STATISTICAL ANALYSIS OF FACTORS IMPACTING HOTEL REVIEW SCORES IN THE LAS VEGAS STRIP

ABSTRACT

Due to the well known nature of Las Vegas as a large hub for gambling and travel, a study was conducted to look at the factors which would best help to predict TripAdvisor ratings for hotels on the Las Vegas Strip. A high hotel rating would likely translate to higher foot traffic in the corresponding hotel and thus a higher income. The data used described stays in the hotels in the year 2015. After fitting a linear regression model, it was found that hotel name, specific types of visitors, and the existence of a pool or free internet were the most significant factors which could help predict the ratings score. Through a knowledge of these factors, hotels on the Vegas Strip could focus resources on the parts of hospitality service which would maximize review scores on average.

INTRODUCTION

The city of Las Vegas welcomes over 41 million visitors every year (Downtown Vegas Alliance). With its thousands of conventions, shows, and sporting events, it is easy to understand why it is such a popular tourist destination. There is also an obvious need to have a place for all of these visitors to stay when they are visiting. Going along with Las Vegas' reputation for being "over the top" in many areas, the city also has over half of the twenty largest hotels in the world (Downtown Vegas Alliance). But what makes some of these hotels more popular than others? Is there a way to predict how much a guest will enjoy a hotel? With thousands of rooms and countless amenities, Las Vegas hotels and casinos are incredibly complicated entities that are constantly competing for visitors. Knowing how to attract visitors and ensure they enjoy their stay is very beneficial to the hotel management and owners as they try to improve guests' experiences, in addition to improving efficiency for which sectors of customer service are best to focus efforts on. It is also helpful to visitors who want to know which amenities to look for when deciding where they want to stay, and which hotels best serve different types of guests.

Previous research on this topic has focused more on the visitors themselves and their attributes and demographics. Specifically, looking at the links between overall rating given to each hotel and what possible predictor variables can be used. A recent research done by Coehlo, Moro, and Rita in 2015 studied predictors which included TripAdvisor users' previous number of hotel reviews, traveler type, user country of origin, and hotel amenities among other factors. As noted by Coehlo, Moro, and Rita, "user features related to TripAdvisor membership experience play a key role in influencing the scores granted." Their analysis found that review scores are much more influenced by the reviewer themselves than by the actual hotels, and that other factors including day of the week of the review had large impacts as well. This led to a discussion about how hotel management can best respond to customers on sites like TripAdvisor to improve their likelihood of receiving a positive review score without much discussion on improvements to the hotels themselves. However, since online reviews tend to be unreliable, characteristics of the hotel are the most likely driving force of customer satisfaction. The goal of this project is to determine the combination of hotel characteristics that provides the highest level of customer satisfaction for visitors to the Las Vegas Strip.

METHODS

The dataset that will be used for this project is a collection of hotel review scores from TripAdvisor's website, and information about each of the hotels on the strip. All reviews were regarding stays in the year 2015, and the data was published online by Sérgio Moro. The data was collected from TripAdvisor, using built-in tools to narrow down review criteria and randomly select reviews that meet the specified characteristics. 24 reviews were randomly selected for each of the 21 hotels, two for each month of the year, giving a sample size of 504. Other information about the characteristics of the hotel is available on TripAdvisor, as well as the individual hotel's website.

The dataset includes 20 variables, some about the TripAdvisor user, and some about the hotels themselves. The variables about the user will not be used since these deal more with the individual reviewer, not the hotels. The main response variable that will be considered is the review score out of 5 given to the hotel, which will be counted as a categorical variable, split into unfavorable (1-3) and favorable (4-5). The quantitative response variables are hotel size (number of rooms in the hotel), and the number of helpful votes given to the review by other TripAdvisor users, which can be used to evaluate the trustworthiness of a review. The categorical variables about the hotel were binary variables relating to hotel amenities. These variables were if the hotel has a pool, gym, tennis court, spa, casino, and free internet. The other categorical variable that were considered for this project are hotel name, to see if some are just better liked than others, hotel stars (out of 5, as assigned by third-party raters), type of travelers (friends, business, couples, families, and solo), and period of stay (specifically what months the travelers' stay was in, which will be adjusted to a categorical variable for each of the four seasons), to analyze potential changes or trends in the overall score based on the time of year.

ANALYSIS

The regression model was created using a binary response variable. Review scores of a 1,2, or 3 out of 5 were given an outcome of "Unfavorable" and reviews with a score of 4 or 5 were given an outcome of "Favorable." Thus, a logistic regression model was used, with outcome as the response variable. The first model included nearly all the characteristics of the hotel. The dataset contains more than 500 observations, so there were enough degrees of freedom that could be sacrificed to create this model including multiple categorical variables. The initial model used pool, gym, tennis court, number of rooms, period of stay, spa, casino, free internet, hotel stars, hotel name, and traveler type as the predictor variables. Several of these variables had very high p values and did not appear significant, so the AIC selection technique was applied. The resulting model removed the variables gym, tennis court, number of rooms, spa, period of stay and casino. However, it left five explanatory variables: pool, free internet, hotel stars, traveler type, and hotel name. The Hosmer-Lemeshow test for this reduced model resulted in a P value of 0.9686, indicating that there was not evidence that the model was an inadequate fit for predicting the outcome of a review score based on the inputs. Since the model had been tested and there was no evidence to suggest that it was not an adequate fit for the data, the ROC curve was graphed.

The ROC curve for the reduced regression model that contained the previously mentioned five explanatory variables is shown in Figure 1. The area under the curve value was 0.7344, and the optimal cutpoint to balance the specificity and sensitivity of the model was roughly 0.75, in the light green region of the scale to the right of the graph. With area under the curve values of 0.5 meaning random classification and of 0.8 meaning the model is an excellent classifier, the value of 0.7344 found here was not excellent but was quite close to being so. Within the final regression equation for the model that determines the probability of a "Favorable" review, the cutpoint of 0.75 means that, on average, probabilities below that value correspond to an "Unfavorable" rating while probabilities above are "Favorable".

The final equation for the model was quite long since there were multiple categorical variables (hotel stars, traveler type, and hotel name) which led to a total of 24 slope predictors:

$$
\begin{aligned}
&ln\left(\frac{\pi(favorable)}{1-\pi(favorable)}\right) = -2.000 + 1.109(Pool) + 1.302(Free\ Internet) + 1.162\ (3.5\ stars) + 0.349\ (4\ stars) + 0.186(4.5\ stars) \\
&+ 0.869\ (5\ stars) + 0.763(Couple) - 0.033(Families) + 0.978(Friends) + 0.577(Solo) - 0.382(Casear's\ Palace) \\
&+ 0.703(Encore) - 0.656(Excalibur) - 0.051(Flamingo) - 0.275(Boulevard) + 1.27(Marriott\ Grand\ Chateau) \\
&- 0.343(Paris) - 0.179(The\ Cosmopolitan) - 0.087(The\ Palazzo) + 1.347(The\ Venetian) - 0.341(The\ Westin) \\
&- 0.188(Treasure\ Island) + 0.352(Trump\ International) + 0.559(Wynn)
\end{aligned}
$$

DISCUSSION

As can be seen from the regression output, there were a few meaningful takeaways from the model. First, there were some key factors that could help predict whether hotels are rated favorably or not. The presence of a pool and free internet both increased the likelihood of a positive review. Though some other features, like tennis courts, that one might think would have an impact on the likelihood of positive favorability rating were included in the initial model, these factors did not have a meaningful impact and were taken out of the final model. This suggests that niche add ons like tennis courts simply are not worth it for hotels on the Vegas Strip to include, and they should instead rely on pools and other staples of hotels to improve popularity. Unsurprisingly, hotels with higher stars were better liked as well, although the difference was not as high as could be expected. For 4&4.5 star ratings, the difference was not very large compared to the baseline 3 star rating, although the difference for 5 star ratings was larger. This indicates that it may be worth it for these nicer hotels to earn 5 stars and get this extra boost instead of settling for 4 or 4.5 stars. The model also indicated that business travelers and families were more likely to give unfavorable reviews. Some hotel amenities like casinos are not conducive to family environments, so hotels could try to target other groups to help boost their favorability. Finally, some hotels were just better liked than others due to factors like reputation that were not captured in this dataset, leading to differences based on hotel names.

One of the limitations to this model was the use of a binary response variable for favorable and unfavorable ratings. Ideally a proportional odds model would have been used as this would have allowed a ranked categorical response variable of review score, however this modeling process was beyond the scope of this project. Another limitation was the potential of correlation between the star rating and the pool and internet variables, as these factors can influence star ratings. Future research could focus on using categorical response variables, as well as empirical ratings for hotel features such as employee service. Location could also be an important factor to consider in future research, so other locations could be studied to see if the same results found here hold true or if the results are isolated to the Vegas Strip. An important thing to note is that the draw of hotels' name recognition and the expectation that comes from that could influence ratings, although that could not be accounted for in an analysis.

In conclusion, this analysis has found evidence that, on average, TripAdvisor reviews for hotels on the Vegas Strip were positively influenced by facts about the hotel like the star rating, the existence of the specific amenities pools and free internet, and hotel name (and likely the name recognition that comes with it). The reviews were also influenced by the type of traveler that stayed at the hotel with solo, couple, and friends being the types that, on average, resulted in higher review scores. However, there are some limitations to this analysis which include: not using a proportional odds model, potential predictor variable correlation between the hotel star rating and amenities, and not knowing hotel customer service data outside of select amenities. Overall however, the data suggests that hotel owners on the Vegas Strip should look into focusing on the areas listed above in order to give themselves the best foot forward to achieve more favorable ratings for their hotel, which will likely correspond to more success for the hotel.

REFERENCES

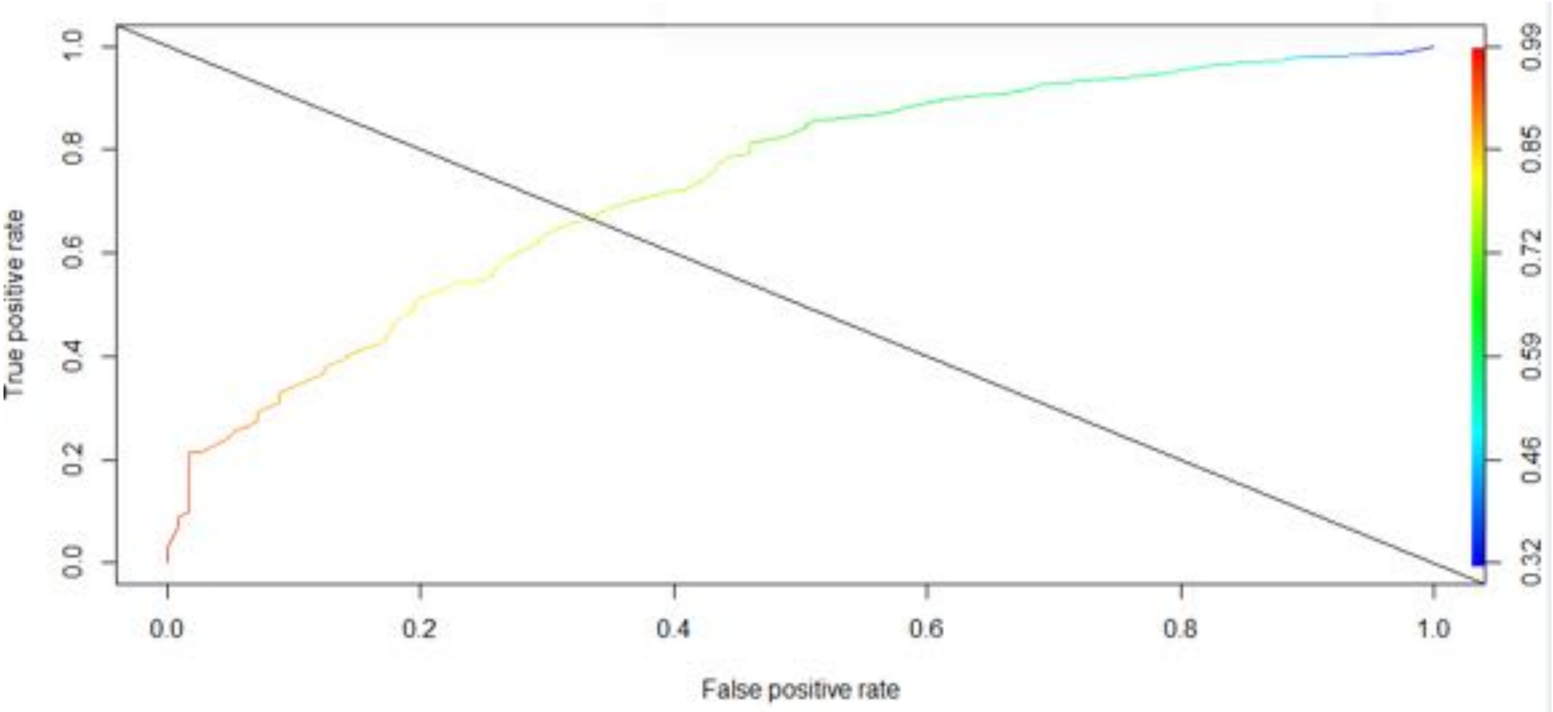
Downtown Vegas Alliance. “Fun Facts.” (https://downtown.vegas/visitors-guide/fun-facts/)

Moro, S., Rita, P., & Coelho, J. (2017). Stripping customers' feedback on hotels through data mining: The case of Las Vegas Strip. Tourism Management Perspectives, 23, 41-52. (https://novaresearch.unl.pt/en/publications/stripping-customers-feedback-on-hotels-through-data-mining-the-ca)

Moro, S., Rita, P. “Las Vegas Strip Dataset.” (2017). University of California Irvine Machine Learning Repository. (http://archive.ics.uci.edu/ml/datasets/Las+Vegas+Strip#)

APPENDIX

Figure 1 (ROC Curve):


True positive rate
False positive rate