Customer Analytics for the Hospitality Industry

Neochoritis Vasileios

SID: 3305170011

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in e-Business and Digital Marketing

DECEMBER 2019

THESSALONIKI – GREECE

Customer Analytics for the Hospitality Industry

Neochoritis Vasileios

SID: 3305170011

Supervisor: Dr Nikolaos S. Thomaidis

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in e-Business and Digital Marketing

DECEMBER 2019

THESSALONIKI – GREECE

Abstract

This dissertation was written as a part of the MSc in e-Business & Digital Marketing at the International Hellenic University. Travel and tourism industry has changed rapidly over the years as customers are using digital channels to make their bookings online. Hotel managers are making efforts to reach possible customers to this new digital market, where the competition is fiercer than ever. One of the most important and effective strategies is the online reputation management. Online reputation consists of the electronic word of mouth created by online reviews and ratings provided by customers.

The first part of this paper examines the impact online reputation has on the decisions of the consumers, as well as the importance of online review and ratings on the performance of a hotel. By investigating previous research, it will explore the ways hotel managers can use the online generated content to improve their hotel's performance and the effective ways they can should interact with it to improve their brand's online presence. On the second part we will perform deep analysis of a dataset containing reviews from Booking.com customers, to discover new patterns and insights. Finally, we will build machine learning explanatory models to determine the most important factors affecting the scores provided by reviewers.

Neochoritis Vasileios

Date 02/12/2019

Contents

1 Introduction

Travel and tourism are already one of the biggest industries in the world. This already huge market is going to keep growing in the future, as the Internet, combined with technologies like machine learning and artificial intelligence have created unlimited potential for growth. Online bookings are now the main option for booking hotel accommodation, flights and even the activities of preference before arriving at the destination. While traditional travel agents and tour operators still exist, most of them have also shifted their focus on building their online presence and conducting business through the internet. The biggest OTAs (booking.com, Expedia etc.) are also offering another option and have gained a significant amount of the Travel and Tourism market share. The final option, for a traveler to book online, is the official hotel websites. Big hotel chains have the brand name and resources to receive an important share of their total bookings through their website, but for smaller or less known hotels this is much more difficult.

This new environment in the hospitality industry has created an abundance of options to consumers, for the same product (i.e. a specific hotel room). Travelers needs to scan and process a huge amount of information before deciding to book, since they not only have to decide on the destination and hotel, but also the channel they will use to book. On the other hand, hotel managers must be able to observe their online presence to all these channels and make sure that it aligns with their brand image. Aside from the information and content hoteliers provide (professional photos, descriptions, amenities information etc.), user generated content plays an integral part on their brand's image. This content comes in many shapes, but the most common are traveler photos, reviews and ratings.

The purpose of this paper is to explore the importance of the online reviews and ratings for the hotels, and which are the factors that play the most prominent role on defining the latter. This process will be conducted in 2 parts. The first part is reviewing the literature around the subject of customer analytics related to the hospitality industry, while the second part is the analysis of a dataset extracted from one of the most prominent

OTAs, Booking.com. The analysis will be also broken down to 2 parts. The first part is the exploratory data analysis and feature engineering, which means that we will transform and summarize the data, and then present them in a visual manner. Additionally, we will explore if there are more ways to transform the existing features, in order to create new ones to use in the next part. The goal of the first part is to explore possible insights and patterns that can be extracted by the dataset. On the second part, we will build two types of machine learning models. The first model is a classification algorithm that predicts if an individual score will be higher than the dataset's average. The second model is a regression algorithm that makes a prediction on the exact score a reviewer will provide. While the predictions are the main functionality of the models, our focus is to explain the predictions and the factors that affect them.

2 Literature review

On this chapter we will explore previous research on the importance of analytics on hotel management. Specifically, we will explore what are the effects of different types of reviews and ratings on the customers' decision to book. Furthermore, we will explore the connection of the eWOM (electronic word of mouth), which is created by the online reviews on hotel websites and third-party websites, with the successful performance of a hotel.

2.1 Online Reviews' effect on travelers' decisions

As the Hospitality industry is leaning on the digital presence and focusing its efforts towards reaching possible customers on the internet, people are bombarded with unprecedented amounts of information. Hotels and third-party sellers are spending billions on digital advertising, in order to reach their customers on every stage of the buying process. This makes it more difficult for possible customers to process all this information and make the final decision to book, since the customers are on a simpler state of mind when they are about to make a purchase, trying to make the decision that best fits their goals and maximize their benefits [1]. eWOM, in the shape of online reviews and ratings, helps simplify this process.

Online reviews are based on the concept of social proof [2]. According to this concept, consumers can significantly affect other consumers by sharing their opinions with them, since people have the tendency to follow the lead of the crowd when they are undecided [3]. Since the internet provides a connection between its users, the effects of social proof can be amplified and reach a consumer on every stage of the online booking process, starting from the awareness until the final decision to book is made [4]. While hospitality companies are spending huge amounts of their budgets on online advertising, Nielsen reports that online reviews and ratings are the most important source of information on a brand quality, apart from the suggestions of friends or family [5]. Moreover, people usually find these sources trustworthy, even though they do not know the actual reviewers. The use of the rating is particularly important on the first stages of the

booking process, since it can help customers filter the possible destinations and accommodations based on a simple criterion [6]. One example of this would be to filter all the hotels, on an OTA website, that have a rating lower than 4 out of 5 stars and then proceed with a closer examination of the presented hotels. It is clear that ratings can help customers save time and make the process more efficient, at least at the first stages of searching for a hotel.

While online reviews and ratings can help customers on their decisions, there are some differences on how positive and negative reviews are perceived. This case was discussed on a research by Papathanassis and Knolle [7]. Customers generally trust negative, or reviews containing both positive and negative elements, much more than only positive reviews. They spend more time reading and interacting with negative or mixed reviews, compared to the positive reviews, and negative (and mixed) reviews usually work as an initial filter to disregard a set of options. Furthermore, the magnitude of the effect varies between the two. Chevalier and Maizline showed that while a higher rating can lead to more online sales, the effect of the low ratings was greater and led to a bigger drop in sales [8].

This difference can be explained by the fact that customers tend to underestimate the possibility of a fake and malicious negative review, while they can label as such a positive review much easier [9]. There is also a connection between the number of reviews and their perceived trustworthiness, although the importance of this is different between good and bad rating reviews. A high number of positive reviews is required to make customers trust them, while even a really low number of negative reviews is enough to achieve high levels of perceived trustworthiness.

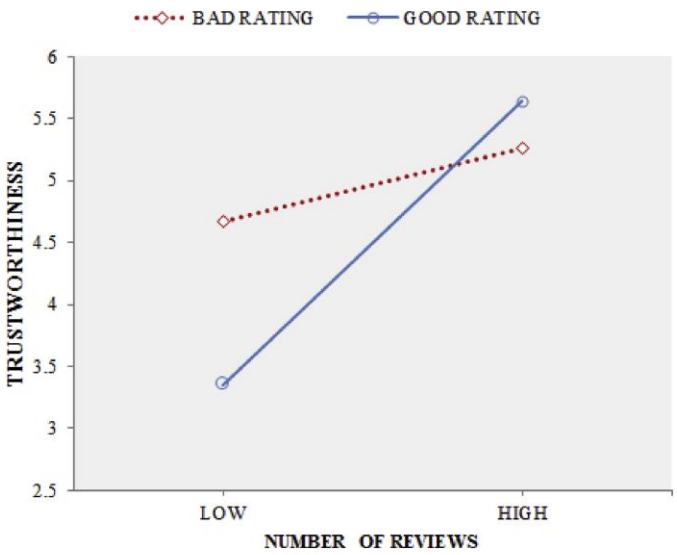

Figure 1: Relationship between number of reviews and trustworthiness

In figure 1 we can see the results of the experiment from Gavilan (2018) [1]. We can see that higher number of reviews has much lower marginal effect on the perceived trustworthiness of negative reviews, while the effect is much higher for the positive reviews.

2.2 Usage of online reviews by hotel managers

We have established that eWOM plays an integral part on the decision of customers that book online, but it is equally important to hotel managers. Since online reviews are, on their majority, honest depictions of the hotel services as experienced by its customers, hotel managers can use this valuable information to improve their services [11]. Using data analytics, hotel managers can even personalize their hotel's services to achieve even higher levels of satisfaction [12]. The importance of this can be better understood by considering the fact that existing reviews can affect the ratings of future ones, creating a bias to reviewers that already know the previous rating of a hotel [13]. Social proof is the main reason for this, as new reviewers will be biased to agree with the existing reviews and ratings, at least on some level. This results in already high rated hotels receiving higher ratings from new customers and retaining their good brand reputation, while it is much more difficult for low rated hotels to improve their rating (and brand image along with it).

While the perceived importance of analyzing and using this type of data is very high among hotel managers, it is not used as widely as one would except. The reason for this is that the complexity and sheer volume of data, that makes analyzing and using the data

to formulate the hotel's strategy very difficult [14]. Considering that user generated content is produced at a much higher pace than it can be analyzed, its actual use by hoteliers remains limited [15]. Advanced technologies like machine learning and artificial intelligence can now fill this gap and can be used to extract important insights from huge amounts of data from various sources. The use of such technologies allows hotels that have adopted it to essentially create tailor made services and experiences according to their customers' needs and preferences [16].

Data are included on the most important resources for hotels, so it is of utmost importance for hotel managers to make sure that they maximize their usage. Advanced data analysis techniques are required to be able to gain even more meaningful insights and support the strategic decisions of managers. To achieve this, hotel managers need to build their teams with people that can understand data and present them on a clear and easily understood format to decision makers.

2.3 eWOM and hotel performance

Online reputation and eWOM not only offers feedback and insights to hotel managers, but they are also connected with the success and profitability of a hotel. In this digital era there are many sources that gather customer feedback and opinions, and then freely circulate this information to all internet users [17]. These sources can be the official website of a hotel, which shares the user generated content, but in most cases, they are third party platforms like TripAdvisor, Booking.com and even Facebook. Data are the most important resource for these companies, as through the monetization of data they make their biggest share of revenue; either by offering a huge audience to advertisers or by getting bookings directly from customers that consume this content. This fact creates a user generated content economy on the hospitality industry where platforms aim to not only receive online bookings, but also receive the customers feedback in terms of reviews, ratings and even photos [18].

The huge amount of data generated by all these sources, and specifically the eWOM created from the customers' input, can work as predictor for various metrics measuring the effectiveness and success of a hotel's operations. Kim and Park (2017) showed that online ratings from different websites can be used to predict occupancy and average daily rates with higher accuracy than internal customer satisfaction data, like surveys and questionnaires [19]. Online ratings and reviews have a significant correlation with a ho-

tel's revenue and occupancy, two of the most important metrics to measure hotel performance and profitability. Furthermore, eWOM plays an important role on customer loyalty and helps hotels create a repeating customer base [20]. This is an important benefit, since repeating customers tend to be the most profitable ones, as hotels need to spend less money to acquire them. Additionally, loyal customers can be the best representatives of a brand by sharing their experiences on social media or travel websites.

Although hotel managers are focusing more on the impact eWOM can have on their hotel's performance, it is not always possible to combine data from social media and travel websites (i.e. TripAdvisor) with the internal data measuring a hotel's successful operation. Meanwhile, research has shown that positive polarity on online reviews has a positive relationship with a hotel's performance [21]. Since the importance of these sharing platforms on the decision of a destination has only grown in recent years, customers consider the eWOM created as a representation of the quality of service provided by a hotel [22]. This association's significance is amplified by the fact that the perceived quality of services strongly affects the price a possible customer is willing to pay for a service [23]. That means hotels with bad online reputation, in terms of low ratings and negative reviews, find it more difficult to offer the highest possible price, since their online reputation is closely connected with their perceived quality of service. As a result, hotels falling in this category have lower potential to generate revenue from bookings compared to high rated hotels [24].

Hotel managers need to specifically manage their brand's online reputation by studying and replying to their customers' feedback. Studies has shown that those hotels that provide replies through their official representatives, tend to receive up to 60% more bookings, compared to hotels that do not follow this practice. As already discussed, the perceived quality of service is very important for possible customers and the managerial responses to reviews is a clear representation of the level of this quality. While managing a hotel's website to improve online reputation is very important, past research has dictated that customers value reviews and ratings on third-party websites even more [26]. This difference in value comes from the fact that third-party website reviews and ratings are perceived more trustworthy, compared to a hotel's website. It should be noted that the actual review can play a more significant part on the customers evaluation of a service, compared to just a numerical rating [27]. Consumers tend to make decisions based on their emotions, and not only on common sense, and previous customers' de-

scription of their experience can have higher impact on the decision making, compared to the ratings alone.

3 Methods and Tools

On this chapter we will explain the methods and tools that will be used on the next steps of the thesis. Those are the Exploratory Data Analysis and the Machine Learning Explanatory models. Additionally, the dataset, programming language and algorithms will be described.

3.1 Dataset description

The data included on the dataset used for the purposes of this dissertation are publicly available and owned by Booking.com; the data were scraped from the official website of Booking.com. This platform was selected because they only allow customers that booked a vacation through them to leave a review, which ensures that the reviews are from actual customers. The dataset includes 515,000 reviews from Booking.com customers and 1,493 reviewed hotels. These hotels are located on the following six European cities: Amsterdam, London, Paris, Barcelona, Milan and Vienna.

Hotel_Address	The address of the hotel	Text
Review_Date	The date the review was submitted	Text
Average_Score	The average score of the hotel	Numerical
Hotel_Name	Name of the hotel	Text
Reviewer_Nationality	The nationality of the reviewer	Text
Negative_Review	The negative review written by a reviewer for the Hotel. If there is no negative review the value is “No Negative”	Text
Review_Total_Negative_Word_Counts	Number of words for the Negative Review	Text
Positive_Review	The positive review written by a reviewer for the Hotel. If there is no positive review the value is “No Positive”	Numerical
Review_Total_Negative_Word_Counts	Number of words for the Negative Review	Numerical
Reviewer_Score	The total score the reviewer attributed to the hotel based on their stay	Numerical
Total_Number_of_Reviews_Reviewer_Has_Given	Total number of reviews the specific user gave on the booking.com system	Numerical
Total_Number_of_Reviews	Total number of reviews for the hotel	Numerical
Tags	Tags assigned to the hotel from the reviewer	Text
days_since_review	Days passed from the time of the review to the time of the dataset creation	Text
Additional_Number_of_Scoring	This column shows how many scores a hotel has received on specific services instead of total score for the Hotel	Numerical
lat	Latitude of the hotel	Numerical
lng	Longitude of the hotel	Numerical

Figure 2: List of the dataset’s columns and description

Figure 2 shows the 17 columns of data included on the dataset, accompanied by their description and the type of data those represent. There are both numerical values like

the ratings, as well as textual ones, like the written reviews provided by users. There is also information on the hotels; specifically, the address, hotel name, average score, total number of reviews, latitude and longitude of the hotel. Moreover, there are columns containing information about the reviewers, their nationality, how many reviews they have provided in the past and tags they have assigned to their trip. Finally, the rest of the columns provide information about the reviews. Those columns are the date of the review, the score accompanying the review, the actual texts of positive and negative reviews, along with their word counts.

3.2 Programming language and environment

The next steps, those of the data analysis and machine learning explanatory models, are accomplished using the Python programming language. Python was released on 1991 by its creator Guido van Rossum [28]. It is now one of the most popular programming languages used for data science projects. Python's syntax is relatively close to human language, so it is easier to learn and easier to understand when reading the code, a fact that led to its widespread usage. But the main benefit is that there are countless open source libraries, created by many collaborators, that help users perform tasks that would require significant effort otherwise. Libraries are collections of code with ready to use commands. Two of the most popular libraries, used for data analysis and machine learning respectively, are Pandas [29] and scikit-learn [30]. Those two will be used for the purposes of this dissertation.

The programming environment used is Google's platform for data science, Kaggle [31]. It is free to use and provides users with computing power and the latest Python libraries used for data science projects, in the form of Jupyter notebooks. Jupyter notebooks allow users to combine code, comments and visualizations in a common environment [32].

3.3 Algorithms

We will use two different types of the XGB algorithm, the classifier and the regressor, to create two different models. The XGB algorithm is a machine learning algorithm, based on decision trees, but its difference is that it is an ensemble method. This means that it uses multiple algorithms to combine their predictions and make a final prediction that is more accurate than each of the individual ones [33].

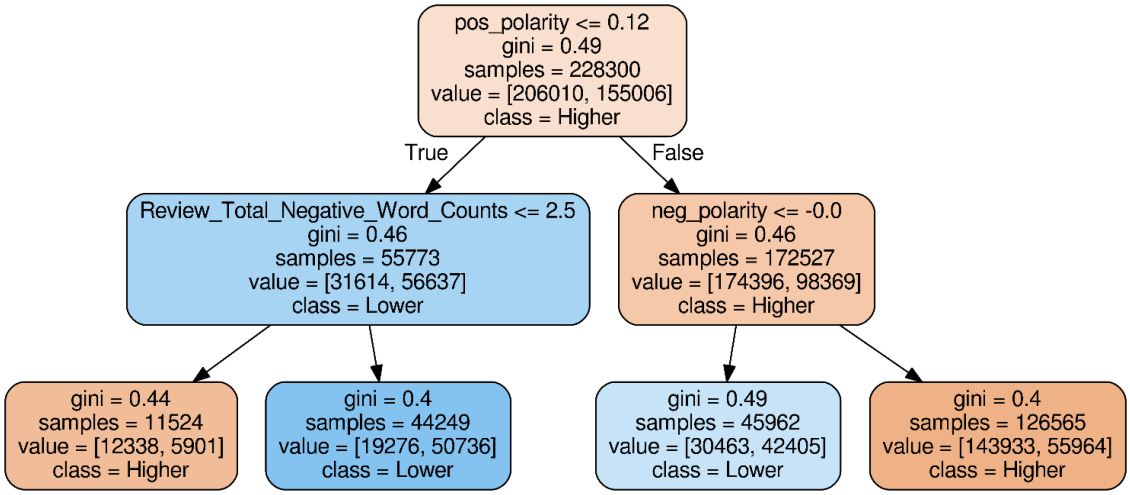

Figure 3: An example of a decision tree

Figure 3 is the visualization of a single decision tree, and shows the process followed until it makes its prediction. For classification purposes, a decision tree makes decisions based on the values of different features until it makes a prediction on whether the variable belongs to a class or not. The difference between a classifier and regressor is that the latter's final prediction can be a real number (i.e. the rating provided by a reviewer). Typically, decision trees like this one do not make very accurate predictions since they are weak learners, meaning that these have only a slight correlation with the predicted outcome [34]. Their prediction on a classification problem, for example, would be only be slightly better at predicting the correct class than random chance would. Boosting is used to transform this weak learner to a strong one, meaning that it improves its accuracy significantly [35]. The increased accuracy is exactly the reason we chose to use the XGB algorithm to create the machine learning models presented on this dissertation.

4 Data Analysis

This chapter contains all the data analysis part of the dissertation. It is separated in two main parts; the exploratory data analysis and the feature engineering part. The goal of the first part is to summarize and present the data using visualizations, in order to extract insights. The second part aims to further explore the existing features, in order to create new ones to be used on the machine learning models.

4.1 Exploratory Data Analysis

We have the description of each variable along with its type. They are divided into two types, text and numerical. We will need to verify that the description is correct so we will use a few commands in Python to examine the dataset. We can also see that some features should be a different type so will make the transformations. The Review Date column should be date type and the days_since_review should be numerical.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 515738 entries, 0 to 515737
Data columns (total 17 columns):
Hotel_Address                               515738 non-null object
Additional_Number_of_Scoring                515738 non-null int64
Review_Date                                 515738 non-null datetime64[ns]
Average_Score                               515738 non-null float64
Hotel_Name                                  515738 non-null object
Reviewer_Nationality                        515738 non-null object
Negative_Review                             515738 non-null object
Review_Total_Negative_Word_Counts           515738 non-null int64
Total_Number_of_Reviews                     515738 non-null int64
Positive_Review                             515738 non-null object
Review_Total_Positive_Word_Counts           515738 non-null int64
Total_Number_of_Reviews_Reviewer_Has_Given  515738 non-null int64
Reviewer_Score                              515738 non-null float64
Tags                                        515738 non-null object
days_since_review                           515738 non-null int64
lat                                         512470 non-null float64
lng                                         512470 non-null float64
dtypes: datetime64[ns](1), float64(4), int64(6), object(6)
memory usage: 66.9+ MB
```

Figure 4: Presentation of all the features

In figure 4 we can see the number of entries on each column, as well as the data type contained on them. All columns are now the proper type based on their data. From this table we can also check if there are any missing values on some of the columns. The only columns we have missing values on , are the latitude and longitude, the coordinates of some hotels. For now, this isn't an issue so we will not fill those two columns.

We will also need to have a look at the values each column has.

:

	Additional_Number_of_Scoring	Average_Score	Review_Total_Negative_Word_Counts	Total_Number_of_Reviews
count	515738.000000	515738.000000	515738.000000	515738.000000
mean	498.081836	8.397487	18.539450	2743.743944
std	500.538467	0.548048	29.690831	2317.464868
min	1.000000	5.200000	0.000000	43.000000
25%	169.000000	8.100000	2.000000	1161.000000
50%	341.000000	8.400000	9.000000	2134.000000
75%	660.000000	8.800000	23.000000	3613.000000
max	2682.000000	9.800000	408.000000	16670.000000

Figure 5: Statistical description of numerical features / first part

Review_Total_Positive_Word_Counts	Total_Number_of_Reviews_Reviewer_Has_Given	Reviewer_Score	lat	lng
515738.000000	515738.000000	515738.000000	512470.000000	512470.000000
17.776458	7.166001	8.395077	49.442439	2.823803
21.804185	11.040228	1.637856	3.466325	4.579425
0.000000	1.000000	2.500000	41.328376	-0.369758
5.000000	1.000000	7.500000	48.214662	-0.143372
11.000000	3.000000	8.800000	51.499981	0.010607
22.000000	8.000000	9.600000	51.516288	4.834443
395.000000	355.000000	10.000000	52.400181	16.429233

Figure 6: Statistical description of numerical features / second part

Figures 5 and 6 provide a holistic view on the data included in each column with some basic statistical description. We already had the count from the previous table, but we can also see the mean, standard deviation, minimum and maximum value of every column. Additionally, we are presented with the percentiles 25%, 50% and 75%.

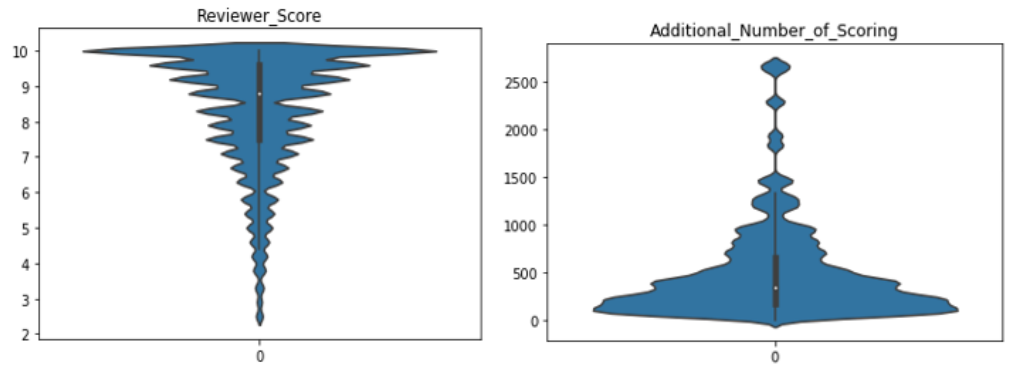

Figure 7: Violin plots, reviewer score and additional number of scoring

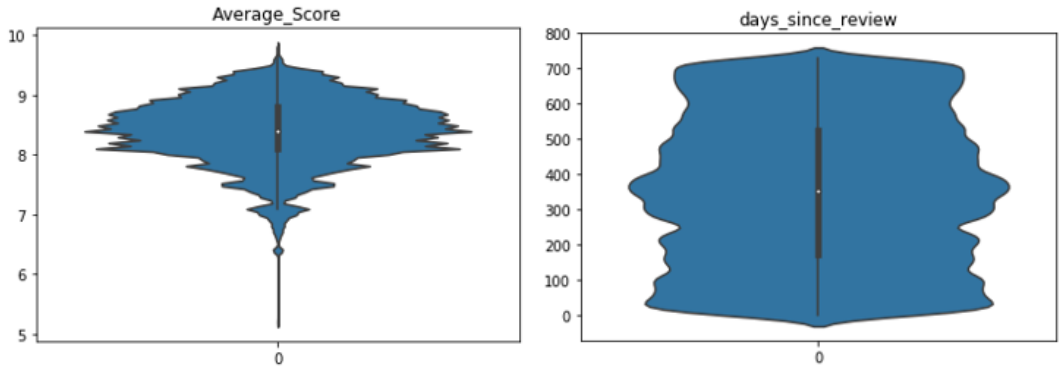

Figure 8: Violin plots, average score and additional number of scoring

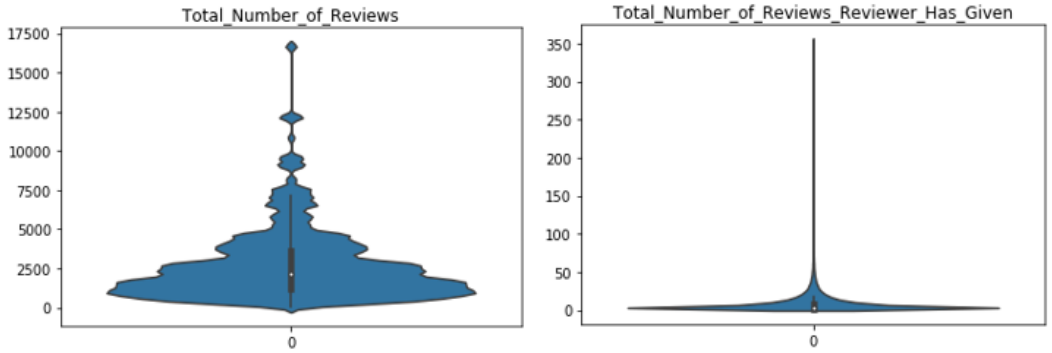

Figure 9: Violin plots, Total number of reviews per hotel (left plot) and per reviewer (right plot)

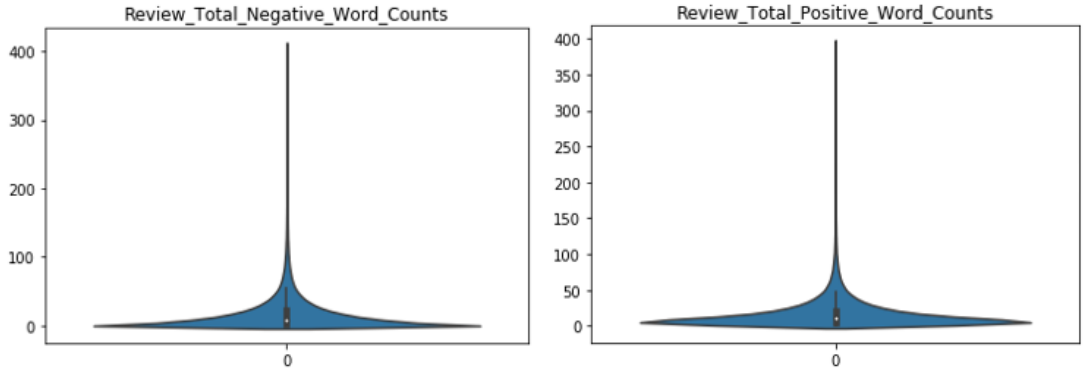

Figure 10: Violin Plots, word counts on negative (left plot) and positive reviews (right plot)

In figures 7,8,9 and 10 we can see the distribution of each numerical feature, in the for of violin plots. The small white dot is the median of the column, the bolder part of the line is the interquartile range (IQR) which includes all the observations between the 25th quartile(Q1) and the 75th (Q3) meaning. The thin line extending further than the bold one, covers the rest of the distribution and collectively includes 50% of the observations [36]. All observations further than the thin line are considered outliers, meaning that they are extremely rare. The outliers are the observations that are either smaller than the number calculated by subtracting the interquartile range, multiplied by 1.5, from the 25th

quartile (Q1 - 1,5xIQ) or bigger than adding the interquartile range, multiplied by 1.5, from the 75th quartile (Q3 + 1,5xIQR) . Essentially, outliers are observations that are far away from the rest of the observations [37].

The violin plots allow us to extract information on the features of the dataset. The positive and negative reviews word counts have similar distribution, something that we can also confirm from the descriptive statistics table. Both features are heavily concentrated between 0 and 50, which means that most reviews have less than 50 words.

The total number of reviews and additional number of scoring columns have also similar distributions. There is an obvious correlation between these two features, since the higher the number of reviews is, the higher is the possibility for additional scorings to be provided. Most reviewers have provided a relatively low number of reviews as the average is 7.16 reviews per user. The interquartile range for the Reviewer Score is between 7.5 and 9.6, which means that half of the scores the reviewers provided are between those two ratings. The highest scored review is 10 (which is the max) and the lowest observation is 2.5. Most of the hotels have an average score between 7.5 and 9, with the lowest average being 5.2 and the highest being 9.8.

The next step in the analysis is to check for correlations between the numerical features.

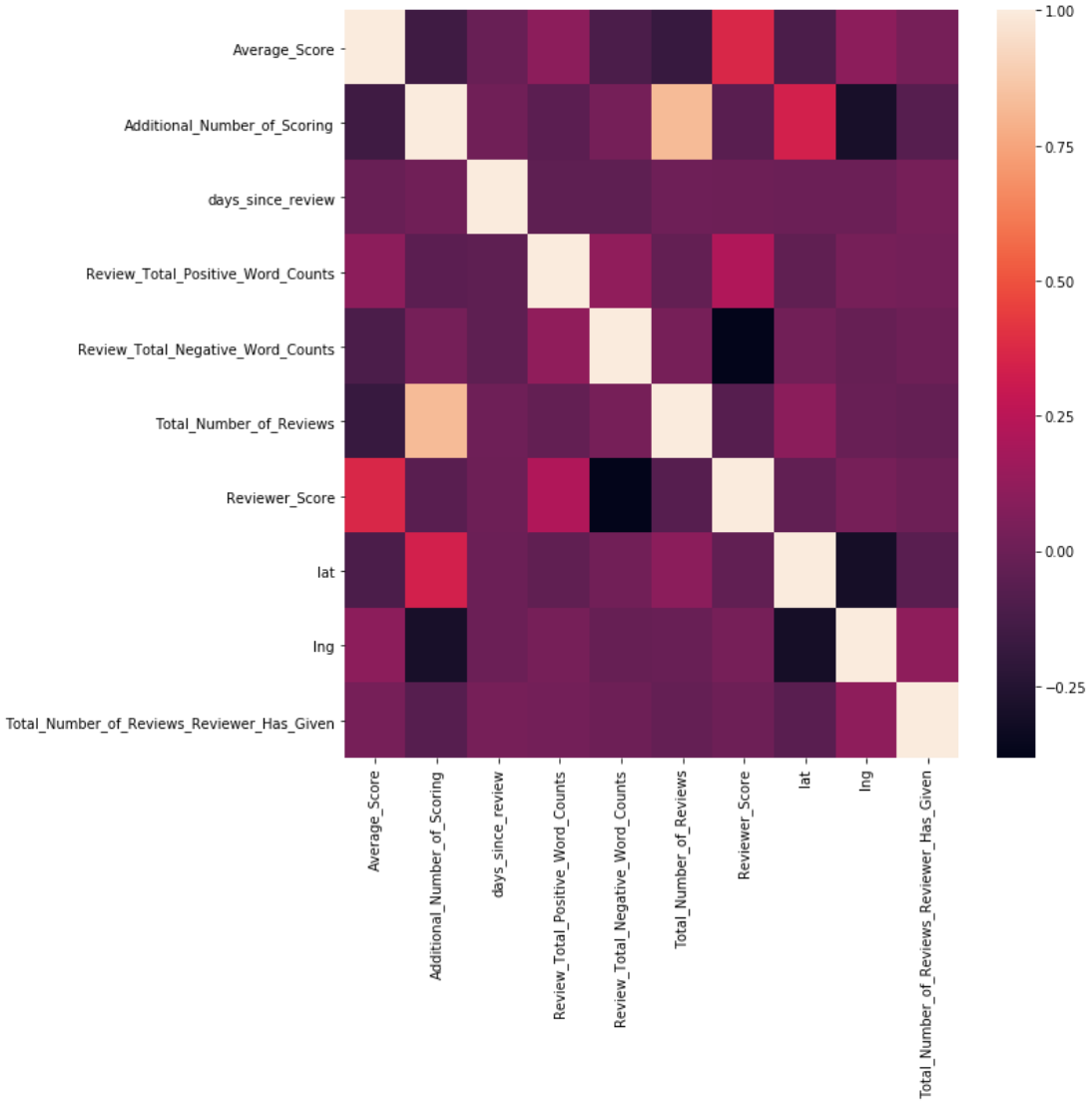

Figure 11: Correlation heatmap of features

The lighter the color block between two columns the stronger the correlation between them. Some pairs that appear highly correlated are the following: Reviewer_Score and Average_Score, Total_Number_of_Reviews and Additional_Number_of_Scoring, lat and Additional_Number_of_Scoring and Reviewer_Score and Review_Total_Positive_Word_Counts. By creating a scatterplot matrix for the features, we could visualize any linear correlation.

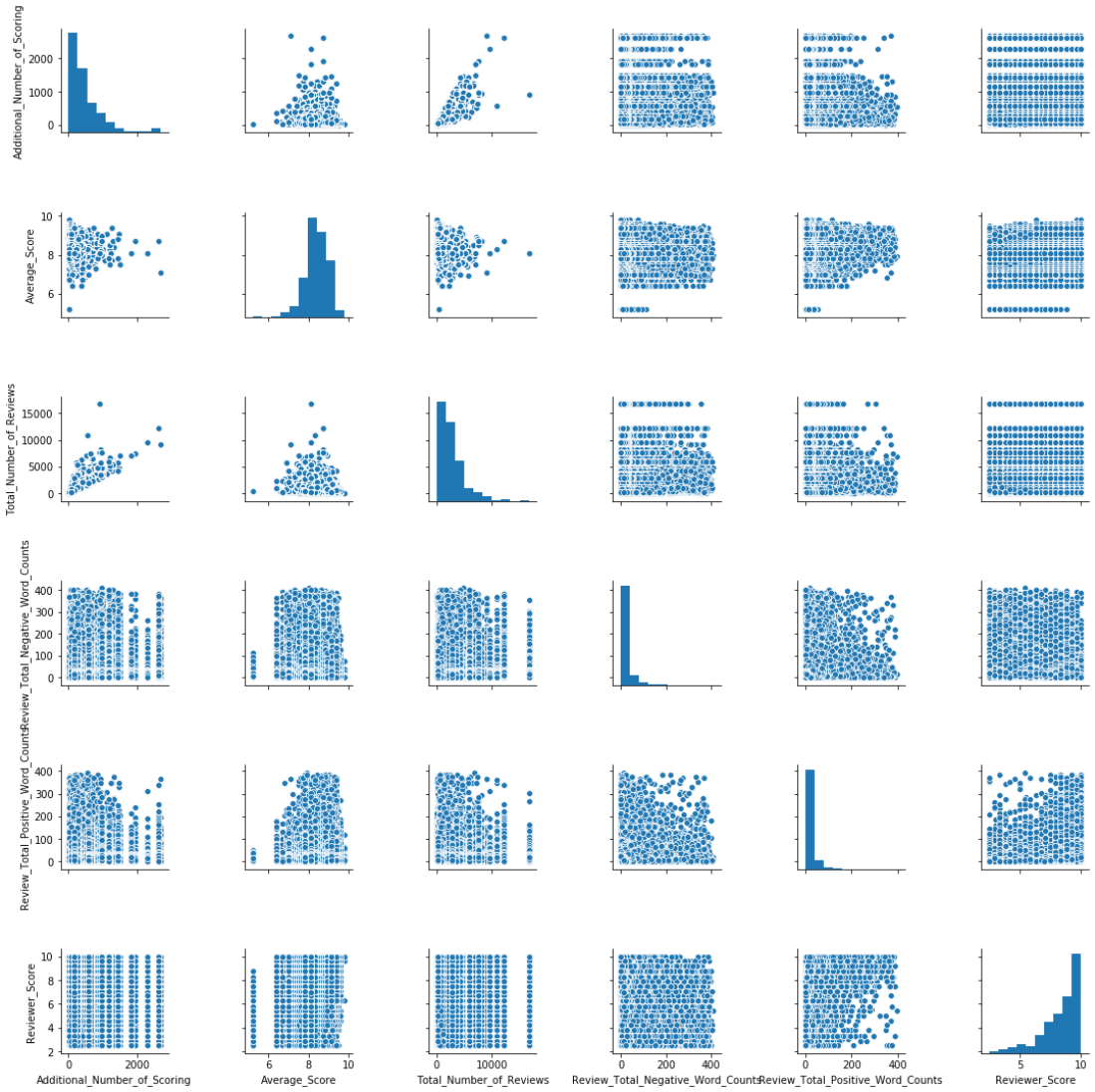

Figure 12: Scatterplot matrix

Figure 12 shows us that on one case there is a linear correlation, between the Additional_Number_of_Scoring and the Total_Number_of_Reviews. As already mentioned, this correlation is natural since there are increased possibilities for more additional scorings when there are more reviews in general. We have explored the distributions of the numerical features, along with their possible correlations. The next step is to explore the categorical features.

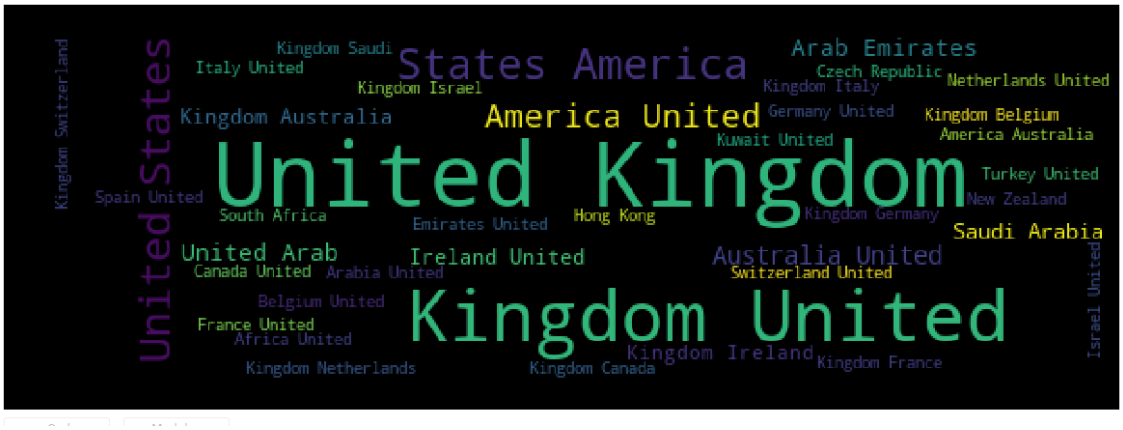

Figure 13: Reviewer nationality word cloud

Figure 13 is a word cloud that was created based on how many times each country appears on the column containing the nationalities of the reviewers. The word cloud makes it possible to find out which are the most common nationalities; most reviewers are from the United Kingdom, followed by the likes of USA, Ireland, Australia and United Arab Emirates. The word cloud does not provide numerical values, so we will create some plots to get more precise answers.

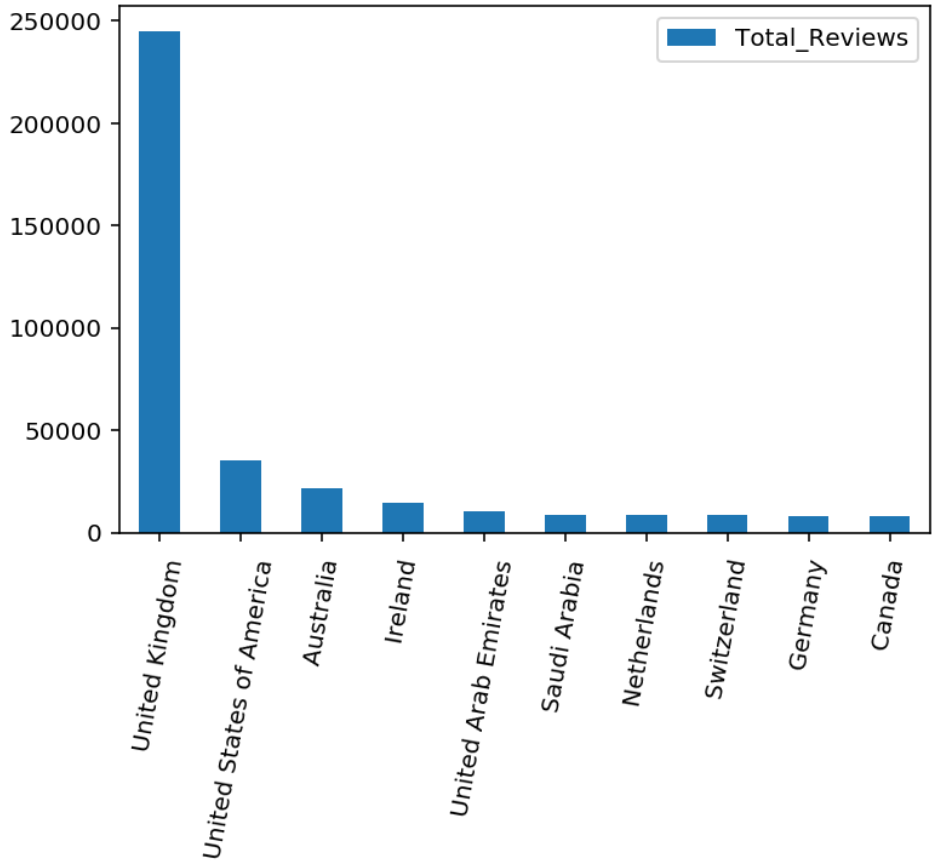

Figure 14: Top 10 nationalities of reviewers

Figure 14 confirms what was already shown on the word cloud; the fact that United Kingdom citizens have provided the most reviews by a wide margin, as they represent almost half of the dataset. The rest of the top 10 most common nationalities can be viewed on the plot and they are more balanced between them. United States citizens are the second most represented nationality at 36.000 reviewers and Australia is third with 22.000 Australian reviewers.

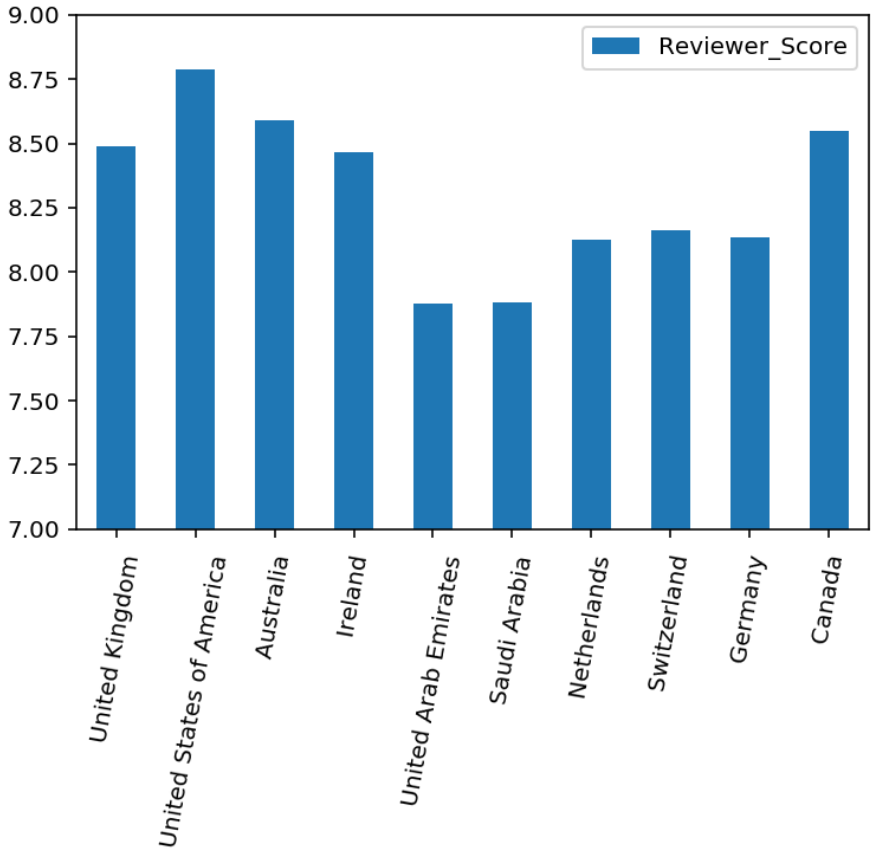

Figure 15: Average score provided by the most common nationalities

Figure 15 shows the average score provided by reviewers from the 10 countries with the most reviews. It is clear from the graph that USA citizens have provided the highest average score, close to 8.8. The next two highest averages have been scored by Australian and Canadian reviewers. This is very interesting as there are some similarities between these three country groups. The distances all three of them had to travel are the highest among the top 10 countries with the most reviews, since all hotels on the dataset are in Europe. We can't say with a high degree of certainty that any of these two cases have a high impact on the reviewer score, but the long-distance travel assumption could be further investigated in future research. While it would be intuitive to think that the higher trouble and fatigue of long-distance travelling would bring the average score these peo-

ple provided down, this is not the case here. Both United Kingdom and Ireland reviewers have provided scores that their average is close to 8.5 and have the 4th and 5th highest average scores respectively. Another interesting observation is the fact that the top 5 nationalities with the highest average scores have English as their mother tongue.

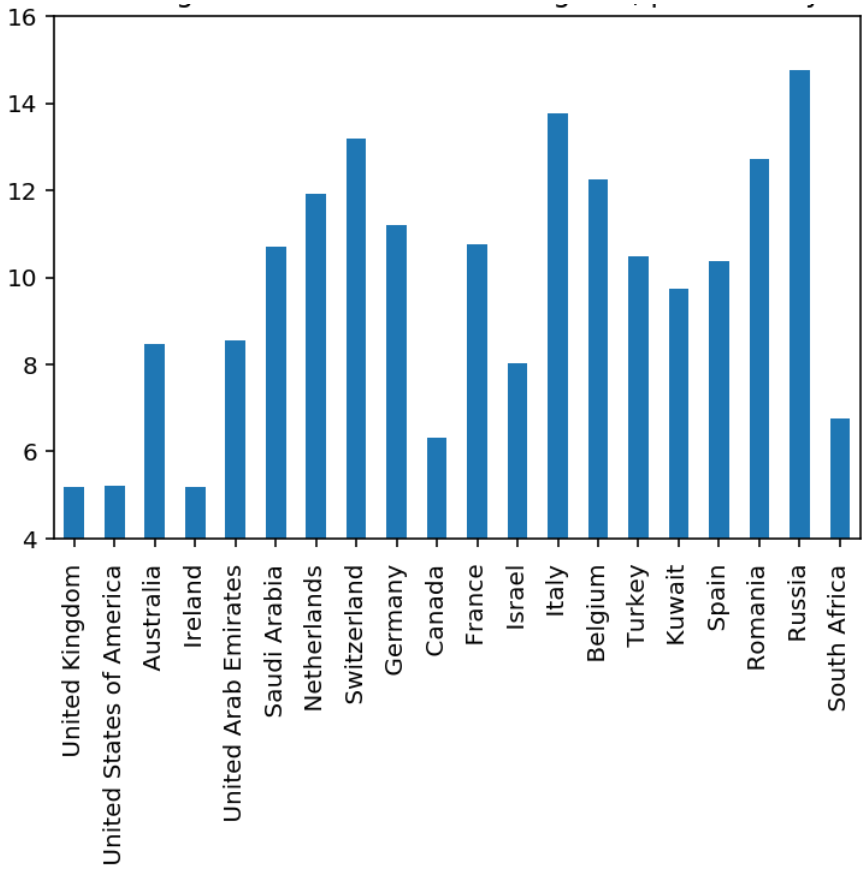

Figure 16: Average number of reviews provided per nationality

In Figure 16 we can see the average number of reviews provided on average by a reviewer, divided per nationality. Russian reviewers top the list with 15 total reviews on average per reviewer; reviewers from Italy and Switzerland come second and third with 14 and 13 reviews per reviewer, respectively. On the other hand, United Kingdom, United States and Irish users have provided close to 5 reviews on average, which is the lowest number among the top 20 nationalities.

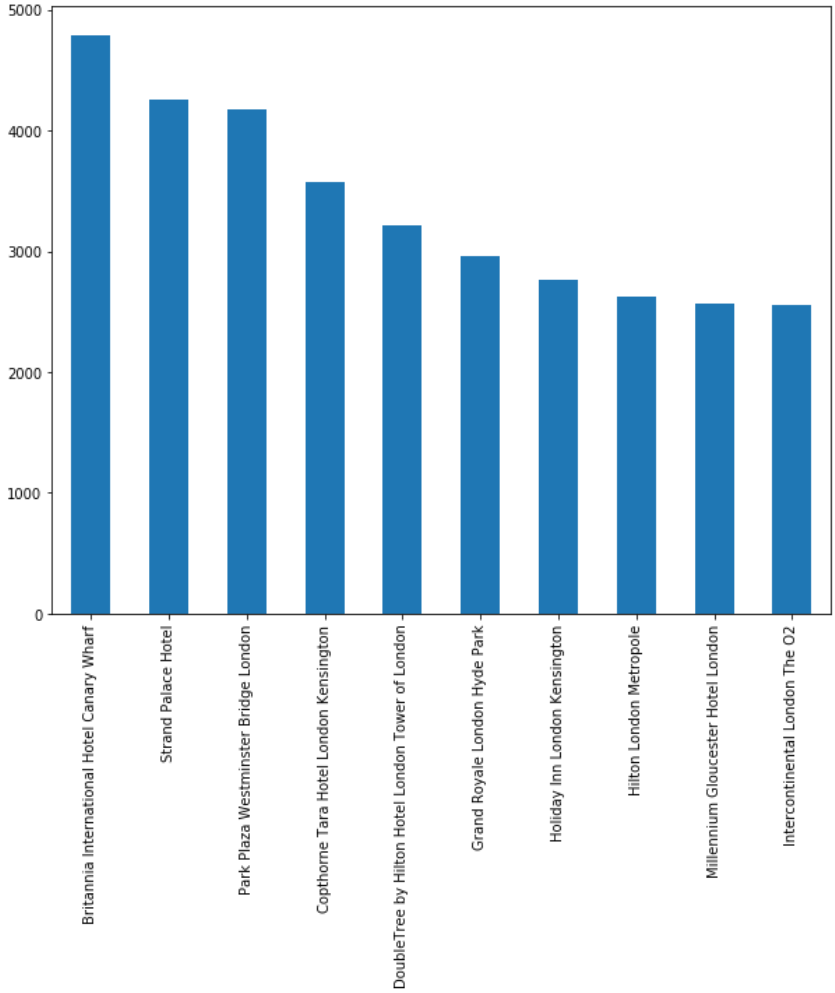

Figure 17: Top 10 most reviewed hotels

Figure 17 presents the most reviewed hotels on this dataset. All the hotels, included on the top 10 most reviewed, are in London. This also explains the fact that almost half of the reviewers are from the United Kingdom. Britannia International Hotel Canary Warf has the most reviews, close to 5,000. Strand Palace Hotel and Park Plaza Westminster Bridge London are the next most reviewed hotels as they have received more than 4,000 reviews each.

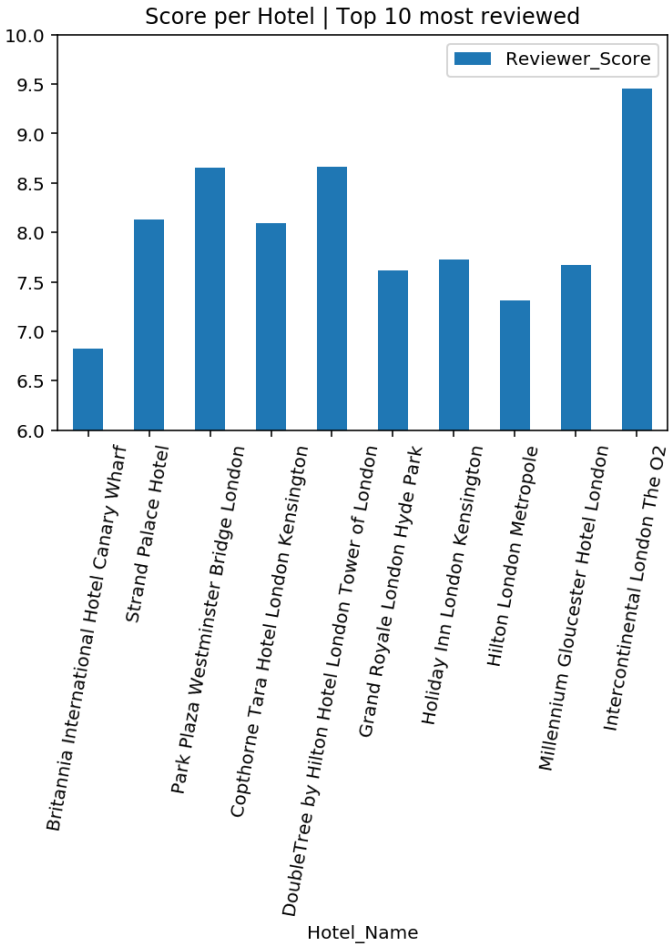

Figure 18: Average score per hotel

In Figure 18 we can see that the most reviewed Hotel, which is the Britannia International Hotel Canary Wharf, has the lowest rating among the most reviewed at 6.7. The average reviewer score is at 8.4 so its rating is well below the average. On this case it could be assumed that the relatively low average score means that reviewers have a higher tendency to provide a review when they are not satisfied with the service. Intercontinental London has the highest rating among the top 10 most reviewed hotels, close to 9.5. Performing a deeper analysis on the hotels would be out of scope, since it is harder to generalize the findings because of the differences between the individual hotels.

4.2 Feature engineering

The first stage, that of the exploratory data analysis has been completed, since we have examined all the existing features. In order to achieve the best results on the machine learning models, we will explore the columns, in order to investigate if it is possible to

extract more information from them and create new features to use the models. Feature engineering is the process of creating new features (in our case new columns) by extracting data from the existing features or combining some of them to create new ones. In this case the column containing the tags will be used, since it contains valuable information provided by the guests. More specifically, the tags column contains tags that were chosen from the reviewers to further categorize their stay. An example of an assigned tag is choosing if the stay was for business or leisure purposes. Feature engineering requires strong knowledge of the industry that the dataset belongs to, in this case Hospitality, to be able to make assumptions and generate ideas [39]. By studying the dataset and the Hospitality industry in general, there are some typical categorizations assigned on online bookings. Those included the type of party staying (couple, family etc.), how many nights the guest stays on the hotel, what is the type of room, the reason for staying (business or leisure) and what device was used to book.

Based on these assumptions, we apply a few functions on the tags column to create new features from the collected information. The function will check if a certain string (essentially piece of text) is included on a row and will create a new column with new categories. Specifically, we can extract the room type each reviewer stayed on, how many days did they stay, if they provided the review using a mobile phone, the type of party (Family, Solo traveler, Couple etc.) and the reason of the trip (Leisure or Business), since there are tags for all the assumptions that were mentioned above .

The new columns created are the following:

- Triptype: The reason of the trip
- Roomtype: The name of the room type
- Pax : The type of party (group, family etc.)
- LOS : The length of stay in days
- Mobile: Checking if the review was submitted using a mobile device

Before examining which feature should be used on the machine learning models, we perform the exploratory data analysis on the new features.

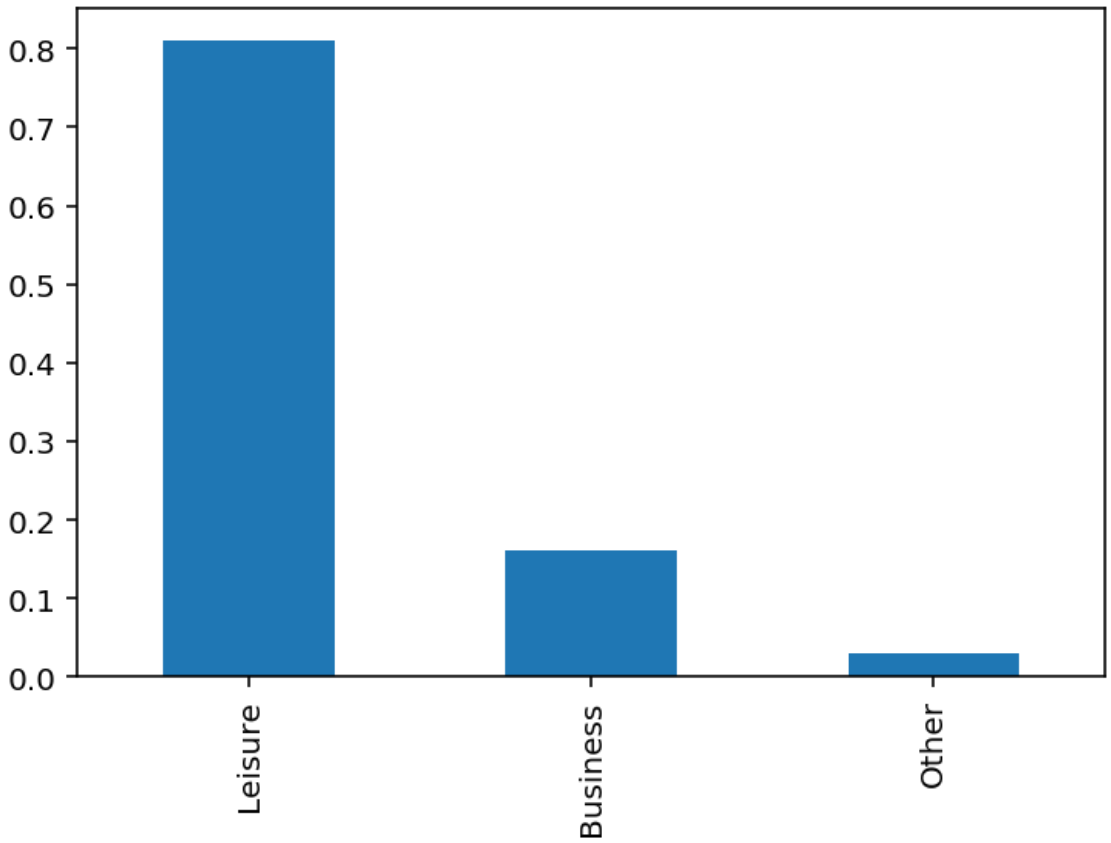

Figure 19: Purpose of the trip

We can see in Figure 19 that most of the reviewers are on a leisure trip, about 80%, while slightly less than 20% are on business; there is also a small number of reviews that are not classified neither as business or leisure and those are labeled as Other.

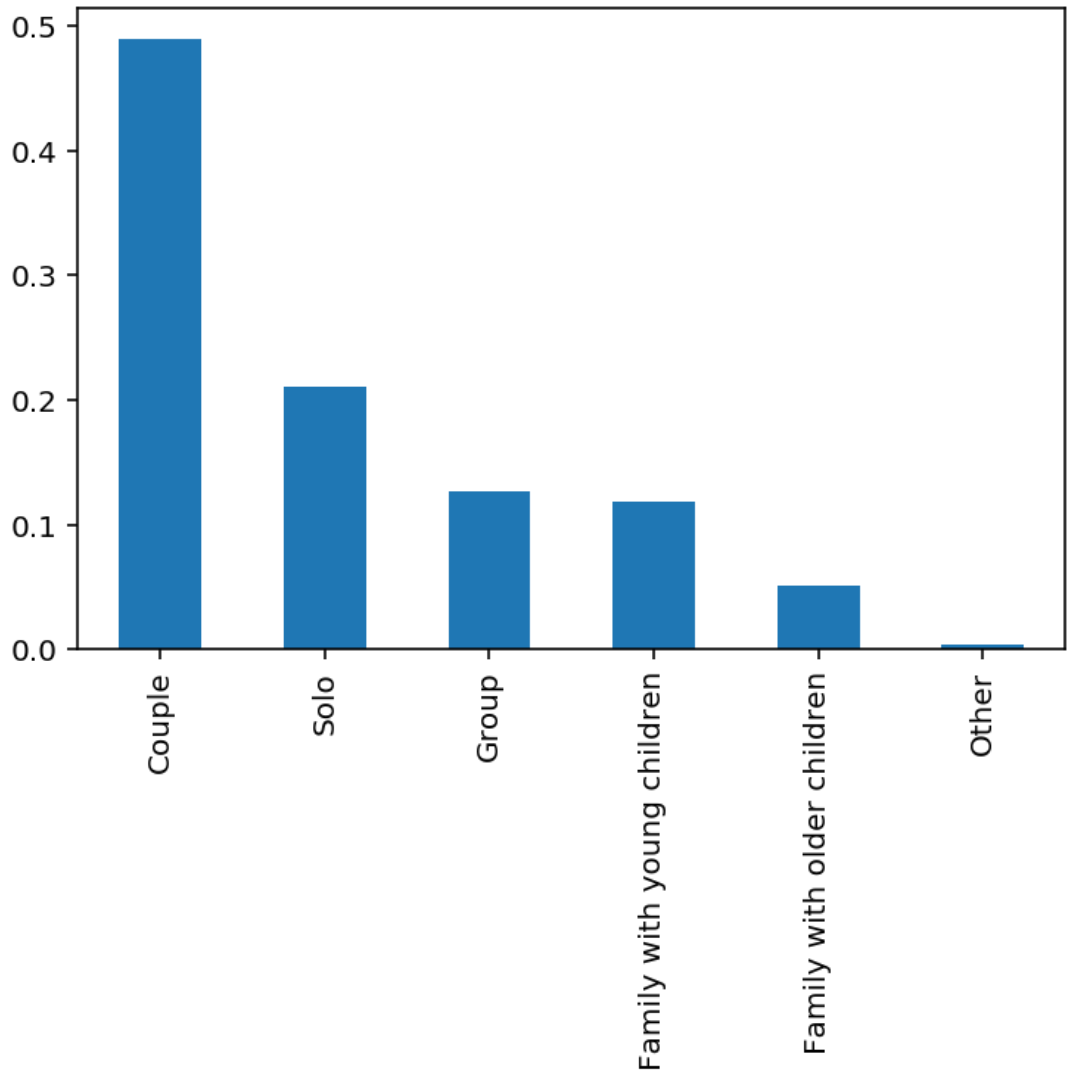

Figure 20: Type of party

As we can see in figure 20, almost half of the reviewers are couples, followed by 20% that are solo travelers; groups represent more than 10% of the total, while families (either with young or older children) cover the rest of the dataset with total percentage of 17%. Finally, there is a small percentage of reviewers that did not provide information on this subject. By combining these two features, the reason of the stay and the type of party, we can better understand the behavior of the reviewers.

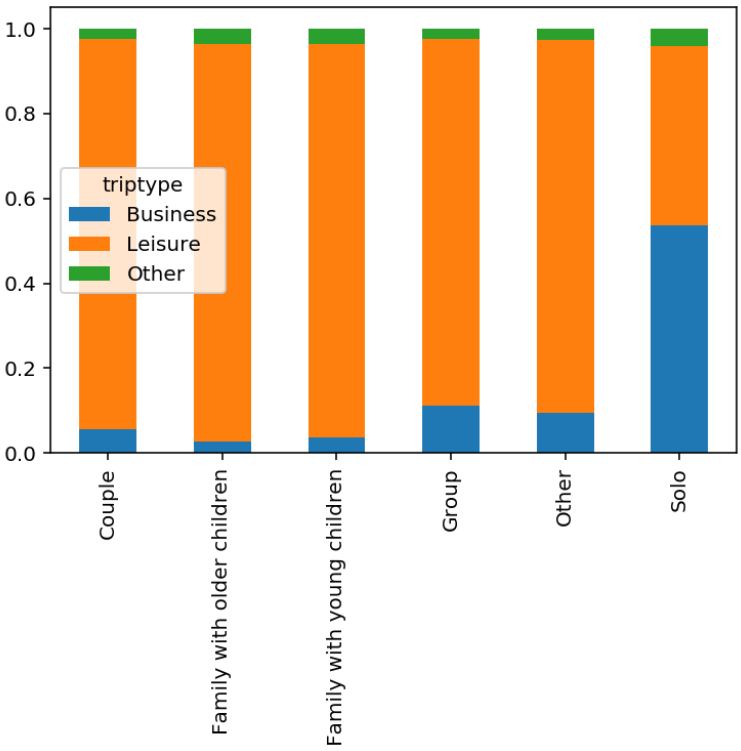

Figure 21: Reason of the stay per type of party

Figure 21 is a stacked bar chart that shows the reason of the stay per different type of party. Most of the reviewers travel for leisure purposes on more than 80% of the time. Specifically, couples and families have leisure as their reason for staying on more than 90% of the cases; this is what common sense would suggest as people usually travel for business mainly alone or, less frequently, in groups. We can see that this is the case on this dataset since the reviewers that travel alone, are travelling for business on more than half of the cases. There is also a significant percentage of solo reviewers, more than 40%, that travel for leisure purposes; this percentage is quite high, considering that travelling alone for leisure is not the most common case.

```
pax
Couple                        8.512303
Other                         8.498833
Group                         8.450558
Family with older children    8.434673
Family with young children    8.303258
Solo                          8.129133
Name: Reviewer_Score, dtype: float64
```

Figure 22: Average score per type of party

Figure 22 is a table generated on the notebook, that shows the average reviewer score per type of party. There is significant variance on average scores between types of parties, ranging from 8.51 for couples, which is 0.11 above the 8.4 average for all reviewers, to the lowest average score of 8.13 that belongs to solo travelers. As discussed already, while for all types of party the purpose of the trip is leisure with a percentage close to 90%, solo reviewers travel for business in 53.7% of the cases. The fact that more than half of the solo reviewers are staying in the hotel for business purposes means that it is harder for them to enjoy their stay, compared to stays connected with leisure. In contrast, couples that travel for leisure purposes in 93% of the cases have the highest average score. Families with young children are also below the average score for the whole dataset, at 8.3; at the same time families with older children and groups are slightly above the average.

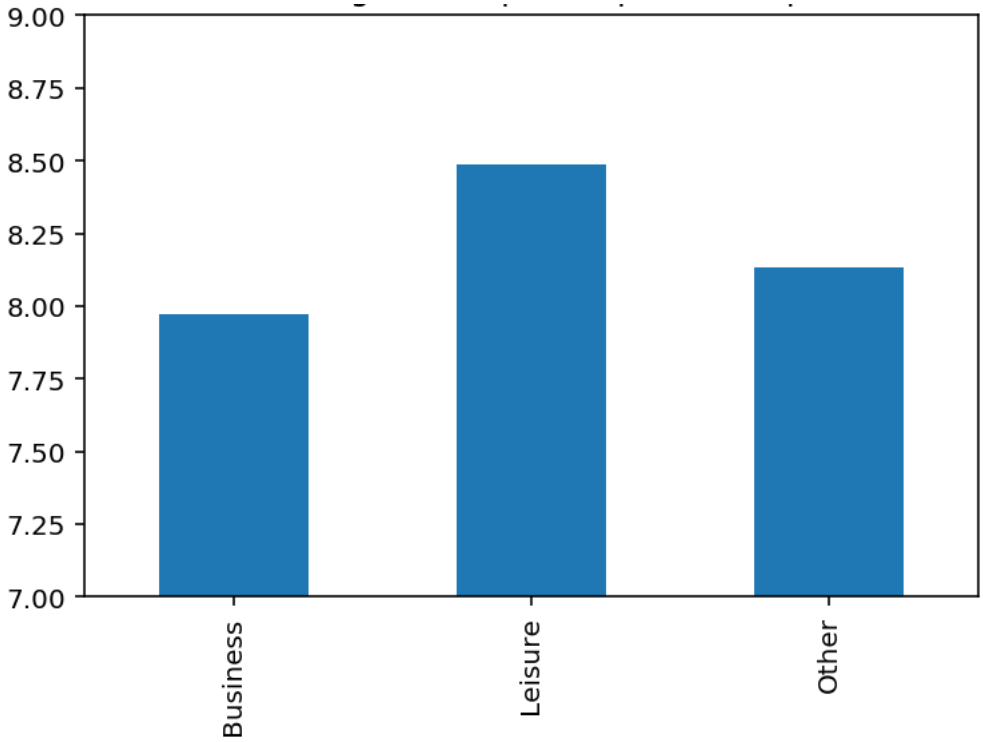

Figure 23: Average score per purpose of trip

Figure 23 shows the average score grouped by the purpose of trip. There is a significant difference between business trips and leisure trips; reviewers travelling for business have an average score that is well below the average, specifically slightly less than 8. On the other hand, leisure travelers have provided scores with an average of 8.5; considering that they cover more than 80% of the dataset, it is normal that the total average (8.4) score is closer to their own. The difference in average scores between business and

leisure stays also explains the low average score of solo travelers, since more than half of them travel for business.

Another feature extracted from the tags column is the room type of each reviewer. This is done by building a function that checks if a specific room type name is included in the tags column for each row and assigns the room type name on a new column called roomtype. Figure 24 shows what percentage of the total is tagged with the specific room type name.

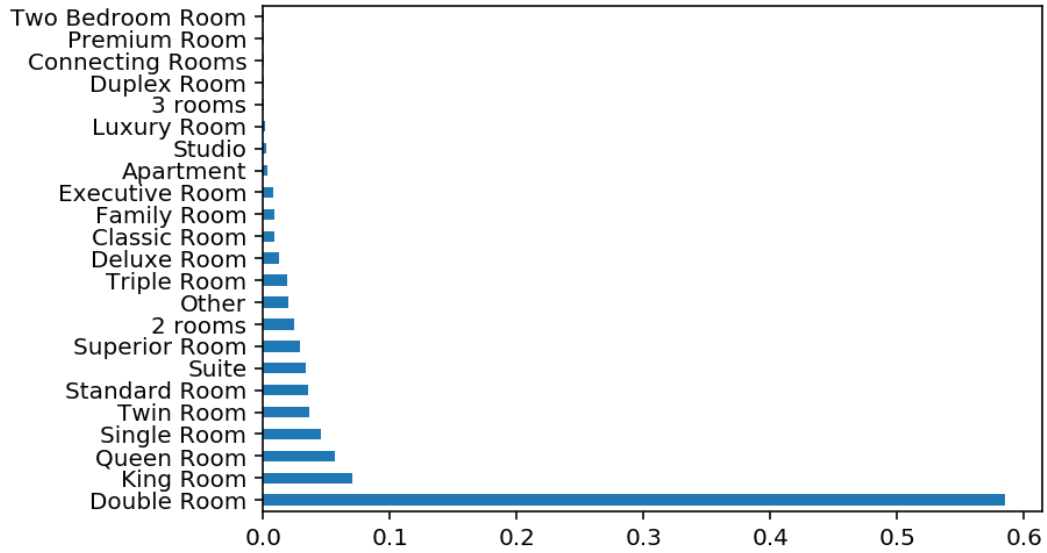

Figure 24: Room types

Close to 60% of all reviews are from guests that stayed in Double Rooms. Considering that almost half of all reviewers are couples, this percentage fits the pattern. King and Queen rooms come second and third with less than 10% each. Single rooms cover a 5% of the total, meaning that a big percentage of the solo travelers stay on other room types, other than the single rooms, as well. It is important to note that the room type names may vary greatly between hotels. There are also instances that the booking includes more than one room type; those are referred as 2 rooms and 3 rooms on this dataset.

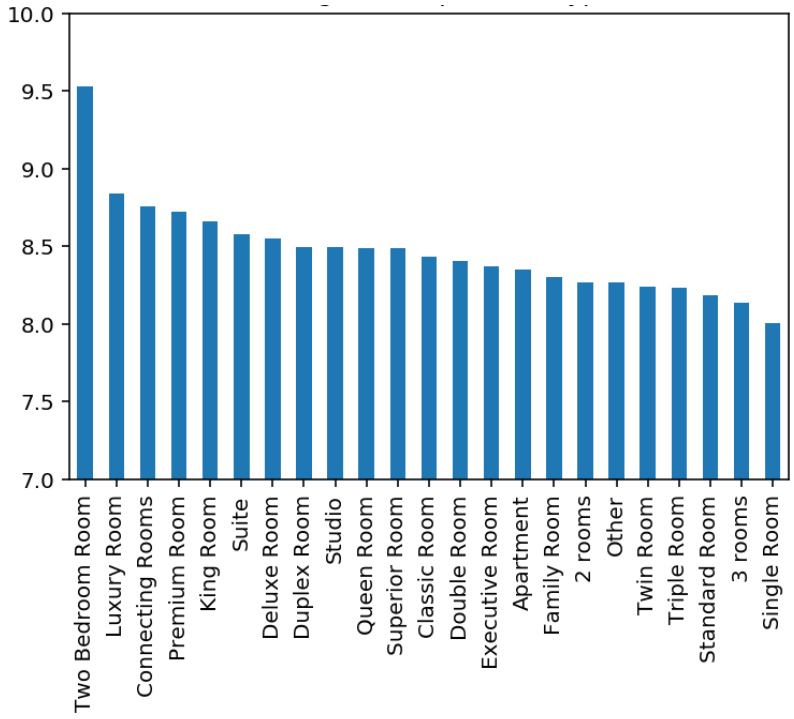

Figure 25: Average score per room type

Figure 25 presents all the room types that were extracted from the tags column, sorted by the highest average score to the lowest. The Two Bedroom room has the highest average score at 9.5; at the same time there is a very low number of such room types on this dataset. The next highest average scores belong to the following room types: Luxury Room, Premium Room, Connecting Rooms, King Room and Suites. These room types are usually more expensive and can host at least 2 people. On the other hand, the single room has the lowest average score, close to 8; we already established that the solo travelers have the lowest averages since they have a strong connection with the business type stays.

The analysis so far has shown that couples and bigger parties tend to give higher average scores, and this is further supported by the higher average score on bigger rooms. While the difference is small, we can see that average scores are lower on simpler and smaller rooms like Standard Room, Single Room and Twin Room. As discussed on the Literature Review part of this dissertation, more expensive services (in this case room types) tend to be perceived as more valuable and of higher quality by customers. This fact can partly explain the higher average scores on bigger and more expensive rooms.

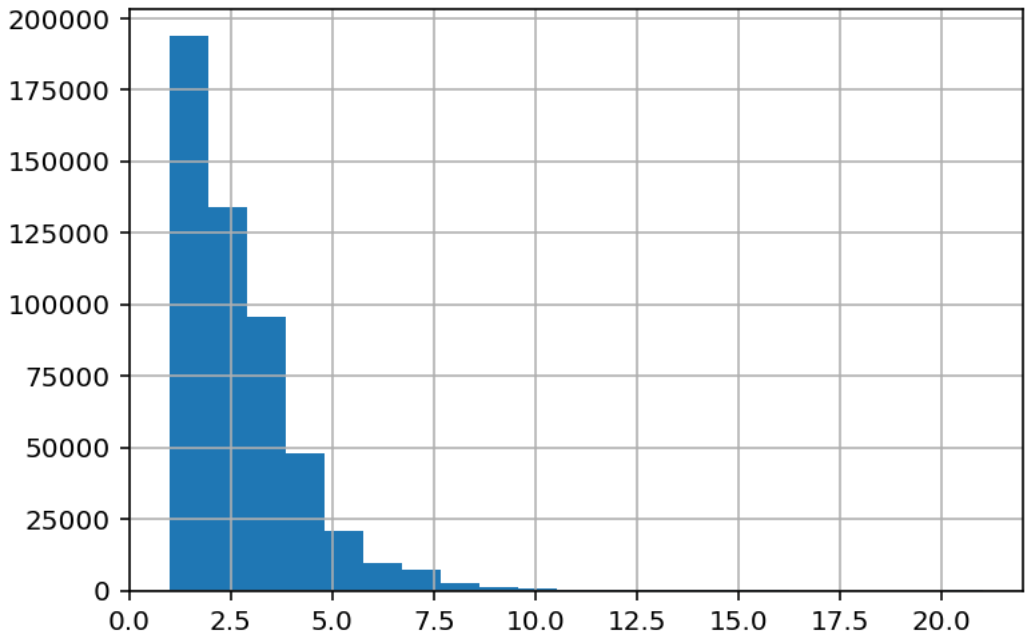

Figure 26: Length of stay distribution

Figure 26 shows the distribution of the length of stay feature. It is clear that the vast majority of reviewers have stayed between 1 and 3 days. There are very few cases that have a length of stay higher than 5 days. The average length of stay for the whole dataset is 2.36 nights. Considering that the hotels are located on big European cities, this is normal, since city hotels tend to have lower stays on average, compared to resorts.

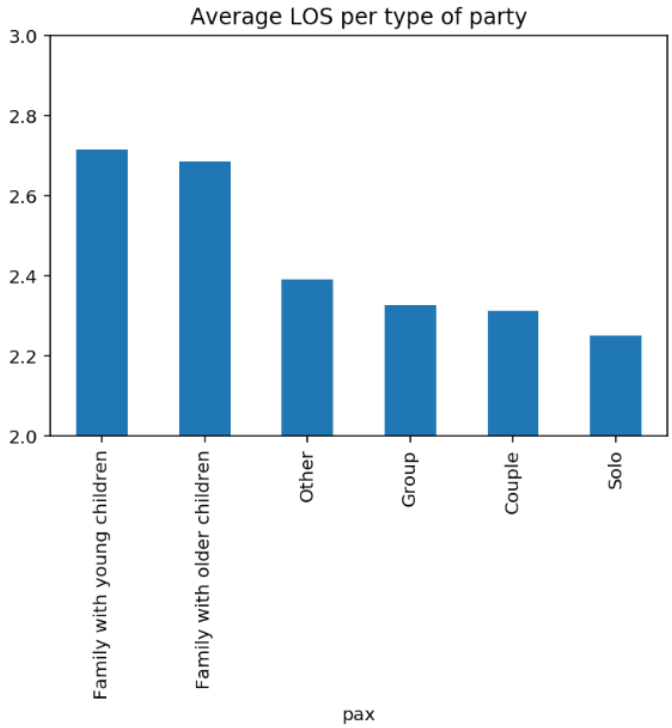

Figure 27: Average length of stay per type of party

In Figure 27 we are breaking down the length of stay by the type of party. Families, with younger or older children, tend to stay longer than the average. Couples and groups have average length of stays that are very close to the total average. Solo travelers have the shortest average stay on the dataset, slightly lower than the average for the whole dataset.

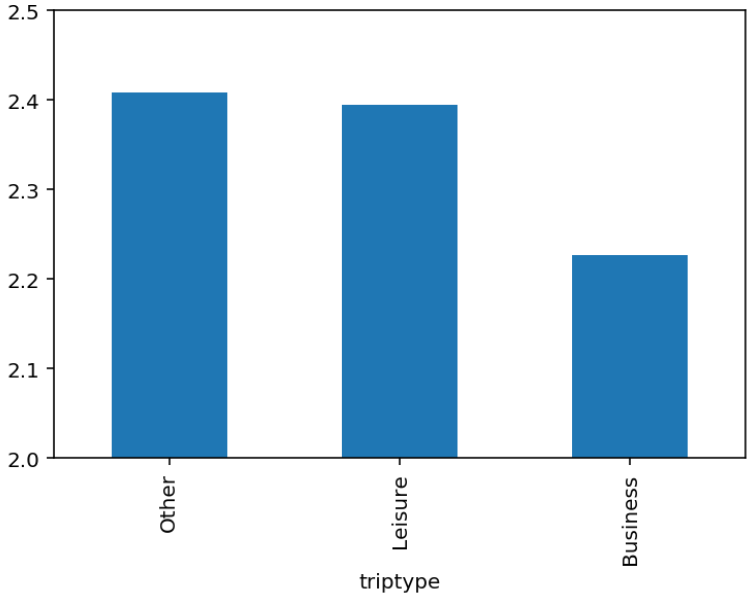

Figure 28: Average length of stay per type of stay

Leisure travelers tend to stay longer than business traveler, as can be seen in Figure 28. We have examined the variance of the length of stay depending on the new features created by analyzing the column containing the tags.

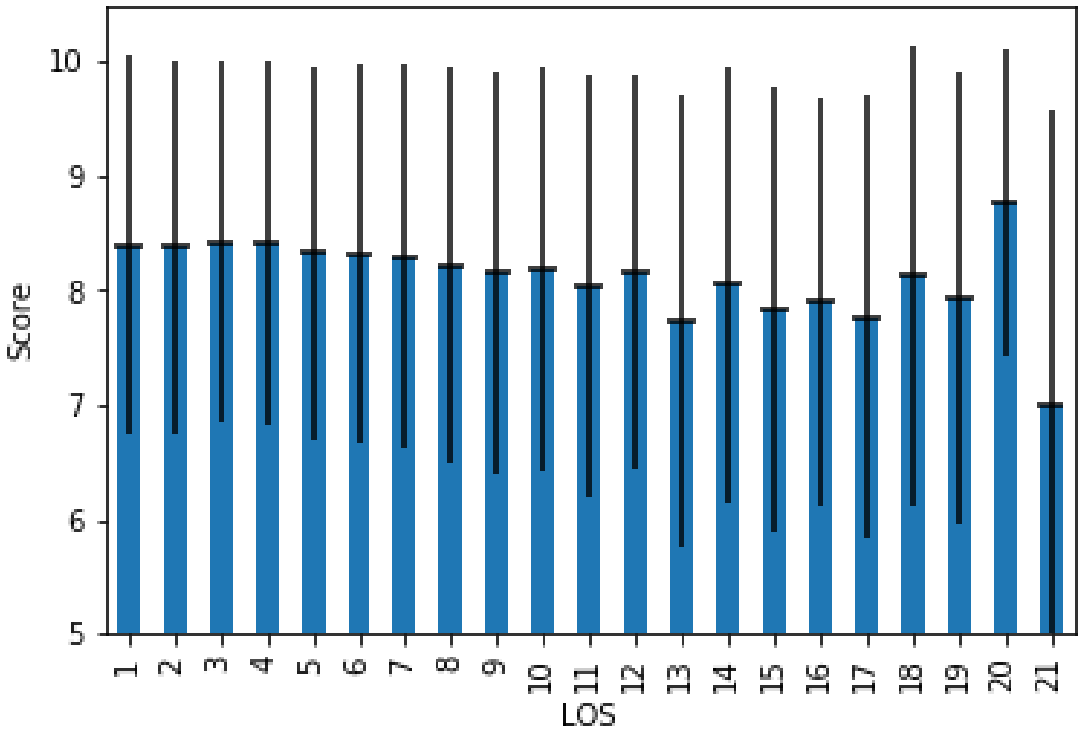

Figure 29: Average score per length of stay

Figure 29 shows the average reviewer score grouped by how many days the reviewer stayed. The black lines ranging above and below the bars, represent the standard deviation, to provide a clearer picture of the distribution on each case. From this figure, there are not any clear patterns uncovered that connect the length of stay with significant differences on the average scores.

Finally, two more features will be created to be used on the models; those are the polarities, for the positive and negative reviews. Polarity shows how positive or negative are the emotions expressed on a review [40], and it can take values between -1 (fully negative) and +1 (fully positive). The polarities can be generated using the TextBlob, which is a Python library providing the tools required to perform natural language processing tasks [41].

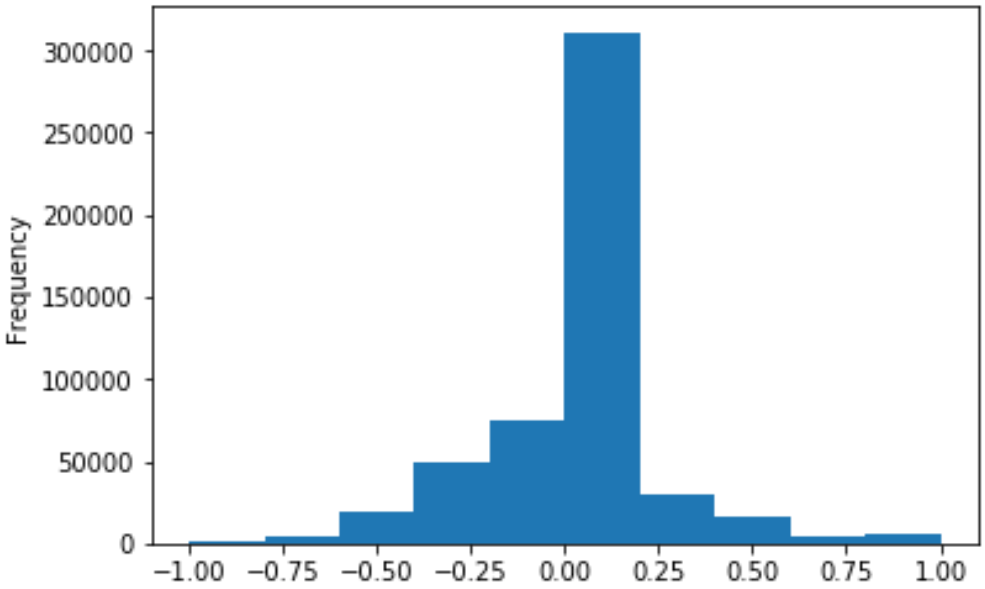

Figure 30: Negative reviews polarity distribution

Figure 30 shows the distribution of the negative polarities for the whole dataset. Most of the reviews are concentrated between 0 and 0.2, which means that they represent either neutral or slightly positive emotions. Many reviews do not have any negative parts and that is the reason for the positive polarities on negative reviews. There is a significant number of reviews that have negative polarities, but mostly less than -0.5; meaning that while the emotions expressed are calculated as negative but not too strong.

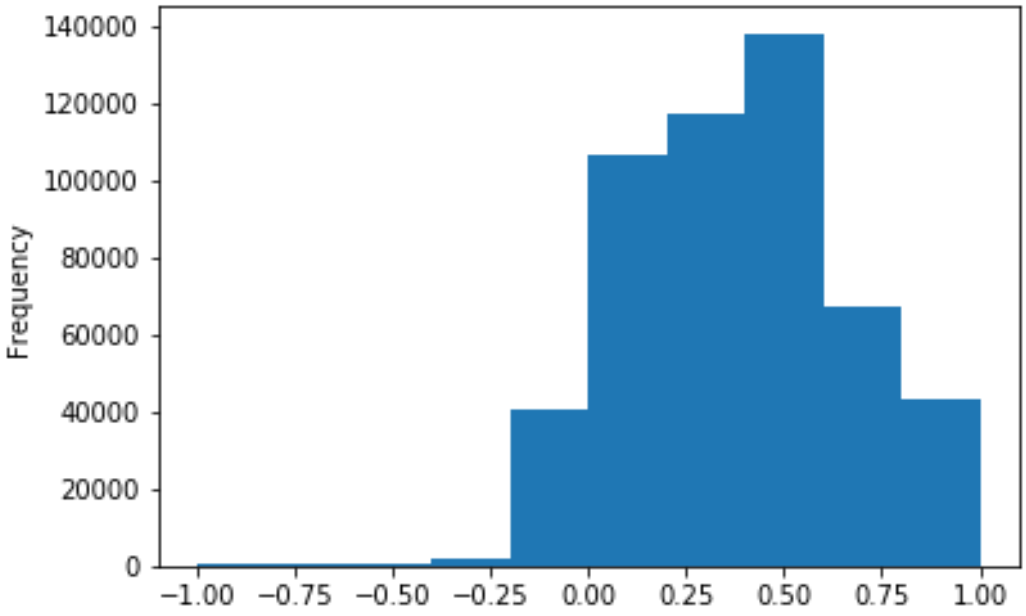

Figure 31: Positive reviews polarity distribution

Positive reviews have strongly positive polarities in most cases, as seen on Figure 31. There is a small number of reviews with neutral or slightly negative polarities, as well.

5 Machine learning models

The exploratory data analysis and the feature engineering parts have been completed, and there is a clearer picture about the dataset and its observations. The next step is the creation of the machine learning models. Two models will be created, a classifier that predicts which reviews have above average score and a regression model that predicts the actual score on a review. The main goal of the models isn't the final prediction, but the insights extracted by explaining how the models reach the prediction. This doesn't mean that the accuracy of the predictions is not important, since the insights should be based on as accurate predictions as possible.

5.1 Classifying reviews with above average score

The goal of this model is to predict if a review will have a score higher than 8.4, which is the average score for all reviewers. As explained on chapter 3.3, the XGB classifier will be used, since it usually performs better in terms of accuracy, compared to Decision Trees and Random Forests algorithms. The algorithm needs to make a prediction on whether the review will have an above average score or not; which means that there are only two classes. In order to select the best features for the model the correlation of each feature with the above average score will be examined.

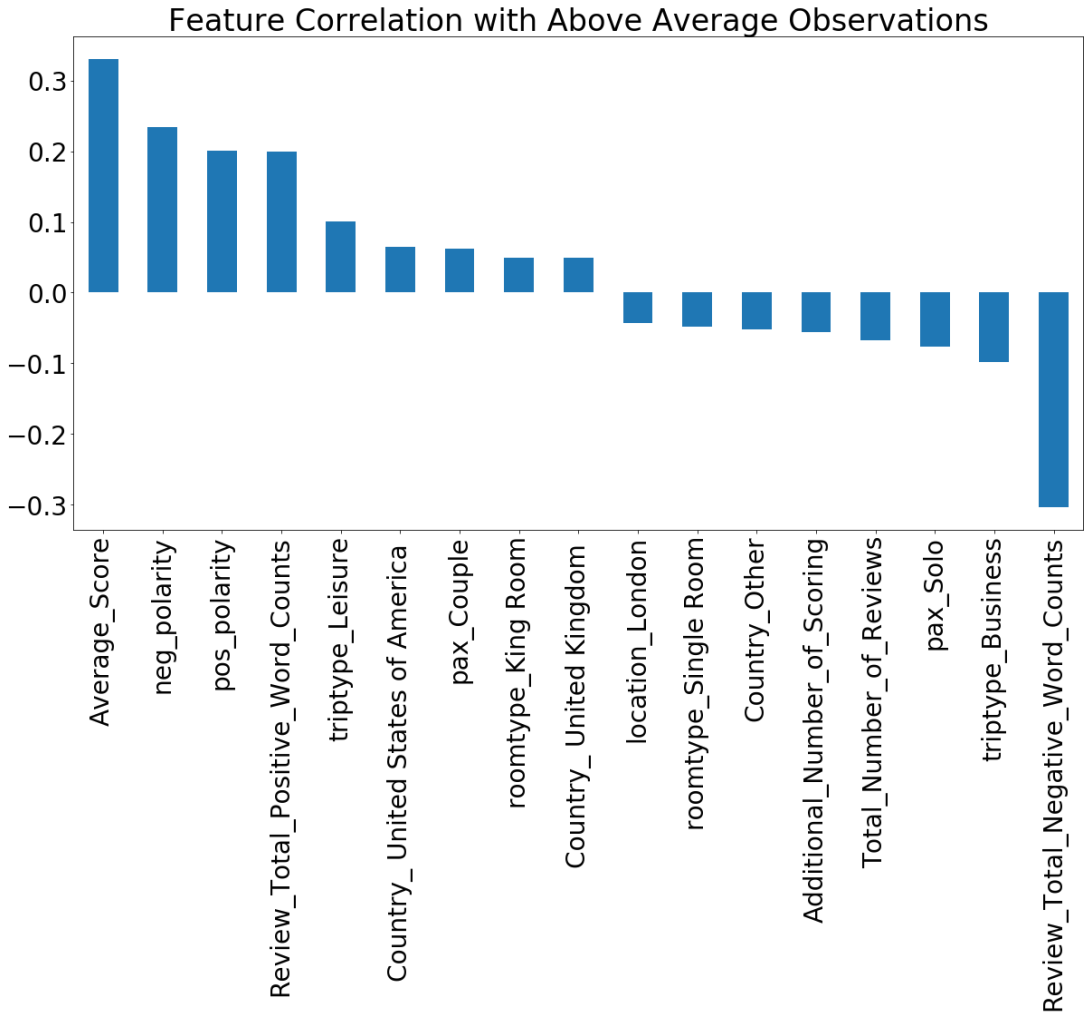

Figure 32: Features' correlation with above average reviews

Figure 32 presents the Pearson correlation coefficient of each feature with the target variable, which is the above average reviewer scores. The Pearson correlation coefficient shows a perfect positive linear relationship when its value is +1 (plus one) and a perfect negative linear relationship when its value is –1 (minus one) [42]. Those two numbers are the maximum and minimum values the Pearson coefficients can receive. All the in-between values show a less than perfect linear correlation, while a value of 0 means that there is no correlation between the two features. The features on the graph are filtered based on their correlation, so any features with correlation coefficient with absolute value lower than 0.04 were removed. This filtering is necessary since having more features with very small correlation (which means they do not contribute much to the final predictions) can have negative effects on the accuracy of the algorithm.

It is important to note that 70% of the selected features data will be used to train the model and the remaining 30% will be used to test the accuracy. The strongest positive

correlation is the average score of each hotel, since this feature is partially dependent on the sum of the scores provided by the reviewers on this dataset. The other most important correlations are between the target and features about the text of the reviews. Polarities created on a previous step; both have positive correlations with the reviewer score. Positive and Negative word counts are also correlated with the score, positively and negatively respectively. There is no point to further explore the correlations, as there are more sophisticated ways to discover what effects, and how strong, each feature has on the predictions and those will be explained after receiving the predictions.

The XGB Classifier has predicted correctly if a reviewer will provide an above average score on 76.50% of the times. The prediction accuracy is high enough to explain the model's predictions, so no further attempts to improve the score will be performed.

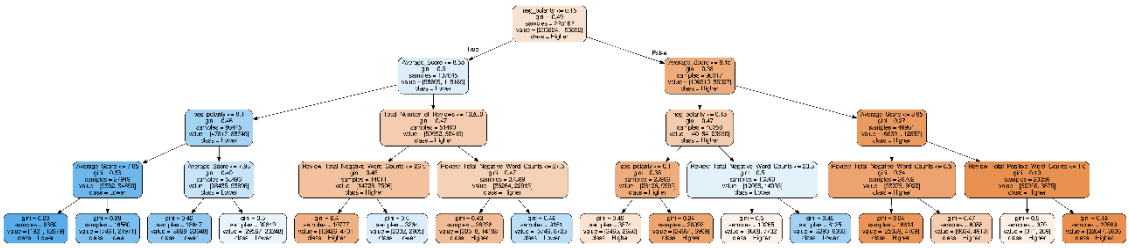

Figure 33: XGB classifier sample decision tree

Figure 33 shows a sample decision tree that was create on this model; it has a max depth of 4, meaning it splits up to 4 levels until it makes a prediction. We have only visualized a decision tree with 4 levels but most of the decision trees split many times more.

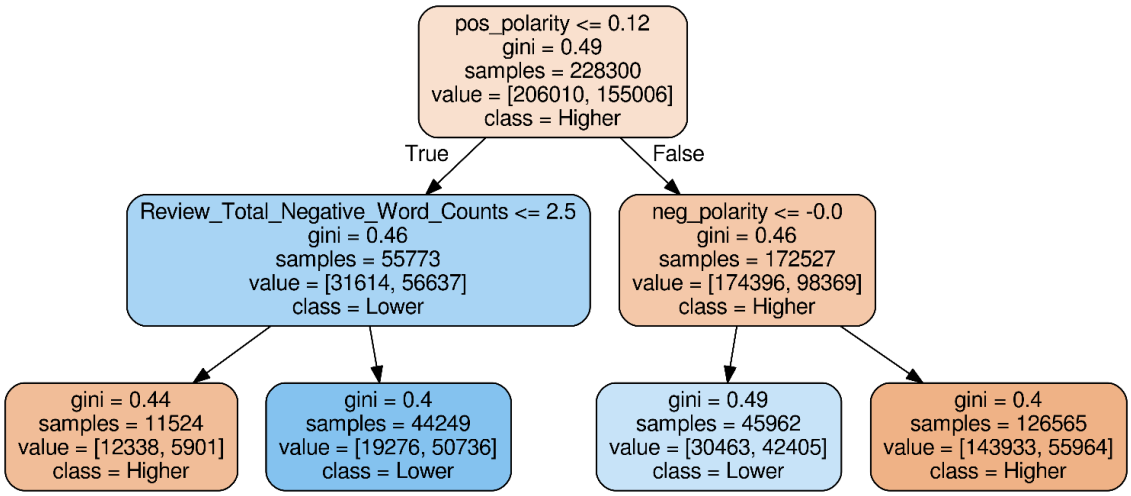

Figure 34: Three-level decision tree

Figure 34 presents a three-level decision tree which allows to understand the process of the algorithm better. Each node represents a feature and the decision is made based on its value; the arrows from one node to the next ones are the decision, whether the condition on the node is true or false and lead to the next decision. The last nodes at the base of the tree are the ‘leafs’ and those represent a prediction of the tree. Classification trees make their decision based on the gini impurity. Gini impurity shows how good the split is; when gini is 0 this means that all the samples on the specific node are from the same class [43]. The maximum value it can take is 0.5, which means that the half the samples belong to one class and the other half belongs to the other one.

We can see that this decision tree has the positive polarity as its root node, and it splits the samples based on their positive polarity. Specifically, it checks if they have a positive polarity higher than 0.12 or not. On each node we can also see the class, which is the decision of the tree if it would be to stop on that level.

There are multiple ways to measure the importance of the features. The first metric we will examine is the Permutation Importance. Permutation importance is calculated by randomly changing a column’s order and observing the effect on the prediction accuracy [44]. This process takes place after an algorithm has been trained and has provided results, which means that the metric shows the real effect on the actual predictions.

Weight	Feature
0.1032 ± 0.0015	Review_Total_Negative_Word_Counts
0.0448 ± 0.0010	Review_Total_Positive_Word_Counts
0.0360 ± 0.0011	Average_Score
0.0178 ± 0.0010	pos_polarity
0.0117 ± 0.0014	neg_polarity
0.0013 ± 0.0006	Country_United Kingdom
0.0010 ± 0.0001	Country_United States of America
0.0004 ± 0.0005	Country_Other
0.0000 ± 0.0003	pax_Couple
-0.0001 ± 0.0002	roomtype_Single Room
-0.0001 ± 0.0004	pax_Solo
-0.0005 ± 0.0003	roomtype_King Room
-0.0006 ± 0.0004	triptype_Leisure
-0.0009 ± 0.0004	triptype_Business
-0.0017 ± 0.0003	location_London
-0.0024 ± 0.0001	Additional_Number_of_Scoring
-0.0025 ± 0.0008	Total_Number_of_Reviews

Figure 35: Permutation Importance

In Figure 35 we can see the results of the Permutation Importance analysis. Apart from the feature name column, there is the Weight of each feature. The weight shows how much the final prediction has changed when the feature's column was shuffled, and the effect is based on the metric we have used to measure the success of the algorithm; in our case it's the accuracy of the predictions. The features are ranked based on the permutation importance, and the number next to the weight (after the plus-minus sign) shows the range of variation for the feature's weight, since it is calculated on the XGB algorithm that is an ensemble model with many different predictions.

The weight represents the loss of accuracy when the feature is shuffled, so higher positive value means that the feature plays an important role on the prediction, while lower positive scores mean a smaller impact. A feature with a negative value weight, means that by shuffling the feature the accuracy of the prediction improved. The reason for this is randomness, and the feature doesn't have a high importance for the predictions [45].

Having explained the meaning of the metrics, we can see that the most important feature in our model number of words on the negative reviews, which decreases the accuracy by 0.1032 on average. The next most important features are the positive review word counts with a weight of 0.0448 and average score for the hotel with 0.0360. Positive and

negative polarities, extracted from the text of the reviews, play a minor role on the accuracy of the predictions. As mentioned, if the weight is 0 or negative, it means that this feature doesn't help to make a better prediction and could possibly be removed from the model.

Having established the importance of each feature on the classification of above average review scores, we have provided some more information on how the algorithm reaches the predictions. In a lot of cases the feature importance isn't enough to disperse the belief that machine learning algorithms are black boxes (meaning that we don't actually know how they work their way to a prediction) and thus create a stronger sense of trust to users. Another side to the solution of this problem is to show how a prediction changes based on the changes of the selected feature's values, what we call the marginal effect. The first way to explore the marginal effect is to plot the Partial Dependence Plots (PDP) for the most important features.

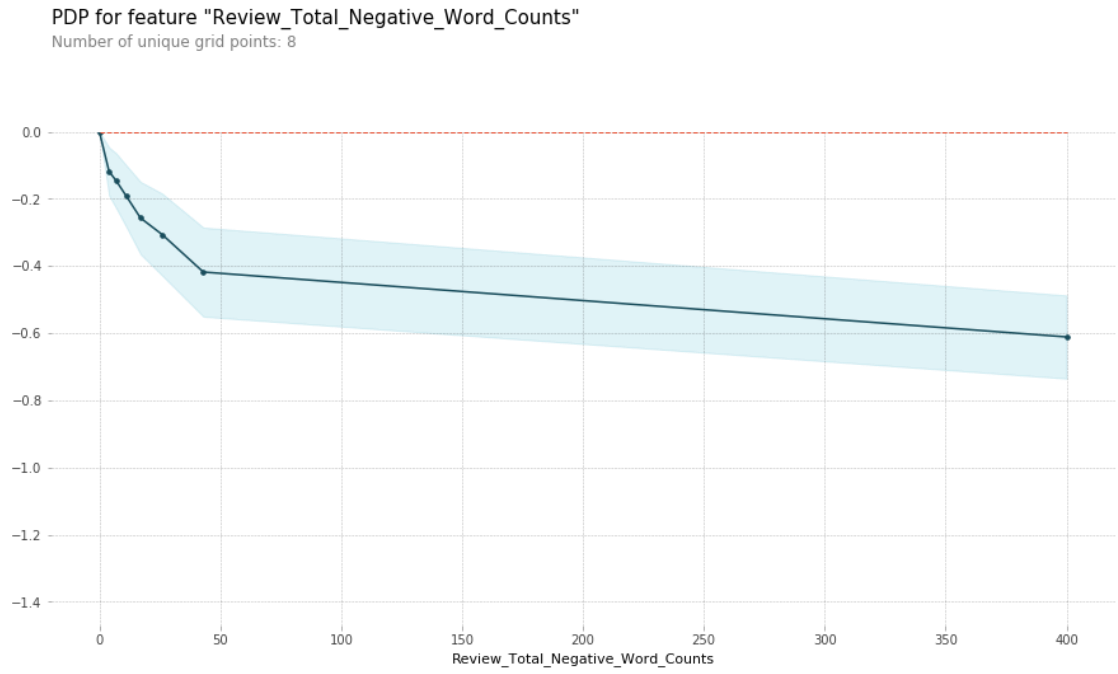

Figure 36: Partial dependence plot for negative reviews word count feature

Figure 36 is the partial dependence plot for the negative reviews word counts feature. In our case, we have a classification problem, so the y axis shows the chances of a positive classification and the x axis shows the values of the selected feature. By keeping all other features at a fixed or average value, we can create a regression model between the

target variable and the one we examine on each case [47]. The shaded space surrounding the line is the confidence level.

We can see that there is a negative linear relationship between the number of negative words and the chances that a reviewer score will be above average. Starting from 0, where the chances of a positive classification are not affected, we can see that increases on the total negative words decreases the chances that the score will be above 8.4. By examining the line, it seems that until reaching 50 words, each small increase has a higher negative effect; after the 50 words mark the line is not so steep, which means that the effect of each additional word is lower after that point.

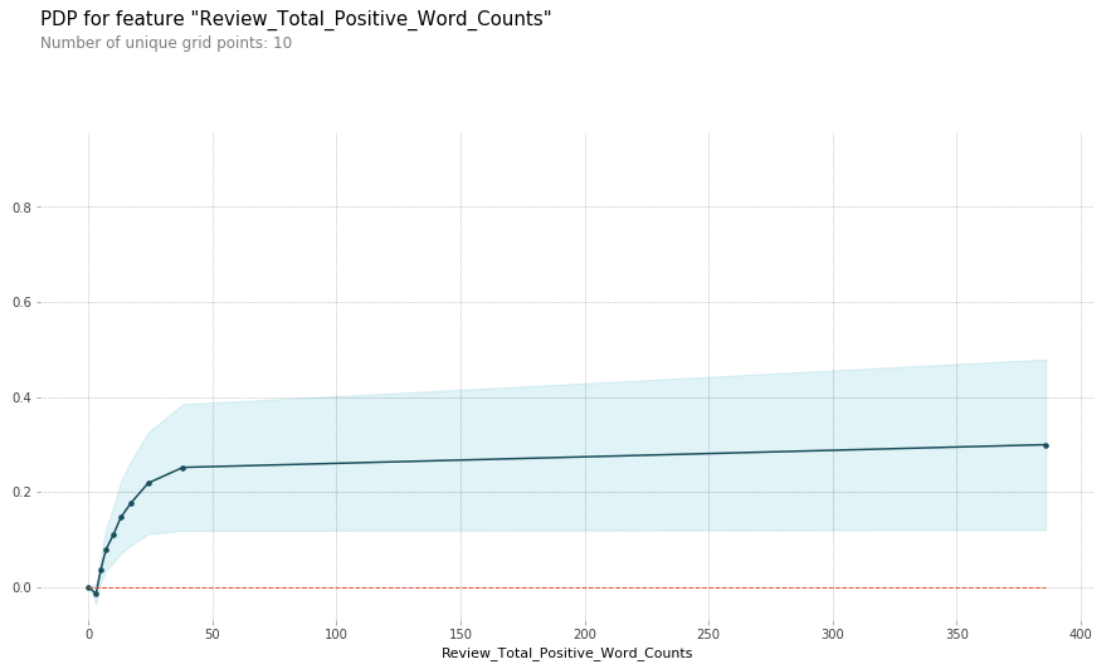

Figure 37: Partial dependence plot for positive reviews word count feature

On the other hand, the positive word counts have the exact opposite effects, as seen in Figure 37. Each additional positive word increases the chances of a positive classification, and the effect starts deteriorating after the first 45-50 words. It should be noted that the maximum increase in chances for an above average score can increase by 30% at most, when there are more than 350 words positive words. At the same time reaching more than 350 negative words decreases the chances for an above average prediction by 60%. An important conclusion can be drawn from these insights, since it shows that people who write a lot of words on a negative review have stronger and maybe more extreme feelings than people who write a lot of words on a positive one.

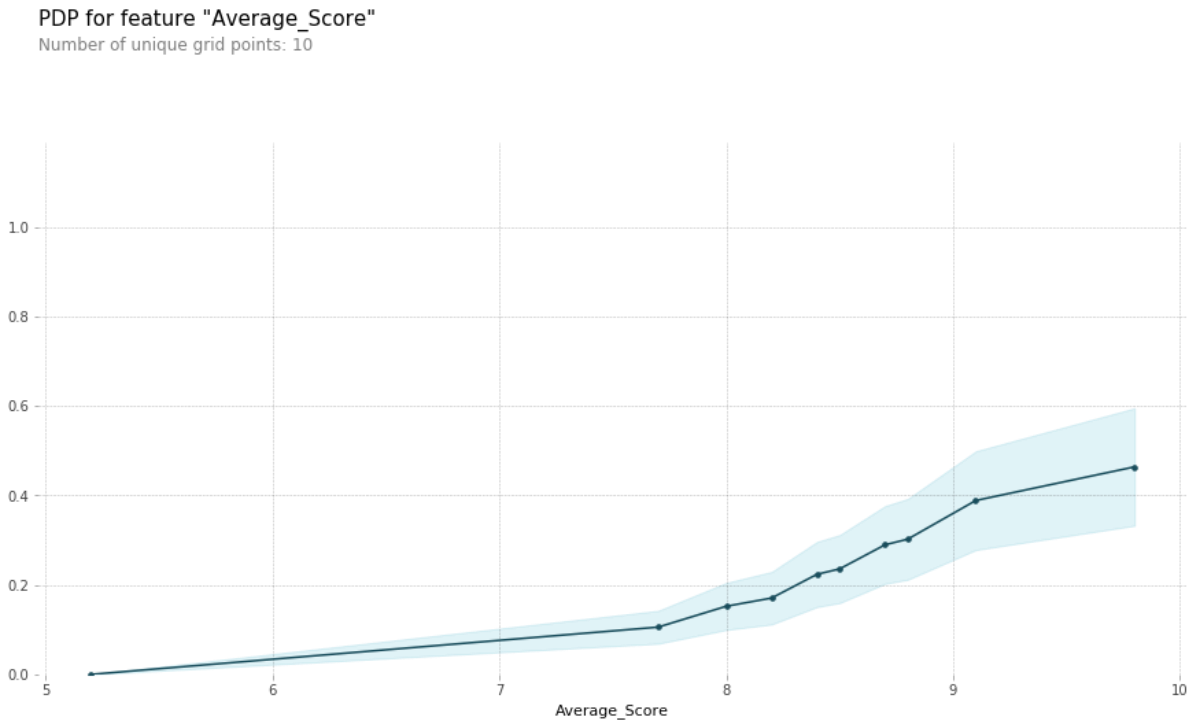

Figure 38: Average hotel score partial dependence plot

Figure 38 shows a clear positive linear relationship between a positive classification's chances and the average score a hotel has achieved. As already noted, the average score is partially dependent on the reviewer scores from this dataset, since the average score is calculated as the average of all reviewer scores for each hotel. The hotel with the highest average score has close to 50% more chances to receive a higher than 8.4 review score than a hotel with average score of 5.

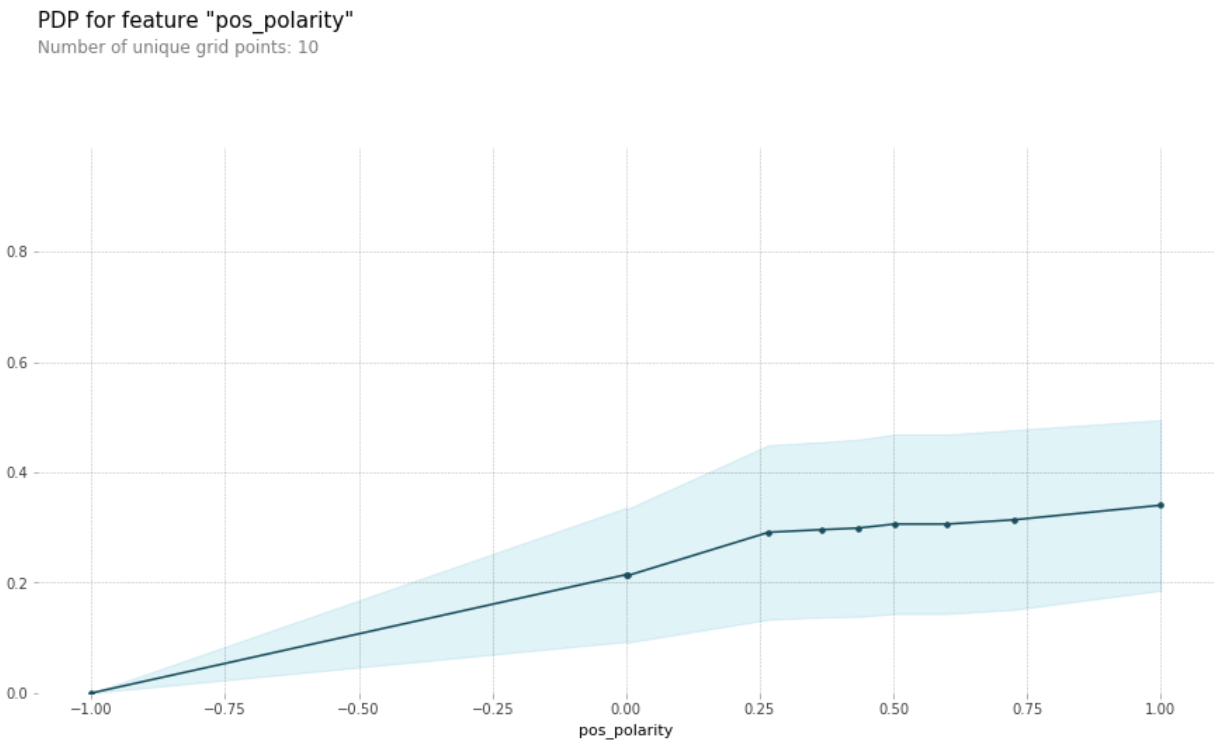

Figure 39: Positive polarity partial dependence plot

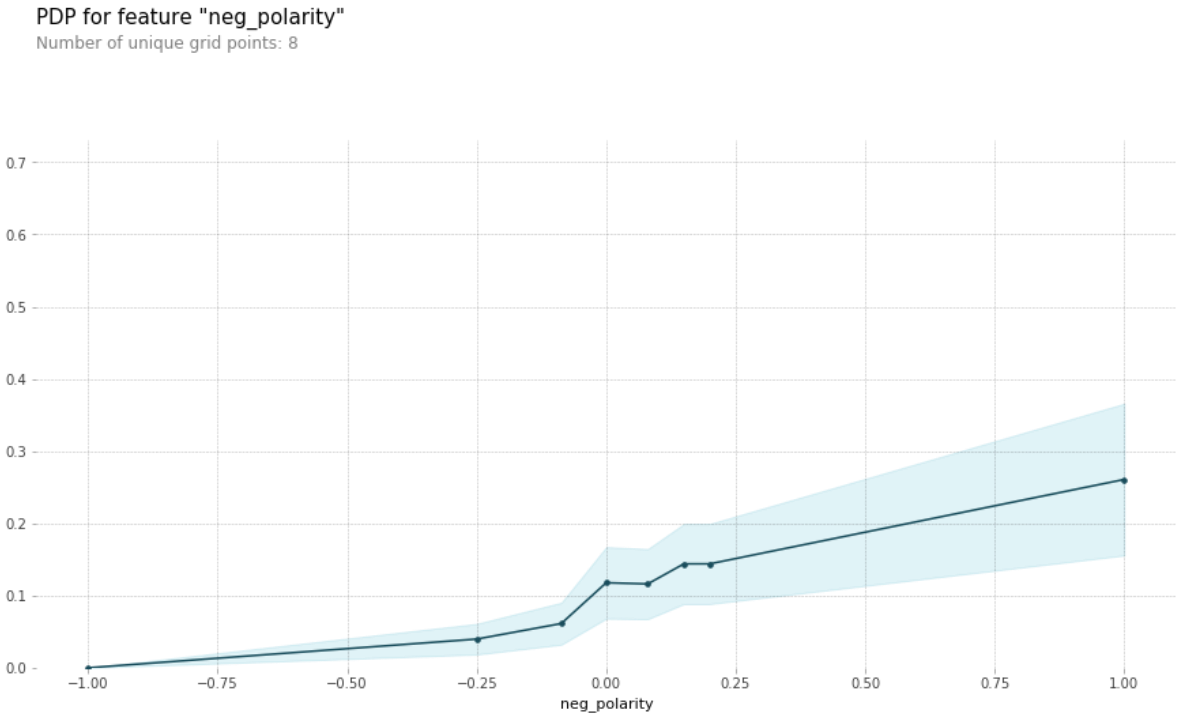

Figure 40: Negative polarity partial dependence plot

Positive and negative polarities have similar plots as can be seen in figures 39 and 40. Polarities are dependent on the word counts, not only quantitively but also qualitatively since the polarity is calculated by examining how strong positive or negative feelings each word expresses and not only how many words are on a review.

Partial dependence plots show the relationship between a feature and the output of the model, but there is another method that provides a clearer picture of the specific effect each feature has on an individual prediction. This method is calculating the SHAP values. The results are presented on log-odds and the connection with the simpler probability interpretation is better understood on the below plot.

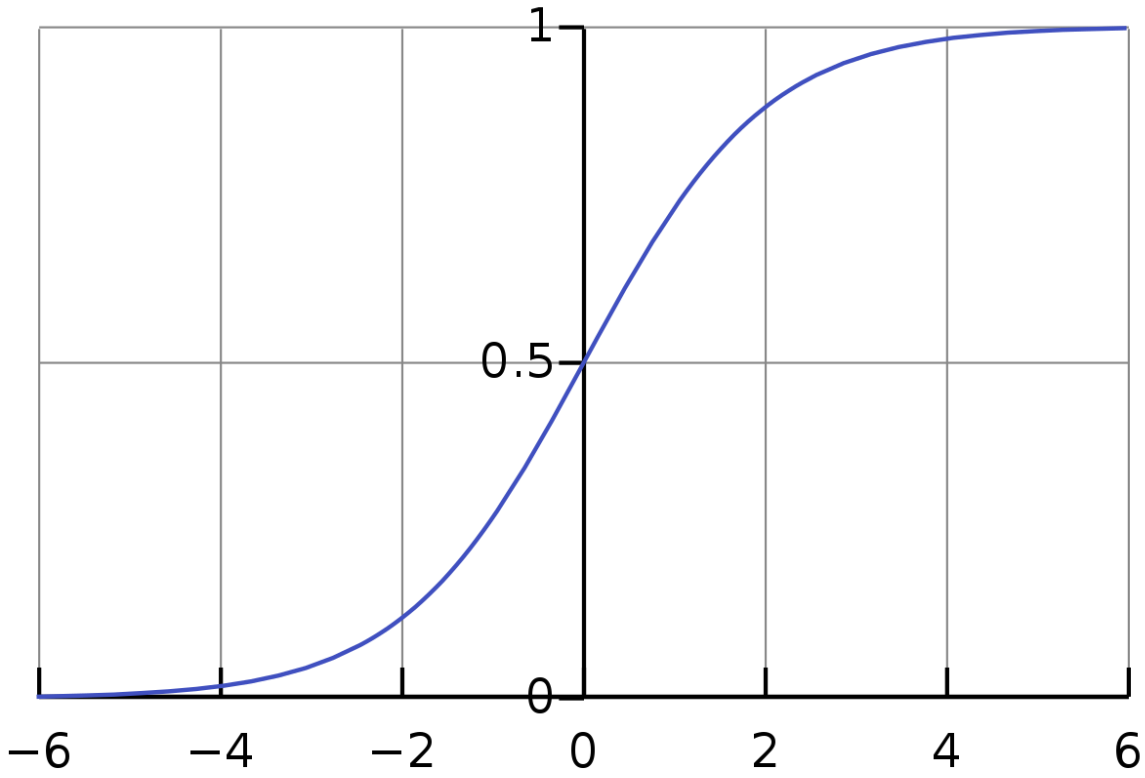

Figure 41: Standard logistic sigmoid function

The y-axis shows the probability, ranging from 0 to 1, while the x-axis shows the log of odds; the line shows the relationship between the two. When the a-axis is 0, this means that there is an even 50% chance for a positive classification (in our case, for the algorithm to predict an above average review score). Positive values on the x-axis mean higher probability for positive classification, while negative values on the log of odds mean lower probability. Further explaining how this relationship is calculated is out of the purposes of this dissertation.

We have explained the relationship between probabilities and log-odds, so the results of the SHAP values for our model should be better understood.

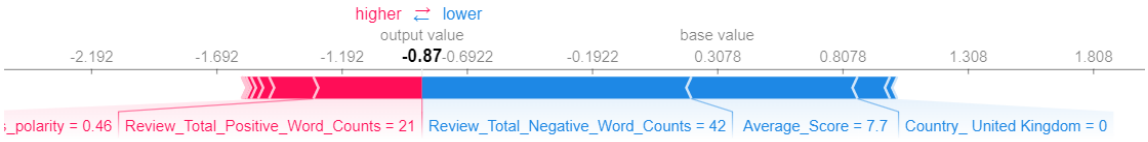

Figure 42: Individual prediction SHAP values

Figure 42 shows the SHAP values for a randomly selected individual prediction. Starting from a base value, which is the average prediction from our model, we can see how the specific prediction is reached [48]. Moreover, we can see the impact each feature had to drive the prediction to the final output value. Features in blue and pointing to the left, decrease the chances of a positive classification (the algorithm predicting that the score of the specific review will be higher than 8.4) and the length of each bar shows by how much. Pink bars represent the positive impact of the features.

In this case we can see that the biggest negative impact is from the negative words which are 42, followed by an average hotel score of 7.7 and the nationality of the reviewer that is not from the United Kingdom, which has a small impact. Positive word counts (21 words in this case) have the most positive effect, followed by positive polarity calculated at 0.46. This method allows us to see exactly why the algorithm made this prediction and how much each feature affected it. In our case we can see that the output value is -0.87, which means that the odds are in favor of a lower than 8.4 review score.

By generalizing this method to include 1,500 predictions of the model, we can see how a feature affects the final prediction over a range of its values.

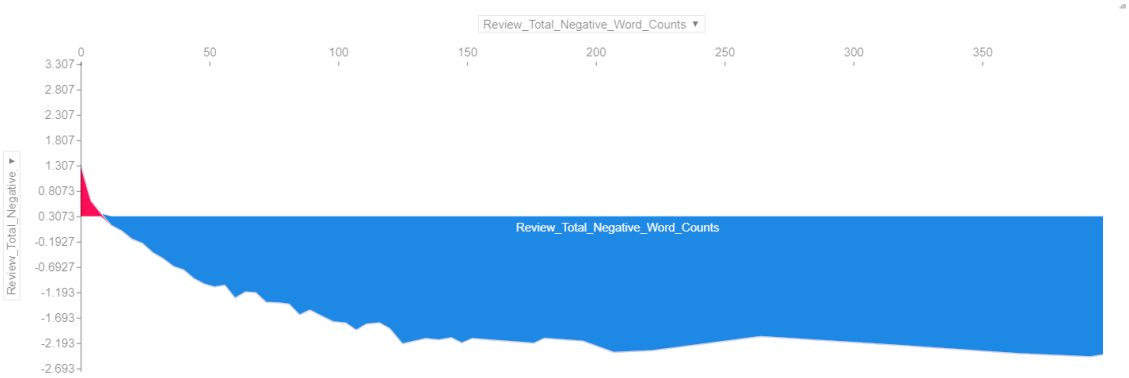

Figure 43: Negative word counts SHAP values

We can see in Figure 43 that the negative word counts affect the prediction towards a positive classification only if the words are less than 10. After that point, there is a rap-

idly increasing negative effect towards a negative classification for up to 120-130 words and then the negative effect keeps increasing but at a slower pace. What that means is that every extra negative word up to 120-130 has a bigger marginal effect on the chances of the review score being higher than 8.4, while additional negative words after that range decrease the chances but at a decreasing rate.

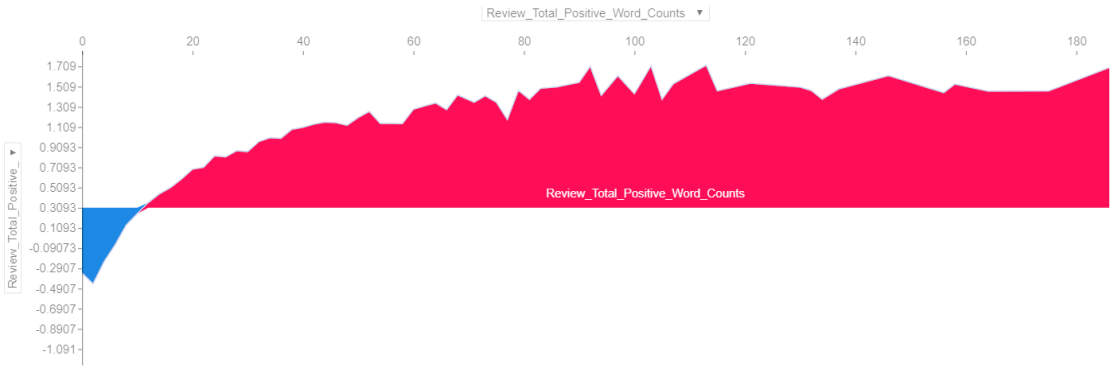

Figure 44: Positive word counts SHAP values

The same plot about the positive word counts (Figure 44) shows an opposite symmetrical pattern, where up to 10 positive words have a negative marginal effect on the chances of a predicted above average score, and after that number there is a positive marginal effect that keeps increasing with every additional word.

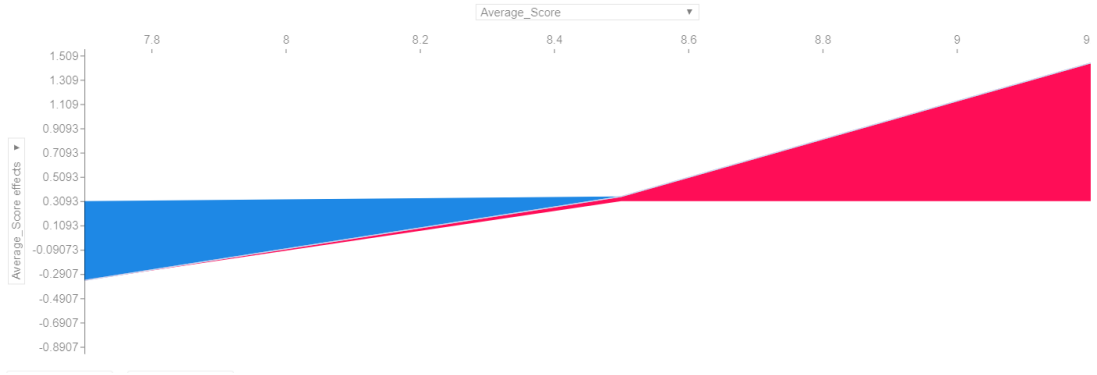

Figure 45: Average Score SHAP values

The Average Score plot shows us that the effect of the feature is decreasing the chances for an above average score up to 8.5; after this mark, the effect shifts to positive and keeps increasing until the higher average score observed (which is over 9).

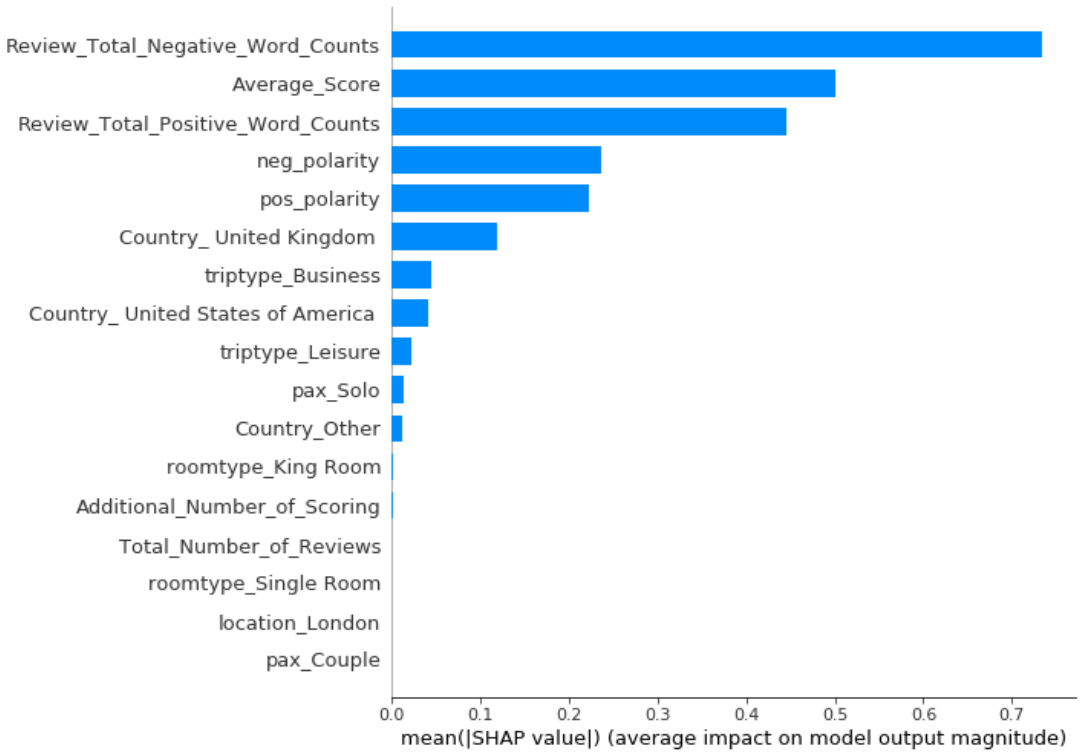

Figure 46: Average SHAP value per feature

In Figure 46 we can see the average SHAP value per feature, which essentially shows the marginal effect (on average) each feature has on the predictions. The pattern is identical to the feature importance we calculated earlier and confirms which are the most important features for this model. There are some features with 0 average SHAP value, which means they do not have any effect on the predictions and could be removed. While the average SHAP value plot shows us the average effect of each feature, it doesn't show us to which direction this effect pushes the prediction.

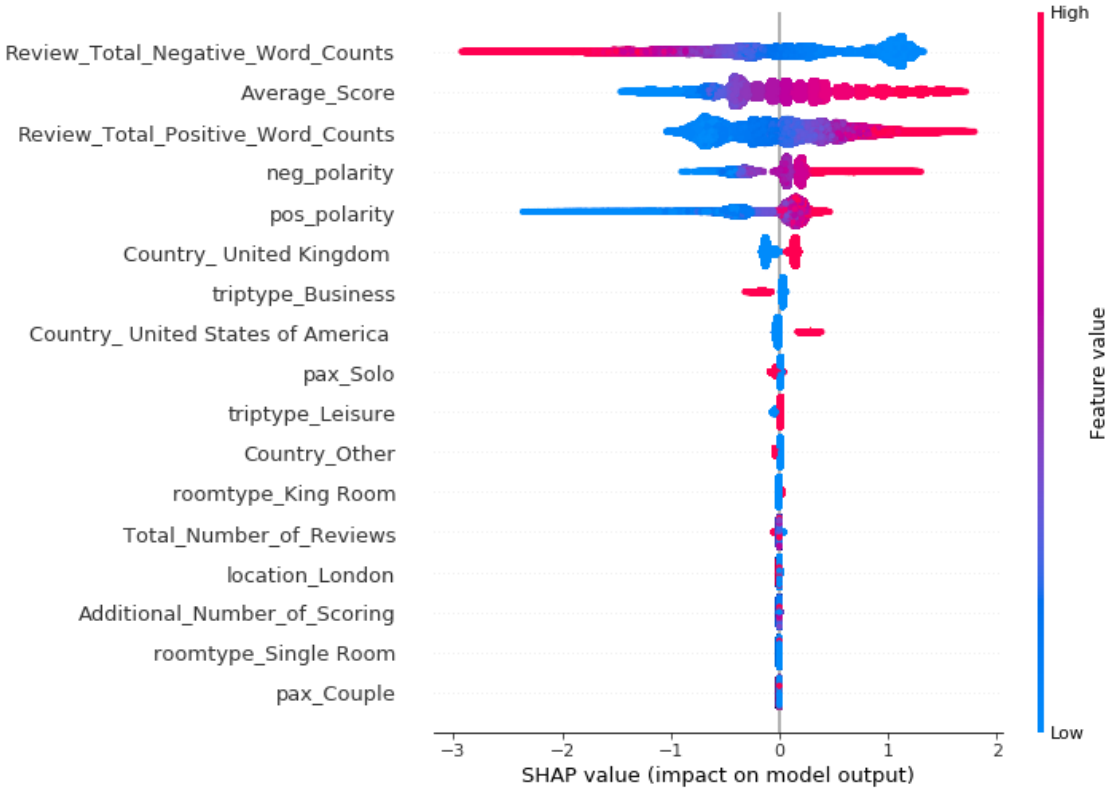

Figure 47: SHAP values summary

The best way to show the complete picture of each features effect, is to plot the summary of all the findings. Figure 47 is the summary of the SHAP values; this plot has a different and a bit more complex approach. The left y-axis has the name of each feature, while the right y-axis shows the range of the values in color ranges where blue is the lower value and red is the higher one (including all the in-between values). The x-axis shows the effect each feature's value has on the prediction. Every data point is represented with a colored dot, so the lines are thicker when there is a high number of points with similar marginal effects.

Using this plot, we can identify some very interesting patterns about the effects of the features and how the model makes its predictions. We can see that the most important feature, the negative word counts, affect the prediction towards a below average score when its values are higher. At the same time, there is a high concentration of observations on the positive side of the predictions. This means that there are a lot of reviews with very low number of negative words (or no negative words at all) and those reviews push the algorithm toward a positive classification. The average score feature is more balanced and its values are concentrated closer to 0 (on a symmetrical pattern both on the negative and positive sides of 0), which means that those values are closer to the av-

erage for the whole dataset and don't have a big marginal effect. Additionally, there are average scores that are higher or lower than the average, and those affect the prediction by a significant margin. The positive word counts have a positive relationship with the predictions, meaning that higher values lead to increased possibility of a positive classification. We can see that a lot of data points affect negatively the prediction, and the reason is that there are a lot of reviews with less than 10 positive words (and on the previous graph about the feature we established that up to 10 words the effect is negative). Polarities below zero, on the positive reviews, significantly decrease the chances of an above average score, while positive values increase the chances by a lower margin. On negative reviews, positive polarities have a bigger (positive) impact on the chances of a positive classification compared to the (negative) impact of a below zero polarity.

5.2 Predicting review scores

While the XGB classifier's goal was to predict if a reviewer will provide a score higher than 8.4, we will now use a model to try and predict the score for each review. The idea behind this model is similar to the classification model we used, but the prediction is different. The XGB regressor is based again on decision trees, but instead of predicting the class of a variable, it predicts the exact score.

Before we build the model, we need to select the features that have high enough correlations with the reviewer score.

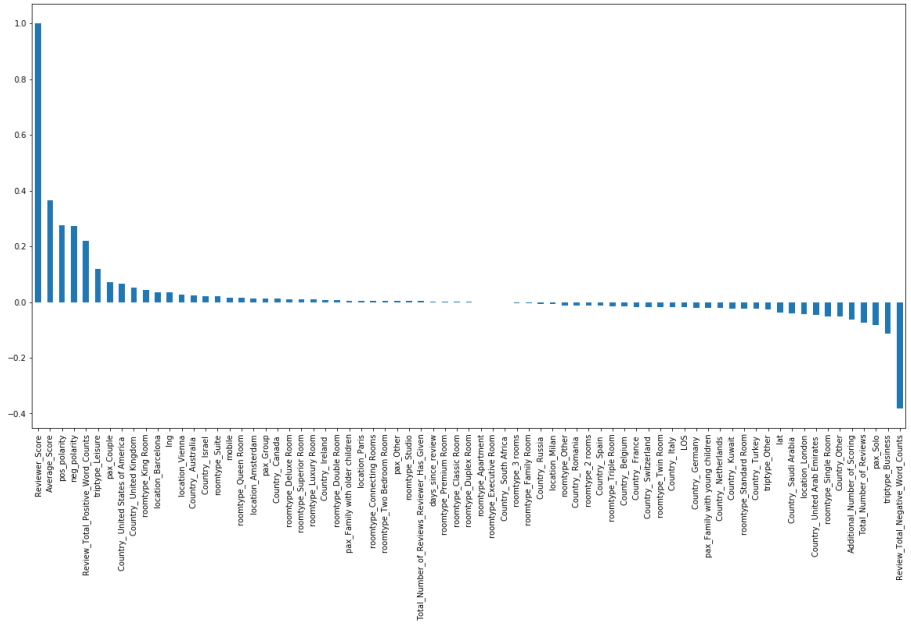

Figure 48: Reviewer score correlations

In Figure 48 we can see the correlation between the reviewer score and every feature on the dataset. We will select all the features that have an absolute correlation coefficient higher than 0.1, since lower correlations usually do not add value on the predictions.

```
0                        Reviewer_Score
1                         Average_Score
2                          pos_polarity
3                          neg_polarity
4     Review_Total_Positive_Word_Counts
5                       triptype_Leisure
71                     triptype_Business
72    Review_Total_Negative_Word_Counts
Name: index, dtype: object
```

Figure 49: XGB regressor selected features

Figure 49 shows a screen with the selected features, according to the filter mentioned. These features will be used to train and test the algorithm.

```
xgb_preds = xgb_model.predict(val_X)
print(mean_absolute_error(val_y, xgb_preds))
```

```
0.8788355779179996
```

Figure 50: Predictions' mean absolute error

Figure 50 presents a snippet of the code producing the mean absolute error of the predictions. Mean absolute error is one of the most common metrics used in machine learning, to measure the accuracy of a model's predictions. In this case, the mean absoluter error is 0.878, which means that the predictions of the model are off by this number, on average. Considering the complexity of the dataset and the number of features, the prediction accuracy is high enough for the purposes of this dissertation.

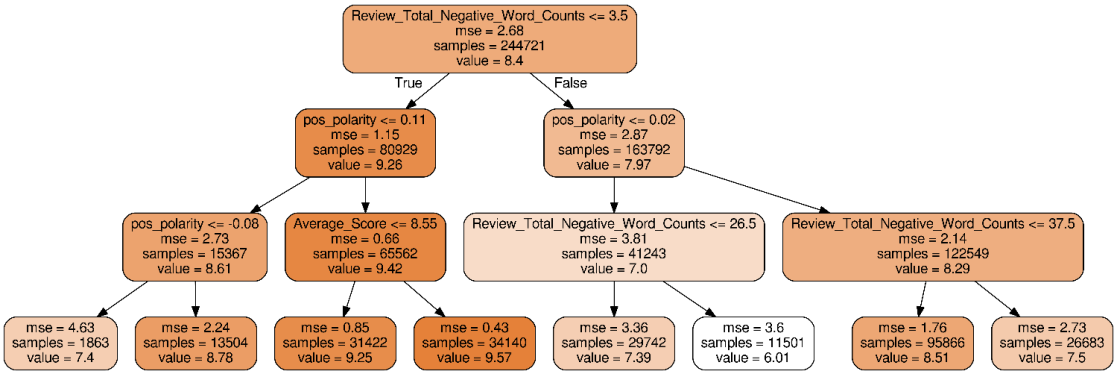

Figure 51: XGB regressor sample decision tree

Figure 51 illustrates a single decision tree used on our model, so we can better understand how the model makes a prediction. This is one of the many trees used and combined for a single prediction. Both the features and their order can be different on each tree but visualizing it can help us better understand the process of making a prediction. In this case the first decision is the number of negative words, specifically if there are less than 4 words on the negative review. The value part of each node shows the predicted score based on this decision and since this is the 'root' node, the value is the average for the whole dataset. The mse value is the mean squared error and it is a metric to measure the accuracy of each prediction, while the samples part is the number of observations on each node. We can extract some insights by examining the tree, for example a lower number of negative words counts predictably results on a higher reviewer score. Based on the condition of the root node, the next steps are differentiated. The most accurate prediction on this tree, is the leaf with lowest mean squared error at 0.43; the predicted reviewer score is 9.57. We can understand the conditions required for this prediction by examining the tree. The negative review should have less than 4 words and the polarity of the positive review should be higher than 0.11; finally, the average score of this hotel should be higher than 8.55. It should be noted that the decision tree checks these conditions is succession, as presented in Figure 51.

Weight	Feature
0.3329 ± 0.0035	Review_Total_Negative_Word_Counts
0.1512 ± 0.0011	Review_Total_Positive_Word_Counts
0.1149 ± 0.0026	pos_polarity
0.1046 ± 0.0015	Average_Score
0.0562 ± 0.0005	neg_polarity
0.0016 ± 0.0001	triptype_Business
0.0013 ± 0.0001	triptype_Leisure

Figure 52: XGB regressor permutation Importance

Figure 52 illustrates the permutation importance of each feature. In this model the accuracy is measured by the mean absolute error, so the weight represents the change of it. The negative word counts have the highest weight at 0.3329, which means that randomly changing the observations on this column, will result in an increase of the metric by this number. Positive word counts, positive polarity, average Score and negative polarity follow in terms of weight, while the type of trip doesn't affect the prediction much. It is important to see which feature has the biggest effect on the predictions, but we also need to see in which direction it affects it. SHAP values can provide a clear image on the marginal effect of each feature.

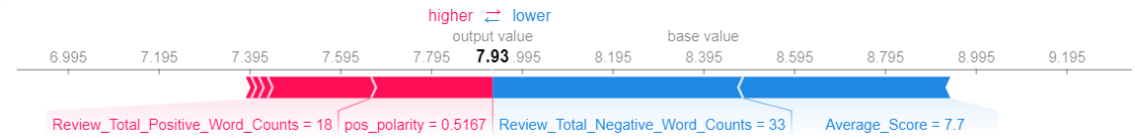

Figure 53: Single prediction SHAP values

Figure 53 shows how a single prediction is affected by the specific values of the features. Starting from the average reviewer score for the dataset (8.395), we can see the marginal effect of each feature. There are 33 negative words in this case which has the biggest negative effect on the prediction, alongside the Average Score of 7.7 that also pushes the predicted score to a lower number. Positive words, which are 18 in this case, alongside Positive polarity of 0.5167, push the final prediction to 7.93. This is an example that helps us see how each feature affects the predicted score. We can visualize the SHAP values for several predictions to better understand the marginal effect changes based on different values of the features.

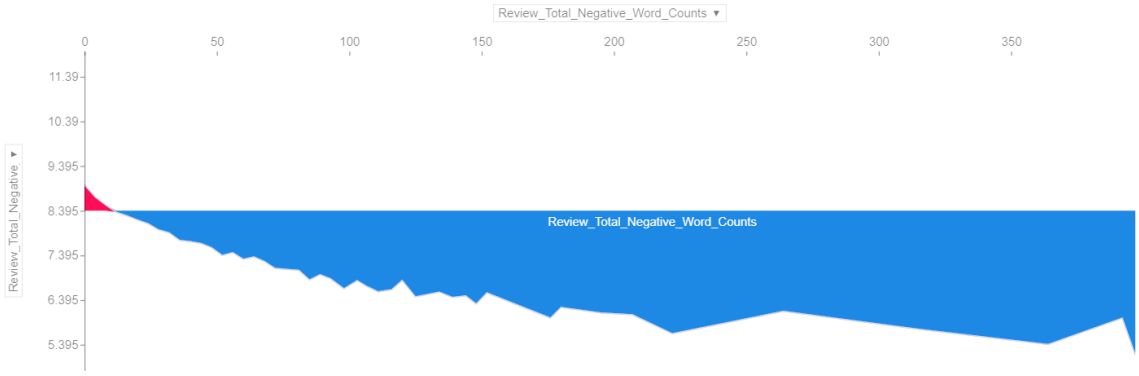

Figure 54: Negative reviews word counts SHAP values

The first feature we will examine is the negative word counts, since it is also weighted as the most important. In Figure 54, we can see that up to 10 words the impact is positive on the predicted score and it is slightly higher than the average. After the first 10 words on negative reviews, the predicted score keeps decreasing and can drop by maximum of 3 points to 5.395 when the negative words are close to 400.

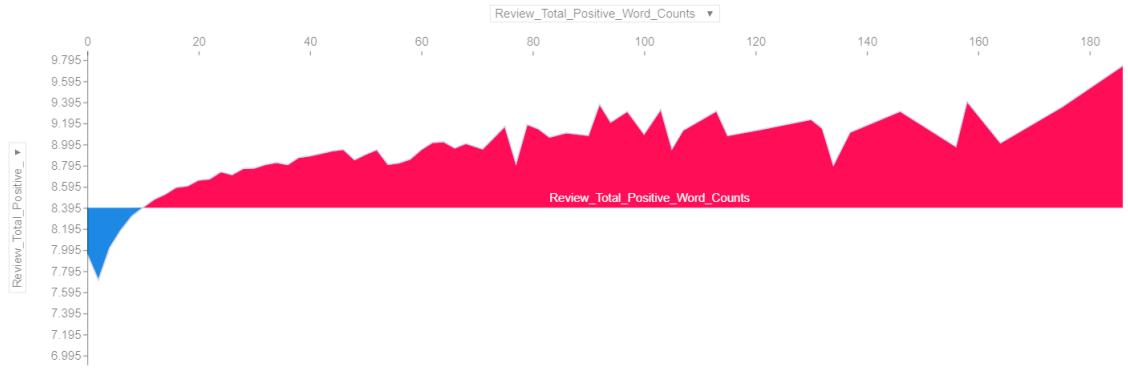

Figure 55: Positive reviews word counts SHAP values

The number of words on positive reviews has the opposite effect, compared to the negative review words. We can see that when the words on a positive review are less than 10, this affects negatively the predicted score and can decrease it down to 7.795. On the other hand, when there are more than 10 positive words, the predicted score increases up to 9.795, in instances that the words on the positive review are more than 180. We can see that the words used on either positive or negative reviews play an important role on the predicted score. Since more words on a review usually show stronger feelings from the reviewer, either negative or positive, this in turn affects the score provided by a reviewer.

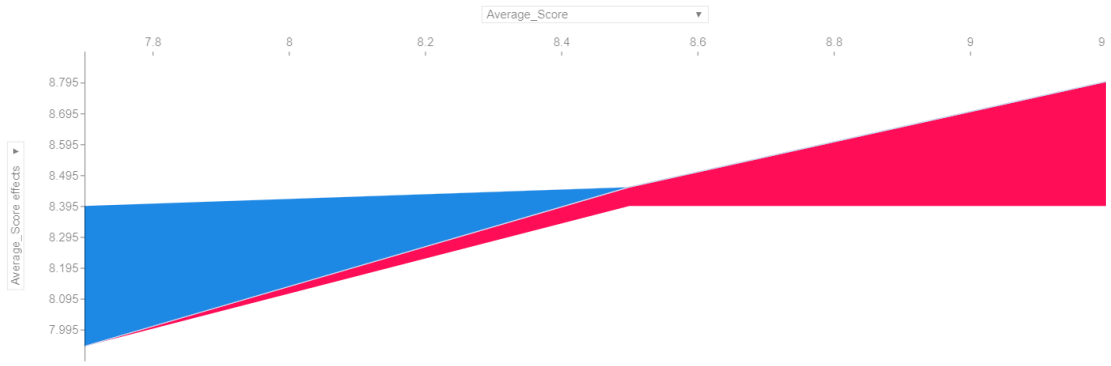

Figure 56: Hotel average score SHAP values

Figure 56 shows the effects of the Average score of the hotels on the predicted reviewer score. When the average hotel score is below 8.5 the predicted reviewer score decreases down to 7.995 at its lowest point. Higher than 8.5 average score has a positive effect on the predicted reviewer score. The highest average hotel score, which is close to 9.2, improves the predicted score by 0.4 compared to the average.

In this case, reviewers are affected about the average score of a hotel when they provide their own evaluation, which means that they are affected by other people's opinions about the hotel. As discussed on previous chapters, social proof has an impact on the minds of the reviewers and the perceived quality of the services they have received. This is an important insight and shows that hoteliers should strive to keep their hotel's average score as high as possible, since a low average can lead to even lower score in the future.

The next features we need to examine are the polarities of positive and negative reviews. Positive polarity refers to the calculated polarity of positive reviews and it shows how strong are the feelings of the reviewer, while the negative polarity examines the negative reviews in the same way.

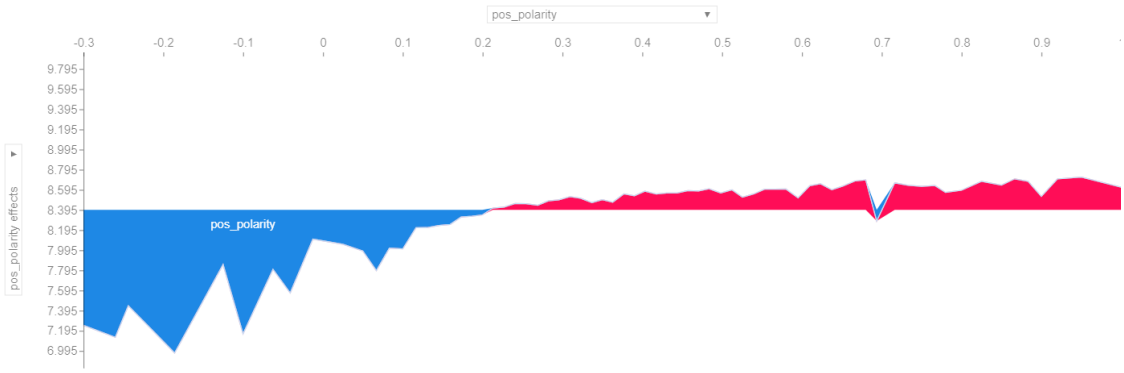

Figure 57: Positive polarity SHAP values

In Figure 57 we can see the positive polarity effects on the predicted score. While there are reviews that are labeled as positive ones, their calculated polarities are slightly negative. Those have a negative effect and can drop the predicted score down to 7, a maximum decrease of more than 1 point. This shows a pretty significant effect, while polarities higher than 0 do not have so strong effects, increasing the predicted score only by 0.4 on the best-case scenario.

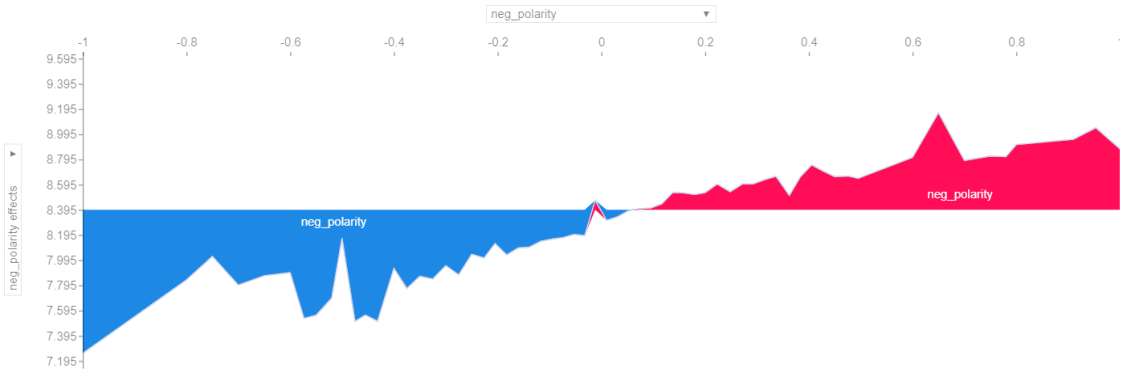

Figure 58: Negative polarity SHAP values

There is a big variation on the polarities of negative reviews, as their observations are covering the full range of polarity scores. A polarity of close to 0 is the point of balance, where there is minimal or no effect on the predicted score. Fully negative polarities (value of -1) drag the predicted score down to 7.2, while positive polarities can improve the predicted score up to 9. Considering that the average reviewer score is 8.4, the potential negative effects are again higher than the potential positive; the same pattern

Polarity is calculated by considering words, exclamation points and context. While these are not perfect metrics, combined with the marginal effect (SHAP values) they can show us that the feelings of the reviewers are strong predictors of the score they will provide.

The last features are the type of trip, whether it is a trip for business or for leisure. Both features are binary (0 if the condition is false and 1 if the condition is true), so there are only two cases to explore. For this reason, these two features will not be explored individually but they will be included in the SHAP values summary.

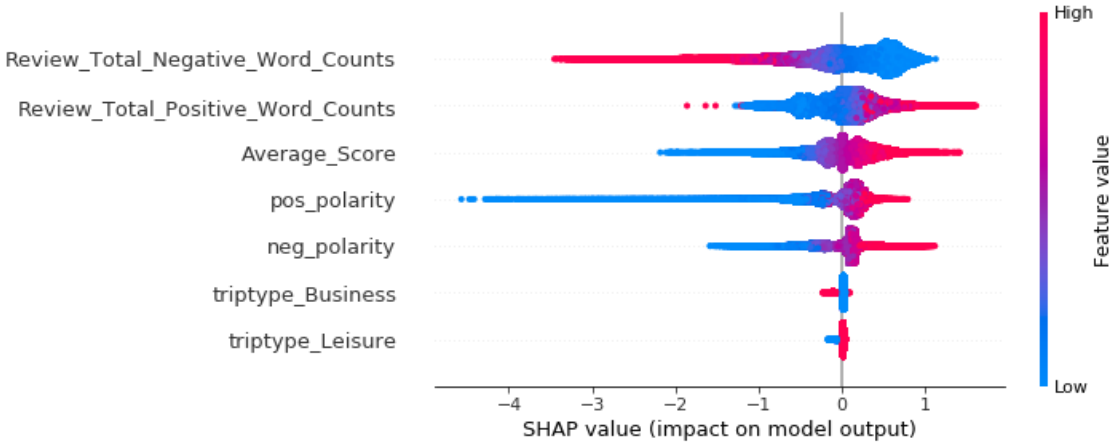

Figure 59: XGB regressor SHAP values summary

Figure 59 illustrates the SHAP values summary. The left y-axis has the names of the features; the right y-axis shows the variation of the values ranging from the highest value (red color) to the lowest (blue color), represent in a color range. The x-axis shows the marginal effect each feature has, in reviewer score points. We can see all the features and their effects, based on their values. The type of trip has minimal impact on our model, since all their points are gathered around 0. We can see that positive reviews' word counts and hotel average score have symmetrical distributions, since their negative and positive effects have similar maximum impact. Negative reviews' word counts have much stronger negative effect but in most cases their effect is positive. This happens because a lot of negative reviews have very few, or no words at all. Positive polarity has a lot of observations with a positive effect, when the emotions of the reviewers are positive; on the other hand, when the calculated polarity score is low, this feature has the highest negative effect of all the features. It can decrease the predicted score by up to 4 points which is the highest predicted decrease, followed by the negative reviews word counts that can cause a decrease of up to 3.5 points.

6 Conclusions

Through the previous chapters of this dissertation, we have explored the importance of online reviews and ratings. Previous studies have deeply discussed the major effect eWOM has on the decisions and behavior of the customers during all stages of the buying process. While the focus of many hotels is mainly on digital advertising, it would very wise to invest more on managing and improving their online reputation; online reviews from previous customers are considered the second most important source of information on a brand, behind only suggestions by friends and family. Consumers have a very high influence on each other, a fact that was further proved by our research on this paper. We showed that the existing average score of a hotel is one of the most important factors determining the score provided by a reviewer. Higher average score positively affects a new reviewer's score, since the heuristic of social proof creates a bias that is hard to overcome.

Positive and negative reviews have different effects on possible customers; the magnitude of the effects is higher on a negative review than a positive one. This difference is caused by the perceived trustworthiness of the negative reviews, since customers place a lot of trust in them and tend to ignore the possibility of a fake. On the other hand, positive reviews are not considered as trustworthy as negative ones, and a significant number of them is needed to affect a customer's decision. Our analysis further proved this point, since the number of words on negative reviews and the negative polarities have higher marginal effects on reviewer scores, compared to the positive reviews and polarities respectively. Furthermore, the negative words are the most important factor and can cause a big decrease on reviewers' scores as we showed while explaining the predictions of our models. This makes clear the fact that managers need to put much effort on improving their services and facing the concerns of their customers, in an attempt to avoid receiving negative word of mouth.

Improving a hotels online reputation can be achieved by focusing on two courses of action; using the feedback of customers to improve the quality of services and responding to the concerns of the customers. Managers can use analytics to extract insights from the

reviews of previous customers to identify common issues and concerns. Reviews are usually the most honest source of feedback and effectively using this feedback to change the operation of the hotel can lead to improved hotel performance and profitability. Additionally, studies have shown that management responses on bad reviews can lead to increased online bookings received. Since addressing the concerns of the customers is important, managers should focus on providing responses that represent the high quality of their brand. Systematically working on these two strategies can lead to improved online brand reputation and increased hotel performance as a result.

Achieving highly positive brand reputation is closely connected with the profitability of a hotel since previous research has proven that there is a correlation between positive online reviews of a hotel and important metrics as the average daily rate and the occupancy rate. An additional benefit of high ratings and positive word of mouth is that customers tend to pay more for high quality services; this means that a hotel can offer higher prices and improve its profitability by managing its online reputation effectively. Loyalty is also connected with positive eWOM, as customers tend to be more loyal to brands that are perceived to be of high quality by previous guests.

Our analysis provided many insights on the behavior of reviewers and the factors that affect their provided scores. These scores tend to reflect the emotions of the reviewers and the satisfaction about their stay on the hotel. As already mentioned, the most important feature in determining good scores is the number of words on negative reviews, followed by the number of words on positive reviews. These features have a strong connection with the emotions of the reviewers, and combined with the negative and positive polarities, determine the chances of a review being above average. Future research could use Natural Language Processing techniques to further investigate the connection between the feelings of the reviewers and their perception of a hotel's quality of services. We also showed that while the nationality, reason of trip (business or leisure), type of party and length of stay are not as important on our models' prediction, there are differences on the average reviewer scores between these categories. For example, we showed that solo travelers tend to provide lower scores, compared to couples, families and groups. This difference is closely associated with the fact that more than half of solo reviewers are staying for business purposes and business trips tend to have lower average scores than leisure trips.

Since internal hotel data like revenue, room nights and occupancy rates were not available for this analysis, future researches could combine public data like those scraped from booking.com with a hotel's sales data. This combination of data could help investigate the correlation between the scores and reviews with the sales data, and provide an even higher motive for hotel managers to use data analytics on their operations.

7 References

[1] Salmon, S. J., De Vet, E., Adriaanse, M. A., Fennis, B. M., Veltkamp, M., & De-Ridder, D. T. (2015). Social proof in the supermarket: Promoting healthy choicesunder low self-control conditions.Food Quality and Preference, 45, 113e120. https://doi.org/10.1016/j.foodqual.2015.06.004

[2] Cialdini, R. (2009). Influence: Science and practice. Boston, MA: Pearson Education. http://refhub.elsevier.com/S0261-5177(17)30237-6/sref11

[3] Jacobson, R. P., Mortensen, C. R., & Cialdini, R. B. (2011). Bodies obliged and unbound: Differentiated response tendencies for injunctive and descriptive social norms. Journal of Personality and Social Psychology, 100(3), 433 - 448. https://doi.org/10.1037/a0021470

[4] Amblee, N., & Bui, T. (2011). Harnessing the influence of social proof in online shopping: The effect of electronic word of mouth on sales of digital micro-products. International Journal of Electronic Commerce, 16(2), 91 - 114. https://doi.org/10.2753/JEC1086-4415160205

[5] Nielsen. (2012). State of the Media: The social media report. Retrieved: 12.07.17 from https://postmediavancouversun.files.wordpress.com/2012/12/nielsen-social-media-report-20122.pdf

[6] Pennington, D. C. (2000). Social cognition. London: Routledge modular psychology series. http://refhub.elsevier.com/S0261-5177(17)30237-6/sref40

[7] Papathanassis, A., & Knolle, F. (2011). Exploring the adoption and processing of online holiday reviews: A grounded theory approach. Tourism Management,32(2), 215e224. http://refhub.elsevier.com/S0261-5177(17)30237-6/sref37

[8] Chevalier, J., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online bookreviews.Journal of Marketing Research, 43(3), 345e354. https://doi.org/10.1509/jmkr.43.3.345

[9] Hu, N., Bose, I., Koh, N. S., & Liu, L. (2012). Manipulation of online reviews: Analysis of ratings, readability, and sentiments. Decision Support Systems, 52(3),674e684. https://dx.doi.org/10.1016/j.dss.2011.11.002

[10] Gavilan, Diana & Avello, Maria & Martinez, Gema. (2017). The influence of online ratings and reviews on hotel booking consideration. Tourism Management. 66. 53-61. 10.1016/j.tourman.2017.10.018.

[11] Carrasco, R. A., & Villar, P. (2012). A new model for linguistic summarization of heterogeneous data: An application to tourism web data sources. Soft Computing, 16(1), 135-151. https://doi.org/10.1007/s00500-011-0740-1

[12] Han, H. J., Mankad, S., Gavirneni, N., & Verma, R. (2016). What guests really think of your hotel: Text analytics of online customer reviews. Cornell Hospitality Report, 16(2), 3-17.

[13] Aral, S., & Walker, D. (2012). Identifying influential and susceptible members of social networks. Science, 337(6092), 337e341. https://doi.org/10.1126/science.1215842

[14] He, W., Tian, X., Tao, R., Zhang, W., Yan, G., & Akula, V. (2017). Application of social media analytics: A case of analyzing online hotel reviews. Online Information Review, 41(7), 921-935. https://doi.org/10.1108/oir-07-2016-0201

[15] Del Vecchio, P., Mele, G., Ndou, V., & Secundo, G. (2018). Creating value from social big data: Implications for smart tourism destinations. Information Processing & Management, 54(5), 847-860. https://doi.org/10.1016/j.ipm.2017.10.006

[16] Buhalis, D., and Leung, R., (2018), “Smart Hospitality – Interconnectivity and Interoperability towards an Ecosystem”, International Journal of Hospitality Management, Vol. 71, pp. 41-50. Buhalis, D., and Sinarta, Y., (2019), “Real-time co-creation and nowness service: Lessons from tourism and hospitality”, Journal of Travel and Tourism Marketing, Vol. 36, No. 5, pp. 563-582.

[17] Yoo, K. H., Sigala, M., & Gretzel, U. (2016). Exploring TripAdvisor. In R. Egger, I. Gula, & D. Walcher (Eds), Open tourism (pp. 239-255). Berlin, Germany: Springer-Verlag Berlin Heidelberg.

[18] Langley, P., & Leyshon, A. (2017). Platform capitalism: The intermediation and capitalization of digital economic circulation. Finance and Society, 3(1), 11-31.

[19] Kim, W. G., & Park, S. A. (2017). Social media review rating versus traditional customer satisfaction: Which one has more incremental predictive power in explaining hotel performance? International Journal of Contemporary Hospitality Management, 29(2), 784-802. https://doi.org/10.1108/ijchm-11-2015-0627

[20] Cantallops, A. S., & Salvi, F. (2014). New consumer behavior: A review of research on eWOM and hotels. International Journal of Hospitality Management, 36, 41-51. https://doi.org/10.1016/j.ijhm.2013.08.007

[21] Phillips, P., Barnes, S., Zigan, K., and Schegg, R. (2017), "Understanding the impact of online reviews on hotel performance: An empirical analysis", Journal of Travel Research, Vol. 56, No. 2, pp. 235-249.

[22] Sparks, B. A., & Browning, V. (2011). The impact of online reviews on hotel booking intentions and perception of trust. Tourism Management, 32(6), 1310-1323. https://doi.org/10.1016/j.tourman.2010.12.011

[23] Israeli, A.A. (2002), "Star rating and corporate affiliation: Their influence on room price and performance of hotels in Israel", International Journal of Hospitality Management, Vol. 21, No. 4, pp. 405-424.

[24] Torres, E.N., Singh, D., and Robertson-Ring, Al. (2015), "Consumer reviews and the creation of booking transaction value: Lessons from the hotel industry", International Journal of Hospitality Management, Vol. 50, pp. 77-83.

[25] Ye, Qiang; Gu, Bin; Chen, Wei; and Law, Rob, "Measuring the Value of Managerial Responses to Online Reviews - A Natural Experiment of Two Online Travel Agencies" (2008). ICIS 2008 Proceedings. Paper 115. http://aisel.aisnet.org/icis2008/115

[26] 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

[27] Sparks, B. A., & Browning, V. (2011). The impact of online reviews on hotel booking intentions and perception of trust. Tourism Management, 32(6), 1310-1323. https://doi.org/10.1016/j.tourman.2010.12.011

[28] Venners, Bill (13 January 2003). "The Making of Python". Artima Developer. Artima. Retrieved 22 March 2007.

[29] https://pandas.pydata.org/

[30] https://scikit-learn.org/stable/

[31] https://www.kaggle.com/

[32] https://jupyter.org/

[33] Rokach, L. (2010). "Ensemble-based classifiers". Artificial Intelligence Review. 33 (1–2): 1–39. doi:10.1007/s10462-009-9124-7.

[34] V. B. Vaghela, A. Ganatra and A. Thakkar, "Boost a Weak Learner to a Strong Learner Using Ensemble System Approach," 2009 IEEE International Advance Computing Conference, Patiala, doi: 10.1109/IADCC.2009.4809227

[35] Zhou Zhi-Hua (2012). Ensemble Methods: Foundations and Algorithms. Chapman and Hall/CRC. p. 23. ISBN 978-1439830031.

[36] Upton, Graham; Cook, Ian (1996). Understanding Statistics. Oxford University Press. p. 55. ISBN 0-19-914391-9.

[37] Zwillinger, D., Kokoska, S. (2000) CRC Standard Probability and Statistics Tables and Formulae, CRC Press. ISBN 1-58488-059-7 page 18.

[38] Maddala, G. S. (1992). "Outliers". Introduction to Econometrics (2nd ed.). New York: MacMillan. pp. 89. ISBN 978-0-02-374545-4.

[39] Jalal, Ahmed Adeeb (January 1, 2018). "Big data and intelligent software systems". International Journal of Knowledge-based and Intelligent Engineering Systems. 22 (3): 177–193. doi:10.3233/KES-180383 – via content.iospress.com.

[40] Allouch, Nada (2018). "Sentiment and Emotional Analysis: The Absolute Difference". Emojics Blog.

[41] https://textblob.readthedocs.io/en/dev/

[42] "SPSS Tutorials: Pearson Correlation". Retrieved 14 May 2017.

[43] Shalev-Shwartz, Shai; Ben-David, Shai (2014). "18. Decision Trees". Understanding Machine Learning. Cambridge University Press.

[44] Breiman L: Random Forests. Machine Learning 2001, 45: 5–32. 10.1023/A:1010933404324

[45] Breiman L: Random Forests. Machine Learning 2001, 45: 5–32. 10.1023/A:1010933404324

[46] https://www.kaggle.com/dansbecker/partial-plots

[47] Jerome H. 2001. “Greedy Function Approximation: A Gradient Boosting Machine.” Annals of Statistics. JSTOR, 1189–1232

[48] Lundberg, Lee 2017 : A Unified Approach to Interpreting Model Predictions