PREDICTING HOTEL BOOKING CANCELLATIONS USING THE XGBOOST ALGORITHM

AN ANALYSIS OF THE TRUE IMPACT OF PSYCHOLOGICAL DISTANCE ON THE PREDICTION OF CANCELLATIONS

SHANSHAN YANG

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

WORD COUNT: 8294

STUDENT NUMBER

2064430

COMMITTEE

prof. dr. Eric Postma
dr. Merel Jung

LOCATION

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

DATE

January 14, 2022

PREDICTING HOTEL BOOKING CANCELLATIONS USING THE XGBOOST ALGORITHM

AN ANALYSIS OF THE TRUE IMPACT OF PSYCHOLOGICAL DISTANCE ON THE PREDICTION OF CANCELLATIONS

SHANSHAN YANG

PREFACE

Dear reader,

Thank you for taking the time to read my thesis covering predicting hotel booking cancellations using the XGBoost algorithm. Working on this thesis has been a challenging and exciting journey for me. Herewith, I would like to express my appreciation and gratitude to the people that helped me during this special time. Firstly, I want to thank my supervisor, Prof. Eric Postma, for all of his invaluable advice and feedback: It helped me to reshape my ideas and construct the content more logically. I also want to express my heartfelt thanks to all of my family and friends, especially to Martin Schäufler, Yanchen Zhu and Qiao Zhang. In the last few weeks, they shared experiences with me, gave me genuine feedback on my thesis, and supported me emotionally. All my thanks to all of you!

I hope you enjoy reading it as much as I do writing it.

Best regards,

Shanshan Yang

CONTENTS

Abstract

Booking cancellations are one of the major issues in the hotel industry. The occurrence of booking cancellations has a great negative impact on hotel revenue. Previous research projects have primarily focused on two research directions. One direction demonstrated theoretical cancellation inducers, namely psychological distance. While the other one demonstrated the development of predictive models for cancellations using machine learning algorithms. However, the association between the two types of research projects remains unclear. This research aimed to investigate the true impact of cancellation inducers when predicting cancellations using the XGBoost algorithm and *Hotel Booking Demand Dataset*. We implemented feature categorization to establish the association between the psychological distance and the features from the dataset. We also studied the impact of the psychological distance using XGBoost feature importance analysis, partial dependency analysis, and correlation coefficient analysis. This research found only one particular distance with great impact on the prediction of booking cancellations. Furthermore, this research also found that the prediction is most accurate when the model is trained using both psychological-distance and non-distance features.

DATA SOURCE/CODE/ETHICS STATEMENT

1. Working on this thesis did not involve collecting data from human participants or animals.

2. This thesis uses two datasets that are made publicly available:

The first one is Hotel Booking Demand Dataset. The data is originally from the article Hotel Booking Demand Datasets, written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019. The Data can be found here: https://www.kaggle.com/jessemostipak/hotel-booking-demand

The second one is ISO-3166 Country and Dependent Territories Lists with UN Regional Codes. This dataset is published on GitHub by Lukes Duncalfe and can be found here: https://github.com/lukes/ISO-3166-Countries-with-Regional-Codes

The original owner of the data and code used in this thesis retains ownership of the data during and after the completion of this thesis.

3. The code developed for this thesis is made publicly available until February 2022 and can be found on GitHub: https://github.com/Sudokuber-/Tilburg_MS_DSS_Thesis

4. All figures and tables used in this thesis are the original works of the author of this thesis.

1 INTRODUCTION

The goal of this research is to investigate the impact of booking cancellation inducers, namely psychological distances, on the prediction of hotel booking cancellations using the XGBoost algorithm. Due to time constraints, we use the *Hotel Demand Booking Dataset* (2019b), which is a published dataset originally designed for tackling hotel booking cancellations using machine learning algorithms.

In the hotel industry, the number of booking cancellations increases significantly during the year 2014 – 2018[1], which leads the global average cancellation rate to almost 40% of total bookings (D-edge, 2019). Hotel revenue managers have been tackling booking cancellations over the years (Weatherford & Kimes, 2003). The occurrence of booking cancellations leads to an increase in the number of unsold rooms and revenue loss. To take control of booking cancellations and minimize their subsequent impacts, hotel revenue managers implement overbooking strategies and cancellation policies (Hu, 2009). However, those tactics are not perfect and can be problematic sometimes. For instance, in October 2018, a passenger was beaten and dragged out of an overbooked airplane at Chicago O'Hare International Airport because he refused to give up his seat (Zdanowicz & Grinberg, 2018). Videos of the scene went viral on social media and the reputation of the Airline company was damaged. Moreover, hotel cancellation policies and cancellation penalties often trigger conflicts between hotels and customers (Lee, Yang, & Chung, 2021).

With machine learning gaining popularity, predicting booking cancellations using machine learning algorithm is considered a better alternative. Predicting cancellations has a positive impact on improving hotel demand forecasting, overbooking strategies, and cancellation policies (Antonio, de Almeida, & Nunes, 2017). Subsequently, predicting cancellations also helps hotels with minimizing potential reputation damage and avoiding conflicts with customers caused by current strategies against cancellations.

Predicting booking cancellations using machine learning algorithms becomes state-of-the-art and has been addressed extensively. Many research projects are conducted to improve the accuracy of predicting canceled bookings.(C. Chen, 2016; Hueglin & Vannotti, 2001; Lemke, Riedel, & Gabrys, 2013; Morales & Wang, 2010; Talluri & Ryzin, 2004). A model developed by Antonio, de Almeida, and Nunes (2019a) using The XGBoost (Extreme Gradient Boosting) algorithm achieved the highest accuracy (94%) among all published research projects that the author can find.

The XGBoost algorithm was first developed by Tianqi Chen in 2016. It takes gradient boosting at its core and pushes the limits of computations

[1] The corona pandemic has had a great negative impact on the hotel industry in 2019 – 2020. The booking cancellation rate is abnormally high during that time, and therefore not considered in this research.

resources (T. Chen & Guestrin, 2016). XGBoost is one of the best-performing supervised learning algorithms. It is often used for classification tasks and structured datasets. The library of XGBoost also provides feature importance analysis which helps with understanding the impact of features (Antonio et al., 2019a).

On the other hand, machine learning algorithms are not widely applied in hotel industry. Many hoteliers try to predict cancellations manually based on their experience and research on cancellation inducers. The most recently published theoretical research, *Out of sight, out of cancellation: The impact of psychological distance on the cancellation behavior of tourists* (Lee et al., 2021), analyses booking cancellation inducers from the customers psychological perspective and presents five distances that have strong impacts on the booking cancellations:

Spatial distance in terms of the distance between customer and reservation behavior is highly influenced by technologies, such as PC-based websites and smartphone applications (Lee et al., 2021). In particular, smartphones enable customers to reserve or to cancel a hotel room without constraints in terms of time and location. Accordingly, we associate spatial distance with devices used for making reservations.

Temporal distance is a perception of time distance which represents how near or far an event is from now (Trope & Liberman, 2010). In the context of this research, we associate temporal distance with lead time, which represents the period between the day of booking and the day of arrival. In line with temporal distance, lead time indicates how soon the check-in time is from now.

Experiential distance emphasizes the knowledge an individual has about certain things (Lee et al., 2021). In the context of this research, we associate experiential distance with frequencies related to reservation behaviors, such as the number of bookings, cancellations, as well as modifications that customers made. As frequency increases, customers would gain more knowledge about the reservation process.

Arbitrary distance measures how freely an individual booker can cancel a booking, and it is measured by the number of companions (Lee et al., 2021). For instance, when the number of travel companions is large and the booker would like to cancel the booking, the number of companions that are forced to cancel with the booker is large as well (Lee et al., 2021). Accordingly, we associate arbitrary distance with the number of companions that each hotel reservation is made for.

Economic distance in terms of different levels of personal wealth can be reflected in the value of the travel product that the customers booked (Lee et al., 2021). In the context of this research, we associate economic distance with all expenses including potential expenses of each reservation.

The result of the research (Lee et al., 2021) is obtained based on statistical inferences, which means the impact of the psychological distance on the

actual predictions remains unclear. Furthermore, psychological distance is a broader concept defined in theory, while predicting hotel booking cancellations requires specific data in practice. The connection between psychological distance and predictions was not addressed in previous research projects, leading to a poor understanding of the true impact of psychological distance on an actual prediction of cancellations. Therefore, this research focuses specifically on the relationship between psychological distance and predictions.

Main Research Question

> *How do cancellation inducers, namely psychological distance, affect the prediction of hotel booking cancellations?*

Three sub-questions are derived from the main research question:

RQ1 *Does psychological distance have an impact on predicting hotel booking cancellations?*

To answer this question, we examine whether a model trained solely with psychological-distance features has sufficient predictive power to perform the task of predicting booking cancellations. The sufficiency of predictive power is measured by comparing the model trained solely with psychological-distance features to a baseline model.

RQ2 *Which distance has the most impact when predicting hotel booking cancellations?*

To answer this question, we need to know which distance is most informative for the model to predict cancellations. The quality of the information can be measured using information gain and the dependence of the predicted outcome towards a distance. Since the model directly deals with features instead of psychological distance, we first need to evaluate the importance of the features based on information gain. After that, we examine how the impact of each feature influences the predicted outcome. Finally, we associate the features and their impact with corresponding psychological distance.

RQ3 *How do psychological-distance features relate to non-distance features with respect to the ability of predicting hotel booking cancellations?*

To answer this question, we need to find the best-performing model and evaluate how much psychological-distance and non-distance features influence the performance of the model using feature importance analysis.

Findings

We developed three models, of which one model is trained using psychological-distance features; one model is trained using non-distance features; one model is trained using both psychological-distance and non-distance features. It is found that the psychological distance indeed has impact on the prediction of cancellations, and the impact is greater than non-distance features. Furthermore, the impact of psychological distance mostly comes from economic distance. In particular, non-refundable booking is the only critical and decisive feature for predicting cancellations. Finally, in order to achieve the best model performance and accuracy, both psychological-distance and non-distance features are needed.

2 RELATED WORK

This section reviews research projects of psychological distance and machine learning-based predictions, in relation to hotel booking cancellations. In the context of this research, the meaning of cancellation inducers and psychological distance are indistinguishable. Therefore, both terms will be used interchangeably in the rest of the report.

Psychological Distance

Trope, Liberman, and Wakslak (2007) propose a Construal Level Theory (CLT) framework which explains how behavior can be influenced by the perceived psychological distance a person has towards someone or something. According to the CLT framework, people tend to think about psychologically far event in an abstract and positive way, therefore, people would like the event to happen. On the other hand, people tend to think about psychologically close event in a complex and objective way and therefore consider their plans and decisions very carefully. A few years later, Lee et al. (2021) employ the CLT framework when studying the relationship between psychological distance and travel product cancellations. Based on the theory of Trope et al. (2007), the closer the psychological distance, the higher the possibility of cancellation. Surprisingly, Lee et al. (2021) found out that closer spatial, arbitrary, and economic distance lead to a higher possibility of cancellation while further temporal and experiential distance lead to a higher possibility of cancellation.

Spatial distance represents how easily a booking can be made or cancelled (Lee et al., 2021). People find both the booking and canceling process on a mobile phone smoother and easier than on a computer (Lee et al., 2021). Therefore, the mobile phone represents a closer spatial distance which has a positive effect on the booking cancellations. This is in line with

Trope et al. (2007): the easier the booking process, the higher the possibility of cancellation. Unfortunately, the dataset used in this research does not contain device information, therefore, the impact of spatial distance will not be evaluated.

Temporal distance refers to lead time, which represents the period from the day of booking to the day of arrival (Lee et al., 2021). Study has found out that hotel booking decisions are time-sensitive (Jang, Chen, & Miao, 2019). An event in the distant future is presented more abstractly, people have less preference for the event to happen (Dhar & Kim, 2007). In line with the notion, Lee et al. (2021) also proves that the longer lead time comes with higher uncertainty about the trip, leading to a higher probability of cancellation.

Experiential distance is highly correlated with familiarity (Massara & Severino, 2013). For instance, people who often make bookings would gain more experience, thus, they are more familiar with the products and the booking process. Therefore, the experiential distance is closer. Lee et al. (2021) found that people with less experience in booking travel products are more likely to make mistakes. To correct the mistakes, people often need to make cancellations first, leading to a higher probability of cancellation. This characteristic is often found in people who booked travel packages and is rarely found in people who booked a single travel product at a time. In contrast, another research (Lee, Chung, & Lee, 2017) suggests that people who are familiar with flight reservation systems are more likely to make cancellations. However, Lee et al. did not mention whether the tickets are a part of a package deal or an independent purchase.

Arbitrary distance represents a sense of separation from others, and it is measured by the number of travel companions (Lee et al., 2021). If the number of companions increases, the sense of separation from others decreases. In the context of booking cancellation, the number of companions decides the extent to which a customer can decide to cancel at will (Lee et al., 2021). According to Lee et al. (2021), as the number of companions grows, the number of people who are forced to cancel grows too, leading to a higher probability of cancellation. On the other hand, group bookings have a relatively large impact on hotel revenue because it occupies 30% - 50% of total bookings on average. Therefore, hotels have a stricter cancellation policy toward groups to protect themselves from major revenue loss (Hu, 2009), which means there is a possibility that the customer will not be able to cancel or receive a full refund. Subsequently, the probability of cancellation decreases.

Economic distance refers to the value of booked travel products (Lee et al., 2021). When the value of the product increases, the economic distance will also increase. In line with this notion, customers often associate quality with price (Kassinis & Soteriou, 2015; Zeithaml, 1988). In terms of travel products, the association between quality and price becomes stronger

(Lichtenstein, Ridgway, & Netemeyer, 1993). Often, this makes people question the quality of cheaper travel products and eventually cancel it without a blink. Furthermore, cancellation fees for travel products are usually associated with prices. Higher cancellation fees make customers hesitate before cancelling (Lee et al., 2017). Hence, closer economic distance leads to a higher probability of cancellation.

Findings of Machine Learning-Based Research

In contrast to the findings of Lee et al. (2021), Antonio et al. (2019a) studies cancellation drivers for 8 individual hotels using the XGBoost algorithm and various data sources. One of the models achieved the highest accuracy (94%) but 26% for precision and 64%for AUC. Another model achieved the highest precision (77%) and AUC (93%), but 85% for accuracy. Furthermore, the study reveals eight features that have the most impact on the prediction of cancellations: *Lead Time, Country, Number of Booking Modifications Made By Guest, Adults, Non-Refundable Booking, Number of Stays Over Weekends, Group, Leisure*, and *Reserved Room Type*. Without any context and proper connection between the features and the psychological distance, we cannot make conclusions on whether or not psychological distance is influential towards the prediction, and even if psychological distance is influential, how much impact it has towards the prediction.

Other machine learning-based research projects are found to be less focused on the features. Satu, Ahammed, and Abedin (2020), analyse and compare the performances of multiple booking cancellation prediction models using the *Hotel Booking Demand Dataset*. The team experiments with multiple feature selection, transformation techniques, and classifiers. In the end, XGBoost algorithm stands out with 79%, 75%, and 78% for accuracy, AUROC, and F score, respectively. Their result shows that the choice of feature transformation techniques and classifiers does not have much impact on the prediction. Although the implementation of feature selections was proven to have a positive impact on the performance, relevant features are not mentioned in the paper. Putro, Septian, Widiastuti, Maulidah, and Pardede (2021) uses the same *Hotel Booking Demand Dataset* and develop several models for predicting booking cancellation using deep neural network and logistic regression algorithms. The result of Putro et al. (2021) shows that the deep neural network with encoder-decoder architecture and Adamax optimizer results in the best accuracy of 87%, which is 8% higher than 79% from Satu et al.

In summary, we first reviewed literature about the psychological distance and whether it has a positive or negative impact on cancellations. Some of the literature presents contradictory impact of experimental distance and arbitrary distance on cancellations. The unsuccessful comparison between the findings of Lee et al. and 'Antonio et al. addressed the prob-

lem caused by the missing connection between the psychological distance and the features from the dataset. To better understand the impact of psychological distance on the predictions, it is necessary to establish the connection. Finally, we reviewed some models that used the same dataset as we used in this research, and Xgboost is proven to be a suitable algorithm for predicting hotel booking cancellations.

3 METHODS

This research consists of three major methods including feature categorization, XGBoost algorithm, and features analysis. In this section, we will introduce the key concept of each method. Subsection 3.1 presents feature categorization. It is the process where we match features with psychological distance. Subsection 3.2 introduces the general concept of the XGBoost algorithm, which is the core of the experimental models of this research. Subsection 3.3 presents the tools we used to conduct in-depth feature analysis.

3.1 *Feature Categorization*

Feature categorization is the process where we categorize features from the dataset to the most relevant distance. As Table 1 presents, features are categorized based on the explanation of the distances (section 1) and the descriptions of features from the *Hotel Booking Demand dataset*. Features, that cannot be connected to any of the distances, are connected to non-distance features. This process is done after data is pre-processed. The complete description of all features from the *Hotel Booking Demand dataset* can be found in Appendix 1.

3.2 *XGBoost Algorithm*

In this research, we use the XGBoost algorithm to perform a binary classification task: whether a booking will be cancelled or not cancelled. XGBoost (T. Chen & Guestrin, 2016) has the gradient boosting-based algorithm at its core and it takes tree structure as the default booster for classification tasks. The algorithm constructs a model of an ensemble of trees with an intuition of the next best tree in combination with previous trees minimizes the prediction error. Unlike random forest building trees in parallel, the XGBoost algorithm builds one tree after another. Predictions will be produced for every tree and compared with the true target.

The actual depth of a single tree is around 25 in our experiment, which is too large to be visualized. Figure 1 illustrates a simpler tree with a depth of 2, using the *Hotel Booking Demand Dataset*. The task is to predict whether a booking will be "cancelled" or "not cancelled".

Table 1: Psychological distance and corresponding features. Note: spatial distance is not in the table because we do not have corresponding feature in *Hotel Booking Demand dataset*.

Distance	Explanation	Feature description	Feature variable name
Temporal	Time distance which represents how near or far an event is from now	1. No. days between the day of booking and the day of arrival	1. "lead_time"
Experiential	Frequencies related to reservation behaviors	1. Whether a guest stayed at the hotel before or not 2. No. previous bookings canceled by guest 3. No. previous bookings not canceled by guest 4. No. adjustment made to the booking	1. "is_repeated_guest" 2. "previous_cancellations" 3. "previous_bookings_not_canceled" 4. "booking_changes"
Arbitrary	The number of travel companions	1. No. adults 2. No. children 3. No. babies 4. Total Number of companions (adults + children + babies)	1. "adults" 2. "children" 3. "babies" 4. "total_num_people"
Economic	The value of booked travel products	1. Average daily rate of the room booked 2. Free stay 3. How is the booking guaranteed, three categories: No deposit – no cancellation fee Refundable – low cancellation fee Non-Refundable – high cancellation fee	1. "adr" 2. "market_Complementary" 3. "deposit_type"
Non-Distance		Features that are not listed above	

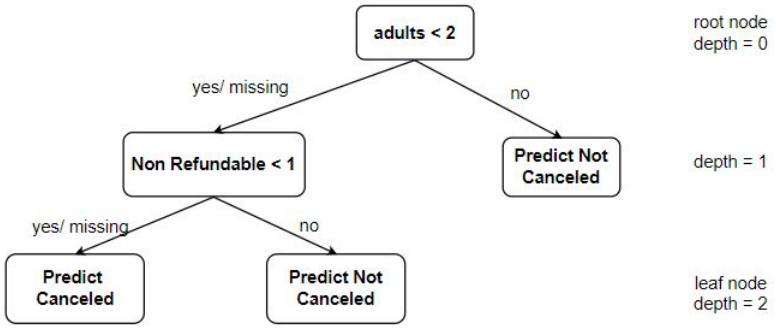


Figure 1: An example of gradient boost tree using *Hotel Booking Demand Dataset*. The node at depth 0 is the root node, all the nodes below the root node are the child nodes of its directly related node from the previous depth. For instance, the node with "Non Refundable < 1" is a child node of the root node. The nodes that make final predictions are called leaf nodes. In this case, the right node of depth 1 and both nodes of depth 2 are leaf nodes.

From top to bottom, the tree selects "adults" as the first feature for splitting. If an observation has "adults" equal or larger than 2, it will go to the right node of depth 1. The tree will predict "Not Cancelled" for that observation; If an observation has "adults" smaller than 2 or has a missing value, it will go to the left node of depth 1, and the tree continues to grow. If an observation has "Non Refundable" smaller than 1 or has a missing value, the observation will go to the left node of depth 2, and the tree will predict "Cancelled" for that observation; if an observation has "Non Refundable" equal or larger than 2, the observation will go the right node of depth 2, and the tree will predict "Not Cancelled" for that observation. Once all observations are predicted or the tree reaches the pre-defined maximum depth, the tree will stop growing. Each tree grows following the same procedures and takes account of the prediction from the previous tree. Once the prediction reaches a certain level or the number of trees reaches a maximum limit, the construction is complete.

3.3 *In-depth Feature analysis*

We implement three feature analysis tools in this research: XGBoost's build-in feature importance analysis, partial dependency plot, and correlation coefficient analysis.

XGBoost python library offers three types of feature importance analysis ("gain", "weight", "cover"). In this research, "gain" is used because it represents the average information gain of splits that use the feature. The outcome of the analysis uses F-score to represent the importance of each feature. Initially, the sum of all F-scores equals 1. For easier interpretation, we transform all F-scores to percentages, and the sum of all F-scores equals 100%.

Partial dependency plot (PDP) (Friedman, 2001) visualizes the impact of the features on the prediction of cancellations. Plot that varies more is associated with important features (Greenwell, Boehmke, & McCarthy, 2018). The PDP also illustrates whether a relationship between an individual feature and the prediction outcome is linear or monotonic.

Correlation coefficient analysis examines the relationships between any two given features. Since Partial dependency plot does not take account of the effect caused by feature interactions, correlation coefficient analysis becomes important because it helps with the interpretation of partial dependency plots. If features are not correlated, a partial dependency plot is more intuitive and straightforward because the plot illustrates how the changes of features influence the prediction. If any two features are highly correlated, the partial dependency plot may not be accurate and need to be interpreted with care.

4 EXPERIMENTAL SETUP

In this section, we present our research experiment procedures: datasets, data pre-processing, experimental implementation, software implementation, and model evaluation.

4.1 *Datasets*

This research primarily use the *Hotel Booking Demand Dataset*. Since the XGBoost classifier can only take in numerical data, all categorical data from the *Hotel Booking Demand Dataset* will be either converted into numerical data or encoded using the dummy coding. Because the categorical feature *Country* includes 125 unique values, using dummy coding will raise the feature dimensionality. To avoid high dimensionality, we include another dataset: *ISO-3166 Country and Dependent Territories Lists with UN Regional Codes Dataset*, which is used to replace feature *Country* with *Subcontinents*. This way, the number of features added after encoding will be reduced from 125 to 18.

Dataset 1. Hotel Booking Demand Dataset

We found the *Hotel Booking Demand Dataset* (2019b) on Kaggle.com. This dataset is collected and published by Antonio et al.. It consists of 119,390 individual observations and 31 variables, including raw data and pre-processed data. Each observation represents a booking made during the time between 1st July 2015 and 31st August 2017. The data is extracted from the property management systems (PMS) of two hotels located in Portugal and it contains information such as bookings change log, meals, distribution channels, transactions, customers profiles, nationalities, and market segments. As mentioned in section 3, a complete feature list is presented in Appendix 1. This dataset was originally collected for tackling the booking cancellation problem and was used for the research projects of Satu et al. (2020) and Putro et al. (2021).

Dataset 2. ISO-3166 Country and Dependent Territories Lists with UN Regional Codes

This dataset is published on GitHub by Duncalfe (2021). It contains 249 countries (rows) and 11 features. The data includes information such as country name, ISO code (3-letter), subcontinent, region code, and so on. The dataset is a result of merging data from two public sources: UN Statistics site and Wikipedia ISO 3166-1 article. For this research, only ISO code (3-letter) and Subcontinent are required. Table 2 presents the most relevant features and corresponding descriptions for this research.

Table 2: Dataset 2 features (used in this research) and corresponding descriptions

Features	**Description**
alpha-3 (Integer)	3-letter ISO code
sub-region (Categorical)	Subcontinent

4.2 *Data Pre-Processing*

We pre-process the dataset to deal with class imbalance, errors, missing data, composite feature, outliers, discarded features, and feature transformation. For simplicity, we use "Dataset 1" to represent *Hotel Booking Demand Dataset*, and "Dataset 2" to represent *ISO-3166 Country and Dependent Territories Lists with UN Regional Codes Dataset* in the rest of the section.

Imbalanced Dataset

For our binary classification task, the target classes are *cancelled* and *not cancelled*. According to our statistics, Dataset 1 is imbalanced with 63% of data labelled as *Not Cancelled* data and 37% of data labelled as *cancelled*. The imbalanced dataset may have impact on the performance of the model as it might predict the majority class by default to reach a 63% accuracy. The negative impact of the imbalanced dataset will be mitigated by assigning different weights to classes that is described in the experimental procedure (section 4.3).

Error Handling

A few errors were found in both datasets and need to be handled before merging data. For *country* in Dataset 1, 3-letter ISO codes are used to represent almost all unique countries except a 2-letter ISO code "CN' is used to represent China. To keep the consistency in the dataset, we replace "CN" by "CHN" for China. Dataset 2 did not contain information corresponding to "TMP" (*Country*) from Dataset 1. Thus, we include a new record of "TMP" in Dataset 2. We also add "Antarctica" (*Sub-region*) to Dataset 2 to match the "ATA" (*Country*) from Dataset 1.

Missing Data Handling

Table 3 presents an overview of missing value counts after we include *Subregions* in Dataset 1. *Company* is removed from the dataset because it has 112593 missing values which takes up 94% of the total number of bookings. As for *children*, *Country*, and *Subregions*, the numbers of missing values are relatively small and can be replaced by the mode (value that appears the most frequently) without affecting the representativeness of

Table 3: Dataset 1 Missing Value Overview

Feature	Number of missing values	Percentage
Children	4	0.003%
Country	488	0.41%
Agent	16340	13.69%
Company	112593	94.31%
Subregions	488	0.41%

the dataset. Finally, *Agent* has 16340 missing values, which takes up nearly 14% of the total number of bookings. In this case, replacing all missing data with a single number may increase model bias. Therefore, we use KNN imputation to predict missing values for *Agent* after data is split. More details on KNN imputation is presented in section 4.3 Experimental procedure.

Composite feature

We add a composite feature to the dataset : *Total Number Companions*. This feature is calculated as the sum of three existing numerical features: *Adults*, *Children*, and *Babies*. According to Lee et al. (2021), the number of companions affect the cancellations, while Antonio et al. (2019a) specifically points out that *Adults* is one of the most influential features for predicting cancellations. Therefore, we add *Total Number Companions* to examine whether the *Adults* have more predictive power than *Total Number Companions*, or the other way around.

Outlier Handling

Dataset 1 contains both numerical and categorical features. For both types of features, we handle outliers.

We use univariate outlier analysis to detect potential outliers from a large number of numerical features. Table 4 presents the univariate outlier analysis result for the numerical features of Dataset 1. The result shows that *Lead Time*, *Adults*, *Previous Bookings Not Cancelled* , *booking Changes*, *Days In Waiting List*, and *ADR* may contain extreme large values and requires further analysis.

We also draw boxplots for the numerical features with potential outliers. Figure 2 illustrates the plots of *Lead Time*, *Adults*, *Previous Bookings Not Cancelled* , *booking Changes*, *Days In Waiting List*, and *ADR*. After carefully considered the uniqueness of the potential outliers and the distances between potential outliers and the rest data points, we remove 6 outliers from Dataset 1: 2 from *Lead Time*, 3 from *Adults*, and 1 from *ADR*.

Table 4: Dataset 1 univariate numerical features analysis result

	count	mean	std	min	25%	50%	75%	max
is_canceled	119390.0	0.370416	0.482918	0.00	0.00	0.000	1.0	1.0
lead_time	119390.0	104.011416	106.863097	0.00	18.00	69.000	160.0	737.0
arrival_date_year	119390.0	2016.156554	0.707476	2015.00	2016.00	2016.000	2017.0	2017.0
arrival_date_week_number	119390.0	27.165173	13.605138	1.00	16.00	28.000	38.0	53.0
arrival_date_day_of_month	119390.0	15.798241	8.780829	1.00	8.00	16.000	23.0	31.0
stays_in_weekend_nights	119390.0	0.927599	0.998613	0.00	0.00	1.000	2.0	19.0
stays_in_week_nights	119390.0	2.500302	1.908286	0.00	1.00	2.000	3.0	50.0
adults	119390.0	1.856403	0.579261	0.00	2.00	2.000	2.0	55.0
children	119390.0	0.103886	0.398555	0.00	0.00	0.000	0.0	10.0
babies	119390.0	0.007949	0.097436	0.00	0.00	0.000	0.0	10.0
is_repeated_guest	119390.0	0.031912	0.175767	0.00	0.00	0.000	0.0	1.0
previous_cancellations	119390.0	0.087118	0.844336	0.00	0.00	0.000	0.0	26.0
previous_bookings_not_canceled	119390.0	0.137097	1.497437	0.00	0.00	0.000	0.0	72.0
booking_changes	119390.0	0.221124	0.652306	0.00	0.00	0.000	0.0	21.0
agent	103050.0	86.693382	110.774548	1.00	9.00	14.000	229.0	535.0
days_in_waiting_list	119390.0	2.321149	17.594721	0.00	0.00	0.000	0.0	391.0
adr	119390.0	101.831122	50.535790	-6.38	69.29	94.575	126.0	5400.0
required_car_parking_spaces	119390.0	0.062518	0.245291	0.00	0.00	0.000	0.0	8.0
total_of_special_requests	119390.0	0.571363	0.792798	0.00	0.00	0.000	1.0	5.0
total_num_people	119390.0	1.968239	0.722394	0.00	2.00	2.000	2.0	55.0

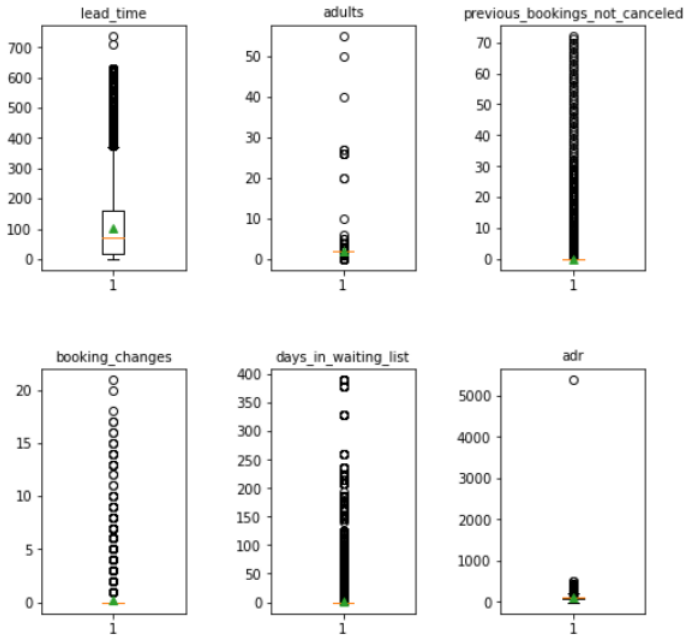


Figure 2: Boxplots of *Lead Time*, *Adults*, *Previous Bookings Not Cancelled* , *booking Changes*, *Days In Waiting List*, and *ADR* from Dataset 1.

Table 5: Dataset 1 categorical feature outliers analysis result

Feature	Unique Value	Counts	Percentage% (Normalized count * 100)
Market Segment	Undefined	2	0.0017%
Distribution Channel	Undefined	5	0.0042%
Reserved Room Type	L	6	0.0050%
Assigned Room Type	L	1	0.0008%
Subregions	Micronesia	3	0.0025%
Subregions	Polynesia	2	0.0017%
Subregions	Antarctica	2	0.0017%
Subregions	Melanesia	2	0.0017%
Total		23	0.0193%
Total (actual)		20	0.0168%

For categorical features, we perform outlier analysis by counting unique values of each categorical feature and normalize the counts of each feature. If the normalized count of any unique feature value is extremely small, this unique value is treated as outlier and removed from the dataset. Table 5 presents our categorical outlier analysis result . *Total* represents the sum of all counts from above cells. *Total (actual)* represents the total number of deleted observations. The difference between *Total* and *Total (actual)* is caused by overlapped observations. In summary, we remove 20 outliers from Dataset 1, which constitute 0.0168% of total observations.

Discarded Features

We remove *Company*, *Reservation Status*, *Reservation Status Date*, and *Country* from the *Hotel Booking Demand Dataset*. *Company* is removed due to large number of missing data. *Reservation Status* is removed due to similar information as what the model is trying to predict. *Reservation Status Date* represents the last update date of a reservation status, it is removed because it is only relevant along with *Reservation Status*. Finally, *Country* is removed because we use *Sub-region* to replace it.

Feature Transformation

After we performed previous steps, Dataset 1 should contain 10 categorical features, of which 4 are ordinal data and 6 are nominal data. Since XGBoost algorithm can only process numerical data we implement feature transformation.

Table 6: Dataset 1 nominal features that require encoding

Nominal features	Number of unique values
Hotel	2
Market Segment	7
Distribution Channel	4
Deposit type	3
Customer Type	4
Subregions	14
Total	**34**

Convert ordinal features to numerical

Reserved Room Type, *Assigned Room Type*, *Month*, and *Meal* are the 4 ordinal features. To convert those features into numerical, we simply replace each unique value with a unique number for each feature.

Encode nominal features

Table 6 presents the 6 nominal features and the corresponding number of unique values they have in Dataset 1. Nominal features cannot be converted directly into numerical due to its lack of mathematical meaning, thus, those features need to be encoded. For this research, we apply the build-in dummies encoding function from Pandas' Python Library (McKinney et al., 2010). After encoding process, the 6 features are replaced by 34 dummy codes, and Dataset 1 now contains 58 features.

4.3 *Experimental Procedure*

After data is pre-processed, we create two additional datasets based on the outcome of feature categorization (section 3). Now we have 3 datasets: one contains only non-distance features, one contains only psychological-distance features, and one contains both psychological-distance and non-distance features. For each dataset, we conduct data split, data normalization, missing data imputation, and DMatrix representation.

Firstly, we separate target variable *Is Cancelled* from Dataset 1. Then, we split 80% of Dataset 1 and targets to the training set, the other 20% to the test set. After that, we standardize the data to speed up learning process and save memory. Next, we use KNN imputation to treat the missing values of *Agent*. KNN imputation implements the K-Nearest Neighbors method with Euclidean distance metric as default to replace missing values with the mean value of k nearest neighbors found in the training set. Missing value imputation is carried out after data split to prevent data leakage and overfitting. Finally, we use DMatrix from the XGBoost Python

Table 7: Models A, B, C and their corresponding feature categories

Model	Features
A	Non-distance features
B	Psychological-distance features
C	Both non-distance and psychological-distance features

Library to represent the data to further optimize memory efficiency and training speed.

After the data is ready, we then develop three XGBoost-based models using the three pre-defined datasets, respectively. We also tune our models individually and we expect to obtain distinctive hyperparameters. Table 7 presents the three models and the corresponding feature categories used for training.

Hyperparameters

The XGBoost Python Library provides a large number of parameters for tuning. In this subsection, we present the most relevant hyperparameters for our classification task.

"eta" is the learning rate that controls the step size in update at each boosting step. A large value indicate a fast update while a smaller value prevents overfitting.

"gamma" is the minimum split loss reduction required in a leaf node in order to make a further partition.

"max_depth" is the maximum depth a tree can grow. A large value indicate a more complex tree which may lead to overfitting and consume more computation power.

"min_child_weight" is the minimum sum of instance weight required in a child node to make a further partition.

"subsample" is the ratio of training samples used in each booster iteration. This value controls the number of instances that is used to build a tree and it prevents overfitting.

"colsample_bytree" is the ratio of features used in each booster iteration that prevents overfitting.

"scale_pos_weight" is a value that controls the weight of classes for imbalanced classes. The *Hotel Booking Demand Dataset* is clearly imbalanced with 63% *Not Cancelled* and 37% Cancelled. In our case, "scale_pos_weight" is set to 1.7 for all models.

"tree_method" is the tree construction algorithm. When using one's own computer for training, the algorithm will choose the fastest method for the user. Since we use Google Colab with an external GPU to train our models, the tree_method need to be set to 'gpu_hist' (GPU implementation of faster histogram optimized approximate greedy algorithm) manually.

"eval_metric" represents the evaluation metric used for monitoring model performances. Error and AUC are the two evaluation metrics specified for our classifier. Error is the default evaluation metric for binary classification task and AUC is recommended in the XGBoost documentation for handling imbalanced dataset.

"random_state" is set to 42 for all models in order to duplicate the same output.

"early_stopping" is set to 5 for all models. Implementing early_stopping prevents overfitting and improves time efficiency.

4.4 *Software*

For our classification task, we use Python (version 3.6.12) programming language (Van Rossum & Drake, 2009) in Jupyter Notebook (Kluyver et al., 2016) and Google Colab (Bisong, n.d.). Pandas (McKinney et al., 2010) and Numpy (Harris et al., 2020) libraries are used mostly for data representation and preparation. Matplotlib (Hunter, 2007) and Seaborn (Waskom et al., 2017) libraries are used for data visualization. The Xgboost Python library (T. Chen & Guestrin, 2016) and XGBoost scikit-learn wrapper (T. Chen & Guestrin, 2016) are both used for model training, cross validation and hyperparameter tuning, prediction and feature importance analysis. We also use sklearn library (Pedregosa et al., 2011) for missing data imputation, data normalization, and model evaluation.

4.5 *Model Evaluation*

The *Hotel Booking Demand Dataset* is imbalanced and the classification task has a focus on cancelled bookings. Therefore, we use the recall, F1-score, and AUC to evaluate and compare the performances of model A, B, and C. The recall measures how well the model captures cancelled bookings. The F1-score measures the performance of an imbalanced dataset using a pre-defined classification threshold (0.5). The AUC score helps with determining the best model. In particular, macro average recall and macro average F1-score are used because it gives even attention to all classes regardless of the proportion of each class in the dataset. Finally, we initialized a baseline model with both recall and the F1 score equals 50%.

5 RESULTS

In this section, we present model performances and in-depth feature analysis which includes feature importance, partial dependency plots, and correlation coefficient. It is important to note that a baseline model has been defined with recall and F1-score all set to 50%. Besides, due to the implementation of early stopping, the "n_estimators" of models A, B, and

Table 8: Best hyperparameter set of model A

max depth	learning rate (eta)	n_estimators (num_boost_round)	min_child_weight	subsample	colsample_bytree
25	0.001	300	0	0.6	0.7

Table 9: Classification report of model A on the test set

	Precision (%)	Recall (%)	F1-score (%)	support
not_cancelled	80	95	87	15525
cancelled	86	56	68	8348
Accuracy			81	23873
Macro avg	83	76	77	23873
Weighted avg	82	81	80	23873

C presented in Table 8, 10, and 13 are not the best parameter but the maximum number of estimators implemented for each boosting iteration.

5.1 *Model A (Non-Distance Only)*

Model A is trained using only non-distance features. Table 8 presents the best hyperparameter set of model A and Table 9 presents the classification report for the test set.

The macro average precision is 83% and the weighted average precision is 82%. The difference between the two averaged values is only 1%. Both values are higher than the best precision (76.99%) achieved by Antonio et al. (2019a). When it comes to the recall, model A classified 95% of all *Not Cancelled* observations correctly as *Not Cancelled*. On contrary, model A classified only 56% of all *Cancelled* observations correctly as *Cancelled*. The F1-scores for model A predicting *Not Cancelled* and *Cancelled* are 87% and 68%, respectively. The difference between the F1-scores is largely influenced by the difference in the recalls. Overall, the macro average F1-score is 77% for model A.

5.2 *Model B (Psychological Distance Only)*

Model B is trained using only psychological-distance features. Table 10 presents the best hyperparameter set of model B and Table 11 presents the classification report of model B for the test set.

The macro average precision is 84%, which is slightly higher than the macro average precision (83%) of model A. When it comes to the recall, model B classified 94% of all *Not Cancelled* observations correctly as *Not*

Table 10: Best hyperparameter set of model B

max depth	learning rate (eta)	n_estimators (num_boost_round)	min_child_weight	subsample	colsample_bytree
25	0.1	300	0	0.4	0.9

Table 11: Classification report of model B on the test set

	Precision (%)	recall (%)	F1-score (%)	support
not_cancelled	82	94	88	15525
cancelled	85	62	72	8348
Accuracy			83	23873
Macro avg	84	78	80	23873
Weighted avg	83	83	82	23873

Cancelled and 62% of all *Cancelled* observations as *Cancelled*. Although the difference (32%) remains large, model B has a higher recall for *Cancelled* than model A. When it comes to F1-score, model B has 88% for predicting *Not Cancelled* and 72% for predicting *Cancelled*. Similar to model A, the difference is largely influenced by the difference in the recall. Overall, the macro average F1-score for model B is 80%.

Feature importance analysis

Figure 3 illustrates the outcome of the feature important analysis of model B. The x-axis represents the F-scores in percentage and the y-axis represents features. The sum of all F-scores equals 100%. As a rule of thumb, a high F-score indicates high importance while a lower F-score indicates low importance. As the result indicates, *Non-Refundable* booking ("deposit_Non Refund") has a F-score of 80.18%, which makes it the most important feature for predicting cancellations for model B. *Average Daily Rate* ("adr") with a F-score of 2.93% and *Lead Time* with a F-score of 2.78% are the second and third importance features, respectively. However, the differences between them and *Non-Refundable* booking is significantly large. In fact, *Non-Refundable* booking is 4 times as important as the sum of all the other features.

Based on the feature categorization in Section 3, we sum up the F-scores of relevant features for each type of distance and we present the result in table 12. As the result indicates, economic distance has the highest importance with an F-score of 86.06%. As previously pointed out, 80.18% of the importance comes from *Non-Refundable* booking. Arbitrary (F-score = 6.96%) and experiential (F-score = 4.19%) distances are corresponding to the second and third important distance, respectively. Temporal is the least important distance with a F-score of 2.78%.

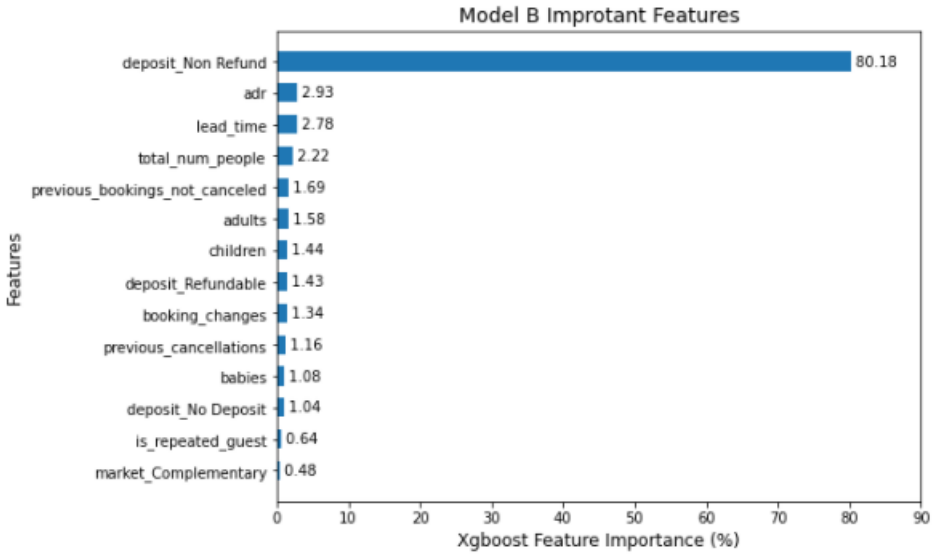


Figure 3: Result of model B feature importance analysis

Table 12: Model 2 F-score per distance

Distance	Importance
Temporal	2.78%
Experiential	4.19%
Arbitrary	6.96%
Economic	86.06%

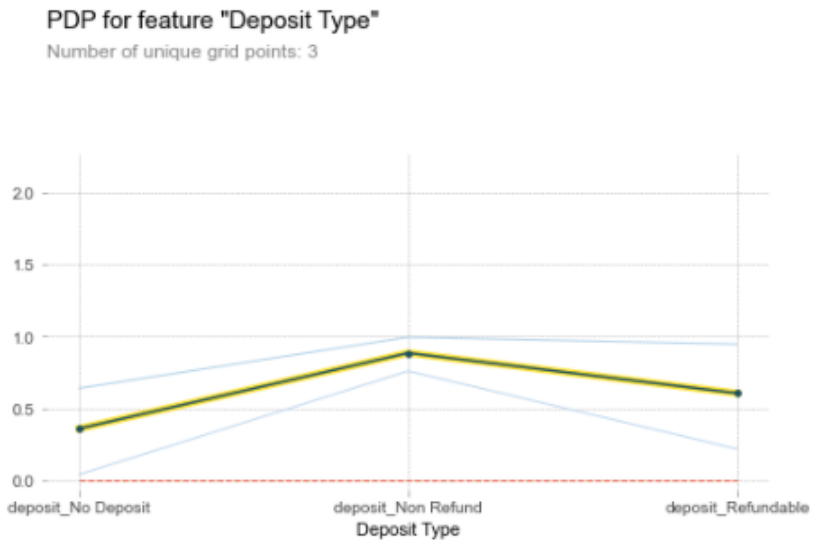


Figure 4: Deposit Types PDP

Partial dependency plots (PDP)

For each partial dependency plot, the x-axis represents a feature and the y-axis represents partial dependence. The value on the y-axis varies from 0 to 1, which is the range of prediction probability for *Cancelled* (1). Taking the default classification threshold (0.5) of the XGBoost, observations with a probability above 0.5 are classified as *Cancelled* (1), else *Not Cancelled* (0). The bold line represents the effect of the feature in terms of changes in prediction probability. The two light blue lines represent the range of confidence intervals.

Since *Refundable* ("deposit_Refundable"), *Non-Refundable* ("deposit_Non Refund"), and *No Deposit* ("deposit_No Deposit") are the unique values (dummies codes) of *Deposit Types* from the original dataset, we draw a partial dependency plots for *Deposit Types* to better understand the impact of the original feature. As figure 4 illustrates, when *Deposit Types* changes from *No Deposit* to *Non-Refundable*, the probability increased from 0.4 to 0.9 which shows a positive impact when predicting *cancelled*. When *Deposit Types* changes from *Non-Refundable* to *Refundable*, the probability decreased from 0.9 to 0.6. However, the PDP value remains above 0.5. The confidence interval is narrow at *Non-Refundable*, which indicates a strong relationship between non-refundable booking and cancelled booking. The curve of the plot indicates a relatively strong nonlinear relationship between deposit and the outcome of the prediction.

Figure 5 illustrates the PDP of *Previous Cancellations*. The shape of the plot indicates a positive linear relationship between *Previous Cancellations* and cancelled booking. As *Previous Cancellations* increases from 0 to 24, the probability for cancelling the booking increases from 0.4 to 0.8. Mean-

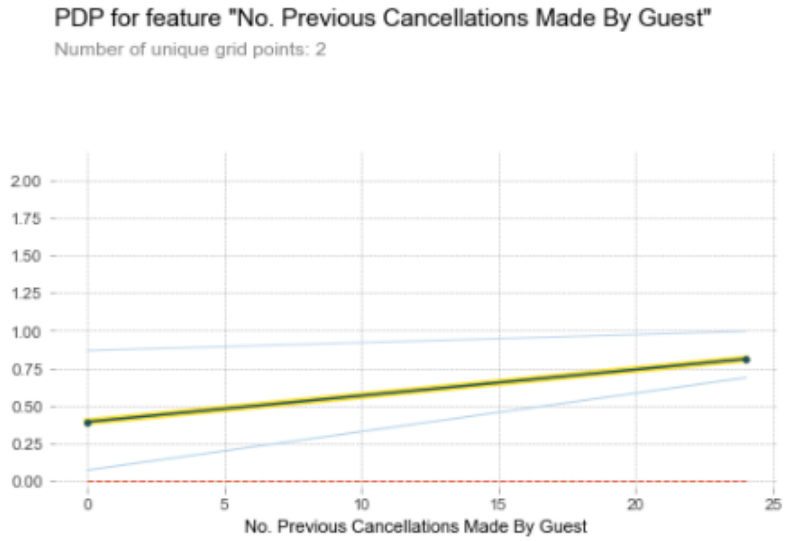


Figure 5: No. previous cancellations made by guest PDP

while, the confidence interval becomes narrower indicating the relationship becomes stronger.

Apart from *Deposit Types* and *Previous Cancellations*, we also draw PDP for all the other features and the plots can be found in Appendix 2. None of the features have positive relationship with cancelled bookings, and therefore, they show little impact when detecting cancellations.

Correlation Coefficient Analysis

Figure 6 illustrates a heatmap of the correlation coefficients analysis. On the right side of Figure 6, the bottom of the color bar indicates a strong negative relationship, the top of the color bar indicates a strong positive relationship, and the middle point 0 indicates no relationship. As a rule of thumb, coefficients with an absolute value equal or less than 0.35 are considered weak relationships, coefficients with an absolute value between 0.36 and 0.67 are considered moderate relationships, coefficients with an absolute value above 0.68 are considered strong relationships (Taylor, 1990).

Since *Total Number Companions* includes *Children* and *Adults*, there is a moderate-to-strong and positive linear relationship between them. Besides, there is a strong negative correlation between two unique values of *Deposit Types*: *No Deposit* and *Non-Refundable*, but they will not occur at the same time, and therefore, their correlation does not affect the partial dependency analysis.

On the other hand, there are a few relationships that may affect the result of the partial dependency analysis: a moderate and negative correlation between *Lead Time* and *No Deposit*; a moderate and positive correlation

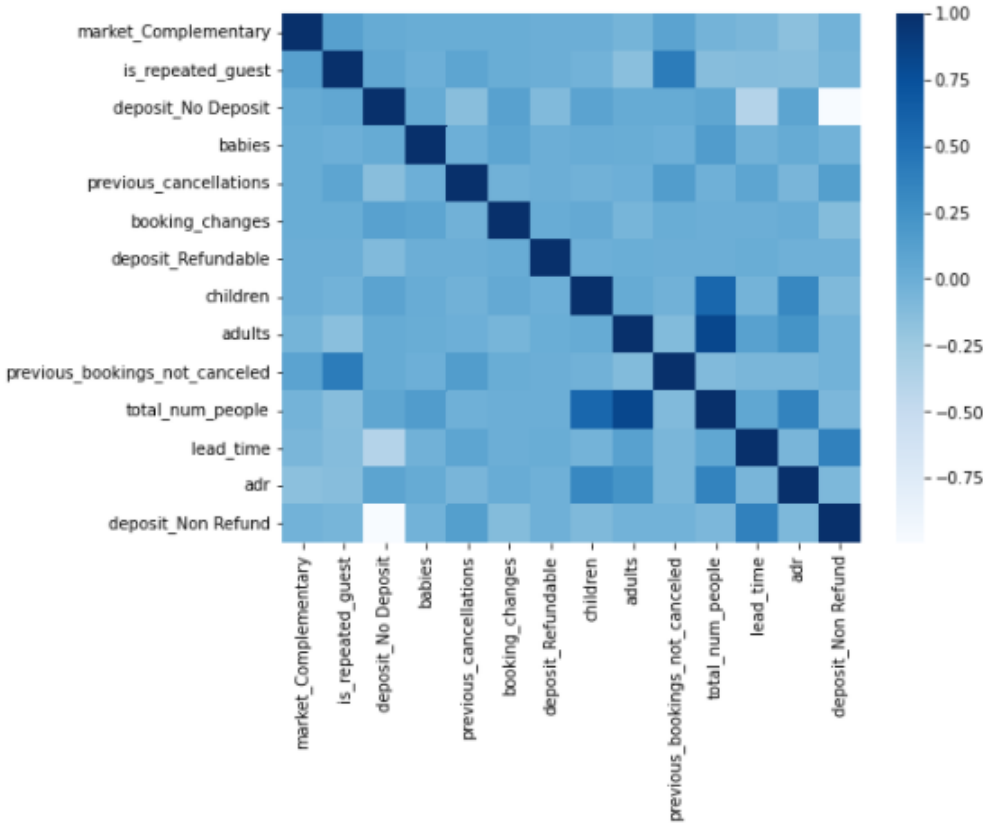


Figure 6: Model B correlation heatmap

between *Lead Time* and *Non-Refundable*; a moderate and positive correlation between *Repeated Guest* and *Previous Bookings Not Canceled*.

5.3 *Model C (All Features)*

Model C is trained using both psychological-distance and non-distance features. Table 13 presents the best hyperparameter set and Table 14 presents the classification report of model C on the test set.

The macro average precision for model C is 91%, which is 9% higher than model A (83%) and 8% higher than model B (84%). Model C classified 97% of all *Not Cancelled* observations correctly as *Not Cancelled* and 75% of

Table 13: Best hyperparameter set of model C

max depth	learning rate (eta)	n_estimators (num_boost_round)	min_child_weight	subsample	colsample_bytree
15	0.1	100	1	0.5	0.6

Table 14: Classification report of model C on the test set

	Precision (%)	recall (%)	F1-score (%)	support
not_cancelled	88	97	92	15525
cancelled	94	75	83	8348
Accuracy			89	23873
Macro avg	91	86	88	23873
Weighted avg	90	89	89	23873

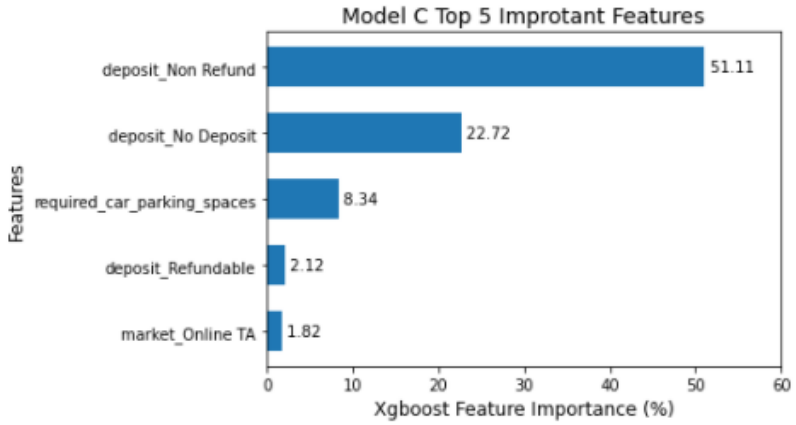


Figure 7: Model C feature importance analysis (top 5)

all *Cancelled* observations as *Cancelled*. When comparing to model B, we see an increase in the recall for both *Not Cancelled* and *Cancelled*. In particular, the increase of 13% in the recall for *Cancelled* is considerably large. When it comes to F1-score, model C has 92% for *Not Cancelled* and 83% for *Cancelled*. Similar to model A and B, the difference is largely influenced by the difference in the recall. Finally, the macro average F1-score for model C is 88%, which is the highest F1-score among all three models.

Feature Importance Analysis

Model C involves 57 features. Most of the features have an F-score less than 1%. For time efficiency, we only focus on the features that have a F-score larger than 1%. Figure 7 illustrates the only 5 features of model C that fall in the criteria: *Non-Refundable*, *No Deposit*, *Parking Space Required*, *Refundable*, and *Online Travel Agency*. The 5 features corresponding to F-scores of 51.11%, 22.72%, 8.34%, 2.12%, and 1.82%, respectively. The sum of the three *Deposit Type*: *Non-Refundable*, *No Deposit*, and *Refundable* reaches a F-score of 75.95%. In particular, *Non-Refundable* has the highest F-score, and therefore, becomes the most important feature for model C. In general, the sum of the 5 F-scores reaches 86.11%, which leaves an F-score of 13.89% to be shared by the rest 52 features.

Table 15: Model C F-score per distance

Distance	Feature	F-score	F-score per Distance
Temporal	Lead Time	0.19%	0.19%
Experiential	Repeated Guest	0.22%	2.14%
	Number Previous Cancellations	1.39%	
	Number Previous Not Cancelled	0.29%	
	Booking Changes	0.24%	
Arbitrary	Adults	0.12%	0.5%
	Children	0.12%	
	Babies	0.12%	
	Total Number Companions	0.14%	
Economic	ADR	0.13%	76.08%
	Free Stay	0.00%	
	Deposit Type	75.95%	
Non-Distance	Features that are not listed above		21.59%

Table 15 presents the feature importance categorized by distances. As we can see, economic distance has the highest F-score (76.08%) among all psychological-distance and non-distance features, and its vast majority of scores come from *Deposit Types*. The second highest F-score (21.59%) is from Non-distance features. In particular, *Parking Space Required* and *Online Travel Agency* are the third and fifth important features for model C. On the other hand, temporal (0.19%), experiential(2.14%), and arbitrary (0.5%) distance are proven to have little impact on the prediction of cancellations.

5.4 *Area Under Curve (AUC)*

Figure 8 presents the AUC scores of all three models: model A, B, and C are corresponding to AUC scores 0.91, 0.82, 0.96, respectively. The result suggests that a model trained using non-distance features has a better overall performance than a model trained using psychological-distance features. In addition to that, a model trained using both psychological-distance and non-distance features has the best overall performance.

6 DISCUSSION

The goal of this research is to investigate the true impact of psychological distance on predicting booking cancellations, using a machine learning algorithm. In particular, psychological distance contains 5 distances: spatial, temporal, experiential, arbitrary, and economic. In order to achieve the goal, we first categorized the features from the original dataset and create

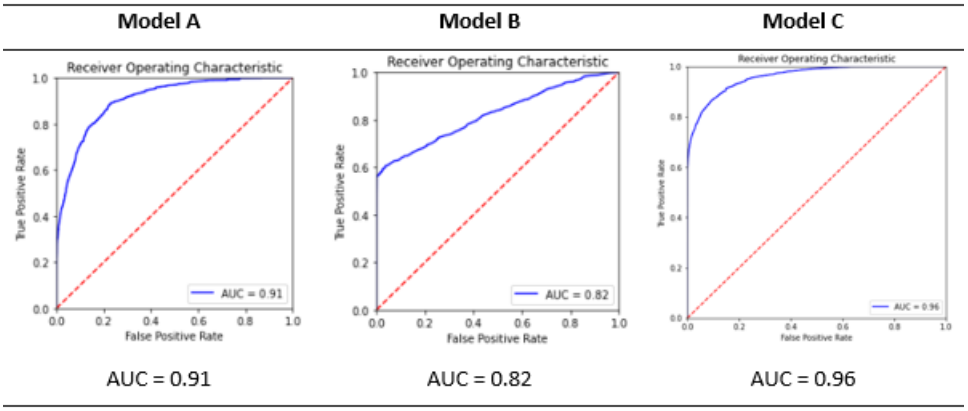


Figure 8: ROC curve & AUC scores of models A, B, and C

three datasets: non-distance dataset, psychological-distance dataset, and all-feature dataset. After that, we apply the XGBoost algorithm and build models A, B, and C using the the three datasets, respectively. In this section, we will briefly discuss the results in combination with literature.

First, we build model A using non-distance dataset. Both the macro average recall (76%) and the macro average F1-score (77%) suggest that model A performs better than the baseline model. However, the recall for bookings that are cancelled is 39% less than the recall for bookings that are not cancelled, which suggests that model A is less sensitive to cancelled bookings.

Then, we build model B using the psychological-distance dataset. The macro average recall (78%) and macro average F1-score (80%) suggests that model B performs better than the baseline model and model A, which implies cancellations can be better identified when using psychological-distance features alone in comparison to using non-distance features alone.

Based on the feature importance analysis, non-refundable booking has the highest F score (80.18%), and therefore, it becomes the key feature for predicting cancellations. Economic distance has the highest F-score (86.06%), and therefore, it becomes the most important psychological distance.

The partial dependency plot of the *Deposit Types* implies that *Non-Refundable* booking is very likely to be cancelled, *Refundable* booking is likely to be cancelled but the chance is lower than *Non-Refundable* booking. *No Deposit* booking is unlikely to be cancelled. In the hotel industry, *Non-Refundable*, *Refundable*, and *No Deposit* bookings are associated with three tiers of cancellation fees: high, low, and zero, respectively. When we apply this relationship to the result of the partial dependency plot, we see that as cancellation fee increases, bookings become more likely to be cancelled. Surprisingly, this finding contradicts the impact of the economic distance presented in the literature and it is against the purpose of cancellation

Table 16: Hotel booking demand dataset - target class by deposit type

Deposit Type	Cancelled	Not Cancelled
No Deposit	29694	74947
Non-Refund	14494	93
Refundable	36	126

fees. To rule out the possibility of mistakes during the experiment, we went back to the original dataset and counted the number of *cancelled* and *Not Cancelled* bookings by their corresponding deposit type. The result presented in Table 16 indicates the number of non-refundable booking that are cancelled is significantly larger than not cancelled. At this moment, we have no other explanation for the contradictory trend.

In addition to *Non-Refundable*, *Previous Cancellations* also has a positive relationship with cancelled bookings. Although this is an interesting finding, the F-score (1.16%) suggests low importance. Hence, the impact of *Previous Cancellations* on the prediction is relatively low. In addition to that, the rest of the features have very little impact on the prediction of cancellations.

The correlation coefficient analysis presents two relationships which may influence the accuracy of the partial dependency plots. The first one is the relationship between *Lead Time* and *Deposit Type*. Second one is the relationship between *Repeated Guest* and *Previous Bookings Not Cancelled*. To better understand the impact of these features on the predictions requires feature interaction analysis, which was not conducted in this research.

Finally, we build model C using all-feature dataset. the macro average recall (86%) and macro average F1-score (88%) suggest that model C outperforms model A and model B. Hence, combining psychological-distance and non-distance features gives the model an advantage when predicting cancellations.

The feature importance analysis presents 5 important features: *Non-Refundable*, *No Deposit*, *Parking Space Required*, *Refundable*, and *Online Travel Agency*. Same as model B, *Non-Refundable* is the most important feature with an F-score of 51.11%. The 5 features contradicts the features presented by Antonio et al. (2019a). However, the research conducted by Antonio et al. (2019a) includes 8 models. Each model uses an unique hotel dataset. The results of the feature importance analysis conducted for the 8 hotels also turn out to be contradictory with each other, which implies that the data used for predicting cancellations may differ according to different hotel situations. Furthermore, psychological distance has a F-score of 79% and non-distance has a F-score of 21%, which indicates psychological-distance features has more impact than non-distance features on the prediction of

cancellations. In particular, economic distance has significant larger impact on predicting cancellations than all the other distances and non-distance add up together.

Finally, we would like to discuss model performance based on the AUC and F1-score. Model A has a higher AUC (0.91) and lower F1-score (77%), while model B has a lower AUC (0.82) and higher F1-score (80%). The conflict between the F1 score and AUC may indicate that the model trained using the non-distance dataset has better overall performance when taking account of all thresholds; while the default threshold (0.5) yields a better prediction for model B. Therefore, the contradiction between AUC and F1 scores can be caused by the choice of the threshold used for classifying prediction probability. Due to time constraints, this research only used the default threshold of 0.5. However, the impact of thresholds on predicting cancellations can vary and is worth investigating. On the other hand, model C has the highest F1 (88%) score and AUC score (0.96). Therefore, we conclude that a model trained using both psychological-distance features and non-distance features is considered to be the best-performing model.

7 CONCLUSION

This research aims to explore the impact of psychological distance towards the prediction of booking cancellations. We expect that this research would reveal important relationships between psychological distance and prediction, and provide answers to the sub-research questions.

Does psychological distance have an impact on predicting hotel booking cancellations?

We found out that the psychological distance certainly have an impact on the prediction. The model trained with only psychological-distance features outperforms the baseline model.

Which distance has the most impact when predicting hotel booking cancellations?

It is proven that economic distance plays an extremely important role when predicting booking cancellations. In particular, it is observed that non-refundable bookings are very likely to be canceled and therefore, needs more attention. Temporal distance, experiential distance, and arbitrary distance did not show a significant impact in this research.

How do psychological-distance features relate to non-distance features with respect to the ability of predicting hotel booking cancellations?

When present alone, psychological-distance features has a greater impact on the prediction of booking cancellations than non-distance features, because the model trained using solely psychological-distance features

has higher recall and F1-score. However, combining both psychological-distance and non-distance features yields the best prediction, because the model trained using both types of features outperforms the other two models. In particular, non-refundable booking and economic distance prove its importance once again, and therefore, the model gives more weight to the psychological distance. Hence, we conclude that the prediction of booking cancellations is influenced by both psychological-distance and non-distance features, of which psychological-distance features have a greater impact on the prediction.

This research has a few limitations and the results of this research should be interpreted with care. First, this research only examines the impact of individual features on the predictions, in spite of the fact that there are interrelated relationships between some of the features. Second, the impact of spatial distance cannot be measured due to the limitation of the original dataset. Finally, due to time constraints, we only tested the default classification threshold. Based on above limitations, we recommend conducting further researches projects on the impact of feature interactions, spatial distance, and classification threshold on the prediction of cancellations.

8 REFERENCES

REFERENCES

Antonio, N., de Almeida, A., & Nunes, L. (2017). Predicting hotel bookings cancellation with a machine learning classification model. In *2017 16th IEEE international conference on machine learning and applications (ICMLA)* (pp. 1049–1054). doi: 10.1109/ICMLA.2017.00-11

Antonio, N., de Almeida, A., & Nunes, L. (2019a). Big data in hotel revenue management: Exploring cancellation drivers to gain insights into booking cancellation behavior. *Cornell Hospitality Quarterly*, *60*(4), 298–319. Retrieved 2021-11-30, from https://journals.sagepub.com/doi/abs/10.1177/1938965519851466?journalCode=cqxb

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

Bisong, E. (n.d.). Google colaboratory. In E. Bisong (Ed.), *Building machine learning and deep learning models on google cloud platform: A comprehensive guide for beginners* (pp. 59–64). Apress. doi: 10.1007/978-1-4842-4470-8_7

Chen, C. (2016). Cancellation policies in the hotel, airline and restaurant industries. *Journal of Revenue and Pricing Management*, *15*(3), 270–275. doi: 10.1057/rpm.2016.9

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *KDD '16: The 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794). ACM. doi: 10.1145/2939672.2939785

D-edge. (2019). *HOW ONLINE HOTEL DISTRIBUTION IS CHANGING IN EUROPE.* Retrieved 2021-11-30, from https://www.d-edge.com/how-online-hotel-distribution-is-changing-in-europe/

Dhar, R., & Kim, E. Y. (2007). Seeing the forest or the trees: Implications of construal level theory for consumer choice. *Journal of Consumer Psychology*, *17*(2), 96–100. doi: 10.1016/S1057-7408(07)70014-1

Duncalfe, L. (2021). *lukes/ISO-3166-countries-with-regional-codes.* Retrieved 2021-11-30, from https://github.com/lukes/ISO-3166-Countries-with-Regional-Codes

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. doi: 10.1214/aos/1013203451

Greenwell, B. M., Boehmke, B. C., & McCarthy, A. J. (2018). A simple and effective model-based variable importance measure. *arXiv:1805.04755 [cs, stat]*. Retrieved 2021-11-30, from http://arxiv.org/abs/1805.04755

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with

NumPy. *Nature*, *585*, 357–362. doi: 10.1038/s41586-020-2649-2

Hu, Z. J. (2009). *Revenue management*. . Retrieved from https://books.google.nl/books?id=iBMjQwAACAAJ

Hueglin, C., & Vannotti, F. (2001). Data mining techniques to improve forecast accuracy in airline business. In *Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 438–442). doi: 10.1145/502512.502578

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in science & engineering*, *9*(3), 90–95.

Jang, Y., Chen, C.-C., & Miao, L. (2019). Last-minute hotel-booking behavior: The impact of time on decision-making. *Journal of Hospitality and Tourism Management*, *38*, 49–57. doi: 10.1016/j.jhtm.2018.11.006

Kassinis, G. I., & Soteriou, A. C. (2015). Environmental and quality practices: using a video method to explore their relationship with customer satisfaction in the hotel industry. *Operations Management Research*, *8*(3), 142–156. doi: 10.1007/s12063-015-0105-5

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., ... Willing, C. (2016). Jupyter notebooks – a publishing format for reproducible computational workflows. In F. Loizides & B. Schmidt (Eds.), *Positioning and power in academic publishing: Players, agents and agendas* (p. 87 - 90).

Lee, H., Chung, N., & Lee, C.-K. (2017). Flight cancellation behaviour under mobile travel application: Based on the construal level theory. In R. Schegg & B. Stangl (Eds.), *Information and communication technologies in tourism 2017* (pp. 417–430). Springer International Publishing. doi: 10.1007/978-3-319-51168-9_30

Lee, H., Yang, S.-B., & Chung, N. (2021). Out of sight, out of cancellation: The impact of psychological distance on the cancellation behavior of tourists. *Journal of Air Transport Management*, *90*, 101942. doi: 10.1016/j.jairtraman.2020.101942

Lemke, C., Riedel, S., & Gabrys, B. (2013). Evolving forecast combination structures for airline revenue management. *Journal of Revenue and Pricing Management*, *12*(3), 221–234. doi: 10.1057/rpm.2012.30

Lichtenstein, D. R., Ridgway, N. M., & Netemeyer, R. G. (1993). Price perceptions and consumer shopping behavior: A field study. *Journal of Marketing Research*, *30*(2), 234–245. doi: 10.1177/002224379303000208

Massara, F., & Severino, F. (2013). Psychological distance in the heritage experience. *Annals of Tourism Research*, *42*, 108–129. doi: 10.1016/j.annals.2013.01.005

McKinney, W., et al. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th python in science conference* (Vol. 445, pp. 51–56).

Morales, D. R., & Wang, J. (2010). Forecasting cancellation rates for services booking revenue management using data mining. *European Journal of*

Operational Research, *202*(2), 554–562. doi: https://doi.org/10.1016/j.ejor.2009.06.006

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, *12*(Oct), 2825–2830.

Putro, N. A., Septian, R., Widiastuti, W., Maulidah, M., & Pardede, H. F. (2021). PREDICTION OF HOTEL BOOKING CANCELLATION USING DEEP NEURAL NETWORK AND LOGISTIC REGRESSION ALGORITHM. *Techno Nusa Mandiri: Journal of Computing and Information Technology*, *18*(1), 1–8. doi: 10.33480/techno.v18i1.2056

Satu, M. S., Ahammed, K., & Abedin, M. Z. (2020). Performance analysis of machine learning techniques to predict hotel booking cancellations in hospitality industry. In *2020 23rd international conference on computer and information technology (ICCIT)* (pp. 1–6). doi: 10.1109/ICCIT51783.2020.9392648

Talluri, K. T., & Ryzin, G. J. (2004). *The theory and practice of revenue management*.

Taylor, R. (1990). Interpretation of the correlation coefficient: A basic review. *Journal of Diagnostic Medical Sonography*, *6*(1), 35–39.

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychol Rev*, *117*(2), 440–463. doi: 10.1037/a0018963

Trope, Y., Liberman, N., & Wakslak, C. (2007). Construal levels and psychological distance: Effects on representation, prediction, evaluation, and behavior. *Journal of Consumer Psychology*, *17*(2), 83–95. doi: 10.1016/S1057-7408(07)70013-X

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

Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Lukauskas, S., Gemperline, D. C., ... Qalieh, A. (2017, September). *mwaskom/seaborn: v0.8.1 (september 2017)*. Zenodo. Retrieved from https://doi.org/10.5281/zenodo.883859 doi: 10.5281/zenodo.883859

Weatherford, L. R., & Kimes, S. E. (2003). A comparison of forecasting methods for hotel revenue management. *International Journal of Forecasting*, *19*(3), 401–415. doi: 10.1016/S0169-2070(02)00011-0

Zdanowicz, C., & Grinberg, E. (2018). Passenger dragged off overbook united flight - CNN. *CNN*. Retrieved 2021-11-30, from https://edition.cnn.com/2017/04/10/travel/passenger-removed-united-flight-trnd/index.html

Zeithaml, V. A. (1988). Consumer perceptions of price, quality, and value: A means-end model and synthesis of evidence. *Journal of Marketing*, *52*(3), 2–22. doi: 10.1177/002224298805200302

Appendix 1

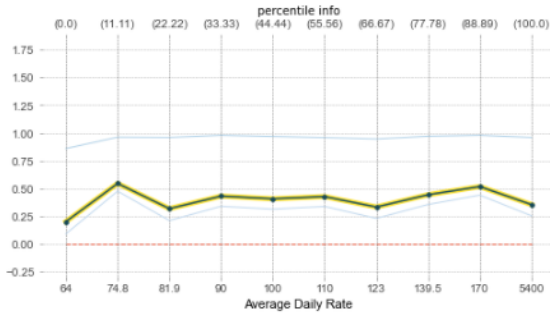
Hotel Booking Demand Dataset Feature Overview

Features	**Description**
adr (Numeric)	Average daily rate
adults (Integer)	No. adults
agent (Categorical)	ID of travel agency that made the booking
arrival_date_day_of_month (Integer)	The date of the month of arrival
arrival_date_month (Categorical)	The month of arrival
arrival_date_week_number (Integer)	Week number of arrival
arrival_date_year (Integer)	Year of arrival
assigned_room_type (Categorical)	Code of room type assigned to guest
babies (Integer)	No. babies
booking_changes (Integer)	No. adjustment made to the booking
children (Integer)	No. children
company (Categorical)	Company that made the booking / pay for the booking
country (Categorical)	ISO code (3-letter). Where is the guest coming from
customer_type (Categorical)	4 categories: contract, group, transient, transient-party.
days_in_waiting_list (Integer)	No. days the guest has waited before booking was confirmed
deposit_type (Categorical)	How is the booking guaranteed.
distribution_channel (Categorical)	Channel in which the booking was made
is_canceled (Categorical)	2 categories: Canceled (1), Not Canceled (1)
is_repeated_guest (Categorical)	2 categories: Repeated guest (1), Not repeated guest (0)
lead_time (Integer)	No. days between the day of booking and the day of arrival
market_segment (Categorical)	2 categories: TA (Travel Agents), TO (Tour Operators)
meal (Categorical)	Type of meal booked. 4 types: undefined/SC (no meal); BB (Bed & Breakfast); HB (breakfast + lunch/dinner); FB (breakfast + lunch + dinner)
previous_bookings_not_canceled (Integer)	No. previous bookings not canceled by guest.
previous_cancellations (Integer)	No. previous bookings canceled by guest.
required_card_parking_spaces (Integer)	No. car parking spaces required by guest
reservation_status (Categorical)	Last status of the reservation. 3 categories: canceled, checked-out, no-show
reservation_status_date (Date)	Date at which the last status was set
reserved_room_type (Categorical)	Code of room type reserved by guest
stays_in_weekend_nights (Integer)	No. weekend nights the guest stayed / booked
stays_in_week_nights (Integer)	No. week nights the guest stayed / booked
total_of_special_requests (Integer)	No. requests made by the customer

Appendix 2

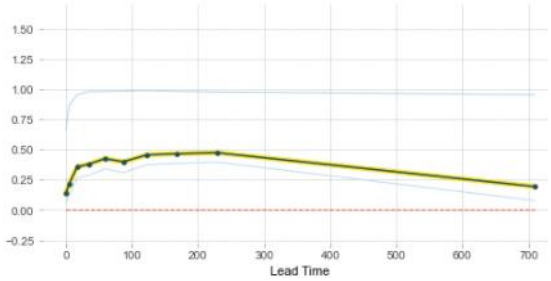
PDP for feature "Average Daily Rate"

Number of unique grid points: 10



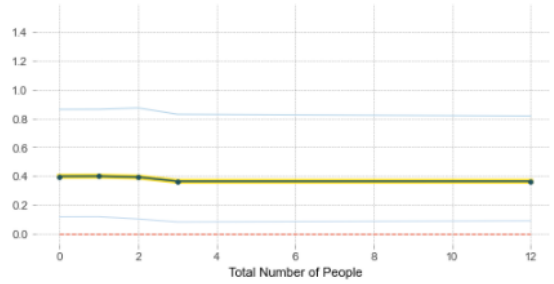
PDP for feature "Lead Time"

Number of unique grid points: 10



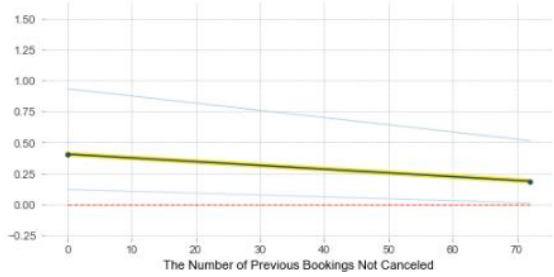
PDP for feature "Total Number of People"

Number of unique grid points: 5



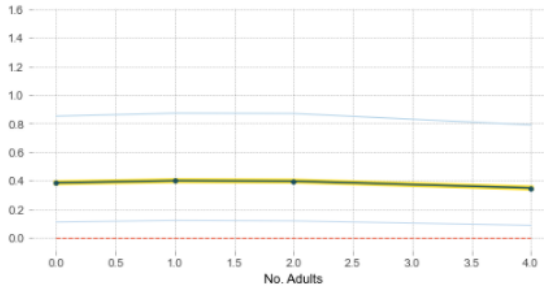
PDP for feature "The Number of Previous Bookings Not Canceled"

Number of unique grid points: 2



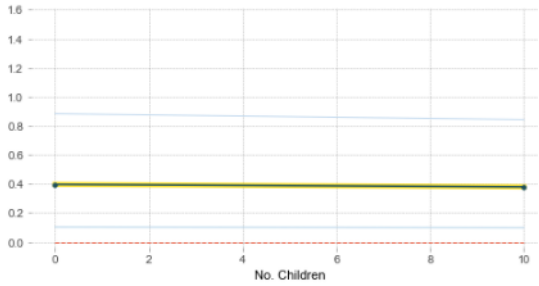
PDP for feature "No. Adults"

Number of unique grid points: 4



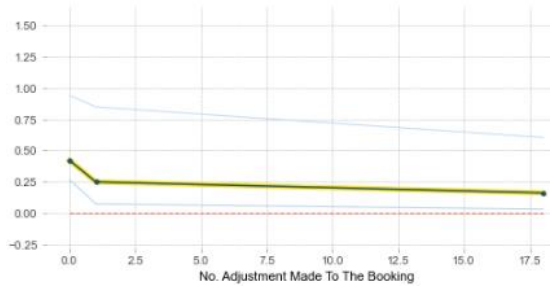
PDP for feature "No. Children"

Number of unique grid points: 2



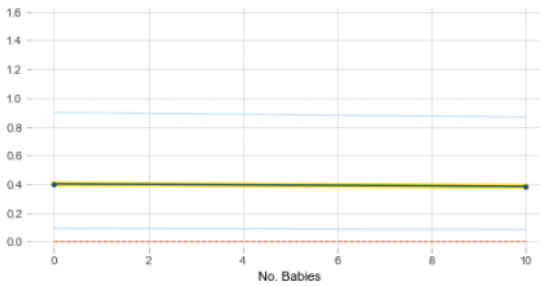
PDP for feature "No. Adjustment Made To The Booking"

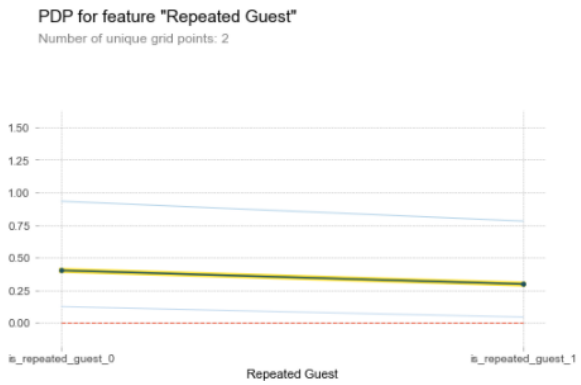
Number of unique grid points: 3

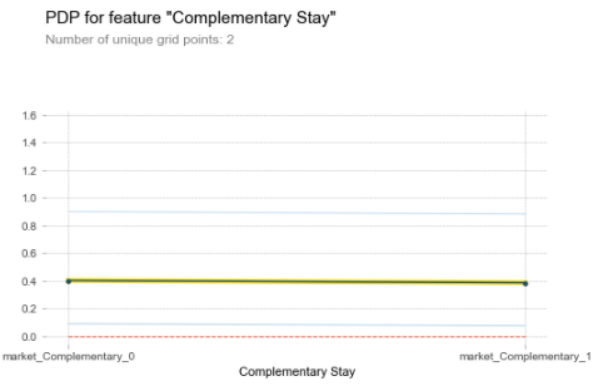


PDP for feature "No. Babies"

Number of unique grid points: 2




PDP for feature "Repeated Guest"
Number of unique grid points: 2
1.50
1.25
1.00
0.75
0.50
0.25
0.00
is_repeated_guest_0
is_repeated_guest_1
Repeated Guest


PDP for feature "Complementary Stay"
Number of unique grid points: 2
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
market_Complementary_0
market_Complementary_1
Complementary Stay