Machine Learning - Final Project

Report on Marriott Hotels

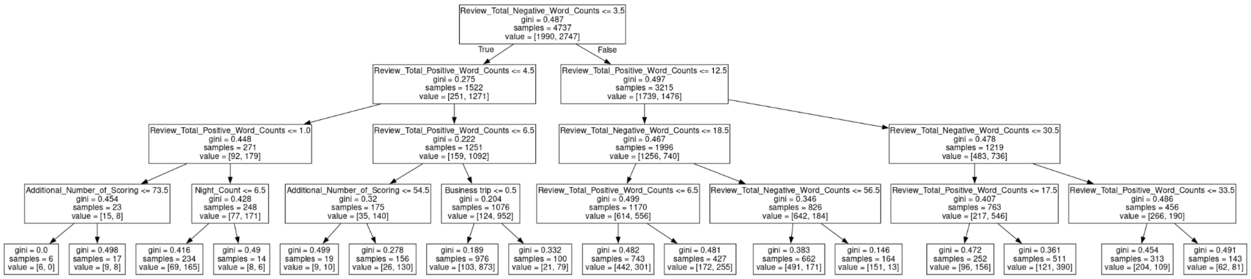
Managing a large, world-renowned hotel chain is most certainly a challenging task. The company has to make sure its customers are satisfied with the service, which includes performing well in most realms of hospitality: offering high-quality, clean rooms, a prestigious restaurant with adequate menu selection, good physical location, and so on. To gain a deeper insight into their customers, the management team of Marriott Hotels & Resorts contacted us to investigate the reviews left online. The reviews contain information on the nationality of each guest, as well as a separate positive and negative review text field. There is also data available on review date, the total number of reviews that customer has left across all hotels, and we also have tags available regarding some customer segmentation. These tags include items such as the number of nights stayed at the hotel, the purpose of the stay (business or leisure), family demographics (solo traveler, couple, or family with children), and whether the guest brought a pet along to the hotel.

For the analysis, we used the file "HotelReviews.csv", which contains reviews on a large number of hotels from numerous countries. As a first step, after using Pandas to read the csv file, we modified the DataFrame to contain only reviews on Marriott hotels, by filtering the variable *hotel_name* to contain the string "Marriott". With this, we have arrived at the core of the data to be analyzed: 6315 records of reviews about 25 of the Marriott hotels, located in London, Paris, and other major European cities.

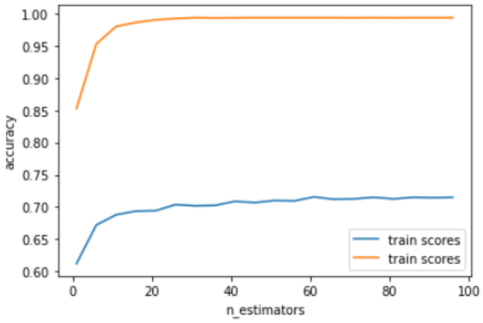
Our investigations were aimed around two central themes: firstly, we set out to examine the individual ratings against the average scores for each hotel to uncover any systematic patterns. Are there any nationality biases present in the reviews received, is there enough evidence to suggest that there are guests from some nationalities who rate our establishments higher than average, while maybe some citizens of some other countries might tend to give lower than average ratings? Using the information we have available on our guests, as mentioned before, such as purpose of travel, traveler demographics, or number of nights stayed at the hotel, which of these metrics are important when predicting how much a guest will enjoy their stay at a Marriott Hotel? Can we track hotel performance over time, and is there a way to cluster hotels into different groups based on whether they are improving their ratings or declining in quality, and would it be possible to pick out points in time where decisions have caused volatility in a hotel's rating? The following sections will present our analyses of the aforementioned problems.

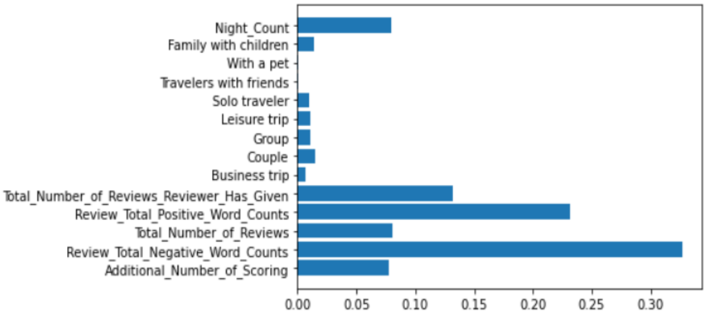
First, we created a new column in our DataFrame, equal to the delta between each reviewer's given score and the reviewed hotel's average rating. This variable *Score_Delta* measures not only whether the reviewer in question rated their stay higher or lower than the average score for the hotel, it also signals how much above or below the average the score was, signaling its strength.

For our first analysis, we combined our customer types with the contents of the reviews to uncover which of the features are deemed important when classifying reviews into those that reward scores above the hotel's average and those which give scores below average. First, we fitted a decision tree algorithm on the data. Following a grid search on the parameter *max_depth*, we set it to be 4 to avoid overfitting the model on the training data. With an accuracy score of 0.73, included below is the visualization of this decision tree. As can be seen from the first two levels of decision nodes, *Review_Total_Negative_Word_Count* and *Review_Total_Positive_Word_Count* were deemed to be the most important determinants of the score given at the end of the review. Analyzing the four-level deep decision tree, we see that features such as *Business_Trip* and *Night_Count* were also involved in the decision-making process, yielding some leaves with gini scores below 0.2.

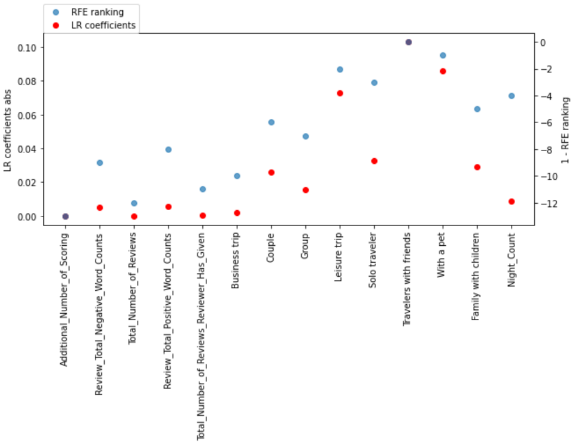


For this classification task, we also used the ensemble method of decision trees by applying a RandomForest classifier on the data. The plot on the left presents the training and test scores, along with the number of decision tree estimators used in the process. The random forest yields a score similar to that of the decision tree, reaching an accuracy of around 0.70 as more trees are included in the model. The bar plot on the right contains the feature importance for each of the features. The random forest algorithm shows *Review_Total_Negative_Word_Count* and *Review_Total_Positive_Word_Count* to be the two most important features when classifying higher/lower than average ratings.

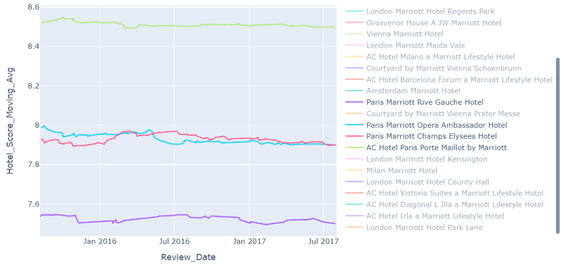


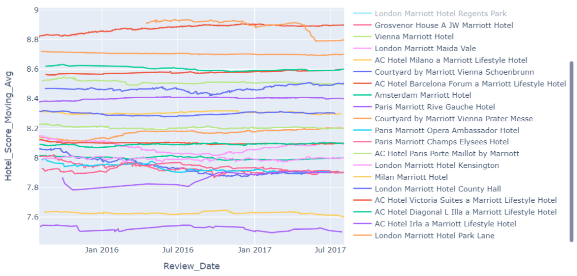


We continued by conducting recursive feature elimination for feature selection. This algorithm fits a model, in this case, a linear regression model, on smaller and smaller sets of features by eliminating the one it deems the weakest predictor at each recursion. Contrary to the binary classifiers, the customer type "Traveler with friends" is determined as the most important feature for predicting *Score_Delta*. The number of words in either the positive or negative review text field rank lower than expected, coming in 8th and 9th place.

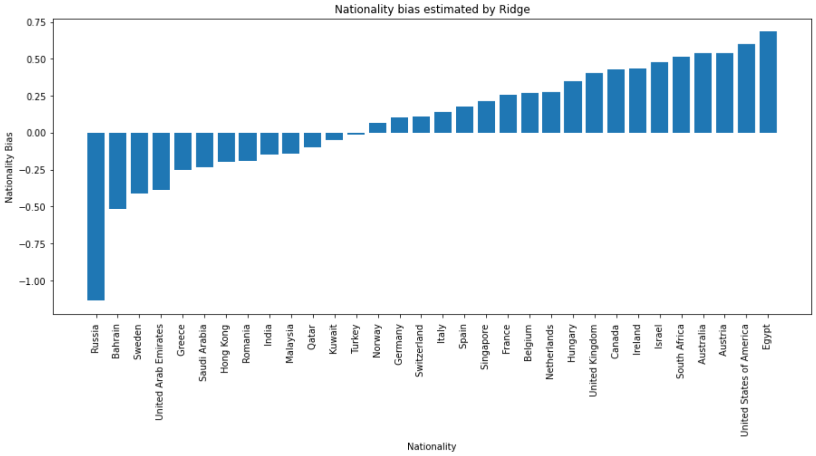


Next, we set out to explore hotel performance with time-series analysis for potential turning points in the average ratings of hotels. When plotting the rolling average ratings for each of the hotels, we see no major changes over the time period investigated. While there are smaller local points of interest, such as a small dip for the Paris Marriott Opera Ambassador Hotel before July 2016, we see that the hotel ratings stay fairly stable in the time period analyzed, but certain establishments can be distinguished from each other based on their ratings.



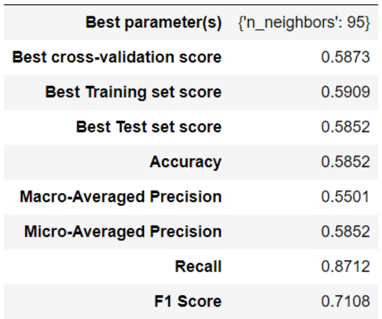


Next, we set out to investigate potential nationality biases. The 6315 reviews investigated were submitted by guests from 126 nationalities in total, however, many of them had sample sizes too small for meaningful analysis. Therefore, we have decided to investigate only those nationalities with a sample size larger than 30 reviews, leaving 32 nationalities in our DataFrame. Examining the 32 nationalities we most commonly host still provides us with a significantly large sample size, as well as enough nationalities to make reasonable deductions. We extracted the nationality data from our original DataFrame into a new nationality DataFrame that contained each nationality in the form of dummy variables, as well as the *Score_Delta* target variable. For this task, we used linear regression, Ridge regression, and Lasso regression. Having applied the regression models onto the nationality data, we plotted the regression coefficients for each nationality. Due to the large number of samples, all three models yielded similar estimates, and we will include the coefficients from the Ridge regression model.



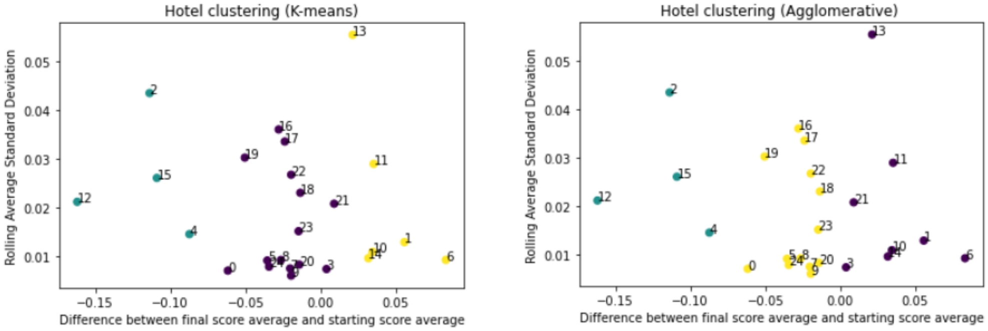
The output shows the estimated coefficients of each of our nationalities with a large enough sample size, in an increasing order. At the bottom of the list, we can see that guests from Russia are estimated to give ratings more than 1 point below the average for a hotel. In the middle, Turkish guests seem to be right on the average, with a coefficient of near 0. The most generous nationality seem to be Egyptians, with Americans as a close second, both of them achieving a coefficient above 0.5.

As a further attempt at classifying nationalities and higher/lower than average rating, we added a binary target variable, *Is_Score_Above_Avg*, based on the value of *Score_Delta.* The table below shows the results from a kNearestNeighbors classifier algorithm, which was applied after a grid search on the parameter of optimal number of neighbors. With an optimal number of 95 nearest neighbors, this classifier model had a test score of 0.58, but more importantly, it had a recall of 0.87, with an F1 score of 0.71. This provided us with some insight into nationalities, and the kNN algorithm seemingly performs relatively well with classifying guests into the category of "lower than average score" based on their nationality.



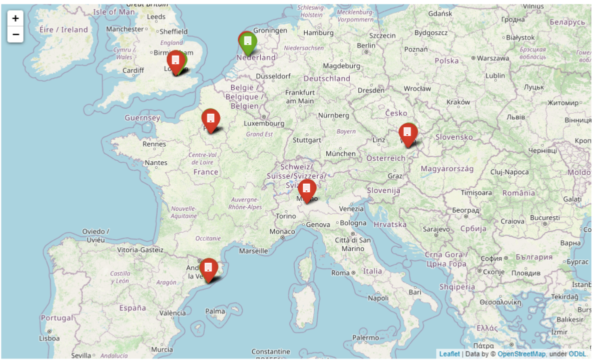
Best parameter(s)	{'n_neighbors': 95}
Best cross-validation score	0.5873
Best Training set score	0.5909
Best Test set score	0.5852
Accuracy	0.5852
Macro-Averaged Precision	0.5501
Micro-Averaged Precision	0.5852
Recall	0.8712
F1 Score	0.7108

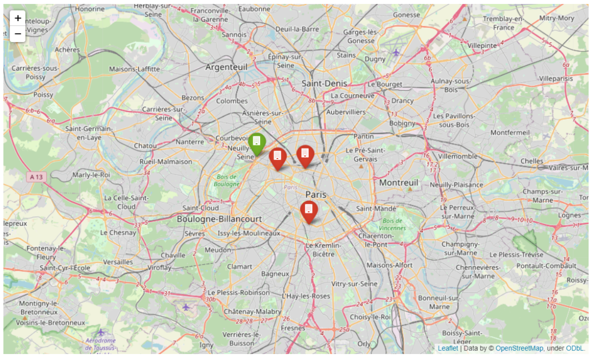
Having undertaken the previous time-series analysis task, we attempted to uncover the slight changes in the rolling average ratings of hotels. To do so, we have created two further numerical measures: one that took the difference between the final average rating of a hotel and the initial average rating of the hotel, and a second one measuring the standard deviation of the rolling average, aiming to show volatility of each hotel's average rating over time. Having created these measures, we applied both K-means and Agglomerative clustering algorithms to explore related hotels. With a chosen *n_clusters* of 3, the K-means algorithm created the three clusters, as can be seen on the plot below. The cluster marked with the color yellow contains only hotels that have final average ratings higher than they started with, signaling that the improvements they made throughout the examined time period were beneficial. These improvements could also possibly be useful to the cluster of hotels marked with blue, which all have a relatively high decline between the starting and final average ratings. In this case, hotels with higher volatility in their average ratings might indicate that at certain points in time, there were changes undertaken that have made a significant impact on their rating, whether that be positive or negative. These highly volatile hotels should be further analyzed to pinpoint times where their ratings increased or decreased, and if the causes behind these can be uncovered, the knowledge can be used in formulating strategy in the future. The Agglomerative method of clustering yielded somewhat similar clusters to K-Means, with the only expectation being that every single Marriott with a score increase over time has been classified into one group by this algorithm.



0	London Marriott Hotel Regents Park
1	London Marriott Hotel Marble Arch
2	London Marriott Hotel Park Lane
3	London Marriott Hotel Kensington
4	Paris Marriott Opera Ambassador Hotel
5	Paris Marriott Rive Gauche Hotel
6	London Marriott Hotel West India Quay
7	AC Hotel Paris Porte Maillot by Marriott
8	Paris Marriott Champs Elysees Hotel
9	AC Hotel Diagonal L Illa a Marriott Lifestyle Hotel
10	AC Hotel Victoria Suites a Marriott Lifestyle Hotel
11	AC Hotel Irla a Marriott Lifestyle Hotel
12	London Marriott Hotel Grosvenor Square
13	Courtyard by Marriott Amsterdam Arena Atlas
14	London Marriott Hotel County Hall
15	Grosvenor House A JW Marriott Hotel
16	Vienna Marriott Hotel
17	AC Hotel Barcelona Forum a Marriott Lifestyle Hotel
18	AC Hotel Sants a Marriott Lifestyle Hotel
19	London Marriott Maida Vale
20	Courtyard by Marriott Vienna Schoenbrunn
21	Amsterdam Marriott Hotel
22	Courtyard by Marriott Vienna Prater Messe
23	AC Hotel Milano a Marriott Lifestyle Hotel
24	Milan Marriott Hotel

The last direction of our analysis aimed at a location-wise exploration of the ratings of the Marriott Hotels, as well as a time-series analysis. We used the Folium library to display information on the map. Looking closely at the map of Europe, we can see that each of the cities Marriott operates in, there are hotels both above and below the continental average rating for Marriott Hotels, with an exception of Milan. Seemingly, both of Milan's hotels are rated below the Marriott-wide average, which should urge the management team responsible for Italy to make improvements. Further, zooming in on Paris, we uncover that three out of the four establishments in the city have an overall guest rating below the average. However, while the red markers could signal problems, we have to remember that at Marriott Hotels, the standard is set fairly high, meaning that, as we will see later, even a hotel rating of 8.0 is considered to be below average in the Marriott family.





Whether guest review scores can be considered a true measure of hotel performance greatly determines the next steps to be taken considering the results of this analysis. Looking back at the nationality bias analysis using regression coefficients, we can see a potential shortcut to higher guest ratings: simply focus on marketing for those nationalities who tend to be more generous with their ratings and try to minimize guests from the low-scoring countries such as Russia or Bahrain. However, we believe that Marriott Hotels & Resorts represents a higher standard, and therefore, we would advise the management team to further investigate the nationalities on the negative end of the scale to uncover any systematic reasons Marriott cannot properly cater to the need of the nationalities.

A quite trivial conclusion can be made about the important predictors of *Score_Delta*. The features *Review_Total_Negative_Word_Count* and *Review_Total_Positive_Word_Count.* Also looking at the random forest model, we uncovered that a higher number of words in the negative review text box greatly increases the probability of leaving a score below the average for a given hotel. Naturally, many individual perceptions and biases go into leaving a detailed review, but by providing guests with an experience that inspires them to leave long positive and short, if any, negative comments, we can inherently increase hotel scores.

Further conclusions can be drawn from the clustering, especially the Agglomerative algorithm. As mentioned before, this method grouped all hotels with an improved average score over time into one cluster, forming a semantically meaningful cluster. This cluster of hotels can be looked at as the best-case for Marriott Hotels & Resorts. Naturally, a main goal of any company is continuous improvement, and therefore, seeing that many hotels experienced declining scores over time needs to be addressed. On one hand, this can be done by pinpointing high-volatility, decreasing-quality hotels and see where score changes occurred over time, and whether any specific causes can be identified that triggered the decreases. If this can be done, the mentioned decisions, causes, should be reversed if possible. Secondly, though there is no one-size-fits-all solution, especially not in the hotel industry, we believe that the cluster with the color purple should be treated as the model, and the improvements made by them should be introduced at the hotels with decreasing ratings, if possible.

Overall, our hope is that this analysis will serve as an appropriate starting point for a company-wide assessment of the hotels operated in order to maximize not only the average ratings for hotels, but of course the guest experience as well.