ERASMUS UNIVERSITY ROTTERDAM

ERASMUS SCHOOL OF ECONOMICS

MASTER THESIS DATA SCIENCE AND MARKETING ANALYTICS

PREDICTING HOTEL RATINGS BASED ON CUSTOMER, TRAVEL CIRCUMSTANCES AND HOTEL CHARACTERISTICS

NAME STUDENT: MARJON VAN DE LINDELOOF

STUDENT ID NUMBER: 504427

SUPERVISOR: M. van Crombrugge

SECOND ASSESSOR: F. Frasincar

DATE: 7-1-2023

Abstract

This research uses the online hotel reviews written by customers on the website of Expedia. There is a high variance in hotel ratings given by customers, but literature didn't find the main determinants of the customers' rating. Therefore, this research focuses on investigating what determines the hotel rating. We derive several characteristics from Expedia's online reviews and these are categorized in three sets of characteristics: customer, hotel and travel circumstances characteristics. The main research question focuses on these sets: *"How can hotel ratings be predicted by different sets of characteristics and which set or combination of sets is best in predicting?"*. The hotel rating is divided in 'High' and 'Low' to be able to classify the rating in the right category. The research question is answered by comparing different models based on four method. We the methods binary logistic regression, decision tree, random forest and support vector machine. Each method creates different models meaning that we use every method for each individual set of characteristics. Afterwards, combinations of sets and interactions will be added in each method. This gives us the possibility to determine which method and which combination of characteristics is best in predicting ratings. It appears that binary logistic regression is the worst performing method and support vector machine is the best performing method. The best performing model is a support vector machine with a radial kernel that uses all sets including interaction. Then, we discuss which variables are most important in the prediction. The variables travel group composition and country are the most important variables in predicting hotel ratings. These variables are both from different sets of characteristics meaning that we can't conclude that there is one set of characteristics that is best in predicting ratings. Concluding, a support vector machine can predict hotel ratings well and the combination of all sets including interaction is best in predicting ratings. There isn't one set which is best in predicting meaning that it is useful to combine sets.

Table of content

1. Introduction

Expedia is an international website where consumers can look for a hotel and book this hotel at locations in the whole world. Expedia offers the opportunity to their customers to write reviews on their website. Other consumers are able to read these reviews to help them in choosing the right hotel. Online reviews are very important for both consumers and companies. The product choices of consumers are for example affected by online reviews (Senecal & Nantel, 2004). The reviews help consumers in a way that they are mostly used for information search or for evaluating alternatives (Baek et al., 2012). Online reviews are also useful for companies since they are a main predictor of sales (Chong et al., 2016). This indicates that companies can use online reviews to predict sales. However, it appears that there can be high variance in hotel ratings. When looking at the reviews of a specific hotel, for example NH Atlanta Rotterdam, some people give this hotel a very high rating while other people give the hotel a very low rating (Expedia.nl). Ratings are very important in influencing the number of bookings (Ye et al., 2009) and they are also an important driver of a customer's hotel choice and hotel performance (Bigné et al., 2020; Xie et al., 2014). Therefore, it is helpful to understand what determines the hotel rating. A high variance in ratings is a problem since ratings are an important driver of a consumer's choice.

1.1 Research question and sub-questions

Reviews can be affected by different characteristics such as consumer characteristics, hotel characteristics and travel circumstances characteristics (Phillips et al., 2015). It is interesting to know which of these characteristics are best in predicting hotel rating or that a certain combination of characteristics is better in predicting. Therefore, the central research question is:

"How can hotel ratings be predicted by different sets of characteristics and which set or combination of sets is best in predicting?"

The ratings of the hotels that will be researched are all NH Hotels. NH Hotels is a company with around 400 hotels over the whole world (nh-hotels.nl).

We investigate the central research question by using different sub-questions. These sub-questions are based on the possible predictive performance of the individual sets of characteristics and possible combinations of sets. The sets of characteristics are consumer, hotel and travel circumstances characteristics. These sets are discussed later on. The sub-questions are:

1. Is there one set of characteristics (customer, travel circumstance and hotel) that is best in predicting hotel ratings?
2. What is the effect on the prediction of hotel ratings when combining different sets of characteristics?
3. How well can hotel ratings be predicted when using all sets as predictors?

1.2 Sets of characteristics

After presenting the central research question and the sub-questions, we will discuss the different sets of characteristics. These sets are the possible drivers of ratings. The first set of characteristics is customer characteristics. The variable travel group composition forms this set. Travel group composition is an interesting driver since different travel group compositions have a different set of determinants of satisfaction (Susilo & Cats, 2014). The second set is travel circumstances characteristics which consists of travel season and length of stay. The weather influences a person's experience of the day, so customers have a different experience when travelling in different seasons (Connolly, 2013). The length of stay can also be a driver of hotel ratings since the experience can be different between a one-day stay and a seven-days stay. The third set is hotel characteristics consisting of the hotel's location and the number of stars. The GDP of the hotel's city has a positive effect on the number of bookings (Ye et al., 2009). Therefore, it is interesting to research if these higher number of bookings also results in higher ratings. The dataset will contain hotels with three, four and five stars. Therefore, it is possible to investigate whether there is a difference in hotel ratings between numbers of stars.

1.3 Motivation

1.3.1 Managerial relevance

It is interesting for hotels to understand why customers give a certain rating. Ratings are an important indicator since research showed that customers give higher priority to ratings than to review texts (Aicher et al., 2016). The customers' attention is driven by characteristics as images or review stars (Pieters & Wedel, 2004). These prior findings show that ratings are more important in attracting attention than the actual review text. Customers will first look at the rating that other customers gave before reading the review text. Therefore, it is helpful for hotel managers to understand what actually drives the hotel rating, and not the review text. When understanding what drives the customers' rating, managers can take some of these drivers into account when adjusting their services. As described before, there are a lot of different characteristics that are related to hotel ratings. The goal of this research is to give managers insight in which characteristics, or combinations of characteristics,

are more important in predicting hotel ratings than other characteristics. Managers could focus more on the characteristics that are better in predicting hotel ratings in order to improve customers' ratings. This could lead to better ratings which will be seen by other customers. These customers will have a more positive view of the hotel and will more tend to book a room in the hotel.

1.3.2 Scientific relevance

Previous research has found that consumers' online product choices are influenced by online reviews (Senecal & Nantel, 2004). Online reviews also help consumers in making their purchase decisions (Baek et al., 2012). Furthermore, online reviews are one of the most important predictors of sales (Chong et al., 2016). All these research show that online reviews are important for both consumers and companies since they help consumers in making choices and are important for the sales of a company. Online reviews are useful in different sectors and for different products and services. Senecal & Nantel (2004) found that reviews for experience products are more influential than for search products. Hotel visit is an experience product since it is a service. Therefore, online reviews are useful for this research. Online ratings are even a major driver of hotel choice (Bigné et al., 2020). Summarizing, researchers investigated the effects of online reviews on consumer's choices and sales often. However, there is not research that combined drivers of hotel rating in different sets of characteristics to investigate which set, or combination of sets, is best in predicting hotel ratings. Previous research is mostly focused on the actual text of the review and not specific on the hotel rating. As discussed before, research found that consumers are more focused on ratings than review texts. Therefore, it is interesting to investigate what the drivers of hotel ratings are and how they can be predicted. In this case, ratings can be improved which can lead to more hotel bookings. This research will contribute to the existing literature since it focuses on the drivers of hotel ratings. The interaction between drivers is also investigated in order to be able to create the best hotel rating prediction as possible.

First, the previous research in this field will be discussed (Chapter 2). Afterwards, we discuss the data and methodology that is used for the analysis (Chapter 3). Then, the data will be analysed and the results will be discussed (Chapter 4). Finally, we answer the research question and sub-question by using the results (Chapter 5).

2. Literature review

The aim of this research is to identify which characteristics can be used to predict hotel ratings. We describe the research on the role of key characteristics in 1) customer satisfaction, 2) hospitality and WOM research, and 3) research on consumer reviews specifically. We organize the discussion of each characteristic in the literature review accordingly and discuss our contribution based on the same structure. Each discussion will lead to a proposition.

First, the individual effects of each set of characteristics will be discussed. There are several sets of characteristics that could be considered when predicting hotel ratings. However, we will focus on three sets of characteristics: (1) Customer characteristics, (2) Travel circumstances characteristics, and (3) Hotel characteristics. These three sets will be used since it covers a broad range of characteristics that could be determinants of hotel ratings. It considers both the customer side as the hotel side.

2.1 Customer characteristics

The first set of characteristics is 'customer characteristics'. This set gives information about the customer or reviewer. The variable that belongs to this set is travel group composition which is unique per customer and doesn't depend on for example the hotel.

Travel group composition refers to the different group sizes or different compositions customers have travelled with. They could for example travel alone, with family or with friends. Research found that different travel groups have a significantly different set of main determinants of satisfaction (Susilo & Cats, 2014). This indicates that customers find other features of the product or service important for their satisfaction when travelling in different group composition. In this section, two causes for differences in satisfaction will be considered: (1) Differences in utility, and (2) Differences in willingness to pay. The first cause of these differences may be due to customers having other needs and wants when travelling in groups. For example, there are significant differences in satisfaction about the access of the hotel across traveller groups (Xu, 2018). For some group compositions, it is more important that the hotel has an easy access than for other groups, for example for families with elderly people. The second cause of differences in satisfaction across travel group compositions may be the fact that people tend to choose the higher price option when they are in groups than when they are alone (Jeong et al., 2019). This is due to the concept of "social surrogate" where individuals go along with the suggestions made by others in their travel group (Stone, 2016). The research above shows that customers have other determinants of satisfaction when travelling in different travel group compositions, which consequently may affect their reviews. Existing research didn't investigate the

actual effect of different travel group compositions on reviews. Therefore, we use the previous work, about the differences in satisfaction across traveller group, to research whether the travel group composition can be used as a predictor of hotel ratings.

Concluding, the previous discussion of the literature shows that the set of customer characteristics can be helpful in predicting hotel ratings. Therefore, the first proposition is as following:

Proposition 1: Customer characteristics help predicting hotel ratings.

2.2 Travel circumstances characteristics

The second set of characteristics in predicting hotel ratings is travel circumstances characteristics. These characteristics provide information about the type of travel. The variables that belong to this set are travel season and length of stay.

The first variable of this set is travel season which indicates what the type of weather was in the country of the journey. First, research has shown the effect of the weather on human and consumer behaviour in general. The weather affects a person's mood and causes mood changes, so it influences human behaviour (Cao & Wei, 2005). This indicates that consumers have different experiences depending on the type of weather since the weather influences a person's experience of the day (Connolly, 2013). Because of the weather, consumers might make other decisions. For example, consumers tend to purchase more products when it is sunny (Murray et al., 2010; Tian et al., 2021). It is therefore no surprise that seasonal or weather differences have made their way into hospitality research as well, since travellers also make other decisions during trips when experiencing different types of weather. Weather affects travel satisfaction since with some types of weather travellers are more satisfied about their trip than with other types (Ettema et al., 2017). For example, rainy weather and snow have a negative influence on a consumer's travel satisfaction (Abenoza et al., 2019). As discussed before, the type of weather influences a consumer's mood and therefore the consumer's review is also influenced by the weather at the moment of writing the review. Rain has a negative influence on the online review rating while a higher temperature leads to a higher online review rating (He et al., 2020). This indicates that, at the moment of writing a review, the consumer's rating is affected by the weather. However, the effect of weather during a trip on the review is not researched yet. Therefore, we contribute to research by investigating this effect.

The next variable of this set of characteristics is the length of stay which indicates what the duration of a tourist's trip was. The effect of length of stay in hospitality research has been investigated before. Research found that the length of stay is a major driver in the tourist's decision-making process (Wang

et al., 2017). The length of stay is a determinant in the decision-making process together with other factors such as accommodation type (Alegre et al., 2011). Additionally, research found significant differences in satisfaction between different length of stays. Short-term visitors experience lower levels of satisfaction with the perceived quality of service (Neal, 2008). Contradictory, other research found that there are no significant differences in tourists' mood between different length of stays (Nawijn, 2010). This could indicate that there are no differences in consumer behaviour between several length of stays. Therefore, it is not yet clear what the effect of length of stay is on consumer behaviour. In the literature there has been no research yet into the effect of length of stay on reviews. Our research aims to solve this contingency by considering the possible effect of length of stay on hotel ratings.

In conclusion, the previous discussion showed that the set of travel circumstances characteristics is useful in predicting hotel ratings. This leads to the second proposition:

Proposition 2: Travel circumstances characteristics help predicting hotel ratings.

2.3 Hotel characteristics

The next set is hotel characteristics which contains information about the hotel itself. The variables that are included in this set are hotel location and number of stars. These two variables give information about whether the hotel is for example in a big city or in a quiet environment and about the hotel's overall quality.

The first variable is the hotel's location. First, research into the role of the hotel's location in the decision-making process will be discussed. The hotel's location is a major driver in the customer's decision-making process (Xiang & Krawczyk, 2016). The location is namely one of the most important attributes that influences a traveller's hotel choice (Mccleary et al., 1993). For example, consumers prefer different locations depending on their trip purpose. Furthermore, customers are highly aware of the location and even associate different values with it, such that the hotel's safety measures increase the consumer's trust and confidence (Lee et al., 2010). Xiang & Krawczyk (2016) also found that multiple value-related factors are associated with hotel location which indicates that customers derive value from different factors that are related to the hotel location. On the other hand, the hotel location is also one of the most important factors for hotel owners since their strategic choices are based on the location (Yang et al., 2014). An example of this is that hotel owners look at the competitors in the neighbourhood and adjust their choices on the competitors' behaviour. Since the hotel location is important for decision making and for deriving value, it could probably also be

important for the customer when he evaluates the hotel service which will be reflected in the review. Therefore, we contribute by investigating the effect of the hotel's location on the hotel rating.

The next variable is the number of stars that the hotel received. These number of stars displays some qualities of the hotel. A five-star hotel is for example more luxurious than a three star hotel since it has more facilities. However, the number of stars doesn't give an indication about the customers' experience since it is based on certain qualities and facilities. Prior research investigated whether there are differences in hotel ratings among the different numbers of stars. Research found that guests are more satisfied when the hotel has a higher number of stars (Bulchand-Gidumal et al., 2013). This may be due to the fact that the hotel's service quality positively influences consumer's satisfaction (Ladhari, 2009). However, there are differences in which performance predictors are most important to customers across different numbers of stars. For example, the accommodation infrastructure and employee expertise are most important in satisfaction for one and two stars, while waiting time is an important predictor of satisfaction for a five-star hotel (Nunkoo et al., 2020). Additionally, research found an effect of the number of stars on the review's helpfulness. The interaction between the hotel's number of stars and the review rating affects review helpfulness (Hu & Chen, 2016). We contribute by researching how the number of stars can help in predicting hotel ratings.

Summarizing, the discussion of literature about the hotel characteristics leads to the next proposition:

Proposition 3: Hotel characteristics are useful in predicting hotel ratings.

2.4 Comparison between sets of characteristics

Finally, we discuss which set of characteristics will be best in predicting hotel ratings. There hasn't been research that explicitly investigated the differences between the effects of the characteristics. However, research found that the quality of the hotel services is an important factor of customer satisfaction (Gundersen et al., 1996). Furthermore, research found that the performance of the reception, food and beverages and the price are the most important factors of customer satisfaction during a hotel visit (Kandampully & Suhartanto, 2000). The quality of hotel services and the other discussed factors are displayed by the hotel's number of stars meaning that this characteristic can be important in predicting hotel ratings. Next, researchers concluded that the convenience of the location and quietness of the surroundings are perceived as positive factors by hotel visitors (Cadotte & Turgeon, 1988). These factors are part of the hotel's location indicating that this can be also important in predicting hotel ratings. The number of stars and the hotel's location are both characteristics belonging to the set of hotel characteristics. Therefore, we can predict that the set of hotel

characteristics is more important in predicting hotel ratings than the other sets. The last proposition is as following:

Proposition 4: The set of hotel characteristics is more important in predicting hotel ratings than the other sets of characteristics.

3. Data & Methodology

3.1 Data collection

The data is collected by using web scraping and is collected from Expedia's website. Expedia is a website for travellers where they can, for example, book a stay or rent a car. There is also the possibility for customers to write about their experience of their hotel stay. Each hotel has its own page on Expedia and the reviews are shown in a specific part of the website. For each review, the required variables are shown so the website is suitable for scraping all variables that will be used in the models. The variables that are scraped from the reviews are the following: the review text, the review rating, the length of stay, the month of stay, the year of stay and the traveller group. Other variables that were collected are: the hotel's name, the hotel's number of stars and the hotel's country. The reviews of 240 hotels are collected which leads to 1200 reviews. The two selection criteria for the hotels are that they are part of the NH Hotel Group and are located in Europe to ensure that the services of the hotels are quite similar since they are part of the same hotel group and located on the same continent.

3.2 Classification task

The variable that has to be predicted is the review's rating. Each reviewer can rate the hotel on a scale from 1 until 5. Expedia describes the five possible ratings as follows: 1) Terrible, 2) Poor, 3) Okay, 4) Good, and 5) Excellent. We will use these five categories in our models. The proportion of each category is displayed in Table 1 of Appendix A. We conclude from this table that the review's rating has a skewed distribution since more than 50% of the observations has '5/5 Excellent' as rating. To solve this skewness the categories will be merged into two categories. The first three categories, 'Terrible', 'Poor' and 'Okay', will be merged in the category 'Low'. The other two categories, 'Good' and 'Excellent', will be merged into the category 'High'. The proportion of each category can be found in Table 2 of Appendix A. It appears that the distribution is still skewed since about 85% of the observations belongs to the category 'High'. An unbalanced dataset is a problem since the effect of variables can be concealed. Furthermore, certain effects can be concluded which are not true (Kitchenham, 1998). The solution of this unbalanced dataset will be discussed later on in paragraph 3.3.5. Our goal is to predict in which category the customer's experience falls into. Therefore, the models will classify a new reviewer in one of the categories based on the other variables. The models will select the variables that are most important in classifying new reviewers.

3.2.1 Predictor variables

We will use predictor variables to predict a review's rating. The data preparation and summary statistics of each predictor variables will be discussed separately.

3.2.1.1 Customer characteristics

First, the set of customer characteristics will be discussed. The variable that belongs to this set is travel group composition. When the customer writes a review on Expedia, he or she has to fill in the travel group composition. There are seven basic options for the traveller group which are: 1) Solo traveller, 2) Business traveller, 3) Travelled with family, 4) Travelled with group, 5) Travelled with partner, 6) Travelled with family and small children and 7) Travelled with pets. Furthermore, the customer has the possibility to choose multiple types of travel group composition. For example, he or she could choose for the combination "Business traveller, Travelled with family". There are some levels that represent the same category, so these are combined into one category.[1] As a result, there are 22 levels found in the data. Table 1 in Appendix B shows the proportion of each travel group composition, so it indicates how often each travel group composition occurs. It appears that almost all combinations of different traveller groups occur less than 1%. This indicates that these groups have a high probability to be too small to be useful in the models.

Therefore, combinations of traveller groups will be combined in more "basic" options of travel group compositions (see Appendix B for more detail on which groups are combined and why), such that the final travel group compositions are: 1) Solo traveller, 2) Travelled with family, 3) Business traveller, 4) Travelled with partner, 5) Travelled with group, and 6) Travelled with family and small children. The distribution of this variable is shown in Table 2 of Appendix B. As a result, the travel group composition leads to six variables, i.e. $SoloTraveller_{ir}$ which equals 1 if traveller *i* travelled solo when staying in hotel *r*; $FamilyTraveller_{ir}$ which equals 1 if traveller *i* travelled with family when staying in hotel *r*; $BusinessTraveller_{ir}$ which equals 1 if traveller *i* travelled for business when staying in hotel *r*; $PartnerTraveller_{ir}$ which equals 1 if traveller *i* travelled with partner when staying in hotel *r*; $GroupTraveller_{ir}$ which equals 1 if traveller *i* travelled with a group when staying in hotel *r*; and $ChildrenTraveller_{ir}$ which equals 1 if traveller *i* travelled with family and small children when staying in hotel *r*.

[1] For example, some customers had "Business traveller" as group while others had "Business traveller". These two levels are combined into one. The same problem occurred for the other options so these are also combined into one level.

3.2.2.2 Travel circumstances characteristics

The second set of characteristics is the set of travel circumstances which contains the length of stay and travel season. First, the length of stay is measured in the number of nights. The reach of this variable is from 1 until 19 nights, but there are no observations with a length of stay of 14 until 18 nights. The summary statistics of length of stay are displayed in Table 3 of Appendix B. This variable is treated as a continuous variable. Therefore, the variable $LengthStay_{ir}$ has a value between 1 and 19 representing the number of nights that traveller *i* stayed in hotel *r*.

Then, the variable travel season will be explained. The data that is scraped from the website related to this variable is the month of stay. Each month belongs to one of the seasons so the seasons will be created by combining the corresponding months. The months December, January and February belong to the season winter. The months March, April and May belong to the spring. June, July and August belong to the summer. Finally, the months September, October and November belong to autumn. The proportion of each month is displayed in Table 4 of Appendix B. The proportion of each season is shown in Table 5 of Appendix B. This indicates that most of the travellers travel during spring and the fewest during autumn. As a result, the travel season leads to four variables, i.e. $Winter_{ir}$ that equals 1 if traveller *i* travelled during winter when staying in hotel *r*; $Spring_{ir}$ that equals 1 if traveller *i* travelled during spring when staying in hotel *r*; $Summer_{ir}$ that equals 1 if traveller *i* travelled during summer when staying in hotel *r*; and $Autumn_{ir}$ that equals 1 if traveller *i* travelled during autumn when staying in hotel *r*.

3.2.2.3 Hotel characteristics

The next set is the set of hotel characteristics containing the number of stars and the hotel's location. The dataset contains hotels with 3, 4 and 5 stars. The largest part of the hotels has 4 stars. The proportion of the number of stars is displayed in Table 6 of Appendix B. Concluding, the number of stars leads to three variables, i.e. $ThreeStars_{ir}$ that equals 1 if traveller *i* stays in hotel *r* with three stars; $FourStars_{ir}$ that equals 1 if traveller *i* stays in hotel *r* with four stars; and $FiveStars_{ir}$ that equals 1 if traveller *i* stays in hotel *r* with five stars.

Then, the hotel's location is represented by the country where the hotel is located. The hotels are divided over 16 European countries. The proportion of each country is displayed in Table 7 of Appendix B.

The summary statistics of all variables can be found in Table 1.

Table 1: Summary statistics all variables

Variable	Mean (continuous) / Percentage (categorical)
Dependent variable:	
High	49,7%
Low	50,3%
Predictor variables:	
SoloTraveller$_{ir}$	36,25%
FamilyTraveller$_{ir}$	22,41%
BusinessTraveller$_{ir}$	17,91%
PartnerTraveller$_{ir}$	15,83%
GroupTraveller$_{ir}$	5,5%
ChildrenTraveller$_{ir}$	2,08%
LengthStay$_{ir}$	2,28
Winter$_{ir}$	22,5%
Spring$_{ir}$	39,33%
Summer$_{ir}$	19,75%
Autumn$_{ir}$	18,42%
ThreeStars$_{ir}$	5,42%
FourStars$_{ir}$	84,17%
FiveStars$_{ir}$	10,42%
Germany$_{ir}$	21,25%
Belgium$_{ir}$	5,42%
France$_{ir}$	2,08%
Hungary$_{ir}$	1,25%
Andorra$_{ir}$	0,42%
Denmark$_{ir}$	0,42%
Ireland$_{ir}$	0,42%
Luxembourg$_{ir}$	0,42%
Poland$_{ir}$	0,42%
Romania$_{ir}$	0,42%
Slovakia$_{ir}$	0,42%
Austria$_{ir}$	2,92%
Portugal$_{ir}$	6,67%
Italy$_{ir}$	22,92%
Spain$_{ir}$	20,83%
The Netherlands$_{ir}$	13,75%

3.3 Methodology

To gain insight into which methods are useful in predicting hotel ratings, the literature about methods that predict ratings will be discussed. There are several methods that are suitable for our research. We will compare several methods to research which method is best in classifying a new customer in the right rating category. First, we discuss how a logistic regression is able to predict ratings. After this, we discuss three machine learning methods, Decision Tree, Support Vector Machine (SVM) and Random Forest, which could be able to predict hotel ratings.

Each method will create different models. First, every method is used for each single set of characteristics. For example, we run a logistic regression with only the consumer characteristics and a logistic regression with only the travel circumstances characteristics. Then, combinations of sets and interactions will be added in each method. The goal is to conclude which method and which combination of variables is best in predicting ratings.

The dataset is split into a train and a test set. The train set is used to train the model. Afterwards, the test data is used to check the performance of each model. 80% of the data belongs to the train dataset and the other 20% belongs to the test set.

3.3.1 Binary logistic regression

A method that is used several times in literature is a logistic regression which used the text sentiment as a predictor variable (de Albornoz et al., 2011; Qu et al., 2010). This method is effective in predicting ratings. In our research, we will use other variables instead of the review sentiment to predict ratings. Since we have a binary variable as a dependent variable, we need to use a binary logistic regression. This type of logistic regression is suitable for a dependent variable that only has two possible outcomes. The binary logistic regression has four assumptions that need to be checked: 1) The dependent variable is binary, 2) Linearity of independent variables and log-odds, 3) No multi-collinearity, and 4) No outliers (Peng et al., 2010). The first assumption is met based on the discussion of the variables earlier this chapter. The second assumption means that there needs to be a linear relationship between the log-odds of the outcome and each continuous variable. As discussed before, we don't use continuous variables which means that this assumption doesn't need to be checked. The third assumption, no multi-collinearity, can be checked with a correlation plot of all the independent variables. This correlation plot is displayed in Figure 1 of Appendix C. As can be seen in this plot, there are no high correlations between independent variables indicating that there is no multi-collinearity. However, a Variance Inflation Factor (VIF) test should also be performed check formally if there is no multi-collinearity. The VIF scores can be found in Table 1 of Appendix C. The rule of thumb is that there is

multi-collinearity if the VIF score is greater than 10. In our case, all the VIF scores are around 1 meaning that there is no multi-collinearity in our dataset. The last assumption is that there are no outliers in the data. Since we have categorical data, outliers are being judged by the degree of imbalance. As discussed before is the dependent variable imbalanced meaning that the last assumption doesn't hold. Therefore, the data needs to be balanced to be useful in this model. The balancing method will be discussed in paragraph 3.3.5.

We have 11 different logistic regressions. Three logistic regression models have only one set of characteristics as predictors, three other logistic regression models have two sets of characteristics, three models have two sets of characteristics including interaction effects, one model has all sets and one model has all sets including interactions between variables. As an example, we show the formula of the logistic regression for customer and travel circumstances characteristics:

$$\begin{aligned} Hotel\ rating = {} & \beta_0 + \beta_1 SoloTraveller_{ir} + \beta_2 FamilyTraveller_{ir} + \beta_3 BusinessTraveller_{ir} \\ & + \beta_4 PartnerTraveller_{ir} + \beta_5 GroupTraveller_{ir} + \beta_6 ChildrenTraveller_{ir} \\ & + \beta_7 LengthStay_{ir} + \beta_8 Winter_{ir} + \beta_9 Spring_{ir} + \beta_{10} Summer_{ir} + \beta_{11} Autumn_{ir} \\ & + \varepsilon \end{aligned}$$

3.3.2 Decision tree

The first machine learning method is a decision tree. Decision trees have the capability to provide a solution for a complex decision-making process which is easy to interpret (Safavian & Landgrebe, 1991). It breaks down the decision-making process into a collection of simpler decisions. The decision tree splits the training set in classes based on the variables (features). At each node, there is a condition on one or more of the features to separate all the classes. The classification error rate is used to make the split in order to grow the decision tree. The classification error rate is the fraction of training observations in a specific region that doesn't belong to the most common class. This rate can be calculated by dividing the number of incorrect predictions by the total number of observations in the dataset. Another measure that can be used to grow to the tree, is the Gini index. This index calculates the probability that an observation is wrongly classified. The value of this index is between 0 and 1. Value 0 indicates that the majority of a node contains observations from a single class. The formula to calculate the Gini index is $1 - \sum_{i=1}^{n} (p_i)^2$. Furthermore, there is one other parameter that has to be tuned when growing the tree. This parameter is the complexity parameter (cp) which determines the number of terminal nodes. The cp determines the optimal size of the tree which will avoid overfitting. It is also easier to interpret when there is an optimal number of terminal nodes.

We have in total 14 decision tree models, seven models the Gini index as a measure and the other seven have the classification error rate as measure. Out of the seven models, three models have one set of characteristics as predictor variables. Three models have two sets of characteristics and one model has all sets as predictor variables.

3.3.3 Support Vector Machine (SVM)

The second machine learning method is Support Vector Machine. This method is used before to predict ratings. The model uses the information on the website, such as photos and the hotel description, and the customer's nationality as predictor variables (Chu & Huang, 2017). The SVM is an algorithm that is able to classify new data points in the right class by finding a hyperplane in an N-dimensional space, where N represents the number of features (Müller et al., 1997). A hyperplane is a decision boundary that can classify new points. Support vectors are data points that influence the position of the hyperplane since they are closer to the hyperplane. The goal of the model is to find the hyperplane with the maximum margin. This means that this plane has the maximum distance between the data points of all classes.

In our research, we have more than two features meaning that we have a high-dimensional space. In this situation, the SVM uses a kernel which is a function that quantifies the similarity of two observations. There are two types of kernels: 1) Polynomial kernel, and 2) Radial kernel. The polynomial kernel is defined as $K(X_1, X_2) = (a + X_1^T X_2)^d$ where d is the degree of the kernel and a is the constant term. Increasing the number of dimensions leads to a much more flexible decision boundary. The second type of kernel is the radial kernel which is defined as $K(x_i, x_{i'}) = \exp\left(-\gamma \sum_{j=1}^{T} \left(x_{ij} - x_{i'j}\right)^2\right)$. In this formula, γ is a tuning parameter that accounts for the smoothness of the decision boundary. This parameter also controls the variance of the model. When γ has a high value, we have a fluctuating decision boundary that accounts for high variance and overfitting. A low value for γ leads to a smoother decision boundary with a low variance. Both the parameters d and γ will be tuned by 10-fold Cross Validation (CV). We explain CV by using the γ parameter as an example. The data will be split into 10 folds. Each time a model with all possible values for γ will be fitted on 9 folds and tested on the remaining fold. The γ -value with the lowest CV-error will be the value of γ in the model.

Then, SVM contains a parameter that needs to be tuned. This is the cost-parameter *C* which indicates how tolerant we are of violations of the maximum margin. When *C* decreases, we become less tolerant of violations meaning that the margin will become smaller. This parameter is tuned by using 10-fold Cross Validation (CV). *C* also indicates how many support vectors are used. When *C* is high, we have a

wide margin meaning that there are more violations of the margin. These observations, which violate the margin, determine the position of the hyperplane so they are support vectors.

In the end, we have 22 SVM models that will be discussed. Out of these models, 11 use the polynomial kernel and the other 11 models use the radial kernel. Out of the 11 models, three models have only one set of characteristics as predictor variables, three other models have two sets of characteristics, three models have two sets of characteristics including interaction effects, one model has all sets and one model has all sets including interactions between variables.

3.3.4 Random Forest

The last machine learning method is random forest which is an algorithm for classification. Random forest is an ensemble classifier which is more accurate than any of the individual classifiers (Pal, 2007). Random forest is an ensemble classifier since it uses an ensemble of decision trees meaning that the random forest consists of a large number of individual decision trees. The reason why an ensemble of trees is used, is to avoid errors of individual trees. At each split, a random sample of predictors from the total set of predictors is used as split candidates. This means that only a subset of predictors is considered at each split. The parameter '*mtry'* has to be tuned while training the random forest. '*mtry'* is the number of variables that will be considered at each split. To determine the best value for '*mtry'*, 10-fold Cross Validation (CV) will be used. The value of '*mtry'* where the accuracy is maximized will be the value that is used in the random forest model.

We have 11 random forest models. Three random forest models have only one set of characteristics, three other random forest models have two sets of characteristics, three models have two sets of characteristics including interaction effects, one model has all sets and one model has all sets including interactions between variables.

The last three methods that were discussed are machine learning methods. The imbalanced data is a problem for the logistic regression, but it appeared that this could also be a problem for machine learning methods. It is a problem based on the fact that the imbalance affects the performance of these methods. The imbalance could cause a bias since the minority class is not often captured by the model (Provost, 2000). Therefore, we need a solution to create a balanced dataset.

3.3.5 Balancing

The dataset can be balanced in several ways. The first choice relates to undersampling and oversampling. Undersampling means that there will be observations removed from the majority class

resulting in a smaller dataset. Oversampling means that observations will be added to the minority class which results in a greater dataset. In our case, there are 1200 observations which is not that much. Therefore, an oversampling-technique is used to balance the data. The technique that is used is called Random Over-Sampling Examples (ROSE). It uses a smoothed-bootstrap approach to produce a synthetic sample of data (Lunardon et al., 2014). ROSE generates new examples by selecting an observation and looking at the neighbourhood of that observation. The new observations are based on the existing observations in the dataset. We aim to get a dataset where both classes are 50% of the distribution. We end up with a dataset of 3000 observations and the new distribution of the two categories is showed in Table 3 of Appendix A. It appears that each category is approximately 50% of the whole data. This balanced dataset is used in the models.

3.4 Performance measures

After running the models with the different machine learning methods, several performance measures will be used to determine the model that is best in predicting a customer's rating.

The first performance measure is accuracy which measures what part of the test data is correctly predicted by the trained model (García et al., 2009). The formula to calculate the accuracy is $\frac{(TP+TN)}{(TP+FP+TN+FN)}$. The formula sums the true positives and true negatives and divides this sum by all observations. The value of the accuracy is a percentage between 0 and 100. A high accuracy indicates that the model predicts well.

The second performance measure is the Area Under Curve (AUC). The AUC indicates how well the model can distinguish classes (Ben-David, 2007). The AUC can be obtained by plotting the ROC curve. The ROC curve is plotted with the false positive rate on the x-axis and the true positive rate on the y-axis. The true positive rate is the same as the sensitivity and can be calculated as $\frac{TP}{TP+FN}$. Then, the false positive rate is the same as (1 – specificity) and will be calculated as $\frac{FP}{TN+FP}$. The AUC is the area under the ROC curve and has a value between 0 and 1. When the value is close to 1, the model predicts well.

The next measure is precision which measures how well the model correctly identifies the positive class (Walther & Moore, 2005). This measure indicates how many of all predictions of a positive class where actually correct. The precision will be calculated as $\frac{TP}{TP+FP}$ and has a value between 0 and 1. When the value is close to 1, the model predicts well. Another measure that is somewhat similar to precision is recall. Recall shows how good the model is at correctly predicting all the positive

observations in the data. It is calculated by $\frac{TP}{TP+FN}$ and it has also a value between 0 and 1 where 1 means that the model predicts well.

The fifth performance measure is the F1 score which is the harmonic mean of precision and recall. The F1 score has a value between 0 and 1. If the score is 1, this means that there is perfect precision and recall. If the score is 0, either the precision or recall is 0. The F1 score can be calculated with the formula $2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$.

The following performance measure is the Kappa statistic. Kappa compares the observed accuracy to the accuracy that is expected from random chance. This statistic shows how the model performs in comparison to a model that randomly classifies observations. Kappa has a value at or below 1 and negative values are also possible. However, there is no agreed standard for interpretation of Kappa's value. Therefore, we use the general interpretation from previous research where a value of 0,21 is seen as a fair agreement (Landis & Koch, 1977). The Kappa statistics is calculated as $(\text{ObservedAccuracy-ExpectedAccuracy})/(1 - \text{ExpectedAccuracy})$.

The last performance measure is Matthews Correlation Coefficient (MCC). This coefficient takes all possible prediction outcomes into account, so it accounts for all imbalances in the classes (Chicco & Jurman, 2020). It is a correlation coefficient between the observed and predicted classifications. The MCC has a value between -1 and 1. If it has value of 1, we have a perfect model. The MCC can be calculated as $\frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}$.

3.5 Global interpretation methods

The described machine learning methods (Decision Trees, SVM and Random Forest) are black box methods. Black box models are difficult to interpret since we know the input and the output of the model, but it is hard to understand what happens between the input and output. Therefore, we use global interpretation which is used for understanding the whole model, which refers to how the model makes decisions. To interpret the results of the best method, two global interpretation methods will be used. Because of these methods, it will be possible to conclude which variables are most important in prediction and what the relationship between these variables and the dependent variable is. The first method is to calculate the input importance of all predictor variables. The input importance shows what variables are most important in predicting the hotel ratings(Cortez & Embrechts, 2013). The input importance is calculated by using a sensitivity method which varies the input variable from the minimum to the maximum value. We use variance as a sensitivity measure so for each value of the

variable, the variance in the output variable is calculated. The variance indicates how much the model depends on the variable. The variables that have the highest variance, are the most important variables in predicting the dependent variable.

The second method is partial dependence plot (PDP) which shows the relationship between a feature and the dependent variable (Elshawi et al., 2019). The plot shows the marginal effect of the feature on the dependent variable. The marginal effect is on the y-axis and the values of the feature are on the x-axis.

4. Results

4.1 Performance of each method

First, the performance of each individual method will be discussed shortly. This is important to gain insight in the performance of different combinations of sets of characteristics for each method.

4.1.1 Logistic regression

The performance of the logistic regression is quite bad. The results can be found in Table 1 of Appendix D. The results are sorted based on the highest accuracy. Out of all the different logistic regression models, we select the model that is performing 'best'. This model is the one that has the highest performance measures in comparison to the other logistic regression models. This model is the one with the set of travel circumstances characteristics based on its performance metrics ($Acc_{LR}^{Travel\ circumstances} = 0{,}457;\quad Prec_{LR}^{Travel\ circumstances} = 0{,}466;\quad Recall_{LR}^{Travel\ circumstances} = 0{,}594; AUC_{LR}^{Travel\ circumstances} = 0{,}57; F1_{LR}^{Travel\ circumstances} = 0{,}522; Kappa_{LR}^{Travel\ circumstances} = -0{,}086; MCC_{LR}^{Travel\ circumstances} = -0{,}09$).

It is remarkable that the accuracy of all models is below 50% and that the Kappa and MCC for all models is negative. This indicates that the logistic regression is not performing well in classifying the customer's rating. The results of the logistic regression models are shown in Appendix E and each table represents a model. The results show the coefficients and its significance in order to conclude whether there are significant effects and whether these make sense. It appears that almost none of the predictor variables in the models with one or two sets of characteristics have significant effects. On the other side, there are quite some variables in the models with interactions that have significant effects. For example, in the model with customer and hotel characteristics including interaction all the individual variables have significant effects and there are also a lot of interactions that have significant effects. The results show that, when considering the interaction terms, that hotels with 4 stars, in combination with the traveller group, have a lower rating than hotels with 5 stars. This makes sense since hotels with 5 stars are in general more luxurious and offer more services. There are also differences in significant effects among the interaction between country and traveller group. This shows that the rating can depend on the combination between country and the traveller group which makes sense since some countries are more suitable for family holidays, for example, than other countries. However, these models with interaction perform worse than the models without interaction.

4.1.2 Decision tree

The performance measures of the decision tree can be found in Table 2 of Appendix D. These results are sorted on the highest accuracy. It appears that the accuracy, precision and AUC are approximately 50% which indicates that the models are not predicting very well compared to the other methods, which will be discussed later on. It is also remarkable that the recall is 1 for almost each model. This indicates that there are no misclassifications in these methods in the category 'High' since there are no false negatives. However, there are quite some false positives meaning that there are observations that are incorrectly classified as 'Low'. Next, we decide which model among all decision tree models is the 'best' by comparing the performance measures of all decision tree models. We conclude that the two best performing models both use the Gini index instead of the classification error rate. The best performing model is a decision tree based on the Gini index with the sets of customer and hotel characteristics ($Acc_{Decision\ Tree}^{Gini-customer\ \&\ hotel} = 0{,}552$; $Prec_{Decision\ Tree}^{Gini-customer\ \&\ hotel} = 0{,}549$; $Recall_{Decision\ Tree}^{Gini-customer\ \&\ hotel} = 0{,}532$; $AUC_{Decision\ Tree}^{Gini-customer\ \&\ hotel} = 0{,}551$; $F1_{Decision\ Tree}^{Gini-customer\ \&\ hotel} = 0{,}54$; $Kappa_{Decision\ Tree}^{Gini-customer\ \&\ hotel} = 0{,}103$; $MCC_{Decision\ Tree}^{Gini-customer\ \&\ hotel} = 0{,}103$).

4.1.3 Random Forest

Compared to the logistic regression and decision tree models that are established above as not good in classifying the customer's rating, the random forest models are performing better. The results can be found in Table 3 of Appendix D and are sorted based on the highest accuracy. Two models have an accuracy, precision, recall and AUC of approximately 70% ($Acc_{RF}^{All\ sets} = 0{,}72$; $Prec_{RF}^{All\ sets} = 0{,}702$; $Recall_{RF}^{All\ sets} = 0{,}745$; $AUC_{RF}^{All\ sets} = 0{,}722$; $Acc_{RF}^{All\ sets-interaction} = 0{,}713$; $Prec_{RF}^{All\ sets-interaction} = 0{,}695$; $Recall_{RF}^{All\ sets-interaction} = 0{,}738$; $AUC_{RF}^{All\ sets-interaction} = 0{,}716$). These values indicate that the models are predicting well since it indicates that the models are better in predicting than when the classification is done randomly. In this case of random classifying, the values will be much lower. Furthermore, four models have a value for kappa which is higher than 0,21 ($Kappa_{RF}^{All\ sets} = 0{,}44$; $Kappa_{RF}^{All\ sets-interaction} = 0{,}427$; $Kappa_{RF}^{Hotel\ \&\ travel-interaction} = 0{,}305$; $Kappa_{RF}^{Hotel\ \&\ travel} = 0{,}282$). Like discussed in the Methodology section, a kappa value higher than 0,21 is seen as a fair agreement (Landis & Koch, 1977). We can conclude that the random forest performs well, but not for all combinations of characteristics. When comparing the performance measures of all random forest models, we conclude that the 'best' performing model is the one that uses all sets as predictors. However, the difference with the least performing model (random forest with only travel circumstances) is high. Therefore, we can conclude that random forest is a good prediction model but not for all combinations of characteristics.

4.1.4 Support Vector Machine (SVM)

The last method is SVM and the performance measures are displayed in Table 4 of Appendix D. There are quite some models that perform well. Seven models even have an accuracy higher than 70% ($Acc_{SVM}^{Radial-all\ sets-interaction} = 0{,}783$; $Acc_{SVM}^{Radial-all\ sets} = 0{,}76$; $Acc_{SVM}^{Polynomial-all\ sets-interaction} = 0{,}743$; $Acc_{SVM}^{Polynomial-hotel\ \&\ travel-interaction} = 0{,}727$; $Acc_{SVM}^{Radial-hotel\ \&\ travel} = 0{,}723$; $Acc_{SVM}^{Polynomial-hotel\ \&\ travel} = 0{,}72$; $Acc_{SVM}^{Radial-hotel\ \&\ travel} = 0{,}717$) which indicates that the models are able to predict well (Deffenbacher, 1980). Out of all the SVM models, the two models, with the highest accuracy, both use a radial kernel instead of a polynomial kernel. Therefore, it appears that a SVM model with a radial kernel can predict customer's ratings better than a model with a polynomial kernel. Then, we conclude which SVM model is 'best' in predicting among all the SVM models. This SVM model is the one with radial kernel and all sets including interaction between all characteristics. The choice for this model is based on the fact that it has the highest performance measures in comparison with the other models ($Acc_{SVM}^{Radial-all\ sets-interaction} = 0{,}783$; $Prec_{SVM}^{Radial-all\ sets-interaction} = 0{,}759$; $Recall_{SVM}^{Radial-all\ sets-interaction} = 0{,}816$; $AUC_{SVM}^{Radial-all\ sets-interaction} = 0{,}784$; $F1_{SVM}^{Radial-all\ sets-interaction} = 0{,}787$; $Kappa_{SVM}^{Radial-all\ sets-interaction} = 0{,}567$; $MCC_{SVM}^{Radial-all\ sets-interaction} = 0{,}569$).The difference with the model without interaction isn't large but all performance measures, except for recall, are higher so therefore it is useful to add interaction to the model. It is also remarkable that the models that are on places 4 until 7 are all models with the combination of hotel and travel circumstances characteristics. This indicates that the combination of these two sets of characteristics is the best combination to predict customer's ratings, besides of the models with all sets.

4.2 Best performing model

After discussing the performance of each individual method, we discuss which model is best in predicting customer's ratings. We will compare all models of all the methods to find which model will be selected as the best performing model. We use this model for further interpretation. The ordered performance of all models can be found in Table 5 of Appendix D. The table is ordered based on accuracy meaning that the model with the highest accuracy is placed on the top and the model with the lowest accuracy on the bottom. To give a better insight, Table 2 below shows the 10 best performing models. It can be concluded from this table that the first six best performing models are all SVM models since the performance measures are the highest. Therefore, we can state that SVM is in general the best performing method, in terms of all performance measures. This makes sense since SVM is a good predictive method in a high dimensional space. In our case, there are quite some

variables, so we have a high dimensional space. It is also remarkable that logistic regression is by far the worst performing method. From this, we can conclude that machine learning techniques are more effective for predicting customer's rating than a logistic regression. This can be caused by the fact that machine learning methods can better model complex relationships.

Table 2: Performance measures top 10 best performing models

	Accuracy	Precision	Recall	AUC	F1	Kappa	MCC
SVM radial all sets interaction	0,783	0,759	0,816	0,784	0,787	0,567	0,569
SVM radial all sets	0,76	0,712	0,823	0,764	0,763	0,523	0,528
SVM polynomial all sets interaction	0,743	0,698	0,84	0,745	0,762	0,489	0,499
SVM polynomial hotel and travel circumstances interaction	0,727	0,695	0,789	0,728	0,739	0,454	0,459
SVM radial hotel and travel circumstances characteristics	0,723	0,699	0,765	0,724	0,731	0,448	0,449
SVM polynomial hotel and travel circumstances characteristics	0,72	0,683	0,799	0,722	0,737	0,442	0,448
Random forest all sets	0,72	0,702	0,745	0,722	0,723	0,44	0,441
SVM radial hotel and travel circumstances interaction	0,717	0,687	0,776	0,718	0,728	0,435	0,438
All sets interaction	0,713	0,695	0,738	0,716	0,716	0,427	0,428
Random forest hotel and travel circumstances interaction	0,652	0,617	0,681	0,653	0,648	0,305	0,306

We conclude that a SVM with a radial kernel that uses all sets including interaction is the best performing method. The performance measures of this model can be found in Table 2. This SVM model has a Cost-parameter of 1000 and a degree of 2.

Table 3: Performance SVM radial kernel all sets with interaction

Performance measure	Value
Accuracy	0,783
Precision	0,759
Recall	0,816
AUC	0,784
F1	0,787
Kappa	0,567
MCC	0,569

This part of the analysis answers two sub-questions, namely: 1) What is the effect on the prediction when combining different sets of characteristics?, and 2) How well can hotel ratings be predicted when using all sets of predictors? The answer to the first question is that combining different sets of

characteristics results, in general, in better performance measures. It appears from the analysis that all of the ten best predicting models contain more than one set of characteristics indicating that the performance improves when combining sets. The answer to the second question is that hotel ratings can be best predicted when using all sets of predictors. The two best performing methods use all sets of characteristics as predictors. The SVM model with a radial kernel that uses all sets including interaction has an accuracy of 78,3%.

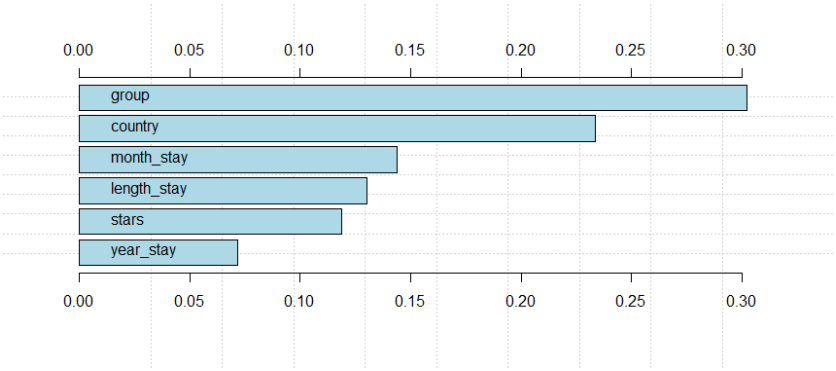
4.3 Interpretation best performing method

The SVM model with a radial kernel, all sets and interaction will be used for interpretation. As discussed in the methodology section, two global interpretation methods are used which we discuss separately.

4.3.1 Input importance

First, the input importance is calculated to determine which variable is most important in predicting the hotel rating. The results are showed in Figure 1. The figure shows the variance between the different inputs of each variable. The effect of all different inputs of an individual variable is calculated. The inputs of a variable mean all the values that the variable can have. For example, the variable traveller group has six values: Travelled with group, Business traveller, Solo traveller, Travelled with family, Travelled with family and small children and Travelled with partner. Afterwards, the variance of these effects is calculated and used to determine the importance. It appears from the figure below that traveller group has the highest importance and year of stay has the lowest importance. This indicates that traveller group is the most important variable in predicting hotel ratings and year of stay is the least important variable. The second most important variable is country since the difference between the scores of country and month of stay is quite high. Therefore, we can conclude that traveller group and country are the most important in predicting hotel ratings. Both variables are from different sets of characteristics, namely customer characteristics respectively and hotel characteristics. This means that there isn't one set that is more important in the prediction than another set.

Figure 1: Variable importance plot

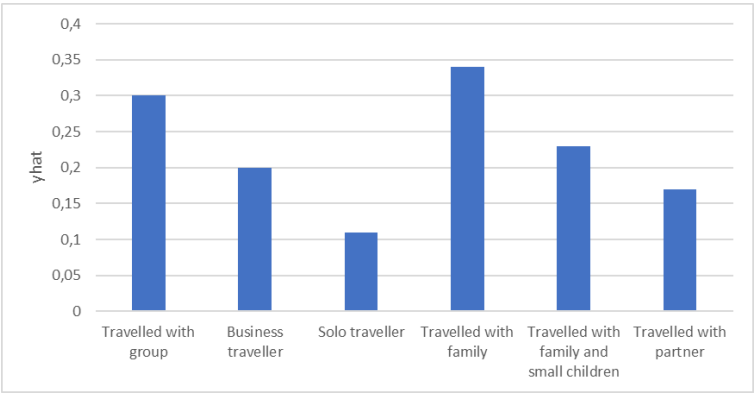


4.1.2 Partial dependence plots

Then, it is interesting to know what the effect of the variables is on the hotel rating. To be able to conclude this, partial dependence plots (PDP) for each variable are used. A PDP shows the marginal effect of a variable on the dependent variable. This means that for each variable the effect on the hotel rating is shown. The y-axis shows the marginal effect, so it indicates the increase or decrease in hotel rating resulting from the variable.

The PDP of traveller group shows that 'Travelled with family' has the highest positive effect meaning that someone that travelled with his or her family is more likely to give a higher rating than the other categories. The category 'Travelled with group' has the second highest effect indicating that someone who travelled in a group is also more likely to rate the hotel high. On the other hand, 'Solo traveller' has the least positive effect, so this category is less likely to give a high rating. Concluding, people who travel with family or a group are more likely to give higher ratings than someone who travels alone.

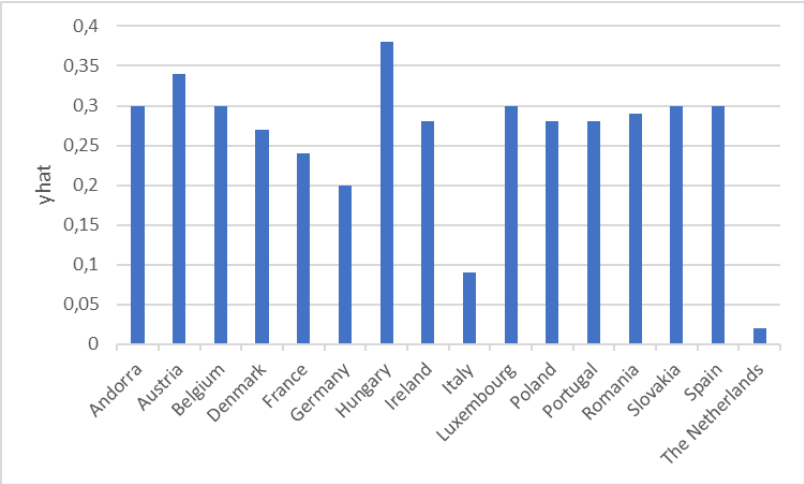
Figure 2: Partial dependence plot traveller group



Then, the PDP of country shows that Hungary has the highest positive effect indicating that people are more likely to give a high rating when the hotel is located in Hungary. Another country that is more likely to gain high ratings is Austria. The other countries have approximately the same ratings, except for The Netherlands and Italy. Hotels that are located in these two countries are less likely to receive

high ratings. In summary, hotels in Hungary and Austria are more likely to receive high ratings while hotels in The Netherlands and Italy are less likely to receive high ratings.

Figure 3: Partial dependence plot country



The next variable that will be discussed is month of stay; the PDP of this variable is showed in Figure 4. It appears that customers are most likely to give a high rating in winter since winter has the highest positive effect on hotel ratings. Customers are less likely to give a high rating during autumn since this has the lowest positive effect.

Figure 4: Partial dependence plot month of stay

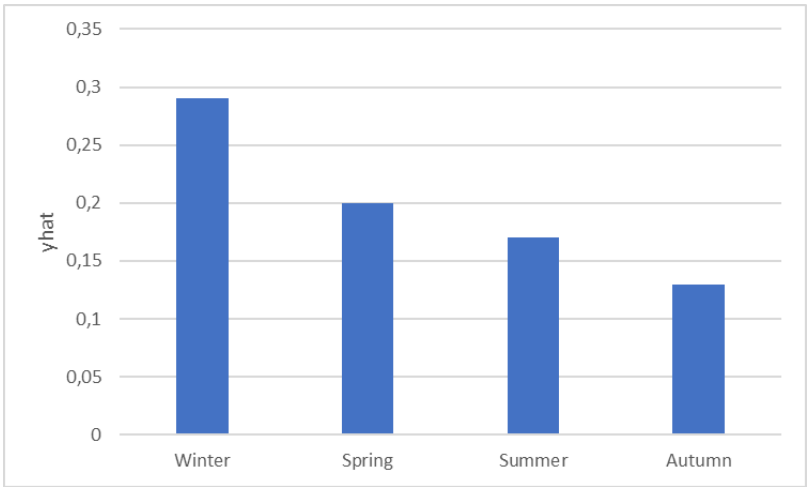
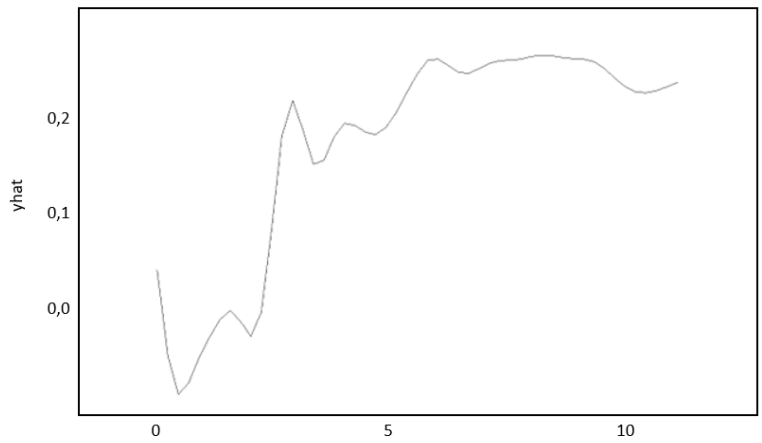


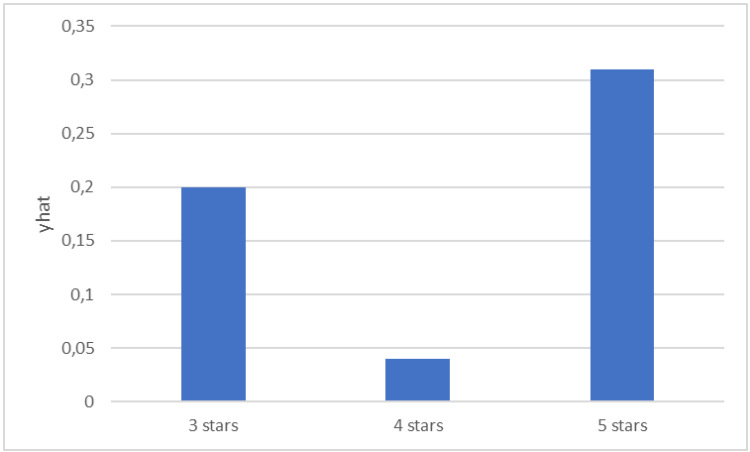
Figure 5 shows the PDP of the variable length of stay. The figure shows that a length of stay between 1 and 3 days results in a lower rating. The effect on the hotel rating is the lowest during these days. The rating increases until 6 days. When the length of stay is longer than 6 days, the positive effect on the hotel rating is equal. This indicates that people who stay between 6 and 10 days are most likely to give high ratings while people staying between 1 and 3 days are less likely to give high ratings.

Figure 5: Partial dependence plot length of stay



The following PDP is showed in Figure 6, this plot shows the effect of the variable number of stars on the hotel rating. The plot shows that hotels with 5 stars have the highest positive effect on hotel ratings meaning that hotels with 5 stars are most likely to receive high ratings. It is remarkable that hotels with 3 stars have a higher positive effect than hotels with 4 stars. This indicates that hotels with 3 stars are more likely to receive high ratings than hotels with 4 stars. It would have made more sense if hotels with 4 stars receive higher ratings since these hotels offer more services and are more luxurious than hotels with 3 stars.

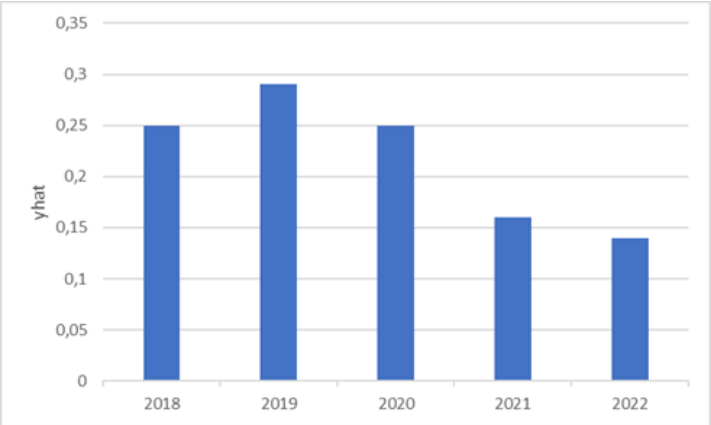
Figure 6: Partial dependence plot number of stars



The last variable is year of stay and the PDP of this variable is shown in Figure 7. The effects on the ratings of the years 2018, 2019 and 2020 are almost the same. The effects of the years 2021 and 2022 are also quite the same. This indicates that people who visited a hotel during 2018, 2019 and 2020 are more likely to give high ratings than people who travelled during 2021 and 2022. The decrease in hotel

ratings after 2020 could be declared by the effects of Covid. It might be that customers have other important factors in determining a hotel rating over the Covid pandemic which results in lower ratings.

Figure 7: Partial dependence plot year of stay



The discussion on the importance and marginal effect of the variables can help answer the third sub-questions, which was: Is there one set of characteristics that is best in predicting hotel ratings? We found that traveller group and country are the most important variables in prediction. These variables don't belong to the same set of characteristics. Out of these results, we can conclude that there isn't one set which performs better in predicting hotel ratings than others. This indicates that we need all sets of characteristics in order to have the best prediction. In this case, all sets of characteristics help predicting hotel ratings. In the literature review, we established that each set of characteristics helps in predicting hotel ratings. These propositions are true since we conclude that there isn't one or two sets that are best in predicting, but all sets are needed to be able to have the best prediction model. Furthermore, we expected that the set of hotel characteristics was more important in predicting hotel ratings than other sets. This proposition isn't true since we concluded that there isn't one set that is best in predicting ratings and that all sets are needed in order to have the best performance. Therefore, our results aren't in line with the discussed theory.

5. Conclusion & Discussion

The goal of this research was to find the best method and combination of characteristics to predict hotel ratings. In order to do this, we discussed three sub-questions and one central research question. First, the sub-questions will be answered. These answers are used to answer the central research question. Afterwards, we give recommendations to hotel owners and future researchers. Finally, we discuss the limitations of this research.

5.1 Sub-questions

The sub-questions are based on which set or combination of sets is best in predicting hotel ratings. To be able to answer these questions, we first need to discuss which method is the best. The performance of the sets when using this method is used in answering the questions. We concluded that SVM is the best method for predicting hotel ratings. This method is on the first five places of all the models.

5.1.1 Is there one set of characteristics that is best in predicting hotel ratings?

The models that are best in predicting hotel ratings are models that use all sets of characteristics as predictors. The models with only one set of characteristics perform worse than models with all sets and models with combination of sets. Therefore, we can't conclude that there is one set of characteristics that is best in predicting hotel ratings. This isn't in line with the theorization since we expected that the set of hotel characteristics would be more important in the prediction than the other sets. We think that this isn't the same as we expected since it is likely that the effect of an individual set is influenced by other sets which results in a better performance when the sets are combined in the prediction model. When looking at the models of SVM, it is remarkable that the set of customer characteristics is on the last two places. This indicates that the customer characteristics are worse in predicting hotel ratings than hotel and travel circumstances.

Another way to answer this question is to use the results from the interpretation of the best performing model. We researched which variables are most important in the prediction. We concluded that the most important variable was traveller group and the second most important one was country. Based on the literature, the importance of traveller group can be declared. Previous research concluded that different traveller groups have different determinants of satisfaction (Susilo & Cats, 2014). Therefore, it makes sense that the ratings differ among traveller group and that this variable is important in the prediction. The effect of country can also be explained since the hotel's location is one of the main attributes in a traveller's hotel choice (Mccleary et al., 1993). Hence, it makes sense that the hotel's country is important in the prediction of hotel ratings. Traveller group belongs to the

set of customer characteristics and country belongs to hotel characteristics. Contrary to the discussion above, this indicates that customer characteristics are important in predicting ratings and may be better in predicting than the other sets. However, we can't conclude that this set is the best since the effect is influenced by the other variables from other sets.

5.1.2 What is the effect on the prediction when combining different sets of characteristics?

Besides models with individual sets of characteristics, we have models with combinations of sets and models that include interaction. The results show that combinations of sets are in general performing better than models with only one set of characteristics. We also conclude that adding interaction effects to the models improves the performance. This is concluded from Table 4 of Appendix D since almost all models with more than one set of characteristics have a higher place based on the performance than the models with a single set. Furthermore, the models including interaction often have a higher place than the models without interaction. Therefore, we conclude that combining different sets of characteristics has a positive effect on the prediction of the hotel ratings since the performance increases.

5.1.3 How well can hotel ratings be predicted when using all sets of predictors?

The last question that needs to be answered is about the performance of models that use all sets of characteristics as predictors. As discussed before, the best performing model is the one with all sets including interaction effects (radial kernel). Besides, the second-best performing model is a SVM with radial kernel that has all sets as predictors. This indicates that using all sets leads to a better performance than using one or two sets. This is in line with the propositions from the literature review since we expected that each set was helpful in the prediction of hotel ratings. Then, we discuss how well hotel ratings can be predicted with this model. All performance measures are quite high, which is showed in Table 2. The accuracy, precision, recall, AUC and F1 are all around 0,8 which indicates that the model is quite well in predicting hotel ratings. Furthermore, Kappa and MCC are above 0,5. Like discussed in the Data & Methodology, a Kappa value of 0,21 is seen as a fair agreement meaning that a value above 0,5 shows a very good performing model. MCC has a value between -1 and 1 so a value of 0,5 is also good. In conclusion, hotel ratings can be predicted quite well when using all sets based on all performance measures.

5.2 Central research question

After answering the sub-questions, the central research question can be discussed. The research question was as following:

"How can hotel ratings be predicted by different sets of characteristics and which set or combination of sets is best in predicting?"

The central research question consists of two parts. The first part focuses on how hotel ratings can be predicted meaning that this part is about which method is best for predicting hotel ratings. We compared four different methods: logistic regression, decision tree, random forest and support vector machine. We found that logistic regression is the worst performing method and SVM is the best performing method. The logistic regression models had, for example, an accuracy around 0,4 while the SVM models had an accuracy around 0,7. This indicates that there is quite a difference between the best and worst method. Therefore, it is useful to choose SVM for prediction in order to have the best performance.

The second part of the question is about which set or combination of sets of characteristics is best in predicting. This part of the question is extensively discussed in the sub-questions. While answering the sub-questions, we found that there isn't one set of characteristics that is best in predicting. Furthermore, we concluded that combining multiple sets resulted in a positive effect on the performance measures. Therefore, it is useful to combine sets. Next, the results showed that adding interaction effects also results in an increase of the performance. Finally, we found that using all sets of characteristics including interaction effects resulted in the overall best performance. In summary, the best way to predict hotel ratings is by using all sets and by adding interaction effects.

5.3 Recommendations to hotel owners

The most interesting part of this research for hotel owners is what variables are most important for predicting hotel ratings. We found that the traveller group is the most important predictor of hotel ratings. People who travel with family or a group are more likely to give a higher rating than people who travel alone. This could indicate that customers are more satisfied when travelling with other people than when travelling alone. The goal of hotel owners is to increase the average hotel rating, so they probably want to focus on customers who are more likely to give low ratings. In this case, the hotel owners could give more attention to single travellers by, for example, offering them a free drink or another act of service. By giving these travellers extra attention, they might be more satisfied and be more likely to give a higher rating. Another important predictor is country. We concluded that hotels that are located in Hungary and Austria are most likely to receive higher ratings while hotels in The Netherlands and Italy are less likely to receive high ratings. Current hotel owners couldn't change the place of their hotel, but hotel owners that have multiple hotels in different countries who would like to open a new hotel in another country, could consider choosing one of the countries that receives

higher ratings. In conclusion, hotel owners could increase their ratings by giving more attention to single travellers since they are more likely to give lower ratings. When an owner considers opening a new hotel in another country, the best choice is Hungary or Austria since hotels in these countries receive the highest ratings.

The effect of the other variables can also improve ratings so these will also be translated into recommendations for hotel owners. The results showed that customers are most likely to give high ratings during winter and are less likely to give high ratings during autumn. This indicates that customers are less satisfied during autumn meaning that hotel owners could give them more attention in order to receive higher ratings during this season. Then, the variable length of stay showed that people are most likely to give high ratings when they stay between 6 and 10 days. People who stay between 1 and 3 days are less likely to give high ratings. We recommend to hotel owners to give people with a short stay more attention by offering them act of services. The extra attention could result in a higher satisfaction among these customers which causes higher ratings. The variable year of stay showed that the ratings decreased after 2020 indicating that customers have other determinants of satisfaction because of the Covid pandemic. It is interesting for hotel owners to know what these determinants are so they could, for example, ask whether customers find hygiene more important than before the Covid pandemic.

5.4 Recommendations to future researchers

In this research, we only investigated which method and which set(s) of characteristics is best in predicting hotel ratings. We also recommended some actions to hotel owners based on the results. However, we couldn't check whether these recommendations will result in higher ratings. A suggestion for further research in this area is to research whether the actions of hotel owners result in higher ratings. Another recommendation for further research is to investigate why hotels in some countries receive higher ratings than hotels in other countries. Based on this, there could be more understanding about what effects ratings. It could, for example, be the case that a certain environment or a type of weather affects customers' satisfaction resulting in a higher or lower rating. These effects could be added to the prediction model as a control variable in order to have a better performance. The last recommendation for further research is to use more variables. We only used variables that were available from Expedia, but there are more variables that could be interesting. An example of a variable that could be interesting is the nationality of the reviewer. We already found that the country of the hotel is important for the prediction, but the reviewer's nationality could also be related to the hotel rating.

5.5 Limitations

Finally, this research has some limitations. The first limitation is the amount of data and the distribution of the data. The dataset was quite small and unequally distributed. The unequal distribution could cause problems in the analysis since the used methods can't handle unbalanced data well. To solve this problem, the data was balanced by oversampling meaning that the dataset has increased. However, the created new observations aren't 'real' observations. It is better to have a larger dataset that is balanced which consists of real observations. In order to create such a dataset, more reviews should be scraped by selecting more hotels or using different websites.

Another limitation is that we used only hotels that are part of NH Hotel. We used one hotel group since we wanted to make sure that the hotels are comparable. When different hotel groups are used, there might be important differences between the hotels that aren't captured by the current analysis. However, each hotel group attracts a certain type of customer based on the price category, services, etcetera. Therefore, the results could be different when running the same analysis for another hotel group. We didn't have the time to analyse another hotel group, but it could be interesting to investigate whether the results differ between hotel groups and what possibly drives these differences. This gives a better impression of the variables that influence the customers' hotel rating.

Bibliography

Abenoza, R. F., Liu, C., Cats, O., & Susilo, Y. O. (2019). What is the role of weather, built-environment and accessibility geographical characteristics in influencing travellers' experience? *Transportation Research Part A: Policy and Practice*, *122*, 34–50. https://doi.org/10.1016/J.TRA.2019.01.026

Aicher, J., Asiimwe, F., Batchuluun, B., Hauschild, M., Zöhrer, M., & Egger, R. (2016). Online Hotel Reviews: Rating Symbols or Text… Text or Rating Symbols? That Is the Question! In *Information and Communication Technologies in Tourism 2016* (pp. 369–382). Springer International Publishing. https://doi.org/10.1007/978-3-319-28231-2_27

Alegre, J., Mateo, S., & Pou, L. (2011). A latent class approach to tourists' length of stay. *Tourism Management*, *32*(3), 555–563. https://doi.org/10.1016/J.TOURMAN.2010.05.003

Baek, H., Ahn, J., & Choi, Y. (2012). Helpfulness of Online Consumer Reviews: Readers' Objectives and Review Cues. *International Journal of Electronic Commerce*, *17*(2), 99–126. https://doi.org/10.2753/JEC1086-4415170204

Ben-David, A. (2007). A lot of randomness is hiding in accuracy. *Engineering Applications of Artificial Intelligence*, *20*(7), 875–885. https://doi.org/10.1016/J.ENGAPPAI.2007.01.001

Bigné, E., William, E., & Soria-Olivas, E. (2020). Similarity and Consistency in Hotel Online Ratings across Platforms. *Journal of Travel Research*, *59*(4), 742–758. https://doi.org/10.1177/0047287519859705

Bulchand-Gidumal, J., Melián-González, S., & González Lopez-Valcarcel, B. (2013). A social media analysis of the contribution of destinations to client satisfaction with hotels. *International Journal of Hospitality Management*, *35*, 44–47. https://doi.org/10.1016/J.IJHM.2013.05.003

Cadotte, E. R., & Turgeon, N. (1988). Key Factors in Guest Satisfaction. *Cornell Hotel and Restaurant Administration Quarterly*, *28*(4), 44–51. https://doi.org/10.1177/001088048802800415/ASSET/001088048802800415.FP.PNG_V03

Cao, M., & Wei, J. (2005). Stock market returns: A note on temperature anomaly. *Journal of Banking & Finance*, *29*(6), 1559–1573. https://doi.org/10.1016/J.JBANKFIN.2004.06.028

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7

Chong, A. Y. L., Li, B., Ngai, E. W. T., Ch'ng, E., & Lee, F. (2016). Predicting online product sales via online reviews, sentiments, and promotion strategies. *International Journal of Operations & Production Management*, *36*(4), 358–383. https://doi.org/10.1108/IJOPM-03-2015-0151

Chu, W. T., & Huang, W. H. (2017). Cultural difference and visual information on hotel rating prediction. *World Wide Web*, *20*(4), 595–619. https://doi.org/10.1007/S11280-016-0404-2/TABLES/7

Connolly, M. (2013). Some Like It Mild and Not Too Wet: The Influence of Weather on Subjective Well-Being. *Journal of Happiness Studies*, *14*(2), 457–473. https://doi.org/10.1007/S10902-012-9338-2

Cortez, P., & Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences*, *225*, 1–17. https://doi.org/10.1016/J.INS.2012.10.039

de Albornoz, J. C., Plaza, L., Gervás, P., & Díaz, A. (2011). A Joint Model of Feature Mining and Sentiment Analysis for Product Review Rating. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6611 LNCS*, 55–66. https://doi.org/10.1007/978-3-642-20161-5_8

Deffenbacher, K. A. (1980). Can We Infer Anything about Their Relationship? *Law & Human Behavior*, *4*(4), 243–260.

Elshawi, R., Al-Mallah, M. H., & Sakr, S. (2019). On the interpretability of machine learning-based model for predicting hypertension. *BMC Medical Informatics and Decision Making*, *19*(1), 1–32. https://doi.org/10.1186/S12911-019-0874-0/FIGURES/48

Ettema, D., Friman, M., Olsson, L. E., & Gärling, T. (2017). Season and weather effects on travel-related mood and travel satisfaction. *Frontiers in Psychology*, *8*(FEB), 140. https://doi.org/10.3389/FPSYG.2017.00140/BIBTEX

García, S., Fernández, A., Luengo, J., & Herrera, F. (2009). A study of statistical techniques and performance measures for genetics-based machine learning: Accuracy and interpretability. *Soft Computing*, *13*(10), 959–977. https://doi.org/10.1007/S00500-008-0392-Y/TABLES/16

Gundersen, M. G., Heide, M., & Olsson, U. H. (1996). Hotel Guest Satisfaction among Business Travellers: What Are the Important Factors? *Cornell Hotel and Restaurant Administration Quarterly*, *37*(2), 72–81.

He, D., Yao, Z., Zhao, F., & Feng, J. (2020). How do weather factors drive online reviews? The mediating role of online reviewers' affect. *Industrial Management and Data Systems*, *120*(11), 2133–2149. https://doi.org/10.1108/IMDS-02-2020-0121/FULL/PDF

Hu, Y.-H., & Chen, K. (2016). Predicting hotel review helpfulness: The impact of review visibility, and interaction between hotel stars and review ratings. *International Journal of Information Management*, *36*(6), 929–944. https://doi.org/10.1016/j.ijinfomgt.2016.06.003

Jeong, J. Y., Crompton, J. L., & Hyun, S. S. (2019). What makes you select a higher price option? Price–quality heuristics, cultures, and travel group compositions. *International Journal of Tourism Research*, *21*(1), 1–10. https://doi.org/10.1002/JTR.2236

Kandampully, J., & Suhartanto, D. (2000). Customer loyalty in the hotel industry: The role of customer satisfaction and image. *International Journal of Contemporary Hospitality Management*, *12*(6), 346–351. https://doi.org/10.1108/09596110010342559/FULL/XML

Kitchenham, B. (1998). A procedure for analyzing unbalanced datasets. *IEEE Transactions on Software Engineering*, *24*(4), 278–301. https://doi.org/10.1109/32.677185

Ladhari, R. (2009). Service quality, emotional satisfaction, and behavioural intentions: A study in the hotel industry. *Managing Service Quality*, *19*(3), 308–331. https://doi.org/10.1108/09604520910955320/FULL/PDF

Landis, J. R., & Koch, G. G. (1977). An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement among Multiple Observers. *Biometrics*, *33*(2), 363. https://doi.org/10.2307/2529786

Lee, K., Kim, H., Kim, H., Tourism, D. L.-J. of H. and, & 2010, undefined. (2010). The determinants of factors in FIT guests' perception of hotel location. *Cambridge.Org*. https://doi.org/10.1375/jhtm.17.1.167

Lunardon, N., Menardi, G., & Torelli, N. (2014). ROSE: A package for binary imbalanced learning. *R Journal*, *6*(1), 79–89. https://doi.org/10.32614/RJ-2014-008

Mccleary, K. W., Weaver, P. A., & Hutchinson, J. C. (1993). Hotel Selection Factors as They Relate to Business Travel Situations. *Journal of Travel Research*, *32*(2), 42–48. https://doi.org/10.1177/004728759303200206

Müller, K.-R., Smola, A. J., Rätsch, G., Schölkopf, B., Kohlmorgen, J., & Vapnik, V. (1997). *Predicting time series with support vector machines* (pp. 999–1004). https://doi.org/10.1007/BFb0020283

Murray, K. B., di Muro, F., Finn, A., & Popkowski Leszczyc, P. (2010). The effect of weather on consumer spending. *Journal of Retailing and Consumer Services*, *17*(6), 512–520. https://doi.org/10.1016/J.JRETCONSER.2010.08.006

Nawijn, J. (2010). The holiday happiness curve: a preliminary investigation into mood during a holiday abroad. *International Journal of Tourism Research*, *12*(3), 281–290. https://doi.org/10.1002/JTR.756

Neal, J. D. (2008). The Effect of Length of Stay on Travellers' Perceived Satisfaction with Service Quality. *Http://Dx.Doi.Org/10.1300/J162v04n03_11*, *4*(3–4), 167–176. https://doi.org/10.1300/J162V04N03_11

Nunkoo, R., Teeroovengadum, V., Ringle, C. M., & Sunnassee, V. (2020). Service quality and customer satisfaction: The moderating effects of hotel star rating. *International Journal of Hospitality Management*, *91*, 102414. https://doi.org/10.1016/J.IJHM.2019.102414

Pal, M. (2007). Random forest classifier for remote sensing classification. *Http://Dx.Doi.Org.Eur.Idm.Oclc.Org/10.1080/01431160412331269698*, *26*(1), 217–222. https://doi.org/10.1080/01431160412331269698

Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2010). An Introduction to Logistic Regression Analysis and Reporting. *Https://Doi.Org/10.1080/00220670209598786*, *96*(1), 3–14. https://doi.org/10.1080/00220670209598786

Phillips, P., Zigan, K., Santos Silva, M. M., & Schegg, R. (2015). The interactive effects of online reviews on the determinants of Swiss hotel performance: A neural network analysis. *Tourism Management*, *50*, 130–141. https://doi.org/10.1016/J.TOURMAN.2015.01.028

Pieters, R., & Wedel, M. (2004). Attention Capture and Transfer in Advertising: Brand, Pictorial, and Text-Size Effects. *Journal of Marketing*, *68*(2), 36–50. https://doi.org/10.1509/jmkg.68.2.36.27794

Provost, F. (2000). *Machine Learning from Imbalanced Data Sets 101*. www.aaai.org

Qu, L., Ifrim, G., & Weikum, G. (2010). *The Bag-of-Opinions Method for Review Rating Prediction from Sparse Text Patterns*. 913–921.

Safavian, S. R., & Landgrebe, D. (1991). A Survey of Decision Tree Classifier Methodology. *IEEE Transactions on Systems, Man and Cybernetics*, *21*(3), 660–674. https://doi.org/10.1109/21.97458

Senecal, S., & Nantel, J. (2004). The influence of online product recommendations on consumers' online choices. *Journal of Retailing*, *80*(2), 159–169. https://doi.org/10.1016/J.JRETAI.2004.04.001

Stone, M. J. (2016). Deciding not to choose: Delegation to social surrogates in tourism decisions. *Tourism Management*, *57*, 168–179. https://doi.org/10.1016/J.TOURMAN.2016.06.002

Susilo, Y. O., & Cats, O. (2014). Exploring key determinants of travel satisfaction for multi-modal trips by different traveller groups. *Transportation Research Part A: Policy and Practice*, *67*, 366–380. https://doi.org/10.1016/J.TRA.2014.08.002

Tian, X., Cao, S., & Song, Y. (2021). The impact of weather on consumer behavior and retail performance: Evidence from a convenience store chain in China. *Journal of Retailing and Consumer Services*, *62*, 102583. https://doi.org/10.1016/J.JRETCONSER.2021.102583

Walther, B. A., & Moore, J. L. (2005). The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography*, *28*(6), 815–829. https://doi.org/10.1111/j.2005.0906-7590.04112.x

Wang, L., Fong, D. K. C., Law, R., & Fang, B. (2017). Length of Stay: Its Determinants and Outcomes: *Https://Doi.Org/10.1177/0047287517700315*, *57*(4), 472–482. https://doi.org/10.1177/0047287517700315

Xiang, Z., & Krawczyk, M. (2016). What Does Hotel Location Mean for the Online Consumer? Text Analytics Using Online Reviews. *Information and Communication Technologies in Tourism 2016*, 383–395. https://doi.org/10.1007/978-3-319-28231-2_28

Xie, K. L., Zhang, Z., & Zhang, Z. (2014). The business value of online consumer reviews and management response to hotel performance. *International Journal of Hospitality Management*, *43*, 1–12. https://doi.org/10.1016/j.ijhm.2014.07.007

Xu, X. (2018). Does traveller satisfaction differ in various travel group compositions?: Evidence from online reviews. *International Journal of Contemporary Hospitality Management*, *30*(3), 1663–1685. https://doi.org/10.1108/IJCHM-03-2017-0171/FULL/PDF

Yang, Y., Luo, H., & Law, R. (2014). Theoretical, empirical, and operational models in hotel location research. *International Journal of Hospitality Management*, *36*, 209–220. https://doi.org/10.1016/J.IJHM.2013.09.004

Ye, Q., Law, R., & Gu, B. (2009). The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, *28*(1), 180–182. https://doi.org/10.1016/J.IJHM.2008.06.011

Appendix A Proportion table

Table 1: Proportion table hotel rating

Hotel rating	Proportion in %
1/5 Terrible	2,75%
2/5 Poor	4,58%
3/5 Okay	7,91%
4/5 Good	26,91%
5/5 Excellent	57,83%

Table 2: Proportion table hotel rating after merging

Hotel rating	Proportion in %
Low	15,25%
High	84,75%

Table 3: Proportion table hotel rating balanced dataset

Hotel rating	Proportion in %
Low	50,3%
High	49,7%

Appendix B Proportion tables predictor variables

Table 1: Proportion table travel group composition

Travel group composition	Proportion in %
Solo traveller	36,17%
Travelled with family	21%
Travelled with partner	15,17%
Business traveller	15,08%
Travelled with group	5,5%
Travelled with family and small children	1,25%
Business traveller, Travelled with family	1,08%
Business traveller, Travelled with partner	0,92%
Travelled with family, Travelled with group	0,67%
Travelled with family, Travelled with small children	0,58%
Business traveller, Travelled with group	0,5%
Travelled with family, Travelled with partner	0,5%
Travelled with partner, Travelled with group	0,5%
Travelled with family, Travelled with pets	0,25%
Travelled with partner, Travelled with pets	0,17%
Business traveller, Travelled with family, Travelled with group	0,17%
Travelled with group, Business traveller, Travelled with pets	0,17%
Travelled with pets	0,08%
Travelled with small children, Business traveller	0,08%
Travelled with small children, Travelled with partner	0,08%
Travelled with small children, Travelled with partner, Travelled with family	0,08%
Travelled with family, Travelled with small children, Travelled with pets, Travelled with partner, Travelled with group	0,08%

All the categories that have “Business traveller” in it will be merged with the category “Business traveller” since this indicates that the type of travel was somehow related to business. The category “Travelled with pets” is merged with “Solo traveller” since this category hasn’t a combination with one that indicates that the customer travelled with other people. Then, the categories that are left and contain “Travelled with family and small children” are merged into one category with this label. The other categories that contain “Travelled with family” are combined under this category. The last combination is of the categories, that are still left, and contain “Travelled with partner”. These are merged into one category with the label “Travelled with partner”. The remaining category is “Travelled with group”. The results are showed in Table 2.

Table 2: Proportion table travel group composition new levels

Travel group composition	Proportion in %
Solo traveller	36,25%
Travelled with family	22,41%
Business traveller	17,91%
Travelled with partner	15,83%
Travelled with group	5,5%
Travelled with family and small children	2,08%

Table 3: Summary table length of stay

Summary statistic	Value
Mean	2,28
Standard deviation	1,75
Minimum	1
Maximum	19

Table 4: Proportion table month of stay

Month of stay	Proportion in %
January	6,33%
February	9,08%
March	8,67%
April	11%
May	19,67%
June	13,75%
July	2,92%
August	3,08%
September	4,83%
October	7,58%
November	6%
December	7,08%

Table 5: Proportion table travel season (combined month of stays)

Travel season	Proportion in %
Spring	39,33%
Winter	22,5%
Summer	19,75%
Autumn	18,42%

Table 6: Proportion table number of stars

Travel season	Proportion in %
3 stars	5,42%
4 stars	84,17%
5 stars	10,42%

Table 7: Proportion table hotel's country

Country	Proportion in %
Germany	21,25%
Belgium	5,42%
France	2,08%
Hungary	1,25%
Andorra	0,42%
Denmark	0,42%
Ireland	0,42%
Luxembourg	0,42%
Poland	0,42%
Romania	0,42%
Slovakia	0,42%
Austria	2,92%
Portugal	6,67%
Italy	22,92%
Spain	20,83%
The Netherlands	13,75%

Appendix C Correlation

Figure 1: Correlation plot independent variables

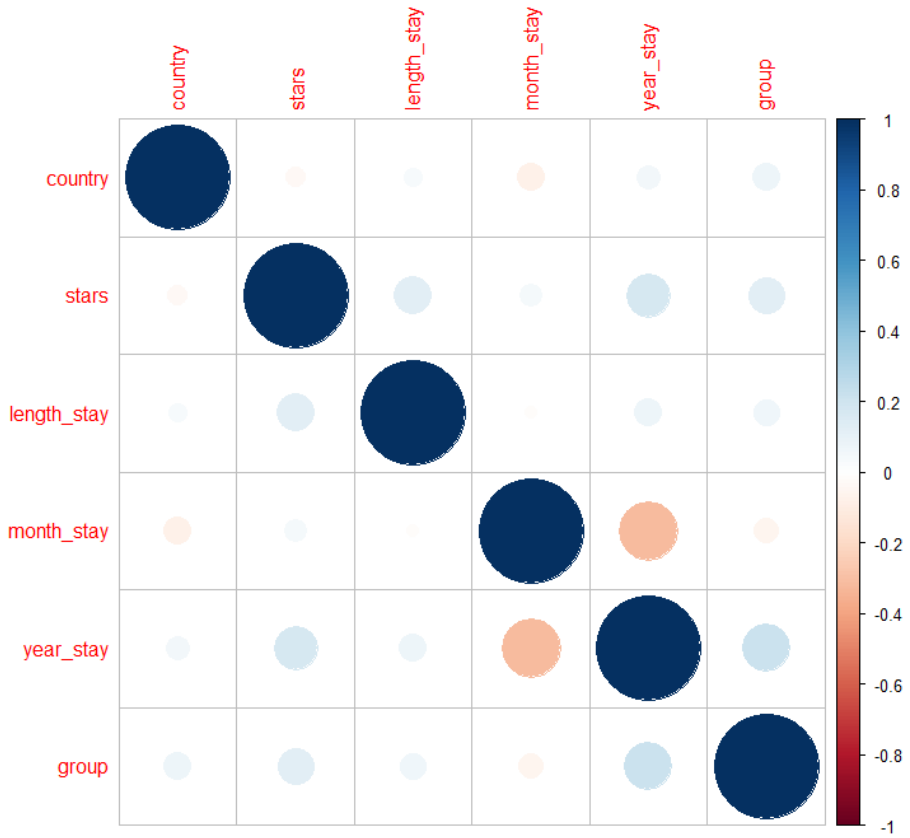


Table 1: VIF scores independent variables

Variable	VIF
Country	1,019
Stars	1,078
Length of stay	1,025
Monty of stay	1,132
Year of stay	1,321
Group	1,201

Appendix D Performance measures

Table 1: Performance logistic regression

	Accuracy	Precision	Recall	AUC	F1	Kappa	MCC
Travel circumstances	0,457	0,466	0,594	0,57	0,522	-0,086	-0,09
Customer characteristics	0,456	0,422	0,238	0,556	0,304	-0,088	-0,098
Customer and hotel characteristics interaction	0,452	0,447	0,405	0,548	0,425	-0,096	-0,096
Customer and travel circumstances	0,445	0,442	0,44	0,584	0,442	-0,111	-0,111
Customer and hotel characteristics	0,427	0,416	0,362	0,615	0,387	-0,146	-0,147
Hotel characteristics	0,426	0,404	0,313	0,602	0,352	-0,149	-0,152
All sets	0,424	0,419	0,397	0,62	0,408	-0,152	-0,152
Hotel and travel circumstances	0,423	0,406	0,337	0,613	0,368	-0,154	-0,157
Customer and travel circumstances interaction	0,404	0,403	0,403	0,636	0,403	-0,193	-0,193
Hotel and travel circumstances interaction	0,383	0,383	0,387	0,617	0,385	-0,234	-0,234
All sets interaction	0,348	0,335	0,298	0,651	0,312	-0,302	-0,305

Table 2: Performance decision tree

	Accuracy	Precision	Recall	AUC	F1	Kappa	MCC
Gini customer and hotel characteristics	0,552	0,549	0,532	0,551	0,54	0,103	0,103
Gini customer characteristics	0,538	0,522	0,808	0,541	0,634	0,082	0,097
CER all sets	0,537	0,517	1	0,541	0,681	0,082	0,206
CER hotel characteristics	0,537	0,516	1	0,541	0,681	0,082	0,206
Gini hotel characteristics	0,537	0,516	1	0,571	0,681	0,144	0,206
CER customer and hotel characteristics	0,537	0,516	1	0,541	0,681	0,082	0,206
CER hotel and travel circumstances	0,537	0,516	1	0,541	0,681	0,082	0,206
Gini all sets	0,505	0,505	1	0,5	0,671	0	-
CER customer characteristics	0,505	0,505	1	0,5	0,671	0	-
CER travel circumstances characteristics	0,495	0,495	1	0,5	0,662	0	-
Gini travel circumstances characteristics	0,495	0,495	1	0,5	0,662	0	-
CER customer and travel circumstances	0,495	0,495	1	0,5	0,662	0	-
Gini customer and travel circumstances	0,495	0,495	1	0,5	0,662	0	-
Decision tree Gini hotel and travel circumstances	0,495	0,495	1	0,5	0,662	0	-

Table 3: Performance random forest

	Accuracy	Precision	Recall	AUC	F1	Kappa	MCC
All sets	0,72	0,702	0,745	0,722	0,723	0,44	0,441
All sets interaction	0,713	0,695	0,738	0,716	0,716	0,427	0,428
Hotel and travel circumstances interaction	0,652	0,617	0,681	0,653	0,648	0,305	0,306
Hotel and travel circumstances characteristics	0,64	0,606	0,67	0,643	0,636	0,282	0,283
Customer characteristics	0,605	0,513	0,745	0,568	0,607	0,135	0,252
Customer and travel circumstances interaction	0,605	0,575	0,6	0,603	0,587	0,205	0,21
Customer and travel circumstances	0,603	0,575	0,596	0,603	0,585	0,205	0,206
Hotel characteristics	0,577	0,555	0,684	0,579	0,613	0,157	0,161
Customer and hotel characteristics	0,563	0,533	0,578	0,564	0,554	0,128	0,128
Customer and hotel characteristics interaction	0,563	0,533	0,578	0,564	0,554	0,128	0,128
Travel circumstances	0,51	0,452	0,496	0,556	0,473	0,113	0,018

Table 4: Performance support vector machine

	Accuracy	Precision	Recall	AUC	F1	Kappa	MCC
Radial all sets interaction	0,783	0,759	0,816	0,784	0,787	0,567	0,569
Radial all sets	0,76	0,712	0,823	0,764	0,763	0,523	0,528
Polynomial all sets interaction	0,743	0,698	0,84	0,745	0,762	0,489	0,499
Polynomial hotel and travel circumstances interaction	0,727	0,695	0,789	0,728	0,739	0,454	0,459
Radial hotel and travel circumstances	0,723	0,699	0,765	0,724	0,731	0,448	0,449
Polynomial hotel and travel circumstances characteristics	0,72	0,683	0,799	0,722	0,737	0,442	0,448
Radial hotel and travel circumstances interaction	0,717	0,687	0,776	0,718	0,728	0,435	0,438
Polynomial all sets	0,688	0,656	0,759	0,69	0,705	0,378	0,383
Radial customer and travel circumstances	0,625	0,612	0,639	0,625	0,626	0,25	0,251
Radial customer and travel circumstances interaction	0,602	0,589	0,619	0,602	0,604	0,204	0,204
Radial hotel characteristics	0,583	0,597	0,459	0,581	0,519	0,163	0,167
Polynomial hotel characteristics	0,583	0,597	0,459	0,581	0,519	0,163	0,167
Radial customer and hotel characteristics	0,58	0,555	0,718	0,583	0,626	0,165	0,172
Polynomial customer and hotel characteristics	0,58	0,555	0,718	0,583	0,626	0,165	0,172
Radial customer and hotel characteristics interaction	0,58	0,555	0,718	0,583	0,626	0,165	0,172
Polynomial customer and hotel characteristics interaction	0,58	0,555	0,718	0,583	0,626	0,165	0,172
Polynomial travel circumstances	0,573	0,579	0,48	0,571	0,524	0,144	0,146
Radial travel circumstances	0,57	0,58	0,446	0,568	0,504	0,136	0,139
Polynomial customer and travel circumstances	0,558	0,547	0,578	0,559	0,562	0,117	0,118
Polynomial customer and travel circumstances interaction	0,555	0,543	0,575	0,555	0,559	0,111	0,111
Radial customer characteristics	0,493	0,488	0,707	0,498	0,578	-0,005	-0,005
Polynomial customer characteristics	0,493	0,488	0,707	0,498	0,578	-0,005	-0,005

Table 5: Performance all models

	Accuracy	Precision	Recall	AUC	F1	Kappa	MCC
SVM radial all sets interaction	0,783	0,759	0,816	0,784	0,787	0,567	0,569
SVM radial all sets	0,76	0,712	0,823	0,764	0,763	0,523	0,528
SVM polynomial all sets interaction	0,743	0,698	0,84	0,745	0,762	0,489	0,499
SVM polynomial hotel and travel circumstances interaction	0,727	0,695	0,789	0,728	0,739	0,454	0,459
SVM radial hotel and travel circumstances characteristics	0,723	0,699	0,765	0,724	0,731	0,448	0,449
SVM polynomial hotel and travel circumstances characteristics	0,72	0,683	0,799	0,722	0,737	0,442	0,448
Random forest all sets	0,72	0,702	0,745	0,722	0,723	0,44	0,441
SVM radial hotel and travel circumstances interaction	0,717	0,687	0,776	0,718	0,728	0,435	0,438
Random forest all sets interaction	0,713	0,695	0,738	0,716	0,716	0,427	0,428
Random forest hotel and travel circumstances interaction	0,652	0,617	0,681	0,653	0,648	0,305	0,306
Random forest hotel and travel circumstances characteristics	0,64	0,606	0,67	0,643	0,636	0,282	0,283
SVM radial customer and travel circumstances characteristics	0,625	0,612	0,639	0,625	0,626	0,25	0,251
Random forest customer characteristics	0,605	0,513	0,745	0,568	0,607	0,135	0,252
Random forest customer and travel circumstances interaction	0,605	0,575	0,6	0,603	0,587	0,205	0,21
Random forest customer and travel circumstances characteristics	0,603	0,575	0,596	0,603	0,585	0,205	0,206
SVM radial customer and travel circumstances interaction	0,602	0,589	0,619	0,602	0,604	0,204	0,204
SVM radial hotel characteristics	0,583	0,597	0,459	0,581	0,519	0,163	0,167
SVM polynomial hotel characteristics	0,583	0,597	0,459	0,581	0,519	0,163	0,167
SVM radial customer and hotel characteristics	0,58	0,555	0,718	0,583	0,626	0,165	0,172
SVM polynomial customer and hotel characteristics	0,58	0,555	0,718	0,583	0,626	0,165	0,172
SVM radial customer and hotel characteristics interaction	0,58	0,555	0,718	0,583	0,626	0,165	0,172
SVM polynomial customer and hotel characteristics interaction	0,58	0,555	0,718	0,583	0,626	0,165	0,172
Random forest hotel characteristics	0,577	0,555	0,684	0,579	0,613	0,157	0,161
SVM polynomial travel circumstances characteristics	0,573	0,579	0,48	0,571	0,524	0,144	0,146
SVM radial travel circumstances characteristics	0,57	0,58	0,446	0,568	0,504	0,136	0,139
Random forest customer and hotel characteristics	0,563	0,533	0,578	0,564	0,554	0,128	0,128
Random forest customer and hotel characteristics interaction	0,563	0,533	0,578	0,564	0,554	0,128	0,128
SVM polynomial customer and travel circumstances characteristics	0,558	0,547	0,578	0,559	0,562	0,117	0,118
SVM polynomial customer and travel circumstances interaction	0,555	0,543	0,575	0,555	0,559	0,111	0,111
Decision tree Gini customer and hotel characteristics	0,552	0,549	0,532	0,551	0,54	0,103	0,103
Decision tree Gini customer characteristics	0,538	0,522	0,808	0,541	0,634	0,082	0,097

Decision tree Gini hotel characteristics	0,537	0,516	1	0,571	0,681	0,144	0,206
Decision tree CER all sets	0,537	0,517	1	0,541	0,681	0,082	0,206
Decision tree CER hotel characteristics	0,537	0,516	1	0,541	0,681	0,082	0,206
Decision tree CER customer and hotel characteristics	0,537	0,516	1	0,541	0,681	0,082	0,206
Decision tree CER hotel and travel circumstances	0,537	0,516	1	0,541	0,681	0,082	0,206
Random forest travel circumstances characteristics	0,51	0,452	0,496	0,556	0,473	0,113	0,018
Decision tree Gini all sets	0,505	0,505	1	0,5	0,671	0	-
Decision tree CER customer characteristics	0,505	0,505	1	0,5	0,671	0	-
Decision tree CER travel circumstances characteristics	0,495	0,495	1	0,5	0,662	0	-
Decision tree Gini travel circumstances characteristics	0,495	0,495	1	0,5	0,662	0	-
Decision tree CER customer and travel circumstances	0,495	0,495	1	0,5	0,662	0	-
Decision tree Gini customer and travel circumstances	0,495	0,495	1	0,5	0,662	0	-
Decision tree Gini hotel and travel circumstances	0,495	0,495	1	0,5	0,662	0	-
SVM radial customer characteristics	0,493	0,488	0,707	0,498	0,578	-0,005	-0,005
SVM polynomial customer characteristics	0,493	0,488	0,707	0,498	0,578	-0,005	-0,005
Logistic regression travel circumstances characteristics	0,457	0,466	0,594	0,57	0,522	-0,086	-0,09
Logistic regression customer characteristics	0,456	0,422	0,238	0,556	0,304	-0,088	-0,098
Logistic regression customer and hotel characteristics interaction	0,452	0,447	0,405	0,548	0,425	-0,096	-0,096
Logistic regression customer and travel circumstances	0,445	0,442	0,44	0,584	0,442	-0,111	-0,111
Logistic regression customer and hotel characteristics	0,427	0,416	0,362	0,615	0,387	-0,146	-0,147
Logistic regression hotel characteristics	0,426	0,404	0,313	0,602	0,352	-0,149	-0,152
Logistic regression all sets	0,424	0,419	0,397	0,62	0,408	-0,152	-0,152
Logistic regression hotel and travel circumstances	0,423	0,406	0,337	0,613	0,368	-0,154	-0,157
Logistic regression customer and travel circumstances interaction	0,404	0,403	0,403	0,636	0,403	-0,193	-0,193
Logistic regression hotel and travel circumstances interaction	0,383	0,383	0,387	0,617	0,385	-0,234	-0,234
Logistic regression all sets interaction	0,348	0,335	0,298	0,651	0,312	-0,302	-0,305

Appendix E Results logistic regression

Table 1: Results logistic regression all variables

	Coefficient
Austria	16,42
Belgium	16,34
Denmark	16,87
France	17,01
Germany	16,64
Hungary	0,335
Ireland	1,064
Italy	16,97
Luxembourg	0,222
Poland	17,00
Portugal	16,42
Romania	0,269
Slovakia	0,292
Spain	16,40
The Netherlands	16,97
4 Stars	-0,223
5 Stars	-0,900***
Length_stay	-0,001
Spring	0,230*
Summer	0,068
Autumn	0,114
2019	16,52
2020	17,04
2021	16,90
2022	16,59
Business traveller	0,584**
Solo traveller	0,493*
Travelled with family	0,224
Travelled with family and small children	0,007
Travelled with partner	0,620**

Note: * = 0,05; ** = 0,01; *** = 0,001

Table 2: Results logistic regression customer characteristics

	Coefficient
Business traveller	0,684**
Solo traveller	0,475*
Travelled with family	0,148
Travelled with family and small children	-0,168
Travelled with partner	0,530*

Note: * = 0,05; ** = 0,01; *** = 0,001

Table 3: Results logistic regression travel circumstances characteristics

	Coefficient
Length_stay	-0,010
Spring	0,259*
Summer	0,037
Autumn	0,147
2019	14,479
2020	15,025
2021	14,853
2022	14,429

Note: * = 0,05; ** = 0,01; *** = 0,001

Table 4: Results logistic regression hotel characteristics

	Coefficient
Austria	15,27
Belgium	15,24
Denmark	15,56
France	15,88
Germany	15,61
Hungary	0,121
Ireland	0,686
Italy	15,89
Luxembourg	0,000
Poland	16,26
Portugal	15,32
Romania	0,000
Slovakia	0,000
Spain	15,35
The Netherlands	15,93
4 Stars	-0,191
5 Stars	-0,877***

Note: * = 0,05; ** = 0,01; *** = 0,001

Table 5: Results logistic customer and travel circumstances characteristics

	Coefficient
Length_stay	-0,011
Spring	0,261*
Summer	0,069
Autumn	0,155
2019	14,531
2020	15,017
2021	14,857
2022	14,460
Business traveller	0,594**
Solo traveller	0,377 .
Travelled with family	0,123
Travelled with family and small children	-0,187
Travelled with partner	0,551*

Note: . = 0,1; * = 0,05; ** = 0,01; *** = 0,001

Table 6: Results logistic customer and hotel characteristics

	Coefficient
Austria	15,253
Belgium	15,244
Denmark	15,628
France	15,949
Germany	15,599
Hungary	0,166
Ireland	0,960
Italy	15,870
Luxembourg	0,128
Poland	16,182
Portugal	15,318
Romania	0,195
Slovakia	0,137
Spain	15,318
The Netherlands	15,921
4 Stars	-0,180
5 Stars	-0,845***
Business traveller	0,630**
Solo traveller	0,522*
Travelled with family	0,233
Travelled with family and small children	0,044
Travelled with partner	0,629**

Note: * = 0,05; ** = 0,01; *** = 0,001

Table 7: Results logistic hotel and travel circumstances characteristics

	Coefficient
Austria	15,43
Belgium	15,33
Denmark	15,80
France	15,93
Germany	15,63
Hungary	0,281
Ireland	0,799
Italy	15,98
Luxembourg	0,112
Poland	16,03
Portugal	15,42
Romania	0,072
Slovakia	0,145
Spain	15,42
The Netherlands	15,97
4 Stars	-0,234
5 Stars	-0,920***
Length_stay	0,000
Spring	0,292*
Summer	0,041
Autumn	0,107
2019	15,46
2020	16,02
2021	15,87
2022	15,54

Note: * = 0,05; ** = 0,01; *** = 0,001

Table 8: Results logistic customer and travel circumstances characteristics with interaction

	Coefficient
Length_stay	-0,612**
Spring	-0,530
Summer	-0,209*
Autumn	14,992
2019	-14,868
2020	1,490
2021	0,893
2022	15,658
Business traveller	-1,704*
Solo traveller	-1,714*
Travelled with family	-2,757***
Travelled with family and small children	14,024
Travelled with partner	0,224
Length_stay * Business traveller	0,704**
Length_stay * Solo traveller	0,612**
Length_stay * Travelled with family	0,664**
Length_stay * Travelled with family and small children	0,245
Length_stay * Travelled with partner	0,337
Spring * Business traveller	0,555
Summer * Business traveller	2,358**
Autumn * Business traveller	-14,552
Spring * Solo traveller	0,645
Summer * Solo traveller	2,568**
Autumn * Solo traveller	-14,739
Spring * Travelled with family	1,489*
Summer * Travelled with family	2,577**
Autumn * Travelled with family	-14,853
Spring * Travelled with family and small children	-16,623
Summer * Travelled with family and small children	-12,091
Autumn * Travelled with family and small children	NA
Spring * Travelled with partner	-0,297
Summer * Travelled with partner	0,360
Autumn * Travelled with partner	-14,328
2019 * Business traveller	14,689
2020 * Business traveller	16,202
2021 * Business traveller	14,716
2022 * Business traveller	NA
2019 * Solo traveller	30,526
2020 * Solo traveller	14,340
2021 * Solo traveller	14,906
2022 * Solo traveller	NA
2019 * Travelled with family	32,177
2020 * Travelled with family	14,242
2021 * Travelled with family	15,782
2022 * Travelled with family	NA
2019 * Travelled with family and small children	NA
2020 * Travelled with family and small children	NA
2021 * Travelled with family and small children	NA
2022 * Travelled with family and small children	NA
2019 * Travelled with partner	13,991

2020 * Travelled with partner	NA
2021 * Travelled with partner	13,923
2022 * Travelled with partner	NA

Note: * = 0,05; ** = 0,01; *** = 0,001

Table 9: Results logistic customer and hotel characteristics with interactions

	Coefficient
Austria	9.048e+15 ***
Belgium	-1.797e+15 ***
Denmark	9.054e+14 ***
France	-1.797e+15 ***
Germany	5.528e+14 ***
Hungary	2.986e+15 ***
Ireland	7.489e+15 ***
Italy	6.415e+15 ***
Luxembourg	2.702e+15 ***
Poland	4.808e+15 ***
Portugal	2.707e+15 ***
Romania	2.707e+15 ***
Slovakia	5.853e+15 ***
Spain	-8.884e+13 ***
The Netherlands	5.852e+15 ***
4 Stars	4.675e+15 ***
5 Stars	-1.083e+14 ***
Business traveller	-3.370e+15 ***
Solo traveller	3.137e+15 ***
Travelled with family	1.416e+15 ***
Travelled with family and small children	-5.440e+15 ***
Travelled with partner	5.233e+15 ***
Austria * Business traveller	-5.168e+15 ***
Belgium * Business traveller	6.948e+15 ***
Denmark * Business traveller	NA
France * Business traveller	3.629e+15 ***
Germany * Business traveller	7.937e+15 ***
Hungary * Business traveller	-1.154e+15 ***
Ireland * Business traveller	NA
Italy * Business traveller	-1.351e+15 ***
Luxembourg * Business traveller	NA
Poland * Business traveller	1.528e+15 ***
Portugal * Business traveller	-2.830e+14 ***
Romania * Business traveller	NA
Slovakia * Business traveller	NA
Spain * Business traveller	6.367e+15 ***
The Netherlands * Business traveller	NA
Austria * Solo traveller	-3.276e+15 ***
Belgium * Solo traveller	4.391e+15 ***
Denmark * Solo traveller	9.887e+14 ***
France * Solo traveller	4.338e+15 ***
Germany * Solo traveller	2.623e+15 ***
Hungary * Solo traveller	-2.946e+15 ***
Ireland * Solo traveller	NA
Italy * Solo traveller	3.592e+14 ***
Luxembourg * Solo traveller	1.837e+15 ***
Poland * Solo traveller	NA
Portugal * Solo traveller	4.556e+15 ***
Romania * Solo traveller	-2.667e+15 ***
Slovakia * Solo traveller	-5.813e+15 ***

Spain * Solo traveller	4.977e+15 ***
The Netherlands * Solo traveller	-4.312e+14 ***
Austria * Travelled with family	-4.504e+15 ***
Belgium * Travelled with family	4.411e+15 ***
Denmark * Travelled with family	3.498e+14 ***
France * Travelled with family	7.842e+15 ***
Germany * Travelled with family	5.697e+15 ***
Hungary * Travelled with family	-2.310e+15 ***
Ireland * Travelled with family	-1.731e+15 ***
Italy * Travelled with family	-1.044e+15 ***
Luxembourg * Travelled with family	1.837e+15 ***
Poland * Travelled with family	NA
Portugal * Travelled with family	3.078e+15 ***
Romania * Travelled with family	-2.666e+15 ***
Slovakia * Travelled with family	-5.812e+15 ***
Spain * Travelled with family	6.238e+15 ***
The Netherlands * Travelled with family	-2.326e+15 ***
Austria * Travelled with family and small children	1.308e+15 ***
Belgium * Travelled with family and small children	7.649e+15 ***
Denmark * Travelled with family and small children	1.202e+16 ***
France * Travelled with family and small children	7.649e+15 ***
Germany * Travelled with family and small children	1.205e+16 ***
Hungary * Travelled with family and small children	NA
Ireland * Travelled with family and small children	NA
Italy * Travelled with family and small children	2.011e+15 ***
Luxembourg * Travelled with family and small children	NA
Poland * Travelled with family and small children	NA
Portugal * Travelled with family and small children	5.719e+15 ***
Romania * Travelled with family and small children	NA
Slovakia * Travelled with family and small children	NA
Spain * Travelled with family and small children	9.801e+15 ***
The Netherlands * Travelled with family and small children	NA
Austria * Travelled with partner	NA
Belgium * Travelled with partner	4.089e+15 ***
Denmark * Travelled with partner	NA
France * Travelled with partner	8.593e+15 ***
Germany * Travelled with partner	5.589e+15 ***
Hungary * Travelled with partner	6.336e+14 ***
Ireland * Travelled with partner	NA
Italy * Travelled with partner	-3.266e+15 ***
Luxembourg * Travelled with partner	1.837e+15 ***
Poland * Travelled with partner	NA
Portugal * Travelled with partner	2.113e+15 ***
Romania * Travelled with partner	NA
Slovakia * Travelled with partner	NA
Spain * Travelled with partner	5.656e+15 ***
The Netherlands * Travelled with partner	NA
4 Stars * Business traveller	-2.594e+14 ***
5 Stars * Business traveller	2.158e+15 ***
4 Stars * Solo traveller	-4.973e+15 ***
5 Stars * Solo traveller	-2.045e+15 ***
4 Stars * Travelled with family	-3.254e+15 ***
5 Stars * Travelled with family	3.145e+14 ***
4 Stars * Travelled with family and small children	-2.210e+15 ***
5 Stars * Travelled with family and small children	NA

4 Stars * Travelled with partner	-7.070e+15 ***
5 Stars * Travelled with partner	-3.152e+15 ***

Note: * = 0,05; ** = 0,01; *** = 0,001

Table 10: Results logistic travel circumstances and hotel characteristics with interactions

	Coefficient
Austria	6.530e+15 ***
Belgium	6.124e+15 ***
Denmark	1.171e+16 ***
France	8.636e+15 ***
Germany	5.317e+15 ***
Hungary	5.230e+15 ***
Ireland	3.586e+15 ***
Italy	6.582e+15 ***
Luxembourg	5.312e+15 ***
Poland	3.877e+15 ***
Portugal	6.983e+15 ***
Romania	-2.172e+16 ***
Slovakia	3.804e+15 ***
Spain	5.026e+15 ***
The Netherlands	6.323e+15 ***
4 Stars	-3.712e+15 ***
5 Stars	-8.285e+15 ***
Length_stay	-5.251e+14 ***
Spring	-3.497e+14 ***
Summer	-2.931e+15 ***
Autumn	-3.784e+15 ***
2019	-3.637e+15 ***
2020	-1.739e+12 ***
2021	2.082e+15 ***
2022	3.495e+15 ***
Austria * Length stay	2.375e+14 ***
Belgium * Length stay	-5.193e+14 ***
Denmark * Length stay	-1.165e+15 ***
France * Length stay	-3.611e+14 ***
Germany * Length stay	-1.856e+14 ***
Hungary * Length stay	-3.408e+14 ***
Ireland * Length stay	-8.907e+13 ***
Italy * Length stay	3.062e+13 ***
Luxembourg * Length stay	-1.574e+15 ***
Poland * Length stay	-2.726e+15 ***
Portugal * Length stay	-1.018e+14 ***
Romania * Length stay	5.304e+15 ***
Slovakia * Length stay	-7.720e+14 ***
Spain * Length stay	-1.540e+14 ***
The Netherlands * Length stay	-1.961e+14 ***
Austria * Spring	-8.957e+15 ***
Belgium * Spring	-1.866e+15 ***
Denmark * Spring	-6.991e+15 ***
France * Spring	-5.983e+15 ***
Germany * Spring	-1.581e+15 ***
Hungary * Spring	-5.288e+15 ***
Ireland * Spring	-4.460e+15 ***
Italy * Spring	-2.533e+15 ***
Luxembourg * Spring	-3.174e+15 ***
Poland * Spring	NA

Portugal * Spring	-4.470e+15 ***
Romania * Spring	NA
Slovakia * Spring	-3.396e+15 ***
Spain * Spring	-1.007e+15 ***
The Netherlands * Spring	-2.634e+15 ***
Austria * Summer	6.774e+14 ***
Belgium * Summer	7.366e+14 ***
Denmark * Summer	NA
France * Summer	-6.573e+15 ***
Germany * Summer	6.520e+14 ***
Hungary * Summer	NA
Ireland * Summer	NA
Italy * Summer	1.008e+15 ***
Luxembourg * Summer	NA
Poland * Summer	NA
Portugal * Summer	-1.856e+15 ***
Romania * Summer	2.054e+16 ***
Slovakia * Summer	NA
Spain * Summer	1.234e+15 ***
The Netherlands * Summer	NA
Austria * Autumn	1.606e+15 ***
Belgium * Autumn	-1.005e+16 ***
Denmark * Autumn	NA
France * Autumn	-6.796e+15 ***
Germany * Autumn	7.621e+14 ***
Hungary * Autumn	-1.626e+15 ***
Ireland* Autumn	NA
Italy * Autumn	5.334e+14 ***
Luxembourg* Autumn	NA
Poland * Autumn	7.187e+15 ***
Portugal * Autumn	3.229e+15 ***
Romania * Autumn	NA
Slovakia * Autumn	NA
Spain * Autumn	-4.499e+14 ***
The Netherlands * Autumn	NA
Austria * 2019	NA
Belgium * 2019	7.142e+15***
Denmark * 2019	NA
France * 2019	NA
Germany * 2019	2.690e+15***
Hungary * 2019	NA
Ireland * 2019	NA
Italy * 2019	-7.836e+13***
Luxembourg * 2019	NA
Poland * 2019	NA
Portugal * 2019	NA
Romania * 2019	NA
Slovakia * 2019	NA
Spain * 2019	7.341e+15***
The Netherlands * 2019	NA
Austria * 2020	NA
Belgium * 2020	1.538e+14***
Denmark * 2020	NA
France * 2020	7.992e+15***
Germany * 2020	6.103e+14 ***

Hungary * 2020	NA
Ireland * 2020	NA
Italy * 2020	-1.944e+15 ***
Luxembourg * 2020	NA
Poland * 2020	NA
Portugal * 2020	NA
Romania * 2020	NA
Slovakia * 2020	NA
Spain * 2020	1.845e+15 ***
The Netherlands * 2020	NA
Austria * 2021	-6.090e+15 ***
Belgium * 2021	5.009e+15 ***
Denmark * 2021	NA
France * 2021	6.727e+15 ***
Germany * 2021	-8.034e+14 ***
Hungary * 2021	-3.087e+15 ***
Ireland * 2021	NA
Italy * 2021	-1.088e+15 ***
Luxembourg * 2021	NA
Poland * 2021	NA
Portugal * 2021	-3.324e+15 ***
Romania * 2021	NA
Slovakia * 2021	NA
Spain * 2021	1.665e+15 ***
The Netherlands * 2021	NA
Austria * 2022	NA
Belgium * 2022	NA
Denmark * 2022	NA
France * 2022	NA
Germany * 2022	NA
Hungary * 2022	NA
Ireland * 2022	NA
Italy * 2022	-1.179e+15 ***
Luxembourg * 2022	NA
Poland * 2022	NA
Portugal * 2022	NA
Romania * 2022	NA
Slovakia * 2022	NA
Spain * 2022	NA
The Netherlands * 2022	NA
4 Stars * Length stay	6.899e+14 ***
5 Stars * Length stay	6.142e+14 ***
4 Stars * Spring	2.919e+15 ***
5 Stars * Spring	7.574e+15 ***
4 Stars * Summer	2.975e+15 ***
5 Stars * Summer	5.696e+15 ***
4 Stars * Autumn	4.909e+15 ***
5 Stars * Autumn	8.366e+15 ***
4 Stars * 2019	4.863e+15 ***
5 Stars * 2019	NA
4 Stars * 2020	3.120e+15 ***
5 Stars * 2020	8.196e+15 ***
4 Stars * 2021	1.721e+15 ***
5 Stars * 2021	1.150e+15 ***
4 Stars * 2022	NA

5 Stars * 2022	NA

Note: * = 0,05; ** = 0,01; *** = 0,001

Table 11: Results logistic all sets with interaction

	Coefficient
Austria	1.086e+16 ***
Belgium	3.488e+15 ***
Denmark	-2.525e+16***
France	2.968e+15 ***
Germany	4.217e+15 ***
Hungary	4.890e+15***
Ireland	2.625e+15 ***
Italy	9.275e+14 ***
Luxembourg	2.526e+15 ***
Poland	8.190e+15 ***
Portugal	-1.366e+15 ***
Romania	1.208e+15 ***
Slovakia	-4.081e+16 ***
Spain	2.578e+15 ***
The Netherlands	7.182e+15 ***
4 Stars	-6.641e+14 ***
5 Stars	-7.025e+15 ***
Length_stay	4.796e+15 ***
Spring	-1.854e+14 ***
Summer	1.285e+15 ***
Autumn	-1.279e+15 ***
2019	-7.516e+15 ***
2020	2.609e+15 ***
2021	-2.186e+15 ***
2022	1.970e+15 ***
Business traveller	-4.822e+15 ***
Solo traveller	-1.905e+15 ***
Travelled with family	-9.873e+14 ***
Travelled with family and small children	-4.208e+14 ***
Travelled with partner	-2.215e+15 ***
Austria * Length stay	-4.486e+15 ***
Belgium * Length stay	-5.324e+15 ***
Denmark * Length stay	-1.284e+15 ***
France * Length stay	-5.566e+15 ***
Germany * Length stay	-5.347e+15 ***
Hungary * Length stay	-4.746e+15 ***
Ireland * Length stay	-5.368e+15 ***
Italy * Length stay	-5.213e+15 ***
Luxembourg * Length stay	-2.870e+15 ***
Poland * Length stay	-8.844e+15 ***
Portugal * Length stay	-5.246e+15 ***
Romania * Length stay	-5.256e+15 ***
Slovakia * Length stay	1.619e+16 ***
Spain * Length stay	-5.271e+15 ***
The Netherlands * Length stay	-5.358e+15 ***
Austria * Spring	-6.848e+15 ***
Belgium * Spring	-1.141e+15 ***
Denmark * Spring	NA
France * Spring	-3.518e+15 ***
Germany * Spring	7.324e+13 ***

Hungary * Spring	-1.231e+14 ***
Ireland * Spring	NA
Italy * Spring	1.277e+15 ***
Luxembourg * Spring	3.347e+15 ***
Poland * Spring	NA
Portugal * Spring	8.904e+14 ***
Romania * Spring	NA
Slovakia * Spring	2.392e+16 ***
Spain * Spring	-1.355e+15 ***
The Netherlands * Spring	NA
Austria * Summer	-3.305e+15 ***
Belgium * Summer	-4.141e+14 ***
Denmark * Summer	NA
France * Summer	-1.143e+16 ***
Germany * Summer	-4.267e+14 ***
Hungary * Summer	NA
Ireland * Summer	NA
Italy * Summer	3.052e+14 ***
Luxembourg * Summer	-5.006e+15 ***
Poland * Summer	-4.423e+15 ***
Portugal * Summer	3.101e+14 ***
Romania * Summer	NA
Slovakia * Summer	NA
Spain * Summer	-1.202e+15 ***
The Netherlands * Summer	NA
Austria * Autumn	-2.400e+15 ***
Belgium * Autumn	-5.960e+15 ***
Denmark * Autumn	NA
France * Autumn	-1.793e+16 ***
Germany * Autumn	8.973e+14 ***
Hungary * Autumn	2.462e+14 ***
Ireland* Autumn	NA
Italy * Autumn	4.897e+14 ***
Luxembourg* Autumn	NA
Poland * Autumn	5.114e+15 ***
Portugal * Autumn	3.374e+15 ***
Romania * Autumn	NA
Slovakia * Autumn	NA
Spain * Autumn	-1.466e+15 ***
The Netherlands * Autumn	NA
Austria * 2019	NA
Belgium * 2019	3.022e+15 ***
Denmark * 2019	NA
France * 2019	NA
Germany * 2019	1.667e+15 ***
Hungary * 2019	NA
Ireland * 2019	NA
Italy * 2019	8.088e+15 ***
Luxembourg * 2019	NA
Poland * 2019	NA
Portugal * 2019	NA
Romania * 2019	NA
Slovakia * 2019	NA
Spain * 2019	1.034e+16 ***
The Netherlands * 2019	NA

Austria * 2020	NA
Belgium * 2020	-1.873e+15 ***
Denmark * 2020	NA
France * 2020	1.217e+16 ***
Germany * 2020	-2.866e+15 ***
Hungary * 2020	NA
Ireland * 2020	NA
Italy * 2020	2.958e+15 ***
Luxembourg * 2020	NA
Poland * 2020	NA
Portugal * 2020	NA
Romania * 2020	NA
Slovakia * 2020	NA
Spain * 2020	-1.629e+15 ***
The Netherlands * 2020	NA
Austria * 2021	-4.230e+15 ***
Belgium * 2021	-3.747e+14 ***
Denmark * 2021	NA
France * 2021	1.437e+16 ***
Germany * 2021	-1.806e+15 ***
Hungary * 2021	NA
Ireland * 2021	NA
Italy * 2021	2.671e+15 ***
Luxembourg * 2021	NA
Poland * 2021	NA
Portugal * 2021	-5.733e+15 ***
Romania * 2021	NA
Slovakia * 2021	NA
Spain * 2021	2.224e+14 ***
The Netherlands * 2021	NA
Austria * 2022	NA
Belgium * 2022	NA
Denmark * 2022	NA
France * 2022	NA
Germany * 2022	NA
Hungary * 2022	NA
Ireland * 2022	NA
Italy * 2022	3.581e+15 ***
Luxembourg * 2022	NA
Poland * 2022	NA
Portugal * 2022	NA
Romania * 2022	NA
Slovakia * 2022	NA
Spain * 2022	NA
The Netherlands * 2022	NA
4 Stars * Length stay	2.384e+14 ***
5 Stars * Length stay	3.428e+14 ***
4 Stars * Spring	1.037e+15 ***
5 Stars * Spring	2.516e+15 ***
4 Stars * Summer	3.313e+12 ***
5 Stars * Summer	1.351e+15 ***
4 Stars * Autumn	3.500e+15 ***
5 Stars * Autumn	4.329e+15 ***
4 Stars * 2019	6.708e+15 ***
5 Stars * 2019	NA

4 Stars * 2020	1.525e+15 ***
5 Stars * 2020	8.226e+15 ***
4 Stars * 2021	1.439e+15 ***
5 Stars * 2021	6.066e+14 ***
4 Stars * 2022	NA
5 Stars * 2022	NA
Austria * Business traveller	-3.867e+15 ***
Belgium * Business traveller	2.318e+15 ***
Denmark * Business traveller	NA
France * Business traveller	-4.684e+15 ***
Germany * Business traveller	2.800e+15 ***
Hungary * Business traveller	-5.169e+15 ***
Ireland * Business traveller	NA
Italy * Business traveller	8.598e+14 ***
Luxembourg * Business traveller	NA
Poland * Business traveller	NA
Portugal * Business traveller	4.021e+15 ***
Romania * Business traveller	NA
Slovakia * Business traveller	NA
Spain * Business traveller	4.289e+15 ***
The Netherlands * Business traveller	NA
Austria * Solo traveller	-4.078e+15 ***
Belgium * Solo traveller	3.957e+15 ***
Denmark * Solo traveller	NA
France * Solo traveller	7.462e+15 ***
Germany * Solo traveller	4.459e+15 ***
Hungary * Solo traveller	-3.297e+15 ***
Ireland * Solo traveller	NA
Italy * Solo traveller	2.405e+15 ***
Luxembourg * Solo traveller	NA
Poland * Solo traveller	NA
Portugal * Solo traveller	8.408e+15 ***
Romania * Solo traveller	-4.841e+13 ***
Slovakia * Solo traveller	NA
Spain * Solo traveller	3.498e+15 ***
The Netherlands * Solo traveller	NA
Austria * Travelled with family	-5.560e+15 ***
Belgium * Travelled with family	2.405e+15 ***
Denmark * Travelled with family	1.338e+16 ***
France * Travelled with family	5.115e+15 ***
Germany * Travelled with family	2.413e+15 ***
Hungary * Travelled with family	2.409e+14 ***
Ireland * Travelled with family	NA
Italy * Travelled with family	1.413e+15 ***
Luxembourg * Travelled with family	NA
Poland * Travelled with family	NA
Portugal * Travelled with family	5.731e+15 ***
Romania * Travelled with family	NA
Slovakia * Travelled with family	4.509e+15 ***
Spain * Travelled with family	4.255e+15 ***
The Netherlands * Travelled with family	NA
Austria * Travelled with family and small children	NA
Belgium * Travelled with family and small children	4.603e+15 ***
Denmark * Travelled with family and small children	3.803e+16 ***
France * Travelled with family and small children	6.295e+15 ***

Germany * Travelled with family and small children	-6.566e+14 ***
Hungary * Travelled with family and small children	NA
Ireland * Travelled with family and small children	NA
Italy * Travelled with family and small children	5.097e+15 ***
Luxembourg * Travelled with family and small children	NA
Poland * Travelled with family and small children	NA
Portugal * Travelled with family and small children	2.041e+16***
Romania * Travelled with family and small children	NA
Slovakia * Travelled with family and small children	NA
Spain * Travelled with family and small children	3.728e+15 ***
The Netherlands * Travelled with family and small children	NA
Austria * Travelled with partner	NA
Belgium * Travelled with partner	6.696e+15 ***
Denmark * Travelled with partner	NA
France * Travelled with partner	1.128e+16 ***
Germany * Travelled with partner	2.083e+15 ***
Hungary * Travelled with partner	NA
Ireland * Travelled with partner	NA
Italy * Travelled with partner	2.708e+15 ***
Luxembourg * Travelled with partner	NA
Poland * Travelled with partner	NA
Portugal * Travelled with partner	6.968e+15 ***
Romania * Travelled with partner	NA
Slovakia * Travelled with partner	NA
Spain * Travelled with partner	5.657e+15 ***
The Netherlands * Travelled with partner	NA
4 Stars * Business traveller	3.070e+15 ***
5 Stars * Business traveller	3.733e+15 ***
4 Stars * Solo traveller	-1.462e+15 ***
5 Stars * Solo traveller	3.720e+15 ***
4 Stars * Travelled with family	-2.088e+15 ***
5 Stars * Travelled with family	1.446e+15 ***
4 Stars * Travelled with family and small children	7.673e+15 ***
5 Stars * Travelled with family and small children	NA
4 Stars * Travelled with partner	-5.150e+14 ***
5 Stars * Travelled with partner	3.426e+15 ***
Length_stay * Business traveller	1.674e+14 ***
Length_stay * Solo traveller	2.214e+14 ***
Length_stay * Travelled with family	2.293e+14 ***
Length_stay * Travelled with family and small children	-1.179e+15 ***
Length_stay * Travelled with partner	-1.723e+14 ***
Spring * Business traveller	-5.127e+14 ***
Summer * Business traveller	-7.705e+14 ***
Autumn * Business traveller	-3.270e+15 ***
Spring * Solo traveller	-6.892e+13 ***
Summer * Solo traveller	-1.128e+15 ***
Autumn * Solo traveller	-1.090e+15 ***
Spring * Travelled with family	6.109e+14 ***
Summer * Travelled with family	-5.873e+14 ***
Autumn * Travelled with family	-1.388e+15 ***
Spring * Travelled with family and small children	-1.120e+16 ***
Summer * Travelled with family and small children	-5.321e+15 ***
Autumn * Travelled with family and small children	NA
Spring * Travelled with partner	-2.872e+14 ***
Summer * Travelled with partner	-9.743e+14 ***

Autumn * Travelled with partner	NA
2019 * Business traveller	5.538e+14 ***
2020 * Business traveller	1.948e+14 ***
2021 * Business traveller	3.116e+15 ***
2022 * Business traveller	NA
2019 * Solo traveller	-1.255e+15 ***
2020 * Solo traveller	-8.642e+14 ***
2021 * Solo traveller	4.428e+15 ***
2022 * Solo traveller	NA
2019 * Travelled with family	9.270e+14 ***
2020 * Travelled with family	NA
2021 * Travelled with family	4.044e+15 ***
2022 * Travelled with family	NA
2019 * Travelled with family and small children	NA
2020 * Travelled with family and small children	NA
2021 * Travelled with family and small children	NA
2022 * Travelled with family and small children	NA
2019 * Travelled with partner	-1.075e+15 ***
2020 * Travelled with partner	NA
2021 * Travelled with partner	2.661e+15 ***
2022 * Travelled with partner	NA

Note: * = 0,05; ** = 0,01; *** = 0,001