Defining Luxury:

Determining What Factors Give Las Vegas Visitors A Good Time

By Peyton Camden, Christopher Barua, and Brandon Cook

Valparaiso University

I have neither given or received, nor have I tolerated others’ use of unauthorized aid

*Peyton Camden, Christopher Barua, Brandon Cook*

**Introduction**

The Las Vegas Strip is the heart of Sin City, with a 4.2 mile stretch full of hotels, casinos, restaurants, shops, performers, clubs, and more. No matter what attracts someone to Vegas, at the end of the night, all travellers need somewhere to stay comfortably and safely. Part of the allure of the entertainment capital of the world is that it is easy to find refuge in hotels within walking distance of all of the attractions to visit. For these hotels, their livelihood is based on customer service and providing a clean and relaxing stay for all guests, no matter where they come from or why. Online review platforms, such as TripAdvisor, Yelp, and Google Reviews, provide a forum for customers to rate their experience at a hotel and give specific comments. Reviews are critical to the likelihood of consumers choosing to book a particular hotel. If a hotel had a one star rating and the reviews say that it was dirty, the door did not lock, and the staff were rude, most people would avoid that hotel at all costs. On the contrary, a hotel with rave reviews and mentions of luxurious amenities and positive experiences will attract more new customers looking for a similar experience. Because of that, it is incredibly important for hotels to be aware of their reviews and ensure that they do everything possible to keep high reviews if they want a successful business. Good reviews tell hotel staff where to continue the practices they use and what customers like, and bad reviews highlight where the hotel can improve customer experience and consequently, ratings.

In the context of existing works relating to online consumer reviews, it is clear that there is a proven relationship between customer reviews and profitability. In a 2019 study of TripAdvisor ratings, it was found that every star in a given hotel’s rating is equal to a $280 per booking transaction (Jenq 2019), which financially rewards the hotel for positive reviews. Each review represented $0.12 per booking transaction (Jenq 2019), which also indicates that not only does the rating add value to a hotel, but the quantity of reviews available adds value as well, regardless of if the reviews are primarily positive or not. It is also found in another 2019 study of Yelp reviews that restaurants, despite being in a different sector of the hospitality industry, have four main categories of guest review content: taste, value, location, and . Each category has associations with either positive or negative reviews. For example, taste is associated with positive reviews and value is associated with negative reviews, as determined by category-specific keywords (Luo 2019). Because of the nature of the restaurant industry, and how reviews have similar effects on restaurant success as to hotel success, a similar conclusion might be able to be drawn about the hotels on the Las Vegas Strip. This leads to the question that this paper aims to address, which asks what the best categories (features) and modeling type are in order to predict Las Vegas hotel ratings.

A model that would predict hotel ratings would be extremely beneficial to a hotel company. By knowing the features most correlated to both high scoring reviews and low scoring reviews, there is a way for hotel staff to know what things matter most to their guests for a fulfilling stay at their hotel. Hotels, investors, and forecasters/analysts alike are able to better understand the likelihood of random negative reviews, and know if the negative reviews a hotel is getting are statistically significant and need to be addressed appropriately. As opposed to a lexicon approach that focuses on the content of the review, this approach focuses more on the presence of hotel amenities and reviewer statistics in numbers. An approach focused on the relationship of quantitative variables and review score of TripAdvisor reviews of Las Vegas hotels was done in a 2017 study by Moro et al, which shows that this question is worth pursuing if others see value in it as well. Their approach showed that the two most important features to determine hotel review score related to the TripAdvisor user, and the next most important feature was length of stay at the hotel (Moro et al 2017). Most studies focused around hotel reviews tend to use a lexicon approach to categorize and display the data in the content of the reviews, but a quantitative approach has more potential to contribute since less has been done with it. Moro et al placed the groundwork for this project, which will continue identifying TripAdvisor features most associated with predicted review scores, but using a more simplistic model, which is hypothesized to create a more accurate prediction.

**Data and Methods**

The dataset used was pulled from the University of California Irvine Machine Learning Repository website, where it was donated by Moro et al 2017 for students to use to explore machine learning techniques. The dataset is composed of 504 total reviews from 21 different hotels on the Las Vegas Strip, extracted from TripAdvisor. There are 20 different attributes that can be analyzed for importance in creating an accurate prediction of TripAdvisor scores based on the attributes of the reviewer and the hotel. The data came within a comma-separated file, which was looked through and one or two obvious errors in the data, such as a negative number in the ‘member year’ column, were manually changed. The comma separated file also had commas instead of decimals when indicating a fraction of a number in the ‘hotel stars’ column, which was manually changed within the comma separated file in order to import it as a .csv type file. The code was written and run in a computer lab in the Gellerson Engineering and Mathematics building on the Valparaiso University campus, with desktop Dell computers. The code was built on a cloud-based platform called Google Colab, which allowed the code to be run from any device with internet connection.

The model was built using python-based programming with many additional modules added to it. These modules include NumPy, pandas, scikit learn, matplotlib, seaborn, SMOTE, RandomOverSampling and RandomUnderSampling, graphviz, and scikit learn modules. The data was split and trained to prepare it, and values of ‘yes’ and ‘no’ were changed to 0 and 1 to make it easier for the machine to comprehend. The first classification ran on the dataset was a decision tree, to use as a simple beginning model to add on to and use to test the validity of the dataset. The features used in the first run of the decision tree were ‘Hotel stars’, ‘Pool’, ‘Free Internet’, ‘Spa’, and ‘Gym’, because those five were found to have the highest correlation to review score using the analysis software R. In R, a linear model was created with each of the 20 variables and a backwards elimination was done to determine the significance of each variable to the model. Each run, the least significant variable was removed and the model was run again, until only five variables remained. All variables were deemed to be significant when p-values were at 0.05 or below. The features used in the base model were applied to other modelling techniques, including random oversampling, random undersampling, a random forest, ada boost classification, and weighting.

For each modelling technique used, there were four measures of success that were calculated for each run of each model: precision, recall, accuracy, and F1 score. These were ratio values that were between a value of 0 and 1, with 1 being perfect and 0 indicating no presence of a measure. The use of confusion matrices were also used when looking at random over- and undersampling to determine how many of the predictions a model made fell into each score category, and ensure that the model was at least trying to predict all five types of scores.

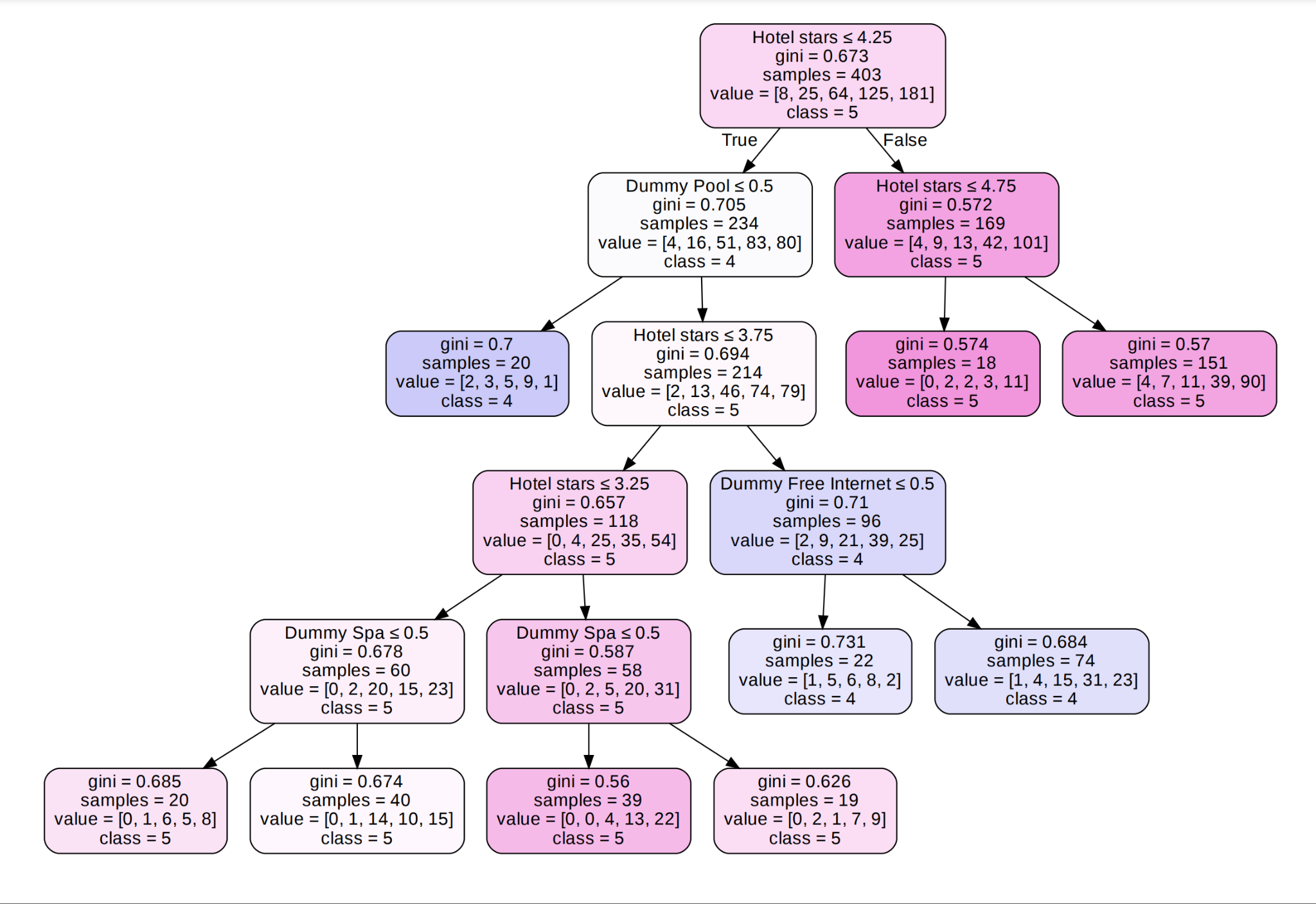
**Preliminary Results**

The earliest finding was that depending on how the data was split, there was a large range of accuracy, precision, recall, and F1 scores for even the most basic model. The decision tree model (Figure 1) proved to be consistently the most accurate and provided the most occurrences of predictions for all five score values. Accuracy was consistently within a range of 0.45-0.65, which far outperformed other modelling techniques that were used.

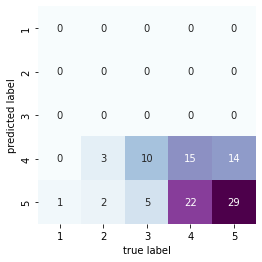
Logistic regression was used and though the accuracy was not the worst of all the techniques, with around a 0.4 accuracy score, it did not compare to the base model. When weighting was attempted, the results were so inaccurate the researchers were unsure if it could be trusted. A similar phenomenon was seen in the undersampling technique, with an accuracy around 0.1. A random forest model was desired, especially considering that the decision tree was the most accurate, but within the features for the model, not enough were included and each of the four scores of success were consistently identical to the results obtained for the base decision tree for any given split of the data. Oversampling was found to be in a range of accuracy values that was much wider than any of the other modelling methods, with an accuracy ranging from 0.15-0.5.

The primary results that have been obtained have led to the following conclusions. The decision tree, though a more simple and basic model, tends to outperform all other attempted modelling techniques in accuracy. However, the confusion matrix consistently showed that the decision tree was not predicting any cases of scores of one, two, or three (Figure 2). Oversampling the decision tree gave much less accurate and much less consistent results than the base decision tree model, but the confusion matrix consistently showed that even when inaccurate, it was predicting more rare cases in lower score values. It is worth noting that while many splits of the data provided very clear evidence that while the model may have predicted lower scores, the predicted scores were often incorrect (Figure 3).

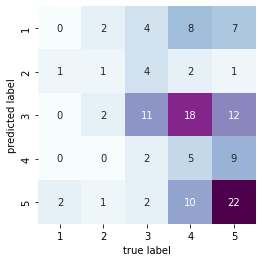
As we refine our model and continue to make choices on how to decide what model serves the interest of both our hypothesis and our potential clients best, it is clear that we need a model that can both predict lower scores but also predict all scores accurately. The initial accuracy of logistic regression suggests there may be a solution to this issue within a logistic type model. Within the dataset of 504 reviews, there are very few that have scores of one, two, or three. While statistically this may suggest that scores of one could be argued to be outliers, when it comes to clientele and the hotels we are conducting this study for, the lower scores will likely be of more value when trying to decide how to improve their business even further. This means the presence of lower scores in the model cannot be ignored, and finalizing the best trade off between accuracy and value of lower scores is where the remainder of the work lies. Cross-validation will help us in this process.



| Figure 1. A decision tree that begins with the requirement ‘Hotel stars <= 4.25’ feeds through the variables ‘Dummy Pool’, ‘Hotel stars’, ‘Dummy Free Internet’, and ‘Dummy Spa’. The ‘Dummy Gym’ variable that should be present is notably absent. The gini values for all results are relatively high, which indicates that the model is likely highly inaccurate. |
| --- |



| Figure 2. A confusion matrix of score predictions for a decision tree model shows twenty-nine accurately predicted scores of five, and fifteen accurately predicted scores of four. Twenty two true four scores were predicted as fives, and fourteen true five scores were predicted as scores of four. All scores of one, two, and three were incorrectly predicted as scores of four or five. |
| --- |



| Figure 3. A confusion matrix for an oversampled model shows predictions for all five types of scores, with a wide spread of accuracy. Twenty two true scores of five were accurately predicted, while only five true scores of four were accurately predicted. Eleven true scores of three were accurately predicted, and one true score of two was accurately identified. No true scores of one were predicted correctly. Many scores that were predicted to be threes were true fours or fives, which is where the bulk of the confusion occurred. |
| --- |

**References**

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*, 357–362. <https://doi.org/10.1038/s41586-020-2649-2>.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science &amp; Engineering*, *9*(3), 90–95.

Jenq, S. (2019) Exploring the Impact of a Hotel’s Rating and Number of Reviews through Online Transactions. *Open Access Library Journal*, 6, 1-15. doi: [10.4236/oalib.1105401](https://doi.org/10.4236/oalib.1105401).

Luo, Y., & Xu, X. (2019). Predicting the Helpfulness of Online Restaurant Reviews Using Different Machine Learning Algorithms: A Case Study of Yelp. *Sustainability*, *11*(19), 5254. MDPI AG. Retrieved from <http://dx.doi.org/10.3390/su11195254>.

McKinney, W., & others. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56).

Moro, Sérgio, et al. “Stripping Customers' Feedback on Hotels through Data Mining: The Case of Las Vegas Strip.” *Tourism Management Perspectives*, vol. 23, July 2017, pp. 41–52., <https://doi.org/10.1016/j.tmp.2017.04.003>.

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Waskom, M. Botvinnik, Olga Kane, Drew Hobson, Paul Lukauskas, Saulius Gemperline, David C Qalieh,Adel. (2017). *mwaskom/seaborn: v0.8.1 (September 2017)*. Zenodo. <https://doi.org/10.5281/zenodo.883859>.