice of which feature to split on and the threshold for the split is determined by a splitting criterion. Common splitting criteria include Gini impurity, entropy, or the mean squared error, depending on whether it's a classification or regression tree. The goal is to select the feature and threshold that maximizes the reduction in impurity (for classification) or the reduction in variance (for regression) at each step. It will continue splitting until one or more stopping

criteria are met. These criteria may include reaching a maximum depth, achieving a minimum number of samples in a node, or having no further improvement in impurity reduction, thus preventing the tree from becoming overly complex and overfitting the data. To make predictions, a sample is passed down the tree, starting at the root node. At each internal node, the sample follows the decision path determined by the splitting criterion (e.g., if feature A < 5, go left; otherwise, go right). When the sample reaches a leaf node, the prediction is made based on the majority class (for classification) or the average value (for regression) of the samples in that leaf node.
Random Forest (Liaw & Wiener, 2002) is an ensemble learning method that combines the predictions of multiple decision trees to improve the accuracy and robustness of a predictive model. It's widely used for classification and regression tasks and is known for its ability to handle high-dimensional data, provide feature importance scores, and mitigate overfitting.

It starts of by building multiple decision trees by sampling the training data with replacement. This process is known as bootstrapping. Each tree is trained on a different subset of the data, making them slightly different from each other. At each node of a decision tree, a random subset of features is considered for the split. This introduces an element of randomness and reduces the chance of an individual feature dominating the decision-making process. The result is an ensemble of decision trees. When making predictions, each tree in the ensemble independently provides a prediction (for classification, this can be a class label vote or for regression, a predicted value). For classification, the final prediction is determined by a majority vote, and for regression, it's often the average of the predictions.

XGBoost, which stands for "Extreme Gradient Boosting" (Chen & Guestrin, 2016), is a powerful and popular machine learning algorithm that excels in various tasks, particularly in structured data and tabular data. It is an ensemble learning method that combines the predictions of multiple weak learners (typically decision trees) to create a strong predictive model. XGBoost is known for its speed, accuracy, and versatility, and it has won numerous machine learning competitions. The key to XGBoost's performance lies in its optimization of both the model structure (the decision trees) and the model weights.

The objective function that XGBoost minimizes consists of two parts: a loss function that measures the model's error and a regularization term that controls the complexity of the model. The objective function to minimize for a simple binary classification problem can be written as the formula 1.1.

$$Obj(\Theta) = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Where n is the number of training examples, yi is the true label of the i -th example, $\hat{y}_i$ is the predicted label of the i -th example, $L(y_i, \hat{y}_i)$ is the loss function that quantifies the error between the true label and the predicted label, K is the number of weak models (trees) in the ensemble, f_k represents the k-th weak model, $\Omega(f_k)$ is the regularization term that measures the complexity of the model. The loss function $L(y_i, \hat{y}_i)$ can vary depending on the specific problem. For example, for binary classification, it might be the binary cross-entropy loss. For regression, it could be the mean squared error. The key idea is to make $\hat{y}_i$ as close as possible to y_i while keeping the model complexity low. The regularization term $\Omega(f_k)$ is designed to prevent overfitting by penalizing the complexity of individual trees.

To make predictions using an XGBoost model, the predictions of all individual trees are combined. For binary classification, a logistic function is applied to the sum of the predictions to obtain the final probability. For regression, the predictions are simply summed up. The final prediction is the result of this aggregation. It also supports multi-class classification, ranking, and other variations, each with its specific loss functions and mathematical formulations.

Neural networks (Günther & Fritsch, n.d.), often referred to as artificial neural networks (ANNs), consist of layers of interconnected nodes, known as neurons or perceptrons. These networks are organized into three main types of layers: Input Layer which receives the initial data, and each neuron represents a feature in the dataset, Hidden Layers which intermediate layers contain neurons that perform complex transformations on the input data, and the Output Layer the final layer that produces the network's predictions or classifications. The core mathematical formula for a neuron's output in a neural network is the weighted sum of inputs passed through an activation function.

$$y = \phi(\sum_{i=1}^{n} w_i x_i + b)$$

Where the y is the neuron's output, $\phi(\cdot)$ is the activation function, $\sum_{i=1}^{n}$ denotes the summation over all input connections. w_i represents the weight associated with the i-th input, x_i is the i-th input, b is the bias term. The activation function ($\phi(\cdot)$) is a key component of a neural network, and there are several options, including the sigmoid function, hyperbolic tangent (tanh), rectified linear unit (ReLU), and many others. The choice of activation function depends on the problem and network architecture.

Training a neural network involves adjusting the weights and biases to minimize a loss function that quantifies the error between predictions and actual target values. This process is often accomplished using optimization techniques like gradient descent. The mathematical details of backpropagation are complex and involve the chain rule for differentiation, which updates weights and biases to minimize the loss function. Understanding the nuances, strengths, and weaknesses of these classification algorithms is vital for selecting the appropriate technique based on the problem at hand. Comparative analysis and evaluation of these algorithms can guide the choice of the most suitable approach for a given dataset and problem domain, a critical aspect for the successful application of machine learning in practice.

2.1.2. Performance Metrics

In the evaluation of predictive models, utilizing appropriate performance metrics is essential to gauge the effectiveness and accuracy of the model. We delve into key metrics that shed light on different aspects of the model's predictive prowess.

Accuracy, a fundamental metric, quantifies the proportion of correctly predicted instances out of the total instances. It offers an overall assessment of the model's correctness, making it a valuable initial evaluation tool. However, it is imperative to exercise caution with accuracy, especially in imbalanced datasets, as it may present an overly optimistic view if one class dominates the dataset.

$$Accuracy = \frac{(True\ Positives\ +\ True\ Negatives)}{Total\ Instances}$$

Precision focuses on the accuracy of positive predictions made by the model. It represents the proportion of true positive predictions in relation to all predicted positives. Precision is vital when the cost of false positives is significant, emphasizing the need for precision in positive predictions.

$$Precision = \frac{(True\ Positives\ +\ False\ Positives)}{True\ Positives}$$

Recall, also known as sensitivity or true positive rate, measures the model's ability to capture all actual positive instances. It signifies the proportion of true positive predictions in relation to all actual positives. Recall is crucial when the cost of false negatives is high, ensuring that the model effectively identifies all positive cases.

$$Recall = \frac{(True\ Positives\ +\ False\ Negatives)}{True\ Positives}$$

The F1 score, a harmonic mean of precision and recall, evaluates the model's performance by combining both precision and recall into a single metric. The F1 score is particularly useful when false positives and false negatives carry varying costs, necessitating a harmonized balance.

$$F1\ Score = \frac{(2 \times (Precision \times Recall))}{Precision + Recall}$$

AUC-ROC metric evaluates the model's ability to discriminate between the positive and negative classes. It quantifies the area under the ROC curve, which plots the true positive rate (sensitivity) against the false positive rate. A higher AUC-ROC indicates a model with better discriminatory power. The AUC-ROC is computed using integration or the trapezoidal rule for the ROC curve.

These performance metrics allow to comprehensively evaluate the predictive models, enabling understanding of their strengths and weaknesses. These metrics can be used to support informed decisions and enhance the overall efficacy of the predictive modelling endeavour.

3. RELATED WORK

The purpose of this chapter is to understand the research that exists on the convergence of Marketing and A.I., where Marketing is the foundation of the object of search "Customer Journey" and AI is the foundation of the object of search "Machine Learning" (Campbell et al., 2020).

3.1. CUSTOMER JOURNEY IN HOSPITALITY

Like the 4 Ps of the Marketing Mix (Product, Price, Promotion, Placement or Distribution) (McCarthy, 1978), the concept of Customer Journey is a foundational aspect of Marketing that is constantly reinterpreted and adjusted based on a company's context (Lee, 2010). It is typically defined as an "*experience with a firm over time during the purchase cycle across multiple touch points*". Conceptually this process is linear and repetitive, but it ends up being iterative and dynamic, as it's very rare that a customer goes through the journey from the beginning till the end in a smooth course (Lemon & Verhoef, 2016).

Touch point is a common term used in CJ. It refers to when there's an "*instance of communication between a customer and a service provider*"(Halvorsrud et al., 2016), which entails three main attributes: Initiator, who is a customer that starts the interaction with a firm, Channel, where the interaction takes place (such as e-mail, website, social media, etc), and a Trace, the consequence of the interaction (Bitner et al., 2008; Halvorsrud et al., 2016).

Within the concept of CJ there's an important sub-concept: Customer Journey Mapping (CJM). This sub-concept stands for "*all the activities performed to analyse an existing service process as is"* (Følstad & Kvale, 2018), with the purpose of understanding which touch points customers are required to go through to successfully complete the CJ process. There's still a lot of missing literature proposing appropriate Key Performance Indicators for each touch point or the entire journey, mainly due to the difficulty of establishing a scale that can fit all journeys, as each customer can have its own unique journey within the same Brand.

The CJ purchase cycle can be defined as a three Stage process (Lemon & Verhoef, 2016), which can be optimised for a hotel business concept while retaining the core of the original cycle:

1. Pre-Purchase or Pre-Service "*encompasses all aspects of the customer's interaction with the brand, category, and environment before a purchase transaction"* (Lemon & Verhoef, 2016). This is widely considered as the longest and hardest stage to measure. Most of the interactions happen outside of the scope of the company which makes it more difficult to assess performance measures. In a hotel context we're considering the navigation of the website and consequent booking of a stay as the pivotal moment that transitions a customer into the next stage (Følstad & Kvale, 2018).

2. Purchase or Service Period "*covers all customer interactions with the brand and its environment during the purchase event itself*" (Lemon & Verhoef, 2016). It is typically viewed as the most time compressed stage, as most transaction tend to take just a few minutes to be completed, although in a hotel business process the length of stay makes this transaction different form the norm. As such it's taken into this stage's scope the entire stay as a

measurement, as well as all purchases of associated services such as Food and Beverage facilities (Følstad & Kvale, 2018).

3. Post-Purchase or Post-Service *"encompasses customer interactions with the brand and its environment following the actual purchase"* (Lemon & Verhoef, 2016). This stage considers all the behaviours post-visit, particularly in the form of feedback through social media or satisfaction questionnaires. It's considered as a measure the positivity index of the review (Følstad & Kvale, 2018).

Several studies have been conducted with the goal of establishing Key Performance Indicators in the Pre-Purchase phase (Trapani et al., 2019) by using traditional research methodologies such as questionnaires. Although some interesting differences in behaviour of Men and Women were found, we believe that a better analysis of patterns in behaviour at any of the purchase stages should be data-driven, using a ML algorithm to detect and predict outcomes of required touch points, to prevent bias.

To be able to properly measure each stage it is necessary to break them down into micro-realities with a specific purpose. The 3-stage cycle can be broken down into four steps, with Step 1 and Step 2 regarding Pre-Purchase stage, Step 3 focusing Purchase and Step 4 concerning Post-Purchase (Wilma & Schrotenboer, 2019):

1. Awareness/Information Search is "*the first time the customer comes in contact with the firm*". In this step the customer is interacting with the brand, through the website for example, to acquire information about the service/product. An appropriate measure for this step is evaluate the activity in the owned website, with KPI's such as Number of Visits, Time Spent on the Website, Number of Pages Visited.

2. Evaluation of alternatives is when "*the customer is evaluating the firm or its products*". The customer will consider the information discovered in the previous step and decide which alternative is better for their needs. A good measure of this step is how long did it take between the last touch point (presumably during step 2) and the beginning of the Purchase, in our case booking a stay.

3. Purchase of product/service includes "*the actual buying of the product*". Because we're focusing particularly on Hospitality the Purchase stage is slightly different, mainly because it's not an instant purchase, but rather a prolonged purchase throughout the stay. It includes not only the purchase of a room for a specific amount of time but also the purchase of additional facilities such as the restaurant. For this step the amount spent on the visit in the room and other facilities is considered as the best measure.

4. Post-purchase review "*is the point at which the customers have used the product and formed their opinions*". At this step the customer has completed their journey and will provide feedback about it. In Hospitality this is normally done via a Satisfaction Enquiry sent by e-mail, although since the proliferation of smartphones and the rise of Social Media a lot of customers prefer to review the service provided via specialised platforms like Tripadvisor. The proper measures of this step are the submission of Feedback and the Positivity rate of the feedback, which measures how many positive words are used in the review.

CJ seems at times more an art rather than a science, especially regarding the development of the mapping and the shaping of its touch points. But what is clear is that a successful management of the CJ increases the satisfaction of customers and clear measures are needed to be properly tracked and managed throughout all the Stages of the journey (Karimi Sahar and Papamichail, 2014; Lemon & Verhoef, 2016; Taheri et al., 2021; Wilma & Schrotenboer, 2019).

3.2. Machine Learning in Hospitality

A lot of research done in Hospitality and Machine Learning revolves around technology that can replace Human interaction, especially Chatbots and Service Robots (He & Zhang, 2022). With more sophisticated algorithms capable of interpreting, learn and act in complex environments, the wave of AI and Robotization is undoubtedly closer than ever and will impact how Hospitality businesses provide value to guests in the future (McCartney & McCartney, 2020).

Chatbots are "*computer programs that simulate human conversations through voice commands or text chats and serve as virtual assistants to users*", such as HiJiffy (Parvez, 2020). They're already a very effective application of AI when interacting with customers, particularly in the Pre-Purchase stage, engaging in the dialogue as successfully as a proficient worker (Luo et al., 2019). With such positive results, the rate of adoption of this technology as an integral part of the customer journey will only tend to increase.

Service Robots are a "*physical embodiment of information technology, providing customized services by performing physical as well as non-physical tasks with a high degree of autonomy*", for example Connie, Hilton's concierge robot (Parvez, 2020). Such technology is still in its infancy, but the potential to add value to all stakeholders is undeniable. The reality of integrating robots into the workforce brings both a sense of dread and excitement, but such reality is inevitable as it will completely redefine the experience hospitality guests are looking for in the marketplace, as these robots take on more and more tasks within the Hotel environment (McCartney & McCartney, 2020).

AI shopping systems are "*systems that replace manpower to provide consumers with shopping support, enabling them to experience a more autonomous shopping process*", for example the Marriot hotel brand in China, which allows customers to Check-In and Out using Facial Recognition (Parvez, 2020). Although customers tend to seek human to human interactions in a hospitality context, no doubt derived from an anthropological need to socialise, there's growing evidence that, if let to their own choice, customers clearly show a preference for innovative AI systems that enhance their experience (Cui et al., 2022). Enhancing is the key factor, as the mal function of Self Check-In and out or lack of simplicity in its deployment play a pivotal role in dissuading customers from using them (de Bellis & Venkataramani Johar, 2020; Meuter et al., 2005).

Even though the application of ML is becoming more common there's still a focus on the operational side of the industry and how much it will shape the Socio-Economic reality in the long term, but the same technology can be applied today in a human-centred framework that can also provide tremendous value to industry (Parvez, 2020).

3.3. Machine Learning to predict Bookings, Revenue and Returns

The available literature focuses on predicting bookings or cancelations, as this is the core objective of any hotel business process and by large the most revenue generator (Andriawan et al., 2020). Accurately predicting bookings gives Revenue Managers greater control over the different metrics, such as Occupancy, Rate, Revenue per Room (Antonio, Almeida, et al., 2017; Antonio, De Almeida, et al., 2017, 2019), helping them to reach business objectives more consistently (Andriawan et al., 2020).

Although the use of statistical forecasting tools has been the norm of hotel business process for several years, from 2017 forward there is an increasing number of studies that suggest ML models can outperform traditional forecasting tools, particularly if applied to datasets with a larger time series (six months to one year at least) (Pereira & Cerqueira, 2022). The usage of Supervised Learning Algorithms such as Classification have been successfully applied in predictive models, achieving rates of Accuracy of 80% or higher (Andriawan et al., 2020; Antonio, Almeida, et al., 2017; Antonio, De Almeida, et al., 2017, 2019; Castelli et al., 2022; Pereira & Cerqueira, 2022; Sánchez-Medina & C-Sánchez, 2020).

One of the drawbacks of ML models is the fact that it become less effective as time goes by. Some research has also addressed this issue by proposing a new Metric: Minimum Frequency, which allows for models to penalise False Positives and False Negatives by redistributing the weight of older data vs new data, thereby correcting for possible misinterpretation of new data (Antonio, De Almeida, et al., 2019).

There's a consensus that there's no need to use very complex datasets with a high number of variables as more variables bring no additional value to the prediction performance (Sánchez-Medina & C-Sánchez, 2020), instead models with as little as thirteen variables have been successfully deployed with the same accuracy as models with thirty-seven or more. This is good news for hospitality professionals as it decreases the complexity of preparing, training, and deploying effective models considerably, thereby making ML solutions more appealing (Sánchez-Medina & C-Sánchez, 2020).

Another use of ML algorithms in hospitality lies in understanding and predicting customer's more complex behaviour such as satisfaction or loyalty, and revenue. By using AutoML, "*an AI tool tailored to consumer behaviour research*", which is easy to use and interpret, businesses can tap into more powerful data-driven tools to measure the success of their business by continuously analysing customer's satisfaction and revenue. Although this research is still in its infancy, there are clear indicators that adopting these tools will increase control of overall business processes (Castelli et al., 2022).

Successfully analysing and predicting bookings, revenue, and behaviour in hospitality is a research area that's growing more interesting with time. Businesses that can introduce Machine Learning models to assist business managers in the decision-making process will no doubt gain a large advantage over their competitors and be able to provide more added value to their guests (António et al., 2019).

Table 1 summarises a series of research endeavours, each focused on harnessing the potential of machine learning techniques to enhance the prediction of hotel booking cancellations. While all studies address hotel booking cancellations, their emphasis varies. For instance, (Sánchez-Medina & C-Sánchez, 2020) aim to predict cancellation rates and identify potential cancelling customers, focusing on customer-centric insights. In contrast, (Antonio, De Almeida, et al., 2017) and (Antonio, Almeida, et al., 2017) are primarily concerned with reducing the financial impact of cancellations and accurately predicting net demand, emphasizing revenue management and forecasting. (Andriawan et al., 2020) focuses on accurate cancellation prediction within the structured framework of CRISP-DM.

While the studies share commonalities in terms of utilizing multi-year data, there are variations in dataset sizes and temporal spans. Most studies involve a split in the data between H1 and H2, due to the fact that they use the same source of data. This approach allows for cross validation and the assessment of model robustness for different purposes.

XGBoost is a common algorithm in three of the studies, reflecting its popularity and effectiveness in predictive modelling tasks. Random Forest also appears in two studies, showcasing its versatility in classification problems.

The choice of evaluation metrics also varies. (Sánchez-Medina & C-Sánchez, 2020) employ accuracy as their primary metric. (Andriawan et al., 2020) consider both accuracy and precision, focusing on the overall correctness of predictions and precision in positive predictions. (Antonio, De Almeida, et al., 2017) incorporates accuracy and the AUC for assessing the model's ability to distinguish between classes. (Antonio, Almeida, et al., 2017) extends the evaluation criteria to include accuracy, precision, and AUC, emphasizing model discrimination.

Reference	Context	Data	Objective	Algorithm	Evaluation
Sánchez-Medina & C-Sánchez (2020)	ML for Forecasting bookings	3 years of data: 10.000 bookings from a single hotel	Predict cancellation rates and likeli**hood** of cancellation	• Random Forest • S.V.M. • C 5.0 • ANN	• Accuracy
Andriawan et al. (2020)	Predicting Hotel Booking Cancellation using CRISP-DM	3 years of Data: 40,060 bookings in H1 and 79,330 bookings in H2 (Antonio, de Almeida, et al., 2019)	To accurately predict cancellations	• Random Forest • Catboost • L.G.B.M. • XGBoost	• Accuracy • Precision
Antonio, De Almeida, et al. (2017)	Predicting Hotel Bookings Cancellation with ML Classification model	3 years of Data: 40,060 bookings in H1 and 79,330 bookings in H2 (Antonio, de Almeida, et al., 2019)	Reduce the negative revenue impact from cancellations	• XGBoost	• Accuracy • Area Under the Curve (AUC)
Antonio, Almeida, et al. (2017)	Predicting hotel booking cancellations to decrease uncertainty and increase revenue	3 years of Data: 20,522 bookings in H1, 9,809 bookings in H2, 9,365 bookings in H3, 33,445 bookings in H4	Accurately predict net demand	• Boosted Decision Tree • Decision Forest • Decision Jungle • S.V.M. • Neural Network	• Accuracy • Precision • Area Under the Curve (AUC)
Antonio, De Almeida, et al. (2019)	Predict Hotel Booking Cancellations	3 years of Data: 40,060 bookings in H1 and 79,330 bookings in H2 (Antonio, de Almeida, et al., 2019)	Accurately predict bookings over time in a real environment	• XGBoost	• Accuracy • Area Under the Curve (AUC)

Table 1 - Summarization of research papers concerning booking cancellation

4. METHODOLOGY

This chapter outlines the structured approach employed to predict variables utilizing the CRISP-DM framework (Chapman et al., 2000). Through an organized progression encompassing business understanding, data preparation, model selection, evaluation, and deployment, we aim to develop accurate and robust predictive models aligned with the objectives of this research.

To be able to develop and present a prototype that provides accurate and relevant predictions across the CJ's touchpoints, two steps are required. Firstly, we start off by obtaining data from a hotel that would allow for a detailed analysis and prediction. Next, we will employ the CRISP-DM methodology, consisting of six steps, which we'll use to build three different models, each targeted at the research questions posed earlier.

In the first step, called Business Understanding, we establish an understanding of the organizational context and the specific objective of our predictive modelling task. We define the research question or business problem we intend to address and align it with the broader organizational goals. Additionally, we evaluate the current situation, including existing knowledge, limitations, resources, and risks, to ensure a well-informed approach.

Data Understanding, the second step, is pivotal for predictive modelling. This step involves data collection from various sources, encompassing the acquisition of relevant datasets and understanding their structures. Through initial data exploration, we grasp the characteristics and quality of the data, laying the foundation for feature selection and engineering.

In the third step, named Data Preparation we prepare the data for modelling. Data cleaning addresses issues such as missing values, outliers, and inconsistencies, ensuring data quality. Feature engineering involves the careful selection and transformation of features, optimizing their representation to enhance the model's predictive capability (Zhang et al., 2003). This step is tailored to the specific variables we aim to predict.

The fourth step, called Modelling, involves selecting appropriate machine learning models aligned with the prediction task. For predicting categorical variables, we employ classification models. Subsequently, we train these models on the prepared data, optimizing their parameters for optimal performance. Comprehensive evaluation of these models is conducted, employing suitable metrics to assess their efficacy.

The fifth step, Evaluation, is critical for validating the predictive capabilities of the models. We rigorously assess model performance on unseen data to ascertain generalizability. Fine-tuning strategies may be employed based on the evaluation results, ensuring that the predictive models are refined and aligned with the defined objectives.

Finally, the sixth step, called Deployment, we transition from experimental validation to practical application. The best-performing model is integrated into the production environment, enabling real-time predictions on new data. Strategies for monitoring, retraining, and updating the model are outlined, ensuring its relevance and accuracy in a dynamic operational setting.

4.1. BUSINESS UNDERSTANDING

The Marketing department in a hotel splits its focus on communication to guests and supporting all other departments. It is expected to assist both the different management departments (Revenue, Reservations, etc.) and operations departments (F&B, Front Desk, etc) and therefore should be considered fundamental in the structure management of a hotel. The proposed model aims to add another tool to the arsenal of a marketing manager by focusing the communication efforts directly to guests who will have been selected by AI with the correct behavioural pattern for each of the possible outcomes of each touchpoint throughout the CJ.

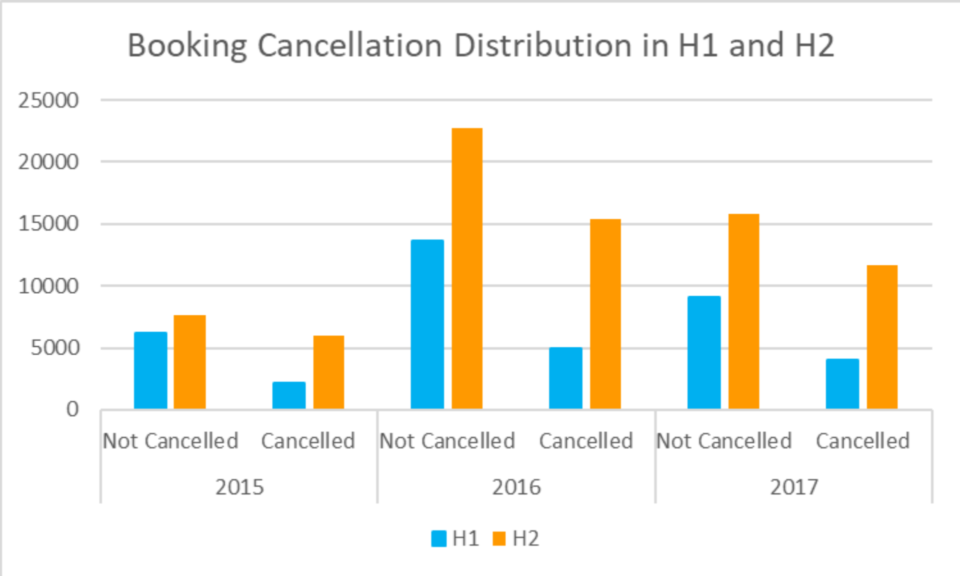


Figure 2 - Booking cancellation distribution

As seen in figure 2, H1 has a total of 11,122 cancelled bookings, an average of 28% over all the years, while H2 has a total of 33,102 cancelled bookings, an average 42% over all the years. H1 has a worrying trend of increasing the proportion of cancelled bookings, growing from a rate of 26% cancellations (2,138 bookings) in 2015, to a 27% cancellation rate (4,930 bookings) in 2016 and 31% (4,054 bookings) in 2017. H2 is worse, starting with an average of 44% cancelation rate (6,004 bookings) in 2015, decreasing to 40% (15,407 bookings) in 2016, and growing back to 43% (11,691 bookings) in 2017. This is a considerable proportion of cancelations, especially given the fact that the data for the years 2015 and 2017 is incomplete, which leads us to conclude that the cancellation rate would likely keep increasing in time.

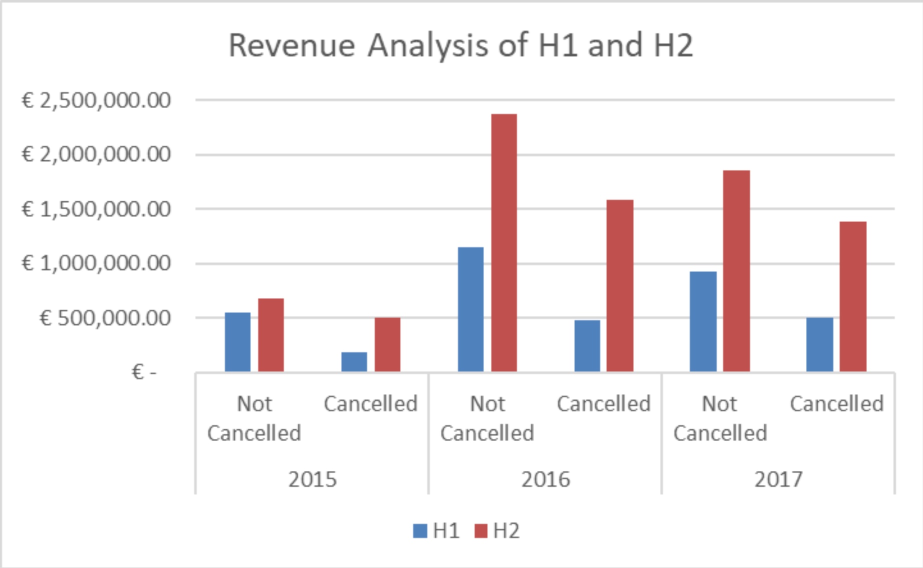


Figure 3 - Revenue distribution of bookings

Although we don't have data regarding occupation, it's acceptable to estimate that a considerable part of cancellations represents a loss of earning, which cannot be recovered. Figure 3 shows us the revenue in euros lost to cancellations, which is in line with the proportion of cancelled booking, in H1 representing 31% of revenue (€1,176,563.13) and in H2 representing 41% of revenue (€3,465,379.54). H1 also shows a concerning increase of loss of revenue over the years, with 2015 losing 25% (€188,530.11), 30% (€484,657.82) in 2016 and 35% (€503,375.20) in 2017. H2 has higher amounts of loss revenue, with 43% (€500,002.11) in 2105, 40% (€1,580,823.17) in 2016 and 43% (€1,384,554.26) in 2017. This is a substantial amount of revenue that is being lost every year, with a pattern leaning towards increasing this amount. This represents an excellent opportunity for communication, identifying those bookings with the potential to not cancel, therefore keeping the lost revenue and potentially increasing guest satisfaction due to the effort of communication with them.

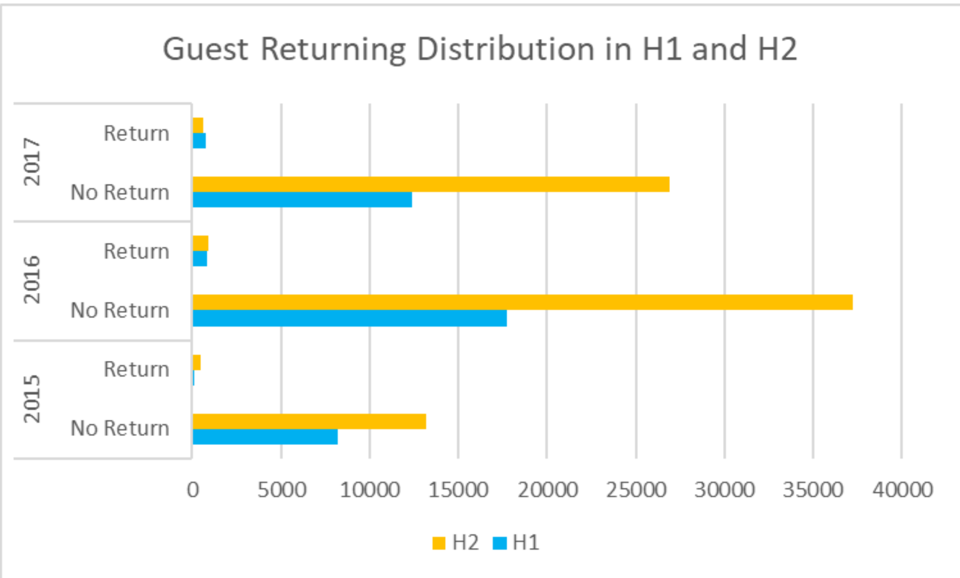


Figure 4 - Distribution of bookings per returning guests

Another issue both H1 and H2 have is their inability to retain guests through time. Figure 4 shows this clearly, with H1 having only 4% of bookings (1778 bookings) made by guests who returned and H2 has 3% (2032 bookings). H1 has identified this issue and it's visible that the proportion of returning guests is increasing each year, going from 2% (134) in 2015, to 5% (863) in 2016 and 6% (781) in 2017. H2, on the other hand, is decreasing the proportion of returning guests, having its peak at 4% (507) in 2015, then going down to 2% (915) in 2016 and keeping the 2% (610) in 2017. Being able to retain guests is one of the most efficient ways of optimising the profit of a hotel, since these guests won't cost more money to acquire, since it increases the lifetime value of each guest.

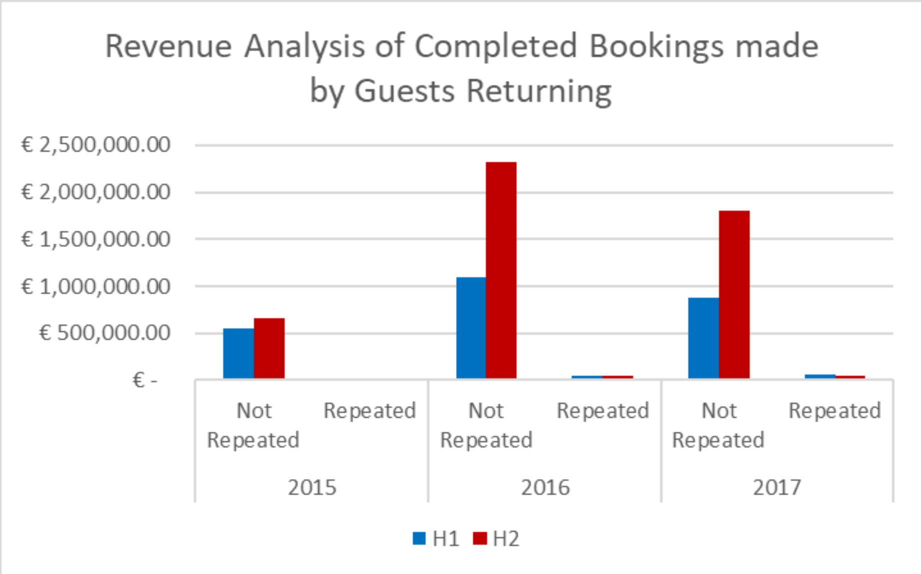


Figure 5 - Distribution of revenue per return guests

Figure 5 clearly translates the impact of the inability to retain guests on revenue. H1, as suggested before, is improving over time increasing both the number of bookings and revenue, capturing 1% (€5,071.28) of revenue from returning guests in 2015, then growing to 3% (€45,910.43) in 2016 and again to 4% (€55,666.35) in 2017. Contrarywise, H2 isn't growing their returning customer base, keeping the same proportion of 1% every year, with the only difference being the amount of revenue captured per year, €9,676.42 in 2015, €47,784.69 in 2016 and €40,652.47 in 2017.

4.2. Data Understanding

One of the limitations of this project is that it uses academic data as source, not a current business scenario. As such, the results obtained will produce a prototype that will have a broader use in hospitality rather than a specific problem of a hotel property or brand. Nevertheless, the dataset obtained possesses all relevant attributes necessary to provide an overall analysis of the operations of a hotel and develop predictive models of the key touchpoints of a CJ.

The dataset is split between two anonymous hotels, a Resort Hotel (H1) and a City Hotel (H2). Both Hotels have 33 features, of which 10 are categorical and 23 are numeric, with the H1 having 40,061 observations and the H2 having 79,331 observations. It keeps track of stays from 2015 till 2017 from the moment a booking takes place (Check-In) to the moment the guest leaves after the visit (Check-Out). In table 2, we describe each of the attributes portrayed in the dataset as obtained from Antonio, de Almeida, et al., (2019).

Attributes	**Description**	**Data Type**
ID	Unique identifier of each observation	Numeric
Hotel	Identifying which hotel the booking belongs to	Categorical
Is Cancelled	If the booking is cancelled or not	Numeric
Stay Duration	How many nights did the guest spend at the hotel	Numeric
Lead Time	Number of days prior to arrival that the booking was placed in the hotel	Numeric
Arrival Date Year	Number of Year of guest arrival	Numeric
Arrival Date Month	Month of arrival date	Categorical
Arrival Date Week Number	Number of week in the year (1 to 52)	Numeric
Arrival Date Day of the Month	Day of month of arrival date (1 to 31)	Numeric
Stays in Weekend Nights	From the total length of stay, how many nights were in weekends (Saturday and Sunday)	Numeric
Stays in Weeknights	From the total length of stay, how many nights were in weekdays (Monday through Friday)	Numeric
Adults	Number of adults	Numeric

Attributes	Description	Data Type
Children	Number of children	Numeric
Babies	Number of babies	Numeric
Meal	Type of meal included during the stay (BB, HB, FB, SC)	Categorical
Country	Country ISO identification of the main booking holder	Categorical
Market Segment	Which market segment does the booking belong to	Categorical
Distribution Channel	Which channel was used to make the booking	Categorical
Is Repeated Guest	Binary value indicating if the booking holder, at the time of booking, was a repeat guest at the hotel (0: no; 1: yes)	Numeric
Previous Cancellations	Number of previous bookings to this booking the guest had that were cancelled	Numeric
Previous Bookings not Cancelled	Number of previous bookings to this booking the guest had that were not cancelled	Numeric
Booking Changes	Created by summing the number of booking changes (amendments) prior to arrival that could indicate cancellation intentions (arrival or departure dates, number of persons, type of meal, ADR, or reserved room type)	Numeric
Deposit Type	What deposit was asked to guest to book (No Deposit, Refundable or Non-Refundable)	Categorical
Reserved Room Type	Room type requested by the guest	Categorical

Attributes	Description	Data Type
Assigned Room Type	Room type assigned to the guest	Categorical
Agent	Which agent id is associated with the booking	Numeric
Company	Which company id is associated with the booking	Numeric
Days in Waiting List	How many days did the booking was hold in waiting	Numeric
ADR	Average Daily Rate	Numeric
Required Car Parking Spaces	How many car parking spaces are required per booking	Numeric
Total of Special Requests	How many special requests were made for each booking	Numeric
Reservation Status	Status of guest after stay (e.g., Check-Out, No Schow, Cancelled)	Categorical
Reservation Status Date	Date of Status of guest after stay	Numeric

Table 2 - Description of attributes of H1 and H2

Upon analysing both datasets more closely, we can see that the most popular month for bookings in both H1 and H2 is August, which coincides with the peak of holidays in Portugal, and the least popular month is December. The majority of guests are Portuguese (44% in H1 and 39% in H2), in which H1 guests stay on average 3 nights during the week and 1 night during the weekend, suggesting an average of 4 nights per booking, while in the H2 guests stay for 2 nights during the week and 0.7 nights during the weekend, suggesting an average of 3 nights per booking. Most bookings are for adults only (91% in H1 and 97% in H2), with a preferred Meal choice of BB (Bed and Breakfast) and guests tend not to have many special requests.

As per the research questions we'll have three target variables for prediction: Is Cancelled, Meal and Is Repeated Guest. We'll analyse each of the target variables regarding their distribution across the Hotels and proportion of observations in each dataset. Figure 6 shows the distribution of cancellations of the H1.

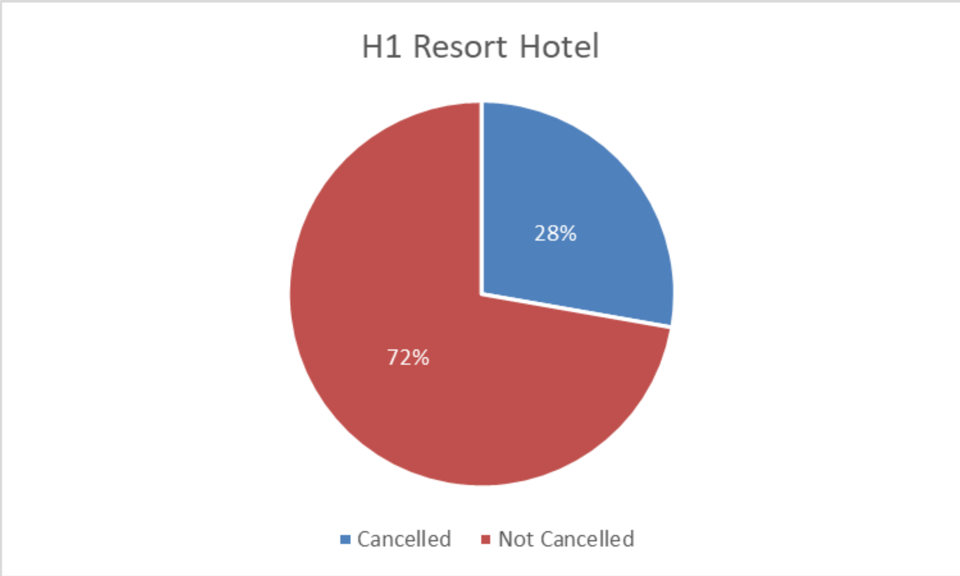


Figure 6 - Distribution of Cancellations in H1

As seen in the figure above the distribution of cancellations is highly imbalanced, with cancellations barely making more than a quarter of observations. While this is good news for the hotel, since it means that it's not losing that much money in cancellations, when it comes to building a predictive model it's preferable to have a balanced dataset, so the algorithms can better interpret and predict the patterns of both outcomes. In this instance the imbalance is not so great as to require any techniques to artificially re-balance the dataset, but it will also mean that the model built for H1 will likely have lower scores than H2 when evaluated. Moving on to the hotel H2, the distribution is as shown in figure 7.

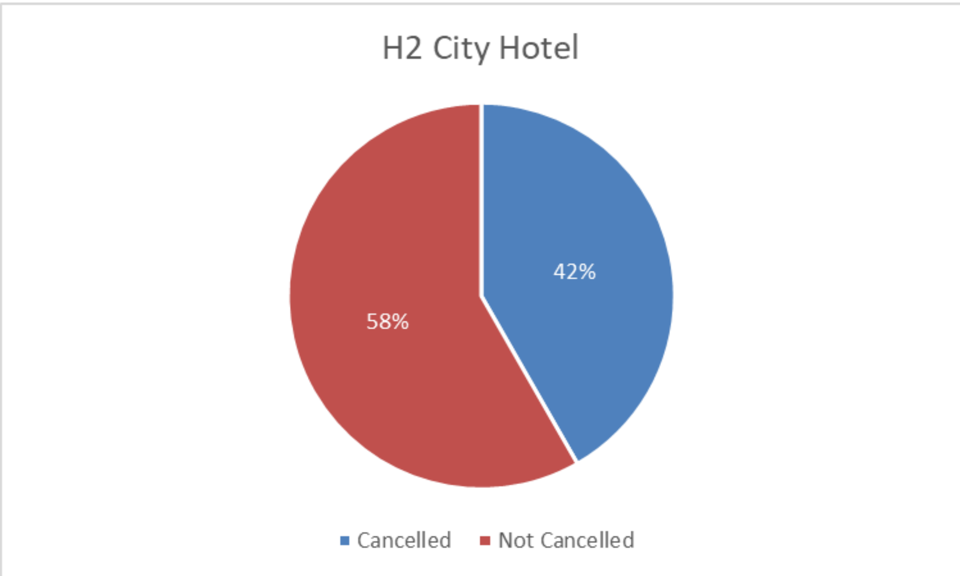


Figure 7 - Distribution of Cancellations in H2

Now this hotel has a much more balanced distribution, with cancellations making up more than 40% of observations. It is clear that cancellations have a big impact on this hotel, and while it may not be good news for the hotel owners, having such a balanced distribution of observations will allow us to produce a model with excellent scores across. Regarding the Meal target variable, which as we've seen before tracks the food packages purchased for the booking, we also found some imbalance, as seen in figure 8.

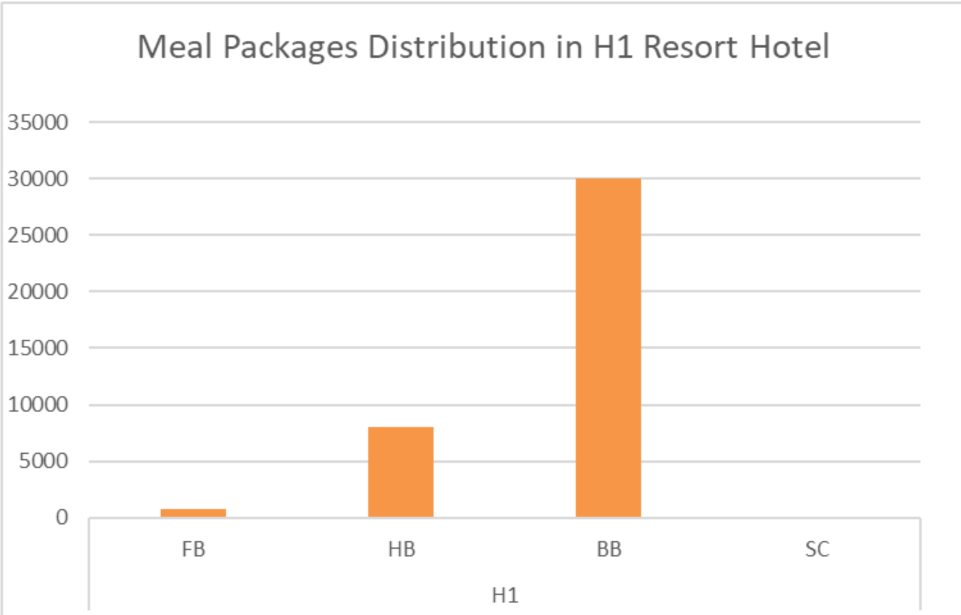


Figure 8 - Distribution of Meal Packages in H1

As it was noted before, the BB option is quite popular when compared to the other options, capturing about 77% of observations. The SC (Self-Catering) and FB (Full Board) are residual, representing 0.2% and 2% respectively which could affect the performance of the model due to their low representation. The HB (Half Board) option, which captures about 20% of observations, is the second most popular option and should be well represented to be interpreted by the algorithms. The H2 the situation is not very different in that there's a big imbalance between the BB and the remaining options, as shown by figure 9.

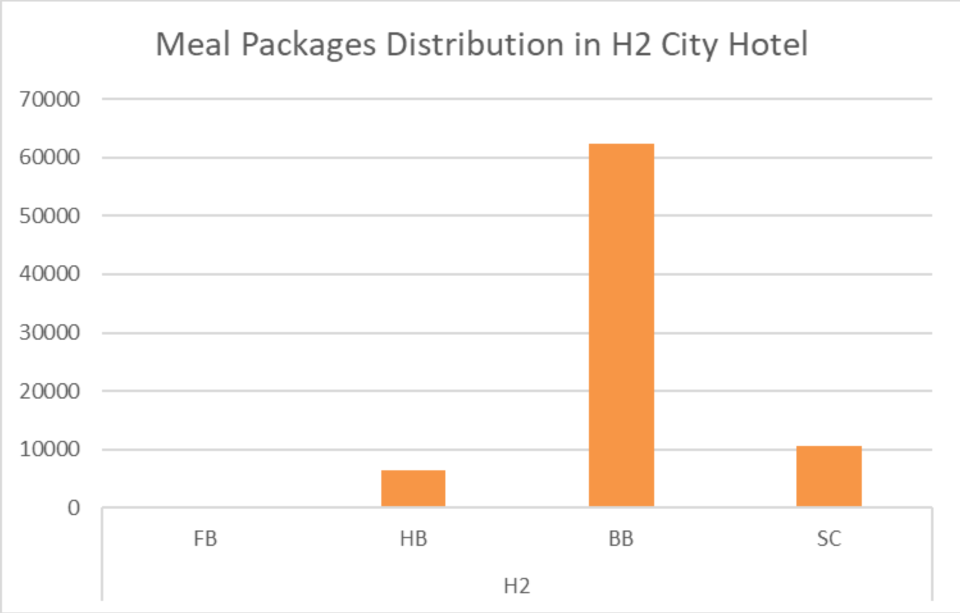


Figure 9 - Distribution of Meal Packages in H2

Here the BB option captures 78% of observations, similarly to H1, however the FB option is essentially non-existent, capturing only 0.06% of observations. Also contrary to H1 the HB option captures fewer observations, about 8%, and the SC captures more observations, about 13%. Despite these differences, the H2 dataset is more balanced than H1, meaning that each of the values in this dataset has more representation in number of observations, which could produce a model with better results. Lastly, we'll analyse the last target variable Is Repeated Guest, which tracks if the guest is new or returning, over both hotels, starting off by the H1 in the figure 10.

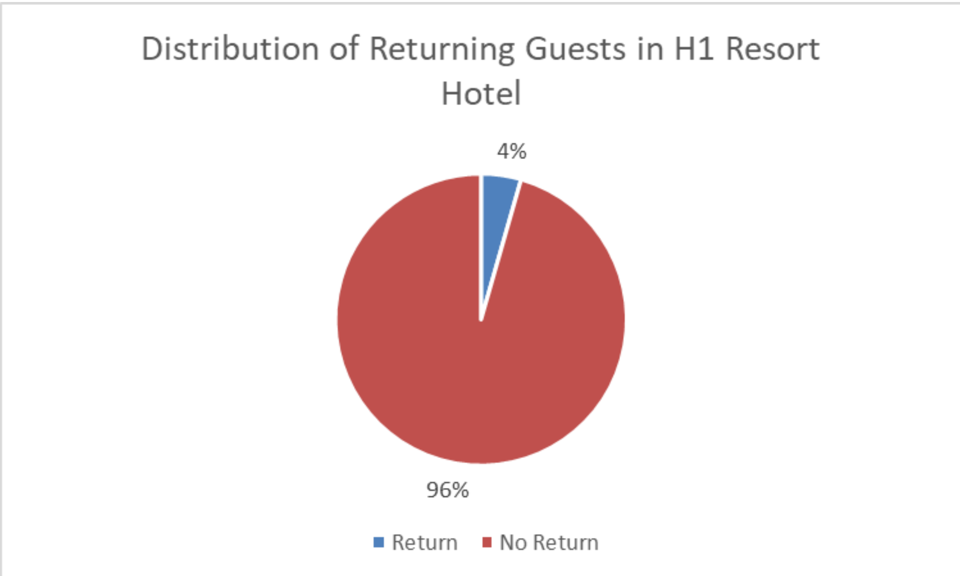


Figure 10 - Distribution of bookings of returning guests in H1

We can see that H1 has a serious problem retaining guests, with only 4% of bookings being made by guests who returned over the course of the better part of three years. Like the previous target variables there's a great imbalance, although this time the imbalance is too severe to have reliable results. Consequently, we will most likely have to employ technique to artificially balance the dataset and have reliable predictions on both scenarios. The H2 has the same problem as H1, seen in the figure 11.

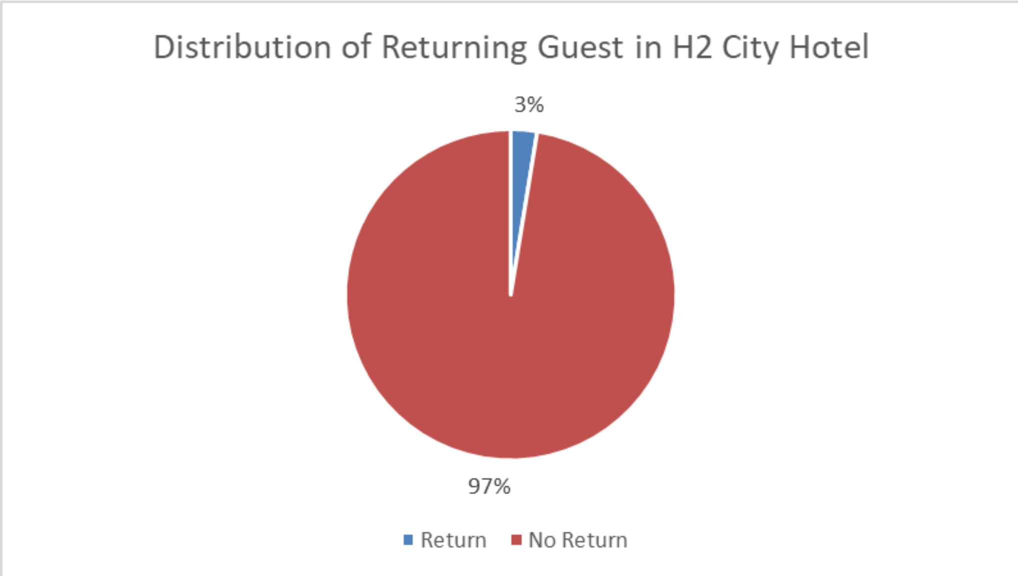


Figure 11 - Distribution of bookings of returning guests in H2

4.3. Data Preparation

After completing the Data Understanding part, we proceed to the Data Preparation phase. In this phase, we are going to perform some useful commands as pre-processing tasks to prepare and clean the dataset to have a better performance of our prediction. Pre-processing typically involves four steps.

Checking and removing Duplicated Data, as it can skew analyses and lead to incorrect conclusions. In this step, we meticulously scan the dataset to detect and remove duplicate entries. The process involves comparing records across relevant attributes, ensuring data integrity, and creating a clean and accurate dataset.

Coherence procedures and Missing Values treatment, where we assess the dataset for inconsistencies, such as conflicting information or discrepancies in variable definitions. Furthermore, missing values pose a common challenge. We employ techniques such as imputation or removal to handle missing data, ensuring that our dataset is complete and coherent, thus preventing biased or erroneous analyses.

Outliers and Skewness, in which we employ statistical methods to detect and address outliers, either by modifying or removing them. Additionally, skewness in the data distribution can affect the model's assumptions, and techniques like log transformation or scaling are employed to address this issue, making the data more suitable for analysis.

Normalizing and Feature Engineering, crucial for ensuring that features are on a consistent scale, preventing some features from disproportionately influencing the model. Feature engineering involves crafting new features or transforming existing ones to better represent underlying patterns in the data. This step goes beyond simple data cleaning; it strives to enhance the dataset's predictive power by extracting meaningful information or creating informative features tailored to the specific modelling task.

4.3.1. Duplicated Data and Missing Values

No Duplicated Data could be found in the dataset and therefore no action was warranted. As we move forward to the Missing Values treatment step, we're able to identify the following attributes with missing values: company, agent, and country in H1 and company, agent, country, and children in H2. The weight of the missing values is shown in the table 3 as percentages.

Attributes	Missing Data in H1	Missing Data in H2
Company	92.24%	95.34%
Agent	20.50%	10.24%
Country	1.15%	0.03%
Children		0.005%

Table 3 - Attributes with missing observations in H1 and H2

Company stands out as having a severe problem with missing data. In such extreme cases, Aguinis et al. (2013) recommends that the attribute be discarded as any attempt to impute missing data will very likely introduce bias in the model, which will affect its performance. The attribute Children, on the other hand, has very few observations with missing data. For this attribute, imputing the missing data with the median for example will very likely have a very small impact, which makes it a very viable solution and was the choice made for this attribute. With regards to both Agent and Country, the first being a numeric attribute and the second a categorical attribute, we need to be more careful with how we approach it. An option for the attribute Agent would be to follow suit with what we've already decided to do about the Children attribute, which is imputing with the mean. For the attribute Country, because it's a categorical attribute, we can't use the same methods as before, instead the most common solution would be imputing it with the mode. This solution would, however, increase the already considerable weight of the value Portugal, which already accounts for 44% of the observations in H1 and 39% of the observations in H2. Imputing the missing values with the mode would then turn this value into 45% of H1 and 40% of H2, clearly introducing bias in our model.

After several experiments with different imputation methods, it was found that simply not using both the attributes Country, also removed by Antonio, De Almeida, et al. (2017), and Agent it would increase the model's performance, leading us to permanently remove both these attributes in addition to the attribute Company.

4.3.2. Outliers and Skewness Correction

We've decided to employ a boxplot analysis, also known as a box-and-whisker plot. A boxplot is a graphical representation of the distribution of a dataset, particularly useful for identifying outliers and understanding the spread of the data. It consists of three components: Box, representing the interquartile range (IQR) which is the range between the 25th and 75th percentiles of the data, Whiskers, extending from the edges of the box to the minimum and maximum values within a specified range, and outliers, data points outside the whiskers representing observations that significantly deviate from the typical values in the dataset.

A boxplot allows us to visually identify observations that are outliers so we can then decide what to do with them. For each feature we can define a point in the distribution outside of the whiskers where we can "slice" the dataset, removing all the observation outside of the slice from the dataset, thus normalising the distribution of the dataset which will help the algorithms read and interpret the data.

We've employed this technique across all the numeric features of the dataset and effectively removed 445 (0.01%) observations from H1 and 1099 (0.01%) observations from H2, thus improving the performance of the models at a very low cost.

Analysing the skewness is the following step in our pre-processing. It is a statistical technique used to assess the symmetry or asymmetry of the distribution of a dataset. It helps in understanding the shape of the data distribution by indicating whether the data is skewed to the left (negatively skewed or below -0.5), skewed to the right (positively skewed or above 0.5), or approximately symmetric (no skew or between -0.5 and 0.5). Skewness analysis may lead to transformation techniques like log transformations, Box-Cox transformations, or power transformations to make the data more symmetric, and consequently improve the performance of our models. We then proceeded to analyse the skewness of all numeric attributes in our datasets in order to identify on which attributes we may need employ a technique to correct the distribution. The results of this analysis can be seen in table 4.

Attributes	Skewness in H1	Skewness in H2
Lead time	1.138102	1.388425
Arrival Date Year	-0.186531	-0.253379
Arrival Date Week Number	-0.10180	-0.007999
Arrival Date Day of Month	0.003808	-0.006743
Stay in Weeknights	0.920129	1.133401
Stay in Weekend Nights	0.585508	0.616243
Adults	-0.582349	-0.424865
Children	3.493425	4.286455
Babies	8.413682	14.738085
Previous Cancellations	21.874668	17.748493
Previous Bookings not Cancelled	10.335358	14.537359
Booking Changes	3.276799	3.950738
Days in Waiting List	15.492367	8.308396
ADR	1.227443	0.779046
Required Car Parking Spaces	2.109080	6.216713
Total of Special Requests	1.212903	1.368866

Table 4 - Skewness of attributes in H1 and H2

Both in H1 and H2, the Lead Time attribute exhibits positive skewness, suggesting that a significant portion of bookings has shorter lead times in both time periods. Arrival Date Year, Arrival Date Week Number, and Arrival Date Day of Month all show slightly negative skewness in both H1 and H2 which indicates a tendency for bookings to be evenly distributed over the corresponding time periods. Both Stay in Weeknights and Stay in Weekend Nights exhibit positive skewness in H1 and H2, which suggests that shorter stays are more common than longer ones. Adults, Children, and Babies all have negative skewness, indicating that the majority of bookings involve fewer adults, children, and babies, with only a small number involving more. Previous Cancellations and Previous Bookings not Cancelled have notably positive skewness values, suggesting that a few bookings account for a large number of cancellations and non-cancellations, respectively. Booking Changes shows slight positive skewness in both H1 and H2, indicating a preference for fewer booking changes. Days in Waiting List has notably positive skewness in both periods, indicating that a few bookings spend a considerable number of days on the waiting list. ADR exhibits positive skewness, suggesting that lower daily rates are more common than higher rates in both time periods. Required Car Parking Spaces and Total of Special Requests both show positive skewness, indicating that most bookings do not require parking spaces or have special requests. However, there is a rightward skew, suggesting that a few bookings require more parking spaces or have multiple special requests.

Upon analysing all the variables we’ve noticed that not all of them have enough skewness to warrant correction. We´ve decided that we would only apply correction techniques to variables that have a skewness below -0.5 or above 0.5, which would mean that we would apply the correction to the following variables: Lead Time, Stays in Weekend Nights, Stays in Weeknights, Children, Babies, Previous Cancellations, Previous Bookings not Cancelled, Booking Changes, Days in Waiting List, ADR, Required Car Parking Spaces, and Total of Special Requests. As it happens, all the variables that need correcting have a positive skew, except for Adults which has a negative skew but because the skewness is so close to the threshold, we’ve decided not to correct this variable. After several tries, we found that including the variables Children, Babies, and Days in Waiting List would create poorer models, despite correcting the skewness, hence we’ve decided to simply eliminate these variables from the model. In this scenario, where all the variables that need correction have a positive skew, the correct approach to employ is to use the function Cube root/Square root and achieve a more normalised distribution, improving the results of our models. The corrected variables achieved a better skewness as shown in table 5. The result of the correction is very positive, particularly in H1 where the majority of the attributes got the skewness to within the -0.5 to 05 range, which will help achieving higher predictive results.

Attributes	Skewness in H1	Skewness in H2
Lead Time Tr	-0.226038	-0.217955
Stays in Weekend Nights Tr	-0.245831	0.109007
Stays in Weeknights Tr	-0.324063	-0.666933
Booking Changes Tr	1.679129	2.346631
Adr Tr	0.199147	-1.197949
Total of Special Requests Tr	0.471647	0.652126
Previous Cancellations Tr	12.469545	4.126289
Previous Bookings not Cancelled Tr	5.630211	9.754246

Table 5 - Attributes with corrected skewness in H1 and H2

4.3.3. Normalization and Feature Engineering

Reaching the final step of the pre-processing, Normalization and Feature Engineering, we start off by encoding the categorical attributes so that these can be better interpreted by the algorithms that we’ve selected for the predictive models. We´ve decided to use the technique One Hot Encoder, which will convert every value in each categorical attribute into several binary attributes, recording the presence or not of the value in each row. The attributes selected for One Hot Encoder are Arrival Date

Month, Meal (except the Meal Package Prediction Model), Market Segment, Distribution Channel, Deposit Type, Reserved Room Type, and Customer Type.

The next step is deciding which attributes won't be used to build the model, by trial and error and comparing the results of each try. After many attempts it was decided that to have the best performing model, we needed to eliminate the following attributes: Arrival Date Year, Is Repeated Guest, Hotel, Country, Children, Babies, Assigned Room Type, Reservation Status, Reservation Status Date, Required Car Parking Spaces. We´re also eliminating original attributes that underwent skewness correction such as Lead Time, Stays in Weekend Nights, Stays in Weeknights, Previous Cancellations, Previous Bookings not Cancelled, Booking Changes, Days in Waiting List, ADR, and Total of Special Requests as we're using the corrected version of these.

We´ve removed the Assigned Room Type, Reservation Status, and Reservation Status Date because these attributes are dynamic, meaning that the values on each attribute will change from the moment of the booking to the moment of Check Out, which means that the values captured in our dataset will only correspond to the moment of Check Out, rendering the model incapable of capturing the reality. The attributes Children, Babies and Is Repeated Guest (in all models except the predictive of Guests Returning) were eliminated because there's little variation in the values, meaning that Children have over 91% of its observations with the value zero in H1 and 94% in H2, Babies has over 99% of its observations with value zero in both H1 and H2, and Is Repeated Guest has 96% of its observations with value zero in H1 and 97% in H2. The attribute Arrival Date Year and Required Car Parking Spaces had no effect on the results of the predictive models, hence it was removed. With the Data Preparation finished for both the H1 and the H2 datasets we're now ready to set up the algorithms, run them over both datasets and analyse their performance.

4.4. Modelling and Evaluating

After all the data analysis and processing we´ve reached the phase of CRISP-DM where the model will be developed. We're focusing on four machine learning algorithms: Decision Tree, Random Forest, XGBoost, and Neural Networks. Random Forest (Andriawan et al., 2020; Sánchez-Medina & C-Sánchez, 2020) and XGBoost (Andriawan et al., 2020; Antonio, De Almeida, et al., 2017, 2019) have been amply used in past research, solidifying them as a choice for the model being developed in this research. Neural Networks and Decision Tree are also used by (Antonio, Almeida, et al., 2017) with good results, and after a series of experiments it was found that both had promising results and kept in this research. We will be evaluating the performance of each algorithm using the K-fold Cross-Validation method (Hastie et al., 2001) and decide which one behaves the best regarding the performance metrics that best suits each specific model.

The K-fold Cross-Validation method (Hastie et al., 2001) is a technique used to evaluate the performance of a predictive model by dividing the dataset into k subsets (or folds). The model is then trained and evaluated k times, each time using a different subset as the test set and the remaining data as the training set, with the results from each iteration being averaged to obtain a more reliable performance estimate. The value of k is a parameter that can be tuned based on the size of the dataset and computational resources, commonly k being 5 or 10, however it can be any other values depending on the specific requirements of the analysis. K-fold is developed in four steps:

- Data Splitting - The dataset is divided into k subsets or folds. Each fold contains an equal or approximately equal proportion of the data.
- Model Training and Testing - The predictive model is trained k times, each time using k-1 folds as the training set and the remaining fold as the test set. This ensures that the model is tested on different subsets of the data.
- Performance Evaluation - The model's performance is evaluated on each iteration using a chosen metric (e.g., accuracy, precision, recall). This provides k performance scores.
- Average Performance - The performance scores from each iteration are averaged to obtain a single performance estimate for the model.

The use of K-fold Cross-Validation has several advantages, namely we can have reduced bias by using multiple subsets for training and testing helps reduce the bias that may result from a single train-test split; creates a more robust performance estimate, as averaging the performance across multiple iterations provides a more robust estimate of the model's performance; also maximizes data usage, since each data point is used for testing exactly once, it maximises the use of available data for both training and testing.

In the context of CRISP-DM, incorporating the k-fold cross-validation method during the model evaluation phase enhances the reliability of the predictive model's performance assessment by aligning with the iterative and thorough nature of the CRISP-DM methodology, it ensures that the predictive model's capabilities are rigorously tested and validated across various subsets of the data. This method contributes to the overall robustness and generalization of the predictive model, making it more applicable to unseen data. In our project, all models will have a k of 10 as it was found to be the best performing one, as was seen in Antonio, Almeida, et al., (2017) and Andriawan et al., (2020).

4.4.1. Hyperparameter Tuning

Hyperparameter tuning involves the optimization of model settings to achieve superior performance, relying on the expertise and intuition of domain experts who possess an intricate understanding of the problem space and the hyper parameters of each algorithm. This process often involves an iterative exploration of hyperparameter values in which experts adjust parameters based on model performance feedback, refining their choices with each iteration. This iterative nature provides great flexibility, enabling practitioners to experiment with unconventional or problem-specific hyperparameter values, depending on the context of the analysis. Despite its merits, manual tuning has challenges, as it is time-consuming, particularly for complex models with numerous hyperparameters, relies on the expertise of the practitioner, and there is a risk of bias or oversight in hyperparameter selection.

Algorithm	Parameter	Values
Decision Tree	**Max Depth**	**5, 10, 15, 20, 25, 30**
Decision Tree	**Max Leaf Nodes**	**2, 4, 6, 8**
Decision Tree	**Splitter**	**Best, Random**

Table 6 - Hyperparameter tuning of Decision Tree

Regarding the parameters of the Decision Tree seen in table 6, we tested three of the parameters, namely the Max Depth which allows the tree to develop until all leaves are pure, Max Leaf Nodes which allows for the tree to limit the number of nodes and the Splitter which is the strategy to split each node. After several iteration the values defined for the best performing parameters are twenty for Max Depth, four for Max Leaf Nodes and Best for the Splitter.

Algorithm	Parameter	Values
Random Forest	Number of Estimators	3, 5, 7, 10, 15, 20
Random Forest	Max Depth	20, 50, 100, 150, 200
Random Forest	Max Leaf Nodes	2, 4, 6, 8

Table 7 - Hyperparameter tuning of Random Forest

The parameters of the Random Forest that were tuned, seen in table 7, are the Number of Estimators which limits the number of trees, Max Depth and Max Leaf Nodes which are the same as for the Decision Tree. The best iteration was developed with the number of estimators as ten, max depth as one hundred, and the max leaf nodes as six.

Algorithm	Parameter	Values
Neural Network	Activation	Regression, Logistic
Neural Network	Solver	Lbfgs, Sgd, Adam
Neural Network	Hidden Layer Sizes	(5,2), (10,2), (15,2), (5,4), (10,4)

Table 8 - Hyperparameter tuning of Neural Network

The tuning of the Neural Network parameters, as seen in table 8, are the Activation which specifies the activation function for the hidden layers, the Solver where it’s selected the optimization algorithm to use for weight optimization during training, and the Hidden Layer Sizes, which sets the number of neurons in each hidden layer. The best results were obtained when the Activation it was set as logistic, the Solver was set as Adam, and the Hidden Layer Sizes were defined as five coma two.

Algorithm	**Parameter**	**Values**
XGBoost	Number of Estimators	50, 100, 200, 300, 400
XGBoost	Max Depth	2, 3, 5, 10
XGBoost	Subsample	0.3, 0.5, 0.7, 0.9
XGBoost	Colsample	0.3, 0.5, 0.7, 0.9
XGBoost	Eval Metric	Error

Table 9 - Hyperparameter tuning of XGBoost

Lastly, in table 9, regarding the hyperparameters of XGBoost, we tuned the Number of Estimators, Max Depth, Subsample which is the ratio of the training instances, Colsample which specify the fraction of columns to be subsampled and the Eval Metric which is a metric to evaluate the data during training. The optimal results were achieved by tuning the Number of Estimators to three hundred, the Max Depth to three, the Subsample, to zero point nine, the Colsample set to zero point nine, and the Eval Metric as error.

5. RESULTS AND DISCUSSION

With the models properly trained and tested using a Decision Tree, Random Forest, XGBoost, and Neural Networks, we're now going to analyse and comment their performance using a confusion matrix, a table that shows a comparison between the predicted classification of the target variable versus the actual classification (Bradley, 1997).

Regarding the Booking Cancellation model and the Guest Return model we're going to pay particular attention to Recall (or Sensitivity), which measures the model´s ability to predict True Positive observations over the total number of Positive observations (Amari & Wu, 1999), and F1 Score, which measures the mean of precision and recall, offering a balanced evaluation of the model's performance (Amari & Wu, 1999). Such evaluation in a booking cancellation model has been successfully tested by Sánchez-Medina & C-Sánchez, (2020).

When evaluating the Food Package model, because it's a multiclass prediction model and we're interest in having a balanced prediction between the different classes, we're going to focus on F1 Score and AUC (Huang et al., 2015). F1 Score and AUC are the most important measures as it allows us to have a clear picture of how accurately the model is predicting each individual class and how balanced the predictions are across the different classes.

Recall is particularly important, from a Marketing perspective, as we're dealing with a cost-inducing set of actions and we're looking to be as efficient as possible with this cost. As such, contacting clients who are predicted to Cancel a booking, purchase a Meal package or Return (Positive Results) will be the preferable objective of our actions, as these are the ones with the potential to improve our business. F1 Score is also very relevant as it includes Precision (accuracy of True Positives), therefore creating a single measure that can evaluate if our model can predict True outcomes accurately, which can then be used to empower the communication strategy of a Marketing department by accurately segment all guests before they've arrived at the hotel.

5.1. Cancelation Predictive Model

The ability to predict cancelations gives hotels the clear possibility to actively manage the Occupancy of the Rooms and therefore have more control of the bottom-line results. From a Marketing perspective, the capacity to actively expect which exact guests are likely to cancel gives the Marketing Manager a superb tool of communication: creating customised offers and communicate directly to these clients. The Marketing Manager can have more clear and tangible goals, by accurately assessing the potential of retaining otherwise lost guests, and budget more appropriately the cost of retaining these guests by carefully building offers that have maximum value for the smallest cost.

The predictive results obtained using a XGBoost algorithm are very positive, although there's a much better performance in H2 than H1 as demonstrated in table 11, particularly regarding the more relevant metrics for this model: Recall at 80% and F1 Score at 78% in H1 and Recall at 83% and F1 Score at 83% in H2. These results reinforce the point made by several papers ((Antonio, Almeida, et al., 2017; Antonio, De Almeida, et al., 2017) that hotels PMS are excellent sources of data and enable hotels to accurately make predictions, particularly using XGBoost. Despite the imbalance of the labels of the target variable, as seen previously, applying a rebalancing technique like SMOTE was found to have little to no impact on results of the model, and as such it was decided to keep the original distribution.

Hotel	Algorithm	Accuracy	Precision	Recall	F1 Score	AUC
H1 Test Results	Decision Tree	0.8088	0.7615	0.7512	0.7560	0.80
	Random Forest	0.7586	0.8750	0.5622	0.5393	0.77
	XGBoost	0.8088	0.7658	0.8039	0.7782	0.88
	Neural Network	0.7889	0.8245	0.6302	0.6452	0.79
H2 Test Results	Decision Tree	0.8218	0.8176	0.8123	0.8145	0.85
	Random Forest	0.7608	0.8543	0.7092	0.7107	0.83
	XGBoost	0.8406	0.8388	0.8296	0.8333	0.91
	Neural Network	0.8089	0.8072	0.7943	0.7988	0.87

Table 10 - Results of Cancellations predictions in H1 and H2 as average of 10-Fold Cross Validation

As shown in the table 10, regarding the H1 model, the Decision Tree algorithm achieved an accuracy of 81%, indicating that it correctly predicted most of the outcomes. The precision, recall, and F1 score are reasonably balanced, all around 75%, suggesting that it has a good ability to correctly identifying positive True Positives (precision) while also capturing True Negatives (recall). The Random Forest algorithm achieved an accuracy of 76%, which is slightly lower than the Decision Tree, however, it has a high precision of 88%, meaning that when it predicts a positive case, it's often correct. The recall is much lower, at 56%, indicating that it struggled to capture all negative cases, which then affects the F1 Score, scoring a low value of 54%. The XGBoost algorithm matches the Decision Tree's accuracy at 81%, although it has a high precision and recall, at 77% and 80% respectively, resulting in an F1 score of 78%, slightly higher than the Decision Tree, which suggests that it captures both positive and negative outcomes quite well, resulting in a balanced predictive ability. The Neural Network model achieved an accuracy of 79%, while having high precision of 82% but a lower recall of 63%. Finally, the F1 score is 65% which suggests a reasonable balance between precision and recall.

Moving on the H2 dataset, the Decision Tree algorithm achieved an accuracy of 82%, which is slightly higher than for H1. The precision, recall, and F1 score are all notably high, at 82%, 81% and 81% respectively, indicating that it performs well in identifying both positive and negative outcomes. The Random Forest algorithm has an accuracy of 76%, which is slightly lower than the Decision Tree but still interesting. Its precision is high at 85.43%, indicating a low rate of false positives while the recall is lower at 71%, and the F1 score is 71%, which suggests that it can captures a significant portion of true outcomes. XGBoost, however, outperforms the other algorithms in H2 with an accuracy of 84% while scoring high precision at 84%, recall at 83%, and F1 score at 83%, indicating a robust and balanced performance. The Neural Network algorithm achieved results close to the Decision Tree algorithm, with an accuracy of 81%, a precision of 81%, recall of 79% and F1 Score of 80%, however it's slightly lower.

In both datasets, XGBoost consistently performs well, achieving the highest accuracy and balanced precision and recall scores. The Decision Tree models also performed admirably, particularly in H2, which further validates its use in predictive models in hospitality. The Random Forest was somewhat disappointing, even though it had the highest precision, but the lower recall an F1 Score compared to other algorithms makes it somewhat unsuitable, contrarily to what was proposed by Andriawan et al., (2020). The Neural Network models offers interesting results, particularly in the H2 model, which suggest that its use should be seriously considered in projects with a similar scope as this one.

To better understand the results of predictions shown previously, focusing on the best performing algorithm, the XGBoost, figure 12 presents the confusion matrix of H1. The effectiveness of the model in predicting True Negatives (Guests predicted to not cancel and didn't) and True Positives (Guests predicted to cancel and did) is evident, with H1 capturing 59% out 72% True Negatives and 22% out of 27% of True Positives, hence the accuracy of 81%. Interestingly, H1 has identified 13% of False Positives (non-cancelled bookings as cancelled bookings), which could constitute a need for improvement should this model be designed for forecasting, as it is overestimating the number of cancelations. The low error rate of 6% in False Negatives (guests predicted to not cancel but have cancelled) is a good result, from a Marketing perspective, as these guests would likely not be contacted, and therefore, we would lose the opportunity to either have a chance to keep them or be able to just confirm they're cancelation.

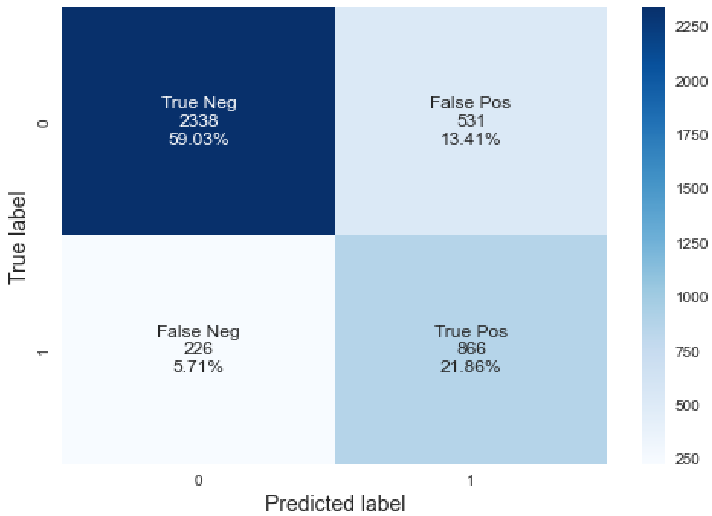


Figure 12 - Cancellation Confusion Matrix in H1

The hotel H2 has a better performance than H1, as reinforced by the figure 13. Upon looking at this confusion matrix, it's clear that the accuracy is higher than H1, capturing 52% out of 59% of True Negatives and 31% out of 41% of True Positives. Contrastingly to H1, this model is much better at avoiding False Positives, scoring a meagre 6%, half of the H1 model, which makes for a much better ability to forecast. Slightly worse than H1 is the scoring of False Negatives at 9%, even though the Recall evaluation is higher. This is because this hotel has on one hand, much more observations (nearly double) and on the other the distribution of Cancelled and Not Cancelled is also much more balanced.

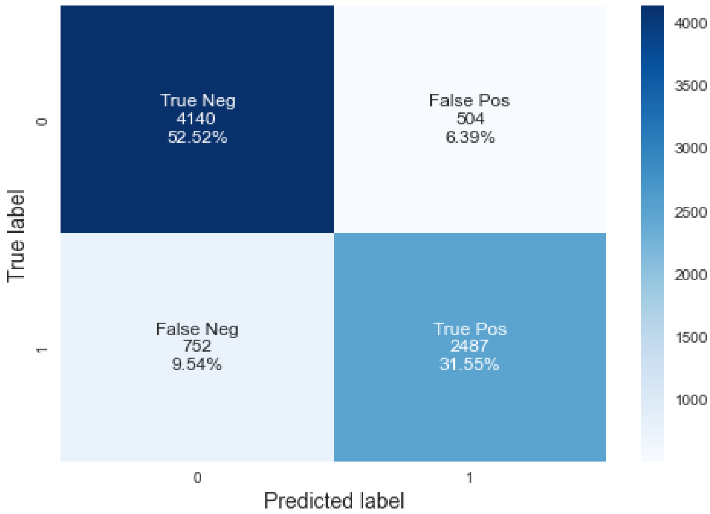


Figure 13 - Cancellation Confusion Matrix in H2

The ability of our model to predict both the cancelation and not cancelation of H1 are further shown when observing the AUC-ROC curve in Figure 14. Here it's visible that the model has the same rate of True Positives over False Positives in predicting both outcomes, scoring 88% in both, which is an excellent result and shows that it can correctly discriminate between both outcomes and have a good prediction.

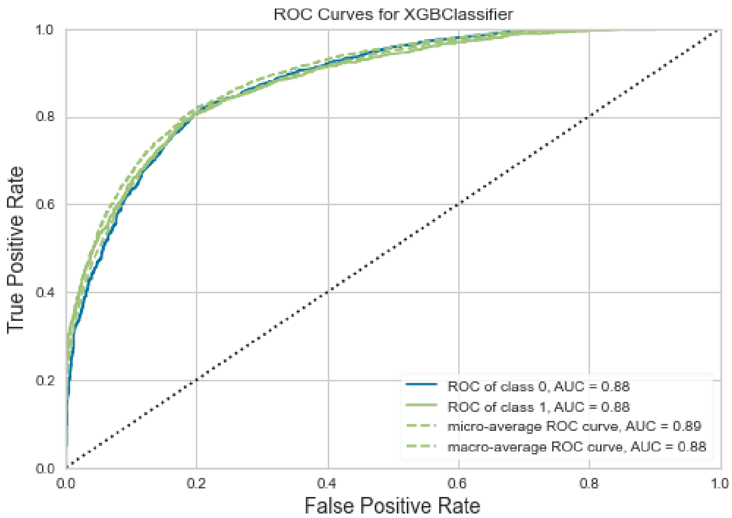


Figure 14 - Cancellation Area Under the Curve of H1

When analysing the AUC-ROC of H2, it's clear that this model has a higher performance, particularly when predicting if the guest is not cancelling, with a score of 91%. This score shows an excellent ability to discriminate both outcomes, although curiously starts off better at predicting the cancellation and, as more predictions occur, it gets better at predicting no cancellation, no doubt influenced by the higher number of train observations of no cancellations.

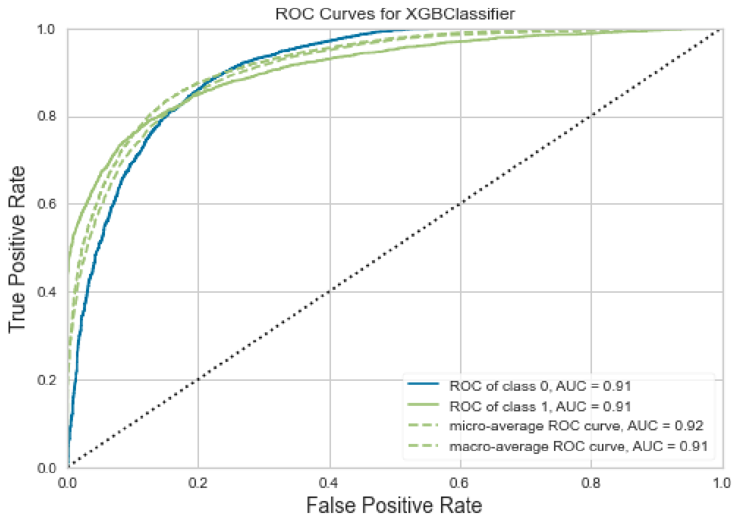


Figure 15 - Cancellation Area Under the Curve of H2

5.1.1. Model comparison with similar papers

Creating predictive models for cancellations in hospitality is not a new idea, with some papers previously published with model proposals with good results. As we're proposing a new one specifically built for marketing and communication, which is new and is evaluated using specific measures as seen previously, we'll now compare our results with Antonio et al. (2019) in table 11.

Hotel	Model	Accuracy	Precision	Recall	F1 Score	AUC
H1	Antonio et al. (2019)	0.8486	0.8205	0.6128	0.7016	0.88
	Proposed Model	0.8088	0.7658	0.8039	0.7782	0.88
H2	António et al. (2019)	0.8563	0.8731	0.7862	0.8274	0.92
	Proposed Model	0.8406	0.8388	0.8296	0.8333	0.91

Table 11 - Comparison of Proposed Model with previous study

The results achieved by this model are similar to the model produced by Antonio et al. (2019), with an improvement in Recall from 61% to 80% in H1 and 78% to 82% in H2. This improvement is no doubt related to the work developed in the Data Preparation step, where the Outlier Treatment and Normalization were explored furthered by addressing a larger number of variables, consequently achieving a more normalised distribution of the dataset. It also reflects the importance of determining the purpose of the model in the early stages of development, as it allows us to focus on specific measures in accordance with the defined purpose. In our case, as the focus is making predictions for marketing communications, such as upselling for example, achieving a higher Recall is fundamental, which is reflected in the results previously discussed.

5.2. Meal Package Predictive Model

When guests make a booking, it's required that a Meal Package is chosen, which then becomes a part of the stay experience and contributes towards the Revenue per Room. Having a Prediction on the what's the most suitable Meal Package for each guest is an interesting method to assess the ability to upsell, as we have a clear indicator of which Meal Package is most suitable for each guest. The Marketing Department can then create tailor-made offers that have a much higher potential of adding value to the guests stay, without increasing the marketing costs.

Table 12 shows the prediction results of all algorithms in both H1 and H2 of the Meal Packages. As we can see the XGBoost is the best performing algorithm in both models, achieving a F1 Score of 66% and AUC of 92% in H1 and an F1 Score of 85% and AUC of 96% in H2.

Hotel	Algorithm	Accuracy	Precision	Recall	F1 Score	AUC
H1 Test Results	Decision Tree	0.8348	0.6599	0.6028	0.6252	0.76
	Random Forest	0.7245	0.3063	0.3175	0.3114	0.73
	XGBoost	0.8551	0.8406	0.5758	0.6561	0.92
	Neural Network	0.7975	0.3578	0.3019	0.3064	0.73
H2 Test Results	Decision Tree	0.9076	0.8249	0.8639	0.8394	0.92
	Random Forest	0.6936	0.4741	0.4403	0.3820	0.84
	XGBoost	0.8961	0.9166	0.8072	0.8508	0.96
	Neural Network	0.8027	0.5168	0.3463	0.3586	0.80

Table 12 - Results of Meal Package predictions in H1 and H2 as average of 10-Fold Cross Validation

Taking a closer look at the performance of each algorithm in H1, starting with the Decision Tree, it achieved a good accuracy of 83%, however the remaining measures have a low score, with precision scoring 66%, recall at 60% and F1 Score at 63%, which means that this algorithm is not great at predicting the meals that guests would choose. This is further confirmed by the AUC of 76%, which shows that it had some difficulty in accurately distinguishing between labels. The Random Forest algorithm, although having a relatively good score of 72% in accuracy, is very poor in all other measures, scoring only a 31% in precision, a 32% in recall and a 31% in F1 Score, which means that it definitely is not suitable for this prediction. Moving on to XGBoost, which had a good performance at accuracy at 86% and precision at 84% but performed poorly at recall with a score of 58%, showing a poor ability to predict negative outcomes. This then influenced the low F1 Score, with 66%, reflecting this issue, although the AUC is very good at 92%. Lastly, the Neural Network, which performed very similarly to the Random Forest, if a little better in its accuracy at 80%. This measure, however, is very misleading, as the remaining measures show, with precision at 36%, recall at 30% and F1 Score at 30%, showing that this algorithm is also unsuitable for this prediction.

With regards to H2, as it happened in the Cancellation model, H2 gives us much better performances overall. The Decision Tree has very good results and it's a close contender to XGBoost as the best algorithm, boasting an impressive 91% at accuracy, 82% at precision, 86% at recall and 84% at F1 Score. It makes for an excellent choice to achieve the desired predictions, reinforced by a 92% score on the AUC. Similarly to H1, Random Forest and Neural Network performed poorly, with the first having 69% accuracy, 47% precision, 44% recall and 38% F1 Score and the second having 80% accuracy, 52% precision, 35% recall and 36% F1 Score. As seen previously both are unsuitable for this prediction. Lastly, the XGBoost, which boasts an accuracy of 90%, a precision of 92%, a recall of 81% and an F1 Score of 85%. Despite the similarity of scores between the Decision Tree and XGBoost, the Decision Tree has a much better recall, although this is not enough to outperform XGBoost in F1 Score, which also has a better AUC at 96% when compared to the Decision Tree which has 92%.

Much like the Cancellation prediction, the algorithm XGBoost is the best performing algorithm. As such we will delve a little deeper into the performance of this algorithm by analysing the Confusion Matrix and the graphical display of AUC-ROC. Figure 16 shows the Confusion Matrix of H1, which demonstrates the cross between predicted label and the true label, allowing us the understand how good the model is at predicting the correct result. Because this is a Multiclass prediction, we need to be careful how we're interpreting the confusion matrix, as the combinations that create False Positives and False Negatives are higher. When analysing the BB option, which is also the one with the highest number of observations, we see that there's an excellent capture rate of True Positives, with 97% of BB observations correctly identified. Regarding the outcome label FB and SC, the labels with the least observations to predict, the capture rate is considerably lower, with only 62% of observations correctly captured of FB and only 25% of observations correctly captured of SC. It seems plausible that a lack of observations has impacted this model's ability to correctly predict these outcomes. Lastly, the model also had some difficulty in predicting the outcome label HB, capturing only 46% of observations correctly. Interestingly, the model misidentified 54% of observations that should have been HB but were instead identified as BB, which suggests that the model has some difficulty distinguishing between BB and HB.

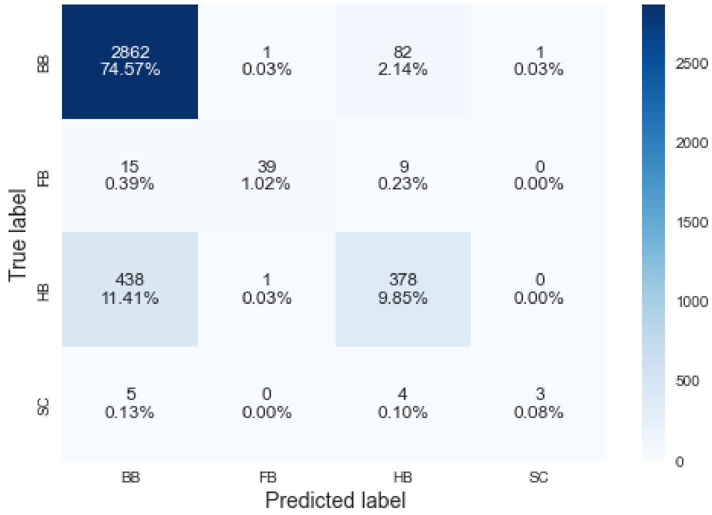


Figure 16 - Meal Confusion Matrix in H1

Overall, the high performance of this model is mostly due to the success of the BB and FB predictions, which represent 76% of true positives. It is interesting to see that the vast majority of observations that were misclassified were labelled as BB, which makes sense since there's a lot more observations BB than all other outcomes combined. The lack of success in SC seems to be mainly due to lack of observations and in HB its inefficacy would need further analysis, so we can understand how to improve this prediction.

Moving on to H2, represented in figure 17, we can see an excellent ability to capture True Positives regarding BB, 97% of observations, similarly to H1. When looking at the FB option, contrary to H1, this model can perfectly capture all observations, despite the extremely low number of observations. Also contrary to H1, SC has much more observations which allows for better predictions, as seen by the capture rate of 60% of observations. Lastly, with regards to HB, there's a much better performance as well, as this model has a capture rate of 66% of observations.

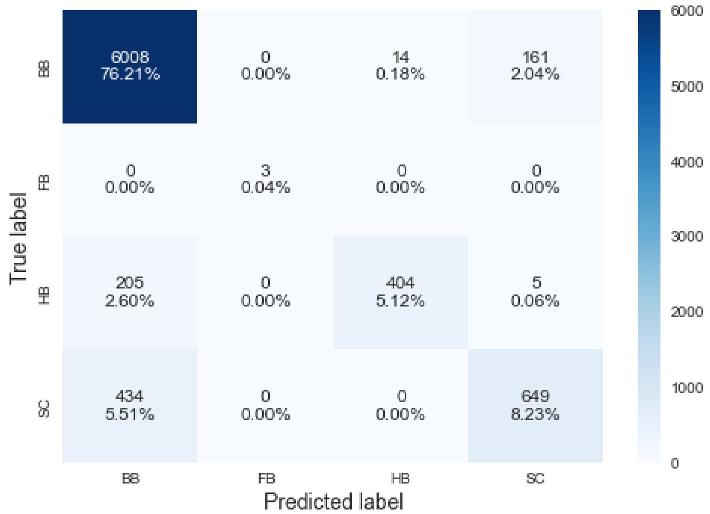


Figure 17 - Meal Confusion Matrix in H2

This confusion matrix helps us understand how the H2 model has a better performance across all four outcomes, particularly when comparing the HB and SC with the H1 model. This may be mostly due to the difference of both the overall number of observations and in particular SC labelled observations. The differences between H1 and H2 can be further analysed by looking at the AUC-ROC of models, in figure 18 and figure 19.

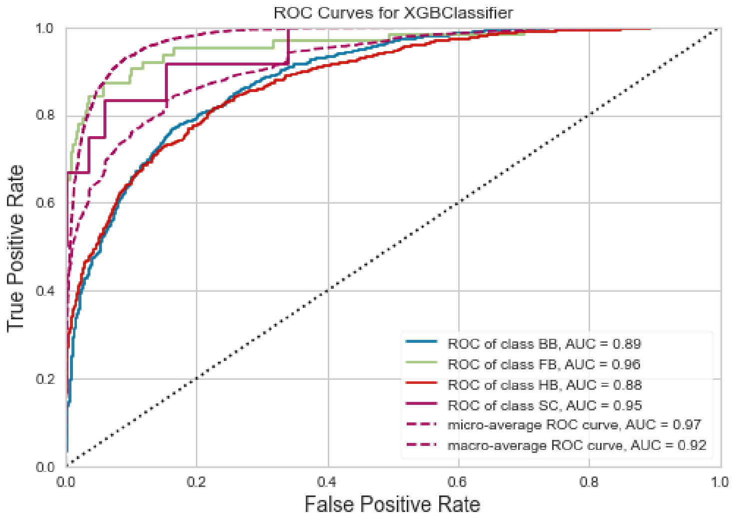


Figure 18 - Meal Area Under the Curve of H1

Figure 18 shows how the reduced number of observations of both FB and SC impacts the model's capacity to predict these outcomes, as the curve appears to be cascading instead of an actual curve. It would be ideal to have more observations with these outcomes in the dataset so we could strengthen the model's ability to predict them, although it's a good sign that, despite this issue, it was able to still detect these outcomes correctly in several instances. The curve of both BB and HB labels looks interesting as it shows a good rate of prediction across number of instances, with the BB being better as it has considerable proportion of observations.

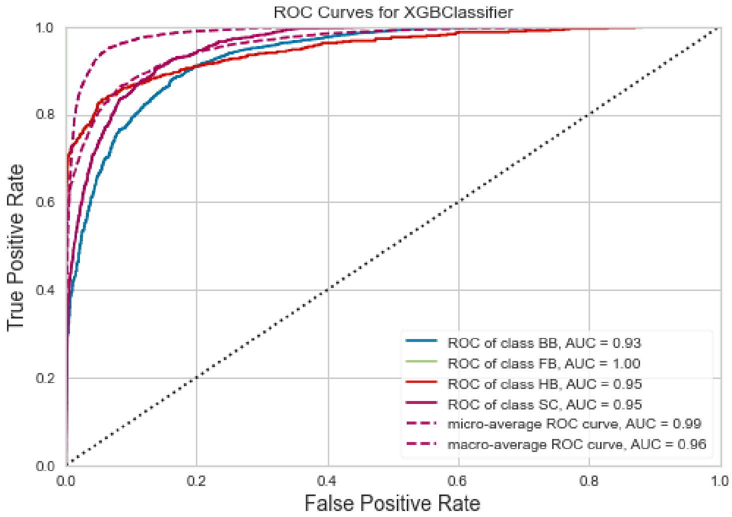


Figure 19 - Meal Area Under the Curve of H2

When analysing the ROC curves of H2, the model behaves better, with proper curves across all outcome labels, even though the label FB has only three observations, which helps explaining the perfect result of 1. With regards to the remaining outcome labels, it's interesting to see that HB starts off as having a superior performance but as more predictions occur it ends up as the least good curve, contrarily to SC and BB, improve with more predictions.

5.3. Guest Returning Predictive Model

Returning guests are the clear objective of many, if not all, hotelier groups, as evidenced by the many loyalty schemes currently in existence. it clearly indicates that the guest was satisfied with the previous stay, which from a marketing perspective means that the customer journey was very successful. Predicting if a guest will return or not is then a very valuable insight for a Marketing Department, as it allows it to understand if there's a good possibility of increasing the lifetime value of a guest without spending more money on guest acquisition, thus improving the return on each guest tremendously.

Table 14 provides the results of the predictions for both H1 and H2, with all algorithms showing very good results, with the exception perhaps of the Random Forest. This much success may be a consequence of the extremely high number of observations registering the label no return, seen previously to be as high as 97%, which may indicate a limitation to this dataset regarding its suitability to train and test a model with the purpose of predicting returns. However, the oversampling smote technique was applied before the k-fold method to split the dataset, thus synthetically increasing the number of observations of the minority outcome, labelled returned, and creating a more balanced dataset. Similarly to the previous models, the algorithm XGBoost was the most successful one, scoring 93% at recall and 84% at F1 score.

Hotel	Algorithm	Accuracy	Precision	Recall	F1 Score	AUC
H1 Test Results	Decision Tree	0.9734	0.8198	0.8775	0.8459	0.86
	Random Forest	0.8745	0.5860	0.7555	0.6113	0.87
	XGBoost	0.9668	0.7785	0.9298	0.8356	0.97
	Neural Network	0.9650	0.7715	0.9084	0.8241	0.96
H2 Test Results	Decision Tree	0.9936	0.9153	0.9613	0.9370	0.96
	Random Forest	0.9604	0.6781	0.8837	0.7388	0.98
	XGBoost	0.9912	0.8745	0.9778	0.9195	1
	Neural Network	0.9936	0.9299	0.9386	0.9342	0.99

Table 13 - Results of Guest Return predictions in H1 and H2 as average of 10-Fold Cross Validation

As table 13 presents, H1's Decision Tree has very good results, with an accuracy of 97%, precision at 82%, recall at 88%, F1 Score at 85%. These results show that this algorithm is suitable for predicting the return of guests, although the score of the AUC is not as good as other algorithms. Random Forest is the worst performing algorithm, despite the impressive accuracy of 87%, which is counterbalanced by a very low precision at 59%, a recall at 76% and a F1 score of 61%. With such a low score at both recall and F1 score this algorithm cannot be considered for this prediction. Regarding XGBoost, the results a very good, if slightly similar to Decision Tree. The scores are 97% at accuracy, 78% at precision, 93% at recall and 84% at F1 score, meaning that XGBoost outperformed the Decision Tree at recall, but underperformed at precision and F1 score, which means that both these models would do well at this prediction. XGBoost is chosen as the preferential algorithm due to increase score at AUC, at 97%, which greatly outperforms the Decision Tree by 10%. Lastly, the Neural Network also performed well, with an accuracy of 97%, precision at 77%, recall at 91% and F1 score at 82%, which puts it at a similar performance level to the Decision Tree and XGBoost, and can therefore be considered a suitable algorithm for this prediction.

H2 results are very good, with all algorithms performing well. The Decision Tree has excellent scores, with accuracy at 99%, precision at 92%, recall at 96% and F1 score at 94%, which confirm that the Decision Tree is an excellent algorithm for this prediction. The Random Forest performed well, scoring 96% accuracy and 88% at recall, although its precision and F1 score are somewhat lower, at 68% and 74% respectively, making this algorithm the lowest performing one. The XGBoost is the best performing again, scoring an impressive 99% at accuracy, 87% at precision, 98% at recall and 92% at F1 score, which similarly to the H1 model, means that this algorithm is very similar to the Decision Tree, outperforming it in recall but underperforming in F1 score. The AUC score is higher in XGBoost, which means that it would be the preferred algorithm, although the Decision Tree would be very suitable as well. Lastly, the Neural Network score is also very similar to the Decision Tree and XGBoost, scoring 99% in accuracy, 93% in precision, 94% in recall and 93% in F1 Score, making it as suitable for as both the Decision Tree and XGBoost for this prediction, although its AUC is slightly better than the Decision Tree's one.

As we've established, XGBoost is the best performing algorithm in this prediction and as such we'll be analysing its confusion matrix and AUC-ROC, starting with the H1 model shown in figure 20. As we can see 93% of true negatives were captured by the model, as well as 3.67% of true Positives, giving us the accuracy of 97%. Interestingly, the model seems to be better at capturing true outcomes, demonstrated by the very low rate of false negatives of 0.46%, while the rate of false positives sits at 2.86%. Nonetheless, these results are excellent, and we can confidently apply this model confidently that it will correctly identify guests who are very likely to return.

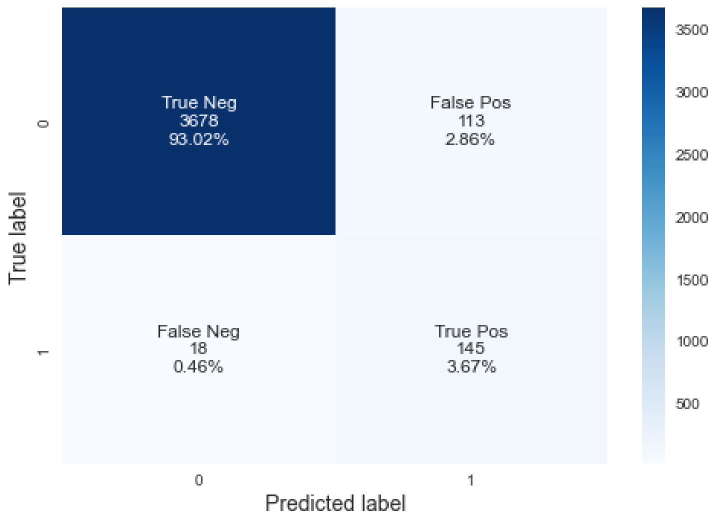


Figure 20 - Guest Return Confusion Matrix in H1

Moving on, figure 21 shows us the confusion matrix of H2, where the model is performing better than H1. In this model, the capture of true negatives is 97% and true positives is 2%, giving us the accuracy of 99%. Interestingly, this model also captures true outcomes better, as evidenced by the extremely low false negative rate of 0.09%, while the false positives has also a very low rate of 0.79%. Similarly to H1, this model is also excellent at identifying guests who will return.

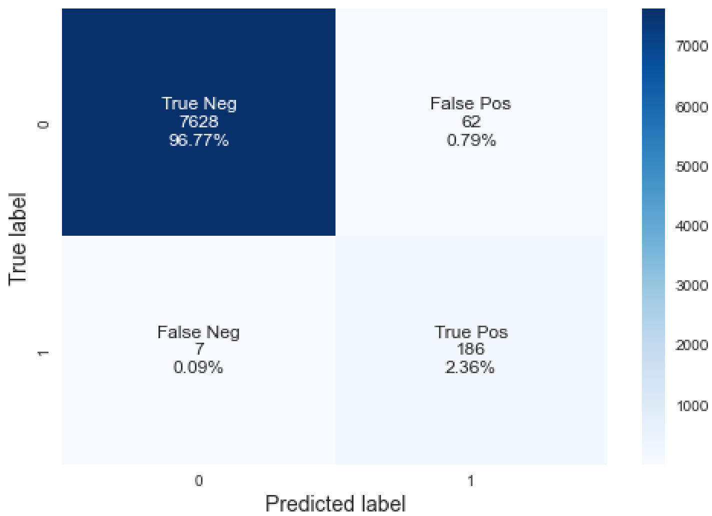


Figure 21- Guest Return Confusion Matrix in H2

Analysing the AUC-ROC of H1 in figure 22, it's interesting to observe that the curve of the label return starts off very well, but its performance decreases as more observations are predicted. Contrastingly, the curve of the label not return starts off slightly worse and improves as more observations are predicted. Overall, the performance of both curves is very good, with a macro average of 97% which indicates an excellent ability of discriminating each outcome label.

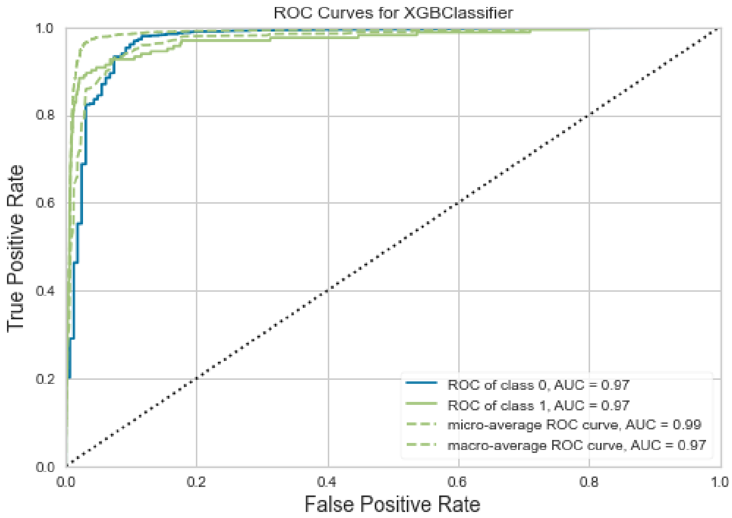


Figure 22 – Guest Return Area Under the Curve of H1

Lastly, the AUC-ROC of H2 shown in figure 23, as the results are very high, with a macro average of 100%. Due to this result, the curves are almost perfect, which implies that this model is excellent at discriminating both outcome labels.

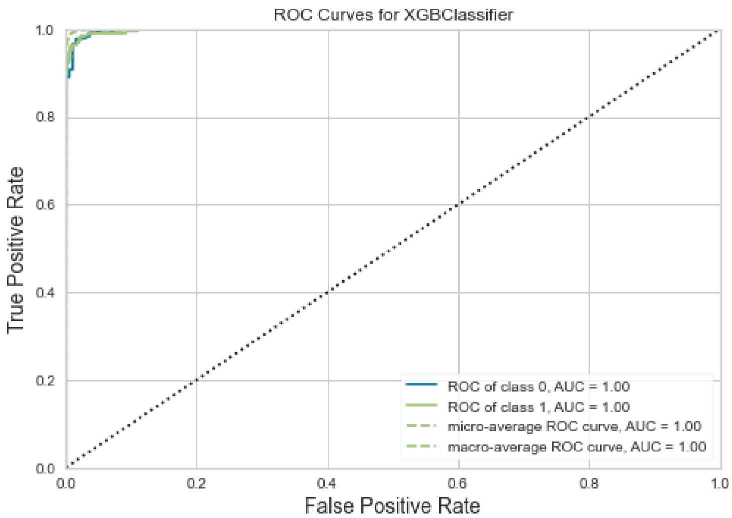


Figure 23 – Guest Return Area Under the Curve of H2

6. CONCLUSIONS, LIMITATIONS AND FUTURE WORKS

This project contributes to the current literature about AI and hospitality by proposing an experimental application of a CJF that integrates CJ and ML to predict key performance indicators throughout the journey of guests at hotels.

The main contribution of this project is to develop a CJF that enables the marketing department of two hotels, a resort hotel (H1) and a city hotel (H2), to predict guest's behaviour using classification models throughout the three phases of their journey, Pre-Service, Service and Post-Service, using data obtained from the reservation management system. The marketing department can use the predictions to improve each guest's stay by anticipating the best outcome at each of these phases for each guest and start a conversation, creating opportunities to improve the service provided to guests while also improving the bottom-line results of the hotel at very little cost.

It expands on work developed by Andriawan et al. (2020) and Antonio, Almeida, et al. (2017), who successfully developed classification models to predict cancellations using tree-based algorithms, by increasing the prediction scope to a full CJ in a hotel, building and testing separate models, divided per hotel, with each model answering each research question.

Data was collected from a reservation management system and the CRISP-DM methodology was then applied applying the following:

1. understanding the business context of each hotel and the impact the target variables have in the bottom line.
2. analysing both datasets, H1 which has 33 variables of which 10 are categorical and 23 are numeric, with 40,061 observations and H2 which has 33 features, of which 10 are categorical and 23 are numeric, with 79,331 observations.
3. then we proceeded to prepare the data so the model can better understand the patterns amongst the datasets and be able to predict the target variables, by addressing missing values, treating outliers and skewness, and normalising the distribution.
4. afterwards we chose the algorithms for the models, namely the DT, RF, XGBoost and NN, manually selecting the best hyperparameters.
5. lastly, we evaluate each model according to the goal of the predictions.

The first models predict the cancellation of bookings, one for H1 and another for H2, particularly identifying guests who will cancel, scoring a recall above 80% in both models, with a score of 80% at H1 and 83% at H2. This demonstrates the ability of both models to identify guests who are likely to cancel, thus permitting the marketing department to act before the guest cancels the booking, preventing the cancellation and retaining the booking or confirming the cancellation with a higher lead time than would otherwise have been possible.

The second models predict the food and beverage package of each booking, one for H1 and another for H2, by successfully estimating each possible package to each booking, scoring a F1 Score of 66% at H1 and a good result of 85% at H2 while also having excellent scores of AUC at 92% in H1 and 96% at H2. These results show a good ability to associate a food and beverage package to each booking, particularly the H2, creating an opportunity to upsell food packages, potentially increasing the revenue per room.

The third models predict which guests will book another stay, one for H1 and another for H2, specifically identifying guests who will book again, scoring an excellent recall above 90% in both models, with a score of 93% at H1 and 98% at H2. This demonstrates the capacity of both models to identify bookings made by guests who will likely book again, so the marketing department can identify guests to target for communication campaigns and ensuring a high success rate of acceptance.

The introduction of these models as part of the discussion and management of guest's journeys reduces behaviour uncertainty, allowing the hotel to adequately prepare and cater for each guest wants and needs, while increasing the effectiveness of the marketing communication, by adequately target guests who will likely be more acceptive towards a given action, such as making another booking for example, increasing the hotel's revenue and forecasting capabilities without increasing the cost of acquisition.

It's interesting to see that different predictive goals achieve different results. We've analysed the previous model built by Antonio, De Almeida, et al. (2019), which focuses on booking cancellations from a Revenue Management perspective centred on the financial viewpoint of a hotel and compared its results to our model. Even though the dataset is the same, there results were not. Our model outperformed the previous one in recall, improving from 61% to 80% in H1 and 78% to 82% in H2, very likely a product of the decision making throughout the data preparation step and hyperparameter tuning which was oriented towards recall. We hope future research carry on comparing the results obtained with previous models to benchmark their work with established models.

It would have been interesting to compare our results with those obtained by Andriawan et al, (2020), however the results presented for comparison with other models appear to be train results, as they're compared with train results obtained by Antonio, De Almeida, et al. (2019). This comparison doesn't seem correct, as train data is not suitable for concluding the efficacy of a model, as there's still a risk of the model to not generalize its training to on unseen data, by underfitting or overfitting during training.

Other authors have also proposed different models to predict cancellations, namely Sánchez-Medina & C-Sánchez, (2020) who proposed a model for predicting hotel booking cancellations using Neural Network. As such, we also found it relevant to train and test a Neural Network algorithm in our project and evaluate if our dataset and prediction goals were better achieved when compared to the tree-based algorithms, specially XGBoost, which was used by Antonio, Almeida, et al., (2017), and RF, which was used by Andriawan et al., (2020). Sánchez-Medina & C-Sánchez, (2020) achieved over 95% across all measures, which exceed the results achieved by our model using Neural Network, where the best performing model achieved 80%. This may be due to several reasons, with the most likely being the fact that we employed different datasets.

One clear limitation of this project is the imbalance of target variables, particularly concerning the Food Package and the Returning Guests, which limits the understanding, and therefore the prediction, of guests who chose the less popular food packages and guest who return. As the hotels increase the number of bookings with underrepresented food packages and bookings from returning guests, the models would need to undergo retraining to relearn patterns and improve the results of the predictions, particularly regarding the Meal Packages predictions.

Another limitation is the absence of business significance of some variables of the bookings, like the monetary value of the food packages for example, as it limits our interpretation of the value of a CJ in these hotels. With more detail about these variables, we could expand the current framework further, by including other predictions, such as regressions, to better estimate the lifetime value of our guests.

Furthermore, this project was developed through manual tuning of hyperparameters, which rely on the developer's technical knowledge of the algorithms. Another option for tuning would be to use automated techniques, such as grid search, which would identify specific parameters settings to each algorithm, potentially improving the performance of the models.

Also, the anonymity of the hotels limits our understanding of the marketing positioning and strategy, irrespective of the type of hotel (resort and city). CJ are heavily influenced by the value proposition of a brand, meaning that whether the brand's positioning is high-end or low-end, it will affect the guest's the guest's expectations of the touchpoints across the three phases of a CJ. With higher-end brands there's a higher demand of service provided, which would potentially translate to more kpi's needing prediction at each phase.

Additionally, this project does not use any data from the marketing department, such as segmentation profiles, interactions on social media, recipients of marketing campaigns, etc. It would enrich the definition of the CJF to relate data from the reservation management system with data from the marketing department and then analyse and predict the CJ kpis with both perspectives.

Further research could include an expansion of the CJF to include a wider variety of sources, such as front desk, marketing, F&B, Services, etc, relating all data to track the behaviour of every guest. It would allow us to map the CJ of the guests and identifying key touchpoints throughout, which would be suitable as target variables.

Also, it would be interesting to develop more research about ML in other hotel departments, such as F&B or Front Desk, to assist decision makers of these departments to better forecast and provide value to guests, by anticipating behaviours and preferences of guests.

However, it is the authors belief that this project contributes to extend the scientific literature on ML and CJ in hospitality. We hope to stimulate further research into these topics, furthering the digital transformation of the hospitality industry and improving the level of service of hotels around the world.

BIBLIOGRAPHICAL REFERENCES

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. In *Organizational Research Methods* (Vol. 16, Issue 2, pp. 270–301). SAGE Publications Inc. https://doi.org/10.1177/1094428112470848

Amari, S., & Wu, S. (1999). Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, *12*(6), 783–789. https://doi.org/https://doi.org/10.1016/S0893-6080(99)00032-5

Ameen, N., Tarhini, A., Reppel, A., & Anand, A. (2021). Customer experiences in the age of artificial intelligence. *Computers in Human Behavior*, *114*. https://doi.org/10.1016/j.chb.2020.106548

Andriawan, Z. A., Purnama, S. R., Darmawan, A. S., Ricko, Wibowo, A., Sugiharto, A., & Wijayanto, F. (2020, November 10). Prediction of Hotel Booking Cancellation using CRISP-DM. *ICICoS 2020 - Proceeding: 4th International Conference on Informatics and Computational Sciences*. https://doi.org/10.1109/ICICoS51170.2020.9299011

Antonio, N., Almeida, A. de, & Nunes, L. (2017). Predicting hotel booking cancellations to decrease uncertainty and increase revenue. *Tourism & Management Studies*, *13*(2), 25–39. https://doi.org/10.18089/tms.2017.13203

António, N., Almeida, A. de, & Nunes, L. (2019). Predictive models for hotel booking cancellation: a semi-automated analysis of the literature. *Tourism & Management Studies*, *15*(1), 7–21. https://doi.org/10.18089/tms.2019.15011

Antonio, N., De Almeida, A., & Nunes, L. (2017). Predicting hotel bookings cancellation with a machine learning classification model. *Proceedings - 16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017*, *2017-December*, 1049–1054. https://doi.org/10.1109/ICMLA.2017.00-11

Antonio, N., De Almeida, A., & Nunes, L. (2019). An automated machine learning based decision support system to predict hotel booking cancellations. *Data Science Journal*, *18*(1). https://doi.org/10.5334/dsj-2019-032

Antonio, N., de Almeida, A., & Nunes, L. (2019). Hotel booking demand datasets. *Data in Brief*, *22*, 41–49. https://doi.org/10.1016/j.dib.2018.11.126

Arco, M. D., Presti, L. lo, Marino, V., & Resciniti, R. (2019). Embracing AI and Big Data in customer journey mapping: From literature review to a theoretical framework. *Innovative Marketing*, *15*(4), 102–115. https://doi.org/10.21511/im.15(4).2019.09

Bitner, M. J., Ostrom, A. L., & Morgan, F. N. (2008). *California Management Review Service Blueprinting: A Practical Technique for Service Innovation*.

Bradley, A. E. (1997). THE EVALUATION OF MACHINE LEARNING ALGORITHMS. In *Pattern Recognition* (Vol. 30, Issue 7).

Campbell, C., Sands, S., Ferraro, C., Tsao, H. Y. (Jody), & Mavrommatis, A. (2020). From data to action: How marketers can leverage AI. *Business Horizons*, *63*(2), 227–243. https://doi.org/10.1016/j.bushor.2019.12.002

Castelli, M., Pinto, D. C., Shuqair, S., Montali, D., & Vanneschi, L. (2022). The Benefits of Automated Machine Learning in Hospitality: A Step-By-Step Guide and AutoML Tool. *Emerging Science Journal*, *6*(6), 1237–1254. https://doi.org/10.28991/ESJ-2022-06-06-02

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., & others. (2000). CRISP-DM 1.0: Step-by-step data mining guide. *SPSS Inc*, *9*(13), 1–73.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 785–794. https://doi.org/10.1145/2939672.2939785

Cui, Y. (Gina), van Esch, P., & Jain, S. P. (2022). Just walk out: the effect of AI-enabled checkouts. *European Journal of Marketing*, *56*(6), 1650–1683. https://doi.org/10.1108/EJM-02-2020-0122

Daqar, M. A. M. A., & Smoudy, A. K. A. (2019). The role of Artificial Intelligence on Enhancing Customer Experience. *International Review of Management and Marketing*, *9*(4), 22–31. https://doi.org/10.32479/irmm.8166

de Bellis, E., & Venkataramani Johar, G. (2020). Autonomous Shopping Systems: Identifying and Overcoming Barriers to Consumer Adoption. *Journal of Retailing*, *96*(1), 74–87. https://doi.org/10.1016/j.jretai.2019.12.004

Følstad, A., & Kvale, K. (2018). Customer journeys: a systematic literature review. In *Journal of Service Theory and Practice* (Vol. 28, Issue 2, pp. 196–227). Emerald Group Holdings Ltd. https://doi.org/10.1108/JSTP-11-2014-0261

Günther, F., & Fritsch, S. (n.d.). *neuralnet: Training of Neural Networks*.

Halvorsrud, R., Kvale, K., & Følstad, A. (2016). Improving service quality through customer journey analysis. *Journal of Service Theory and Practice*, *26*(6), 840–867. https://doi.org/10.1108/JSTP-05-2015-0111

Hastie, T., Tibshirani, R., & Friedman, J. (2001). The elements of statistical learning. Springer series in statistics. *New York, NY, USA*.

He, A. Z., & Zhang, Y. (2022). AI-powered touch points in the customer journey: a systematic literature review and research agenda. In *Journal of Research in Interactive Marketing* (pp. 1–20). Emerald Publishing. https://doi.org/10.1108/JRIM-03-2022-0082

Huang, Z., Xu, W., & Yu, K. (2015). *Bidirectional LSTM-CRF Models for Sequence Tagging*. http://arxiv.org/abs/1508.01991

Karimi Sahar and Papamichail, K. N. and H. C. P. (2014). Purchase Decision Processes in the Internet Age. In J. E. and Z. P. and L. S. and R. R. and D. B. and P. J. Dargam Fátima and

Hernández (Ed.), *Decision Support Systems III - Impact of Decision Support Systems for Global Environments* (pp. 57–66). Springer International Publishing.

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review*, *26*(3), 159–190. https://doi.org/10.1007/s10462-007-9052-3

Lee, G. (2010). Death of 'last click wins': Media attribution and the expanding use of media data. *Journal of Direct, Data and Digital Marketing Practice*, *12*(1), 16–26. https://doi.org/10.1057/dddmp.2010.14

Lemon, K. N., & Verhoef, P. C. (2016). Understanding customer experience throughout the customer journey. *Journal of Marketing*, *80*(6), 69–96. https://doi.org/10.1509/jm.15.0420

Liaw, A., & Wiener, M. (2002). *Classification and Regression by randomForest* (Vol. 2, Issue 3). http://www.stat.berkeley.edu/

Luo, X., Tong, S., Fang, Z., & Qu, Z. (2019). Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science*, *38*(6), 937–947. https://doi.org/10.1287/mksc.2019.1192

McCarthy, E. J. (1978). Basic marketing: a managerial approach. *No Title*.

McCartney, G., & McCartney, A. (2020). Rise of the machines: towards a conceptual service-robot research framework for the hospitality and tourism industry. *International Journal of Contemporary Hospitality Management*, *13*(12), 3835–3851. https://doi.org/10.1108/IJCHM-05-2020-0450

Meuter, M. L., Bitner, M. J., Ostrom, A. L., & Brown, S. W. (2005). Customer Trial of Self-Service Technologies / 61. In *Journal of Marketing* (Vol. 69).

Parvez, M. O. (2020). Use of machine learning technology for tourist and organizational services: high-tech innovation in the hospitality industry. *Journal of Tourism Futures*, *7*(2), 240–244. https://doi.org/10.1108/JTF-09-2019-0083

Pereira, L. N., & Cerqueira, V. (2022). Forecasting hotel demand for revenue management using machine learning regression methods. *Current Issues in Tourism*, *25*(17), 2733–2750. https://doi.org/10.1080/13683500.2021.1999397

Sánchez-Medina, A. J., & C-Sánchez, E. (2020a). Using machine learning and big data for efficient forecasting of hotel booking cancellations. *International Journal of Hospitality Management*, *89*. https://doi.org/10.1016/j.ijhm.2020.102546

Sánchez-Medina, A. J., & C-Sánchez, E. (2020b). Using machine learning and big data for efficient forecasting of hotel booking cancellations. *International Journal of Hospitality Management*, *89*. https://doi.org/10.1016/j.ijhm.2020.102546

Taheri, B., Prayag, G., & Muskat, B. (2021). Introduction to the special issue: Consumer experience management and customer journeys in tourism, hospitality and events. In

Tourism Management Perspectives (Vol. 40). Elsevier B.V. https://doi.org/10.1016/j.tmp.2021.100877

Tanveer, M., Khan, N., & Ahmad, A. R. (2021). AI Support Marketing: Understanding the Customer Journey towards the Business Development. *2021 1st International Conference on Artificial Intelligence and Data Analytics, CAIDA 2021*, 144–150. https://doi.org/10.1109/CAIDA51941.2021.9425079

Trapani, D. G., Presti, L. L., Editors, S. P., & Proceedings, B. (2019). *New frontiers in the tourism and hospitality industry: digital, social and economic transformations*.

Wilma, D., & Schrotenboer, D. (2019a). *12 th IBA Bachelor Thesis Conference*.

Wilma, D., & Schrotenboer, D. (2019b). *12 th IBA Bachelor Thesis Conference*.

Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, *17*(5–6), 375–381. https://doi.org/10.1080/713827180

APPENDIX A (FINAL DATASET)

A1. Finalised dataset for model training and testing

Attributes	Description	Predictive Model
Is Cancelled	Keeps track of whether the booking was cancelled	Cancellation Prediction
Arrival Date Month July	Keeps track whether the month of arrival is July or not	Cancellation, Food Package and Return Prediction
Arrival Date Month August	Keeps track whether the month of arrival is August or not	Cancellation, Food Package and Return Prediction
Arrival Date Month September	Keeps track whether the month of arrival is September or not	Cancellation, Food Package and Return Prediction
Arrival Date Month October	Keeps track whether the month of arrival is October or not	Cancellation, Food Package and Return Prediction
Arrival Date Month November	Keeps track whether the month of arrival is November or not	Cancellation, Food Package and Return Prediction
Arrival Date Month December	Keeps track whether the month of arrival is December or not	Cancellation, Food Package and Return Prediction
Arrival Date Month January	Keeps track whether the month of arrival is January or not	Cancellation, Food Package and Return Prediction
Arrival Date Month February	Keeps track whether the month of arrival is February or not	Cancellation, Food Package and Return Prediction
Arrival Date Month March	Keeps track whether the month of arrival is March or not	Cancellation, Food Package and Return Prediction
Arrival Date Month April	Keeps track whether the month of arrival is April or not	Cancellation, Food Package and Return Prediction

Attributes	Description	Predictive Model
Arrival Date Month May	Keeps track whether the month of arrival is May or not	Cancellation, Food Package and Return Prediction
Arrival Date Month June	Keeps track whether the month of arrival is June or not	Cancellation, Food Package and Return Prediction
Arrival Date Week Number	Number of week in the year (1 to 52)	Cancellation, Food Package and Return Prediction
Adults	Number of Adults	Cancellation, Food Package and Return Prediction
Meal BB	Keeps track of whether the meal plan was Bed and Breakfast or not	Cancellation, Food Package and Return Prediction
Meal FB	Keeps track of whether the meal plan was Full Board or not	Cancellation, Food Package and Return Prediction
Meal HB	Keeps track of whether the meal plan was Half Board or not	Cancellation, Food Package and Return Prediction
Meal SC	Keeps track of whether the meal plan was Self-Catering or not	Cancellation, Food Package and Return Prediction
Market Segment Direct	Keeps track of whether the Market Segment was Direct or not	Cancellation, Food Package and Return Prediction
Market Segment Corporate	Keeps track of whether the Market Segment was Corporate or not	Cancellation, Food Package and Return Prediction
Market Segment Online TA	Keeps track of whether the Market Segment was Online TA or not	Cancellation, Food Package and Return Prediction

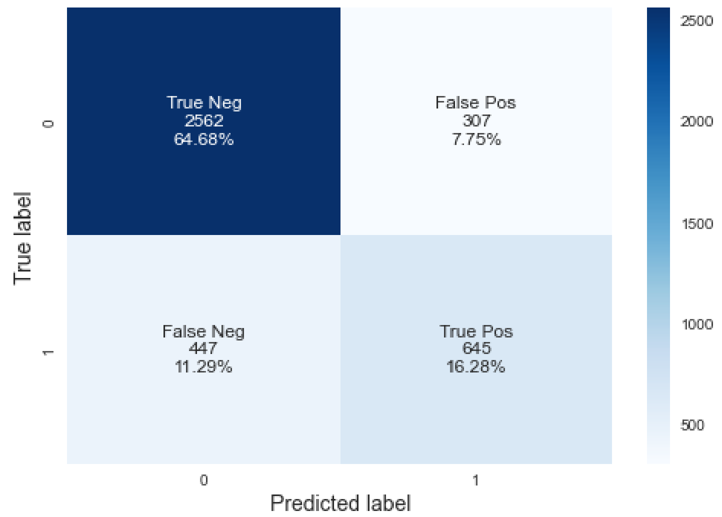
Attributes	Description	Predictive Model
Market Segment Online TA/TO	Keeps track of whether the Market Segment was Online TA/TO or not	Cancellation, Food Package and Return Prediction
Market Segment Complimentary	Keeps track of whether the Market Segment was Complimentary or not	Cancellation, Food Package and Return Prediction
Market Segment Groups	Keeps track of whether the Market Segment was Groups or not	Cancellation, Food Package and Return Prediction
Distribution Channel Direct	Keeps track of whether the Distribution Channel was Direct or not	Cancellation, Food Package and Return Prediction
Distribution Channel Corporate	Keeps track of whether the Distribution Channel was Corporate or not	Cancellation, Food Package and Return Prediction
Distribution Channel TA/TO	Keeps track of whether the Distribution Channel was TA/TO or not	Cancellation, Food Package and Return Prediction
Distribution Channel Undefined	Keeps track of whether the Distribution Channel was Undefined or not	Cancellation, Food Package and Return Prediction
Reserved Room Type C	Keeps track of whether the Reserved Room Type was C	Cancellation, Food Package and Return Prediction
Reserved Room Type A	Keeps track of whether the Reserved Room Type was A	Cancellation, Food Package and Return Prediction
Reserved Room Type D	Keeps track of whether the Reserved Room Type was D	Cancellation, Food Package and Return Prediction
Reserved Room Type E	Keeps track of whether the Reserved Room Type was E	Cancellation, Food Package and Return Prediction

Attributes	**Description**	**Predictive Model**
Reserved Room Type G	Keeps track of whether the Reserved Room Type was G	Cancellation, Food Package and Return Prediction
Reserved Room Type F	Keeps track of whether the Reserved Room Type was F	Cancellation, Food Package and Return Prediction
Reserved Room Type H	Keeps track of whether the Reserved Room Type was H	Cancellation, Food Package and Return Prediction
Reserved Room Type L	Keeps track of whether the Reserved Room Type was L	Cancellation, Food Package and Return Prediction
Reserved Room Type B	Keeps track of whether the Reserved Room Type was B	Cancellation, Food Package and Return Prediction
Reserved Room Type P	Keeps track of whether the Reserved Room Type was P	Cancellation, Food Package and Return Prediction
Deposit Type No Deposit	Keeps track of whether the Deposit Type was No Deposit	Cancellation, Food Package and Return Prediction
Deposit Type Non-Refund	Keeps track of whether the Deposit Type was non-refund	Cancellation, Food Package and Return Prediction
Deposit Type Refundable	Keeps track of whether the Deposit Type was Refundable	Cancellation, Food Package and Return Prediction
Days In Waiting List	How many days did the booking was hold in waiting	Cancellation, Food Package and Return Prediction
Customer Type Transient	Keeps track of whether the Customer Type was Transient	Cancellation, Food Package and Return Prediction

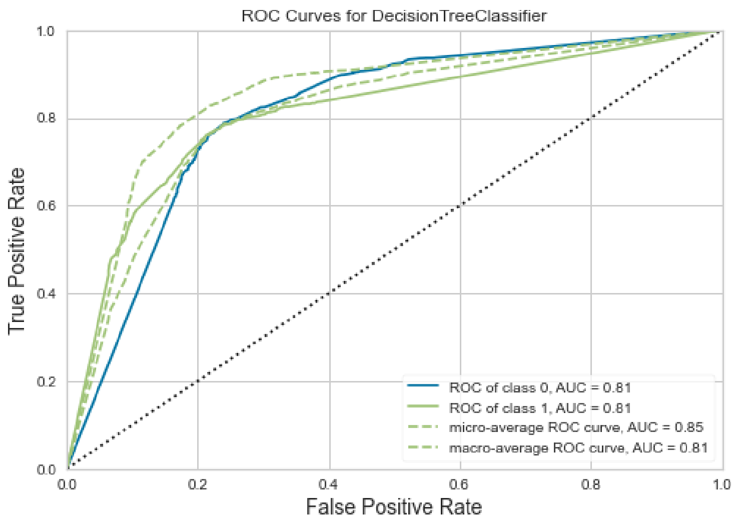
Attributes	Description	Predictive Model
Customer Type Contract	Keeps track of whether the Customer Type was Contract	Cancellation, Food Package and Return Prediction
Customer Type Transient-Party	Keeps track of whether the Customer Type was Transient-Party	Cancellation, Food Package and Return Prediction
Customer Type Group	Keeps track of whether the Customer Type was Group	Cancellation, Food Package and Return Prediction
Lead Time Tr	Transformed attribute of number of days prior to arrival that the booking was placed in the hotel	Cancellation, Food Package and Return Prediction
Stays In Weekend Nights Tr	Transformed attribute from the total length of stay, how many nights were in weekends (Saturday and Sunday)	Cancellation, Food Package and Return Prediction
Stays in Weeknights Tr	Transformed attribute from the total length of stay, how many nights were in weekdays (Monday to Friday)	Cancellation, Food Package and Return Prediction
Booking Changes Tr	Transformed attribute of the number of booking changes prior to arrival that could indicate cancellation intentions (arrival or departure dates, number of persons, type of meal, ADR, or reserved room type)	Cancellation, Food Package and Return Prediction
ADR Tr	Transformed Average Daily Rate	Cancellation, Food Package and Return Prediction
Total of Special Requests	Transformed attribute of how many special requests were made for each booking	Cancellation, Food Package and Return Prediction
Previous Cancellations Tr	Transformed attribute of number of previous bookings to this booking the guest had that were cancelled	Cancellation, Food Package and Return Prediction
Previous Bookings Not Cancelled Tr	Transformed attribute of Number of previous bookings to this booking the guest had that were not cancelled	Cancellation, Food Package and Return Prediction

APPENDIX B (MODELS)

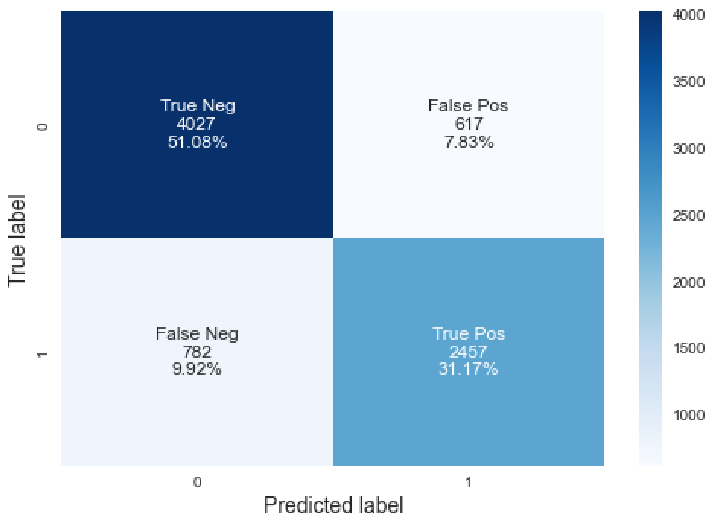
B.1 Confusion Matrix of Cancellation Model using Decision Tree of H1



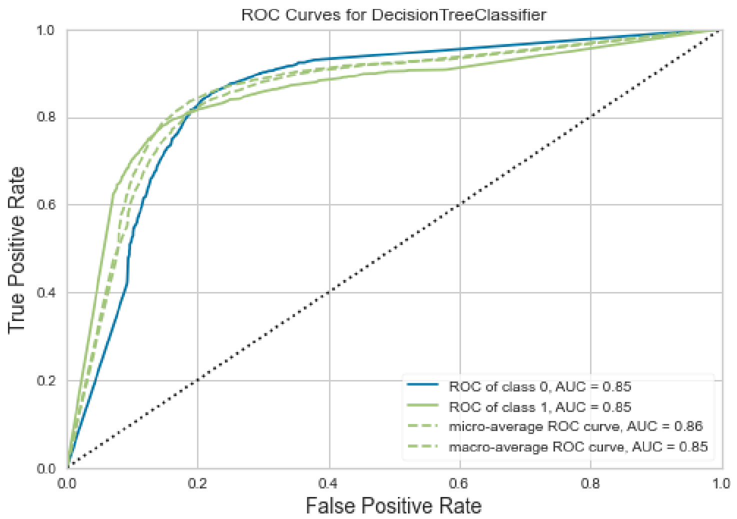
B.2 AUC-ROC graph of Cancellation Model using Decision Tree of H1



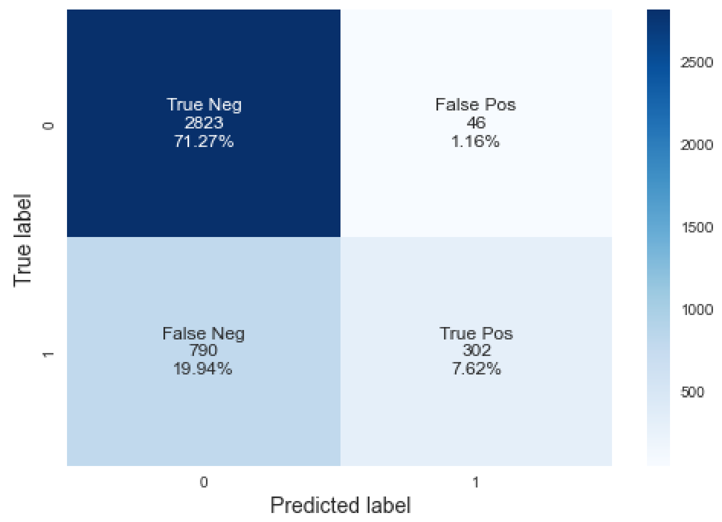
B.3 Confusion Matrix of Cancellation Model using Decision Tree of H2



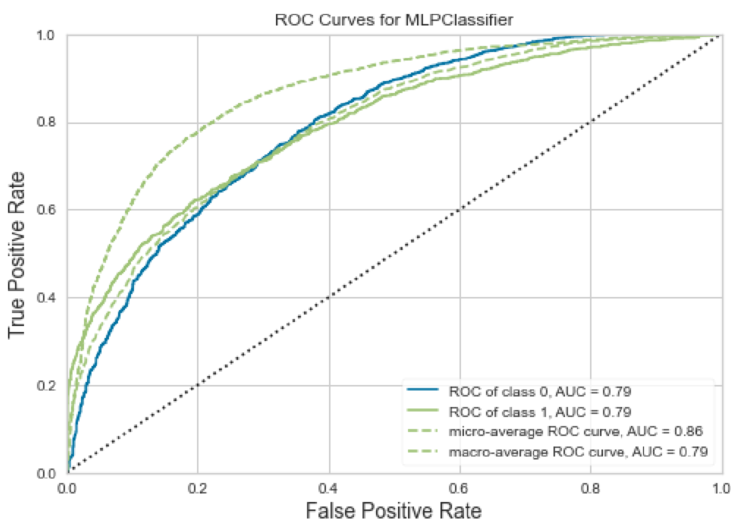
B.4 AUC-ROC graph of Cancellation Model using Decision Tree of H2



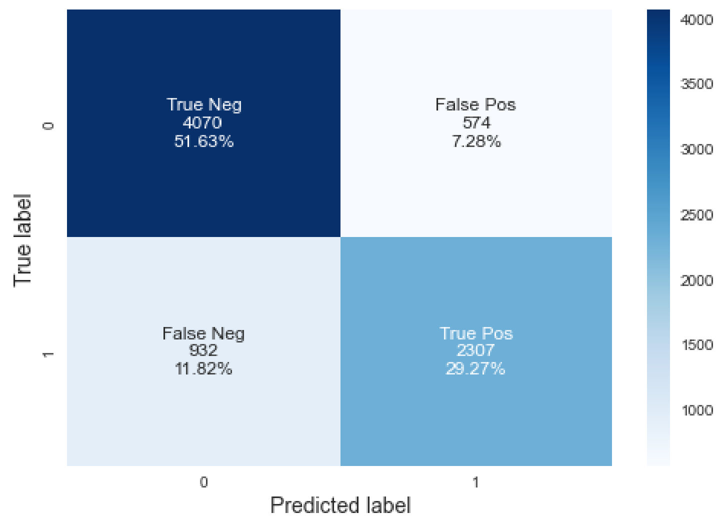
B.5 Confusion Matrix of Cancellation Model using Neural Network of H1



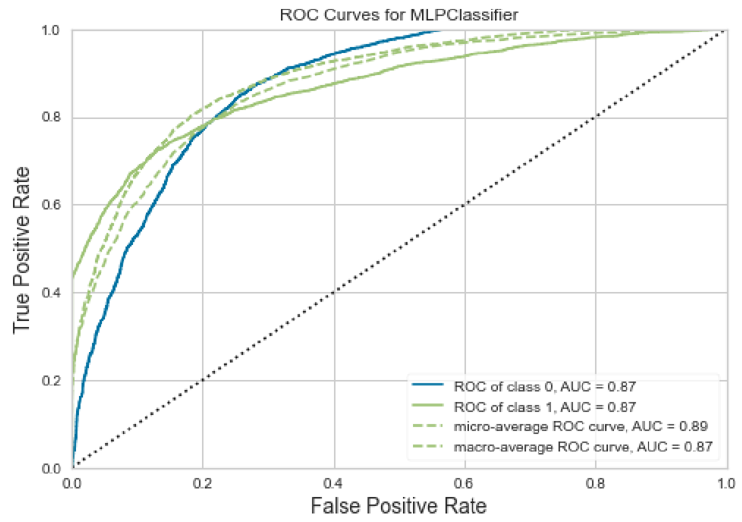
B.6 AUC-ROC graph of Cancellation Model using Neural Network of H1



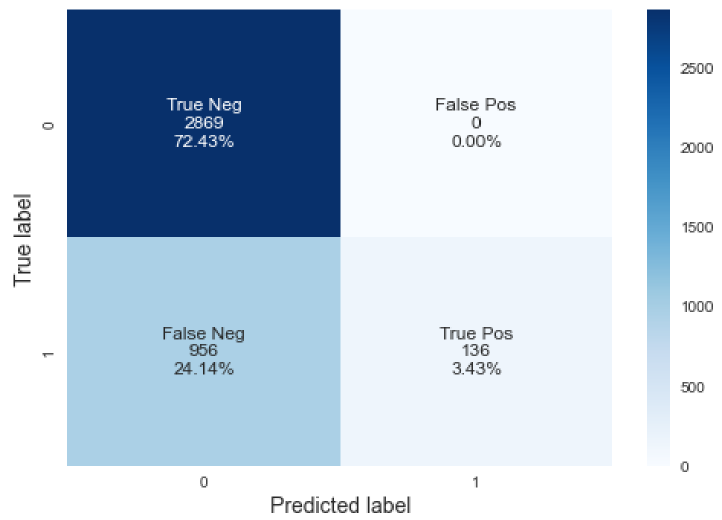
B.7 Confusion Matrix of Cancellation Model using Neural Network of H2



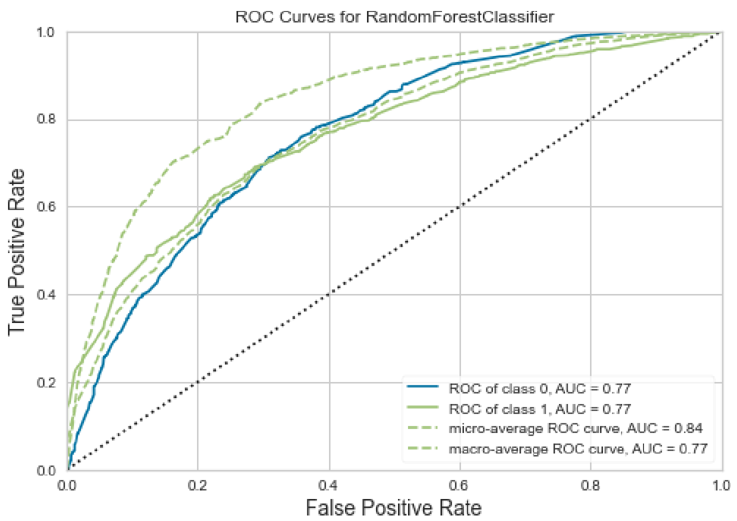
B.8 AUC-ROC graph of Cancellation Model using Neural Network of H2



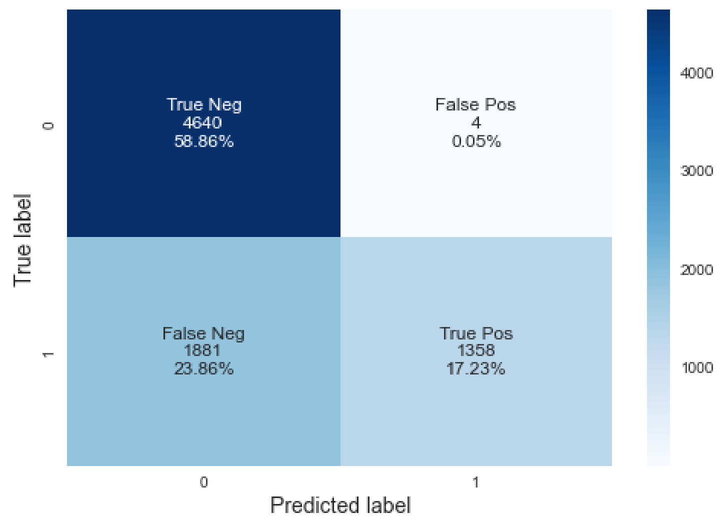
B.9 Confusion Matrix of Cancellation Model using Random Forest of H1



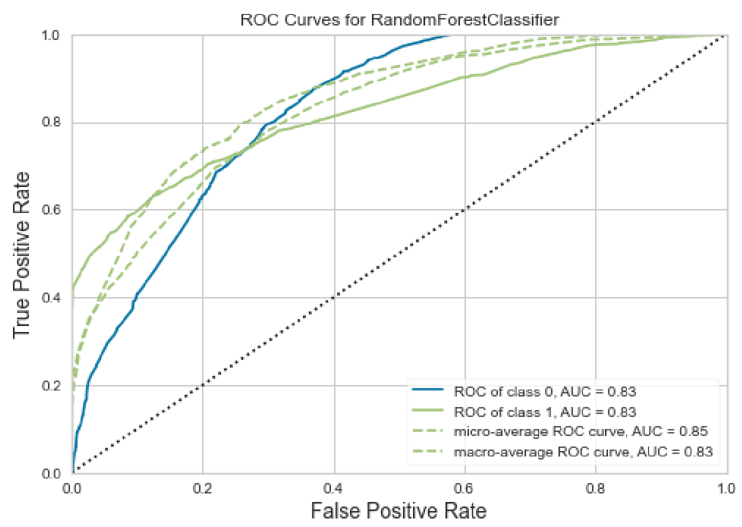
B.10 AUC-ROC graph of Cancellation Model using Random Forest of H1



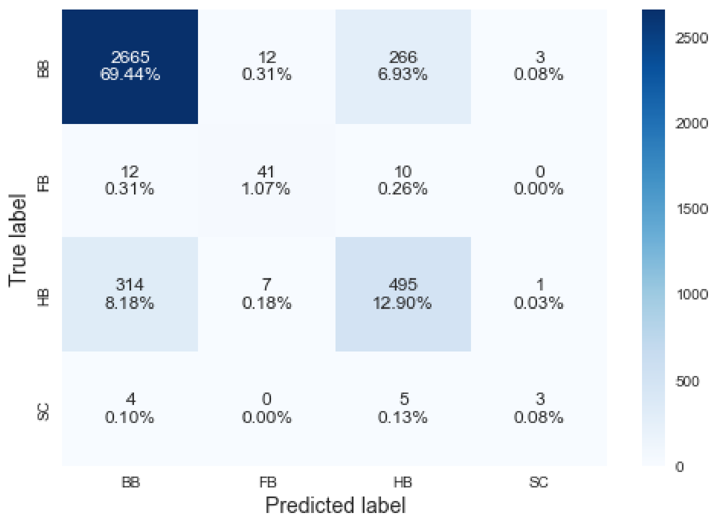
B.11 Confusion Matrix of Cancellation Model using Random Forest of H2



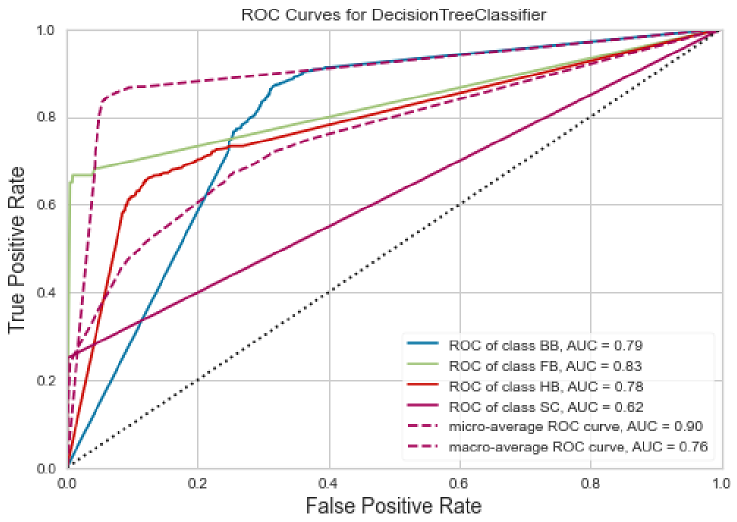
B.12 AUC-ROC graph of Cancellation Model using Random Forest of H2



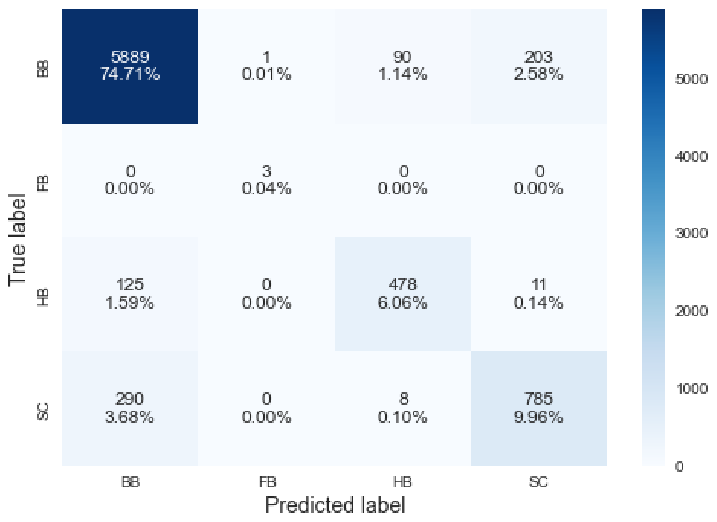
B.13 Confusion Matrix of Meal Model using Decision Tree of H1



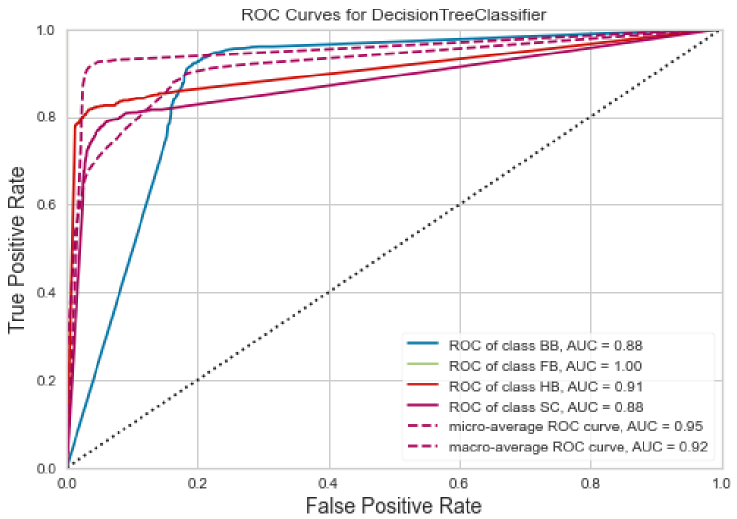
B.14 AUC-ROC graph of Meal Model using Decision Tree of H1



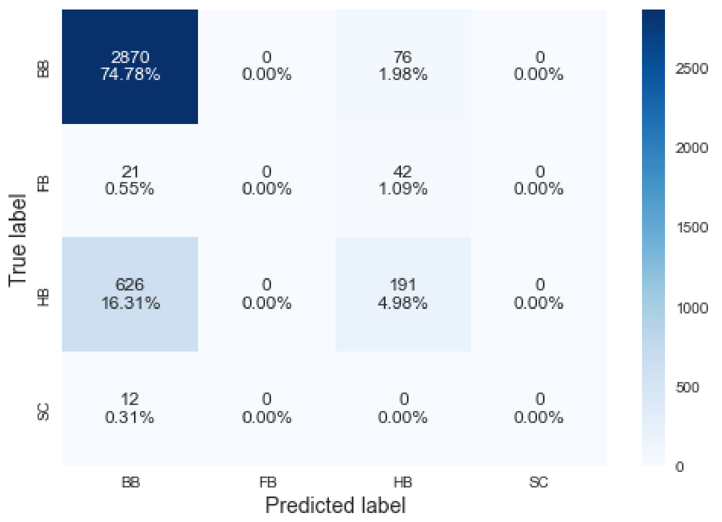
B.15 Confusion Matrix of Meal Model using Decision Tree of H2



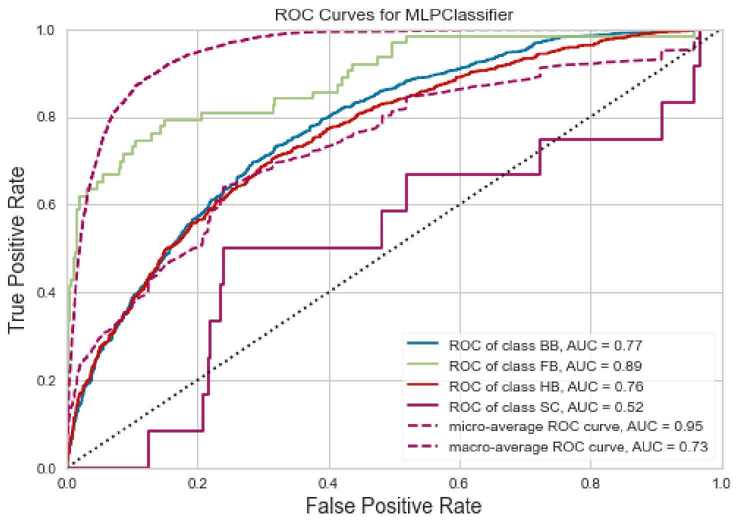
B.16 AUC-ROC graph of Meal Model using Decision Tree of H2



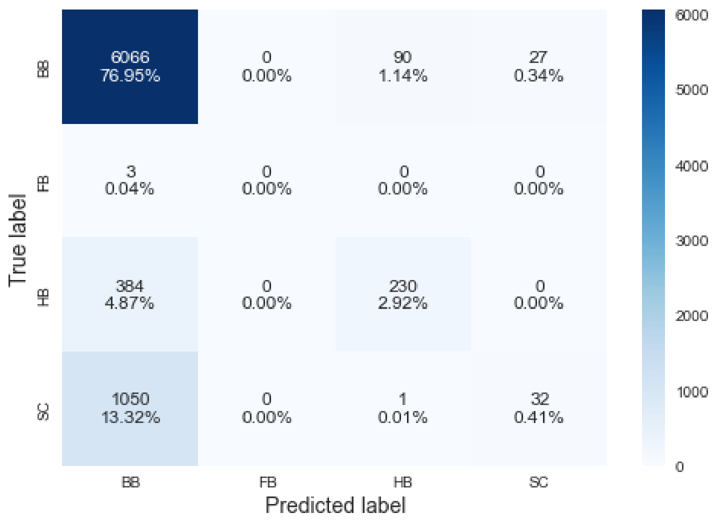
B.17 Confusion Matrix of Meal Model using Neural Network of H1



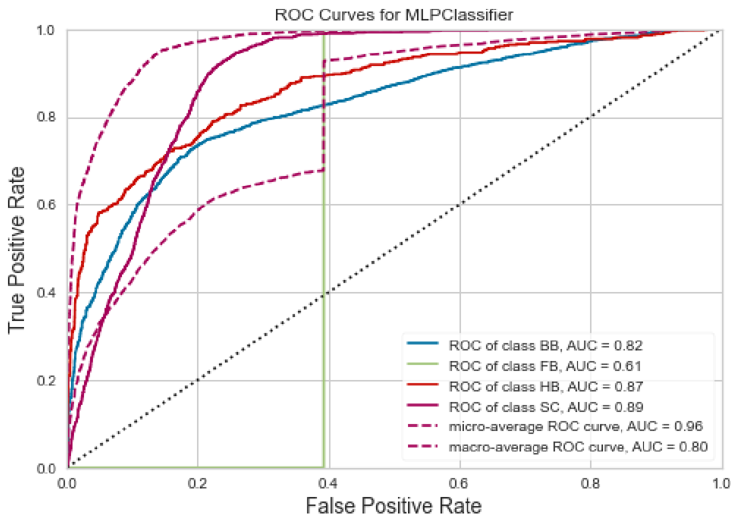
B.18 AUC-ROC graph of Meal Model using Neural Network of H1



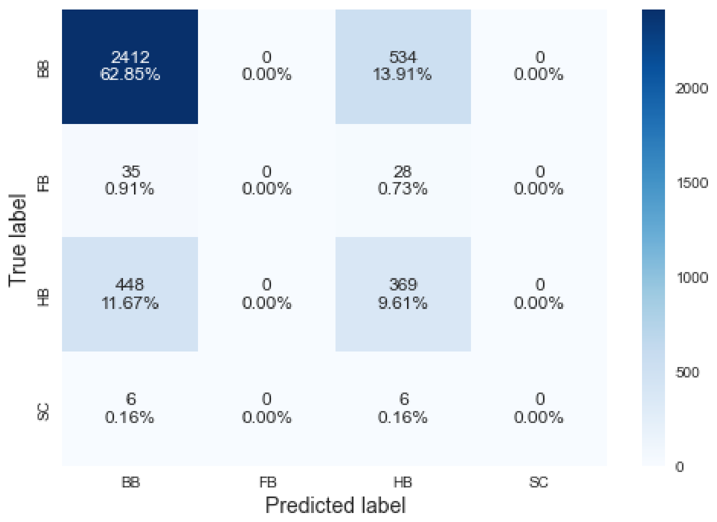
B.19 Confusion Matrix of Meal Model using Neural Network of H2



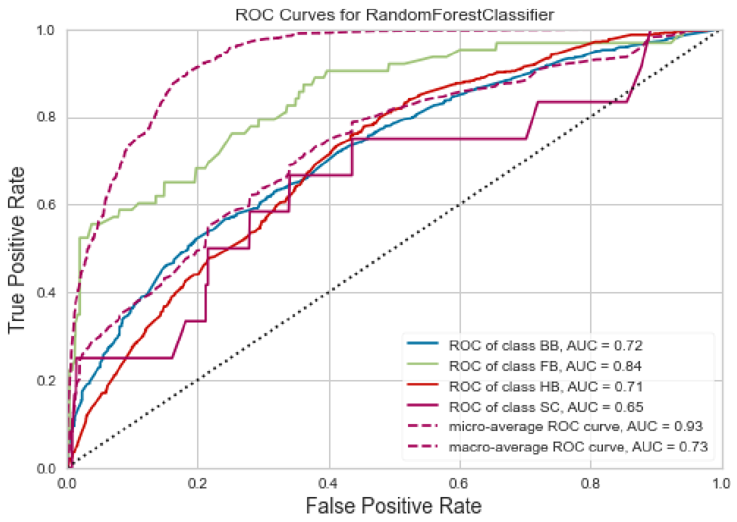
B.20 AUC-ROC graph of Meal Model using Neural Network of H2



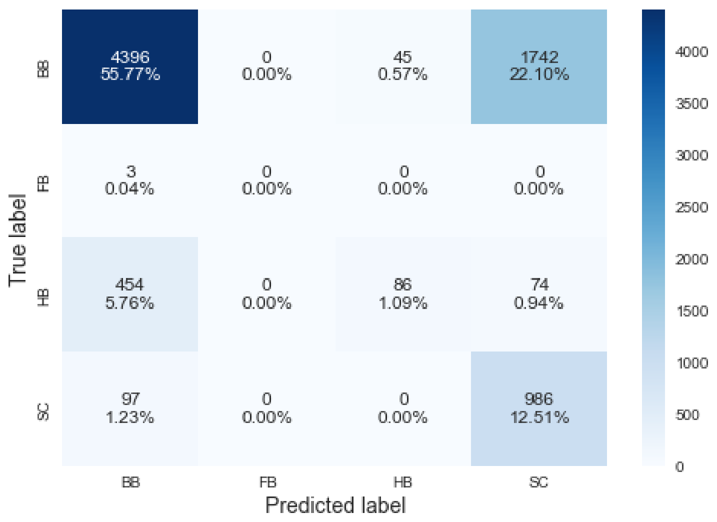
B.21 Confusion Matrix of Meal Model using Random Forest of H1



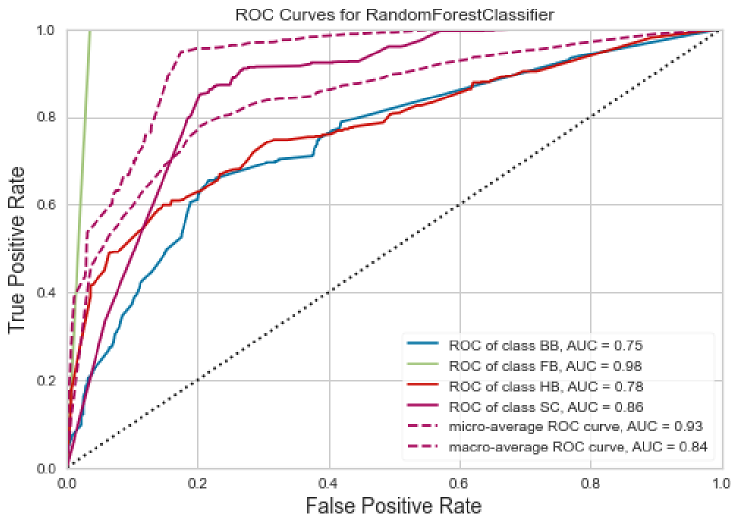
B.22 AUC-ROC graph of Meal Model using Random Forest of H1



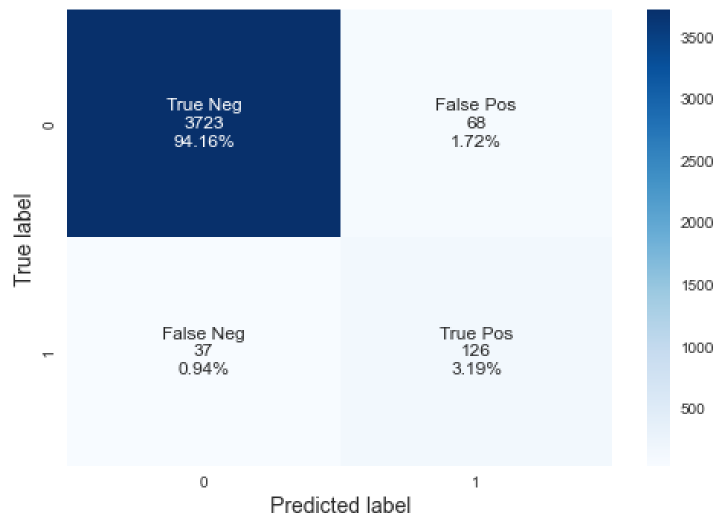
B.23 Confusion Matrix of Meal Model using Random Forest of H2



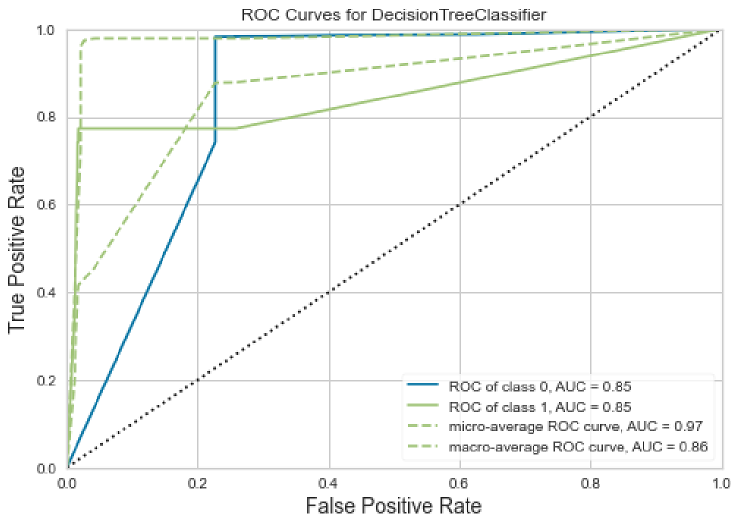
B.24 AUC-ROC graph of Meal Model using Random Forest of H2



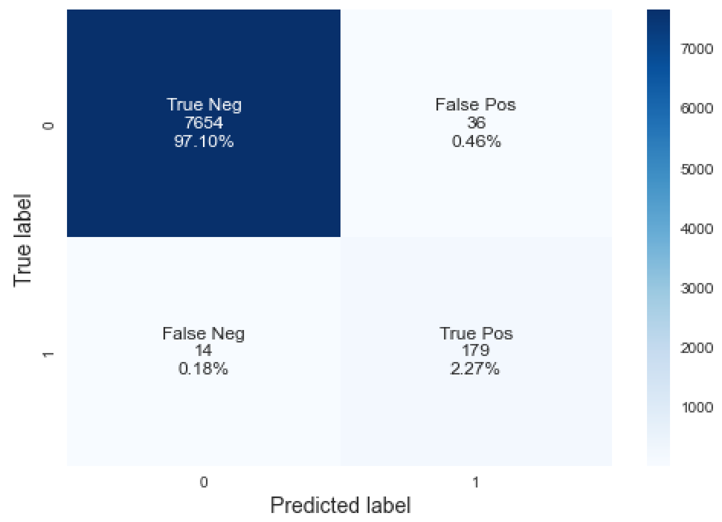
B.25 Confusion Matrix of Return Model using Decision Tree of H1



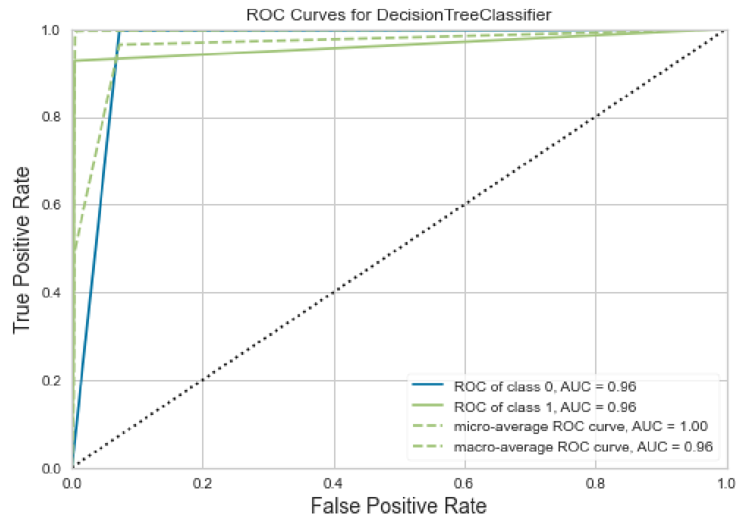
B.26 AUC-ROC graph of Return Model using Decision Tree of H1



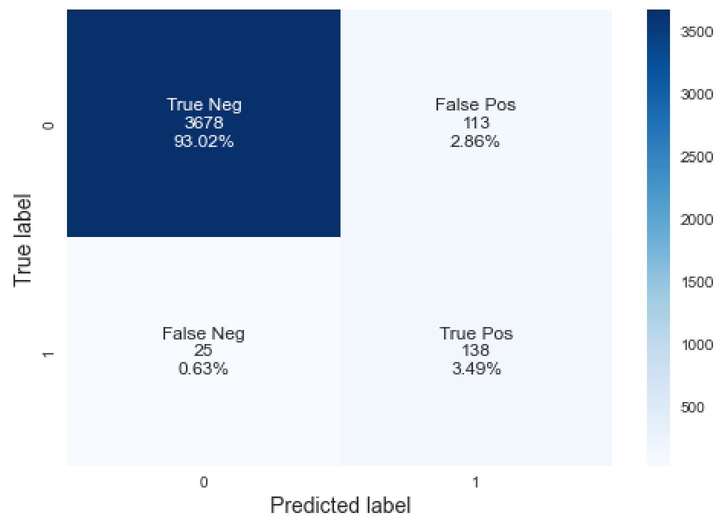
B.27 Confusion Matrix of Return Model using Decision Tree of H2



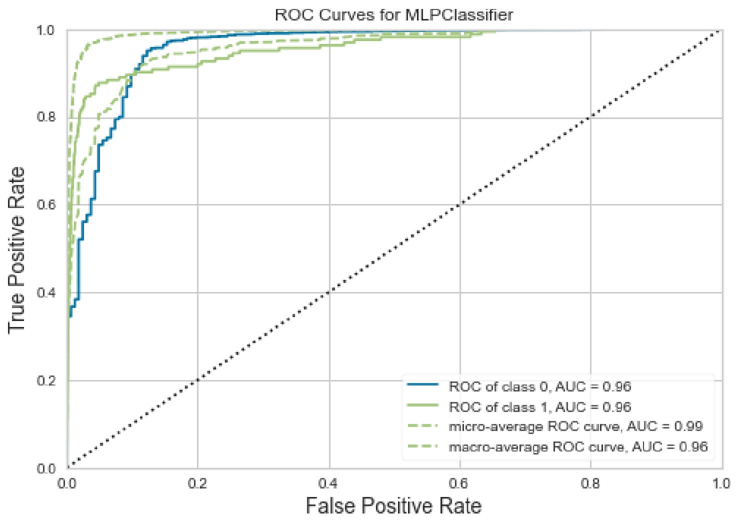
B.28 AUC-ROC graph of Return Model using Decision Tree of H2



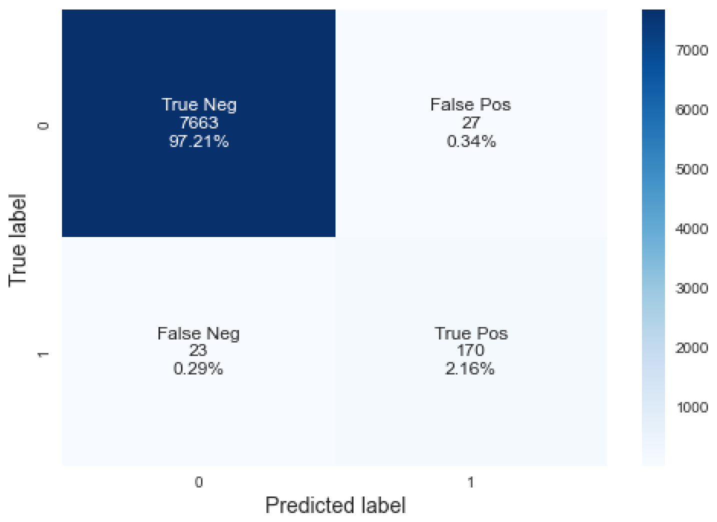
B.29 Confusion Matrix of Return Model using Neural Network of H1



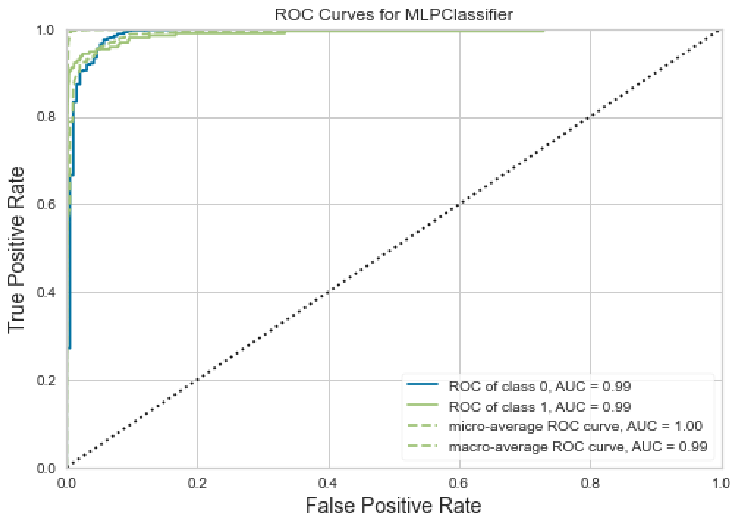
B.30 AUC-ROC graph of Return Model using Neural Network of H1



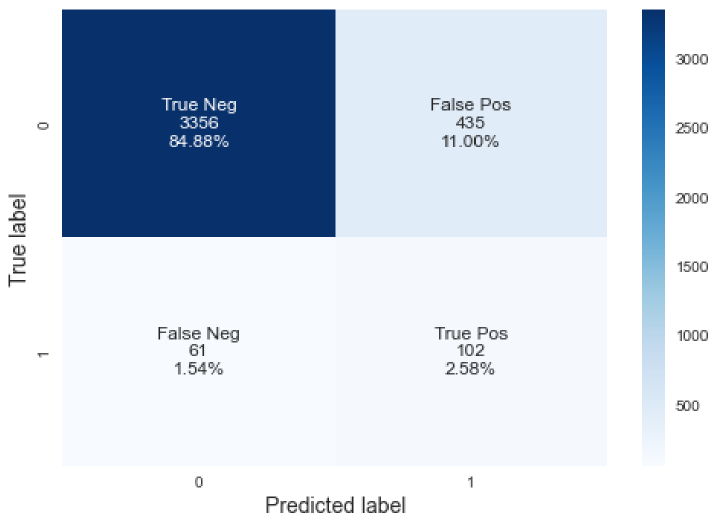
B.31 Confusion Matrix of Return Model using Neural Network of H2



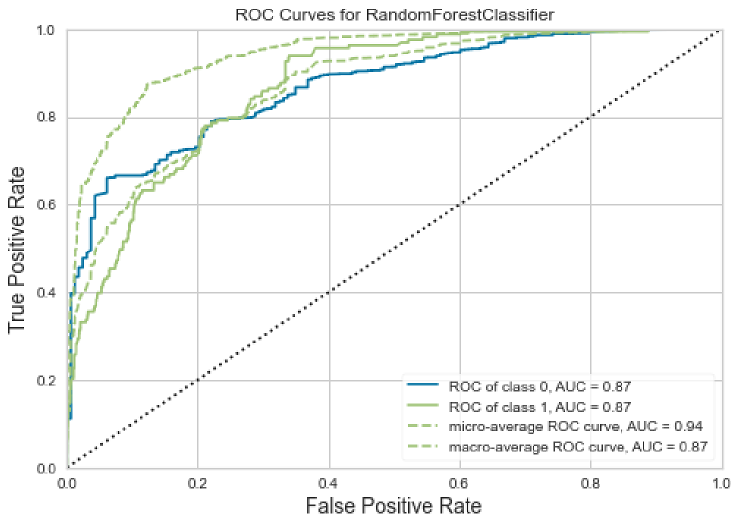
B.32 AUC-ROC graph of Return Model using Neural Network of H2



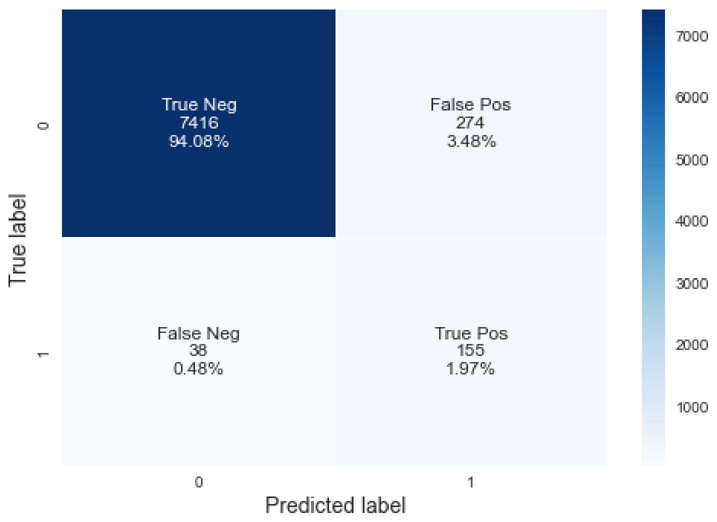
B.33 Confusion Matrix of Return Model using Random Forest of H1



B.34 AUC-ROC graph of Return Model using Random Forest of H1



B.35 Confusion Matrix of Return Model using Random Forest of H2



B.36 AUC-ROC graph of Return Model using Random Forest of H2