Defining Luxury:

Determining What Factors Give Las Vegas Hotel Guests 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

This report details all elements of a semester-long project completed by Christopher Barua, Peyton Camden, and Brandon Cook that aimed to find the best predictors and predictor model for hotel scores using a dataset that collects TripAdvisor reviews from Las Vegas Strip hotels. The process is documented through code, presentations, and reports. The process documents the predictive power of three models, which are the decision tree classifier, the random forest classifier, and the linear regression model. The models were run through three different variations of the data, which were the original dataset, an oversampled version of the dataset, and a weighted version of the dataset.

Ultimately, this project found that the random forest classifier was the best model to use to predict the hotel scores, and the original data or the weighted data was the best balance to use by the highest accuracy and F1 score metrics. The features that were found to best predict the hotel scores were ‘Dummy Pool’, ‘Dummy Free Internet’, ‘Saturday’, ‘China’, ‘Business’, and ‘Excalibur Hotel and Casino’. However, it was made clear throughout the project that the dataset was not complete enough to translate the model results reliably to reality.

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: food/taste, value, location, and experience. Each category has associations with either positive or negative reviews, as determined by category-specific keywords. Though each category contained keywords that were both positive and negative, such as 'expensive' for a negative keyword and 'reasonable' for a positive keyword in the value category. The taste category is more highly associated with positive reviews and the value category is more highly 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 that a hotel gets 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 to make successful predictions.

One such 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 were related to 1) the statistics of the TripAdvisor user's helpfulness and longevity, and 2) the 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 numerical 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 accessible prediction. By examining different classification and regression models and comparing their cross validation and F1 scores, this project will identify the most effective machine learning algorithm and features to predict hotel scores for Las Vegas hotels by TripAdvisor users.

Data

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 between January and August of 2015. There are 20 different attributes (Figure 1) 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 in row 77, were manually changed to zero to represent less than one year. 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. After the data was found to be very imbalanced with a large skew towards hotel scores of four and five(Figure 2), a methodology was built to test if the imbalance was significant and how to account for it in the modeling.

Methodology

The model was built using Python-based programming with many additional modules added to it. These modules include NumPy, Pandas, Scikit-Learn, and Imbalanced Learn modules. The data were split and trained to prepare it, and values of ‘yes’ and ‘no’ were changed to zero and one to make it easier for the machine to comprehend. All code was built and run on a cloud-based platform called Google Colab, which allowed the code to be run from any device with internet connection.

After initial data cleaning, results were obtained that made it clear the data was not properly cleaned for the purposes intended. A dictionary had been made and had edited the data, which did not provide results for categorical variables. It was decided to use the label binarize function from Scikit-Learn on the original dataset, before it was cleaned and adjusted for negative values and zeroes. The original dataset was decided to be used for this to avoid using the dictionary made for the original methodology, which put weight on certain values based on their representative numerical value. Each column with categorical values were binarized and added to the original dataset, and the non-binarized column was dropped. For example, the ‘User continent’ feature had six different values throughout all rows, and once the values were binarized the original ‘User continent’ column was dropped and was replaced with the six categorical values as the feature name of the new columns. Because this was done with each categorical feature, the resulting dataset had 116 columns. The revised dataset had to undergo cleaning again, since the version of the data that had been binarized still had commas instead of decimals in the ‘Hotel stars’ column and a negative value in the ‘Member years’ column. Instead of manually changing all of the data, R was used to make these changes and the desired data was exported back into Colab.

In R, a linear model was created with each of the 20 columns and a backwards elimination was done to determine the significance of each variable to the model. A linear model was chosen due to restraints in the toolset of the researchers. Each run of the backwards elimination, the variable with the highest p-value was removed and the model was run again. All variables were deemed to be significant when p-values were at 0.05 or below, when only five variables remained. This method was used to find the best features for modelling, which resulted in the feature matrix including ‘Dummy Pool’, ‘Dummy Free Internet’, ‘Saturday’, ‘China’, ‘Business’, and ‘Excalibur Hotel and Casino’ being the most significant features, which was an increase from five to six features to use in training and testing the data. The binarized features ‘China’, ‘Business’, and ‘Excalibur Hotel and Casino’ come from the original columns named ‘User country’, ‘Traveler type’, and ‘Hotel name’.

An alternative feature matrix was made with more resulting features, which was also ran and compared to the first decision tree model. The features resulting from this matrix were ‘Dummy Pool’, Dummy Free Internet’, ‘Business’, ‘Families’, ‘China’, ‘Scotland’, ‘Spain’, ‘Dec-Feb’, ‘Excalibur Hotel and Casino’, ‘Hilton Grand Vacations at the Flamingo’, ‘Paris Las Vegas’, ‘The Westin las Vegas Hotel Casino and Spa’, ‘Treasure Island TI Hotel and Casino’, ‘December’, ‘February’, ‘January’, ‘Friday’, and ‘Saturday’. Many of these features are included in the original features selected, and classification metrics were run to compare the two sets of features. It was determined that the original six features yielded better accuracy and overall fit the data better than the alternative features, so the original features were used in conjunction with the cleaned dataset two for the remaining modelling.

The cleaned dataset two was then split, trained, tested, and validated. The data was also oversampled and weighted for use in the models. The values of the weight for each hotel score were a weight of 8.06 set for the hotel score value one, a weight of 4.2421 for a hotel score of two, a weight of 1.4392 for the hotel score of three, a weight of 0.6014 for the hotel score of four, and a weight of 0.4380 for a hotel score of five. The models run from the resampled cleaned dataset two were the decision tree classification, linear regression, logistic regression, Ada boost classification, random forest classification, and a K-means clustering method. For each model, the metrics precision, recall, accuracy, F1 score, Mean Absolute Error, Root Mean Squared Error, Mean Squared Error, and R^2 were calculated and compared. Confusion matrices were also used for each model to determine the success of the model in accurately predicting all five types of scores. Finally, each of the aforementioned models were cross-validated with the SVM module from Scikit-Learn for precision values to draw conclusive results to compare each model.

Results

As mentioned in the methodology section, the linear regression backwards elimination model was used to find the features that best predicted hotel scores. A p-value less than 0.05 was the threshold for finding these features significant. These features were found to be ‘Dummy Pool’, ‘Dummy Free Internet’, ‘Saturday’, ‘China’, ‘Business’, and ‘Excalibur Hotel and Casino’, as previously mentioned. This is a worthwhile result because the basis of the remaining results and scores are contingent upon these findings. Part of the goal of this study was to find the best prediction features from TripAdvisor reviews, which were not dependent upon a single categorical predictor.

The three models with the highest accuracy, F1 score, and mean absolute error (MAE) values were the decision tree, random forest, and linear regression models. These models were run on the original imbalanced data, an oversampled dataset, and a weighted data set. Each balance of the dataset used was modelled from the same split of the original data.

For the imbalanced dataset, the decision tree (Figure 3) had a cross-validated accuracy score of 0.503, a cross-validated F1 score of 0.25, and a MAE value of 0.729. For the oversampled dataset, the decision tree model had a cross-validated accuracy score of 0.397, a cross-validated F1 score of 0.399, and a MAE of 1.517. For the weighted dataset, the decision tree model had a cross-validated accuracy of 0.481, a cross-validated F1 score of 0.409, and a MAE of 0.729. A visual comparison of these metrics can be seen in Table 1. For the decision tree model of the imbalanced dataset, very few scores of two and three were predicted, and no scores of one were predicted. For this same model of the oversampled dataset, there were values of one and three predicted, but not two. None of the lower predicted hotel score values were accurately predicted for the oversampled dataset. For the weighted dataset, the hotel score predictions had correctly forecast values of two, three, four, and five, and incorrectly predicted some scores of one.

For the random forest model, the imbalanced dataset had a cross-validated accuracy score of 0.491, a cross-validated F1 score of 0.399, and a MAE of 0.746. For the oversampled dataset, the random forest model had a cross-validated accuracy score of 0.404, a cross-validated F1 score of 0.335, and a MAE of 1.472. The weighted dataset had a cross-validated accuracy of

0.481, a cross-validated F1 score of 0.412, and a MAE of 0.893. A visual comparison of these metrics can be seen in Table 2. For the imbalanced dataset random forest model, again there were no hotel scores of one predicted, and very few of two or three predicted (Figure 4). With the oversampled dataset, there were many correct predictions of fives, but very few correctly predicted four values, and no values predicted correctly lower than four. There were attempted predictions of one and three, similar to the decision tree model, but no predictions of scores of two, and none of the predictions for scores of one or three were correct. For the weighted data set, values had been predicted for all five possible scores, but scores of one and three were not accurately predicted (Figure 5).

For the linear regression model, only the MAE was calculated for the imbalanced and oversampled data sets. The imbalanced dataset had a MAE of 0.759, and the oversampled data set had a MAE of 1.592. The weighted data set was unable to have a MAE calculated for the linear regression model because the code cell was unable to be run. This is due to the code used not being able to use the weighted data to run a linear regression model. All of the predicted values were predicted to be decimals for the imbalanced and oversampled datasets. The linear regression model for the imbalanced dataset again only predicted a few values of scores below four. For the oversampled data, the linear regression similarly predicted mostly higher values, and not always correctly. For the weighted data, the linear regression model did not work very well, and no predictions were able to be made.

One issue that was prominent throughout the course of the study was that models did not frequently predict hotel scores lower than four correctly, especially when using the imbalanced data. The imbalanced dataset tended to not predict lower values at all, whereas the weighted or oversampled datasets tended to just inaccurately predict lower scores.

Of the three main models used, the random forest model tended to have the highest cross-validated accuracy scores for all three balances of data. The imbalanced dataset had a score of 0.491, while the weighted data set had a score of 0.481. These two values are close enough that the difference is not very significant, and thus both the imbalanced and weighted data sets could be used interchangeably to predict hotel scores with similar accuracy results.

Discussion

When it comes to the features that were best for the predictor values, it may have been beneficial to use more than just linear regression to perform the backwards elimination, like perhaps a decision tree backwards elimination for features of the decision tree models. As mentioned in the results, the features were not exclusively one type of categorical variable, such as the existence of certain amenities or different traveler types, which means a mix of different factors were considered to be significant to making the most accurate predictions possible. However, there may be bias present inherently within each of the features. For example, using Saturday as a predictor variable could mean there is a very rude staff member who consistently works Saturdays or a really close-knit team that works at the Excalibur Hotel and Casino.

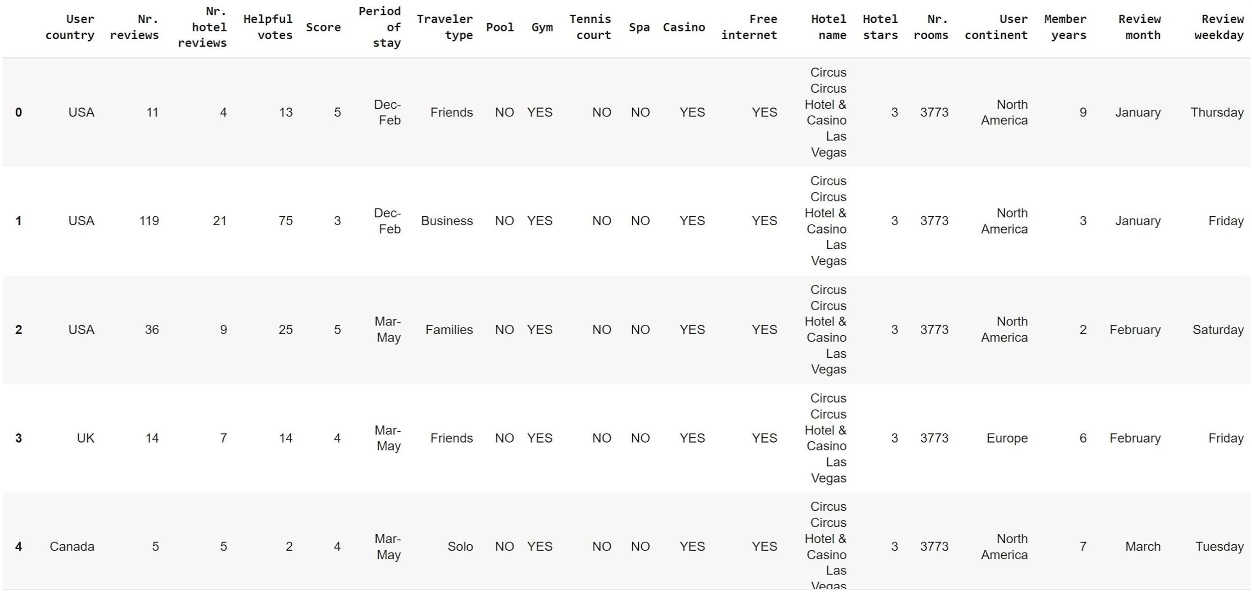
Overall, the random forest model performed most accurately compared to the decision tree model and linear regression, despite what the balance of the dataset was. All of the metrics used to compare the models were not very impressive, and this is likely due to how small of a data set the models were trained on and the limited amount of time the data truly encompassed. Accuracy is the metric most commonly used for the purposes of this study, but F1 scores are included to consider that the best model ultimately depends on the intended use. A model like this will help a potential consumer, in this case a Las Vegas Strip hotel, most effectively increase their ratings by knowing what factors contribute most to a lower score, which is not necessarily depicted by an accuracy score. However, the F1 scores were generally close enough to the same to consider the difference almost negligible.

What seemed to be the recurring theme and issue that was found while comparing confusion matrices for each balance of the data and each model is that the scores below four were predicted infrequently, and predicted correctly almost never. Ultimately, the metrics that represent these models are not very good and show that the issue likely lies within the dataset used. Only twenty two reviews for each of the twenty one hotels were in the entire dataset, and were collected only from one site and for seven months. There is likely so much inherent bias within the dataset that the most effective model, the random forest model, would have to be used on a different or much larger dataset to truly be able to represent what the hotels experience in reality.

Conclusion

In conclusion, three different classification/regression models were analysed for the Moro et al dataset on Las Vegas Strip hotel reviews from TripAdvisor. Firstly, the features were binarized and run through a backwards elimination to find the most significant predictors for the models. A decision tree, random forest, and linear regression model of the original dataset, oversampled dataset, and weighted dataset were created to compare nine different results of predictions of hotel scores by users. The results were compared using accuracy, F1 score, and mean absolute error. The final results found that a random forest classification model was the most accurate and had the highest F1 score, and that either imbalanced or weighted datasets could be used with this model with similar end prediction results. This study concludes that though an interesting and potentially effective methodology was created to use TripAdvisor reviews for hotels to predict their scores, the original dataset was likely too limited and thus too biased to truly be applicable in this moment to the studied hotels.

Figures



	User country	Nr. reviews	Nr. hotel reviews	Helpful votes	Score	Period of stay	Traveler type	Pool	Gym	Tennis court	Spa	Casino	Free internet	Hotel name	Hotel stars	Nr. rooms	User continent	Member years	Review month	Review weekday
0	USA	11	4	13	5	Dec-Feb	Friends	NO	YES	NO	NO	YES	YES	Circus Circus Hotel & Casino Las Vegas	3	3773	North America	9	January	Thursday
1	USA	119	21	75	3	Dec-Feb	Business	NO	YES	NO	NO	YES	YES	Circus Circus Hotel & Casino Las Vegas	3	3773	North America	3	January	Friday
2	USA	36	9	25	5	Mar-May	Families	NO	YES	NO	NO	YES	YES	Circus Circus Hotel & Casino Las Vegas	3	3773	North America	2	February	Saturday
3	UK	14	7	14	4	Mar-May	Friends	NO	YES	NO	NO	YES	YES	Circus Circus Hotel & Casino Las Vegas	3	3773	Europe	6	February	Friday
4	Canada	5	5	2	4	Mar-May	Solo	NO	YES	NO	NO	YES	YES	Circus Circus Hotel & Casino Las Vegas	3	3773	North America	7	March	Tuesday

Figure 1. This table is the original dataset as it was downloaded from the UCI Machine Learning Repository, with 20 columns. There are numerical, categorical, and yes/no values in each column, instead of a consistent data reporting type. In the ‘Review Weekday’ column on the far right, it can be seen that not all of the categorical variables are even the same, which shows that there are many nuances to this dataset that will need to be worked on before any significant analysis or machine learning can begin.

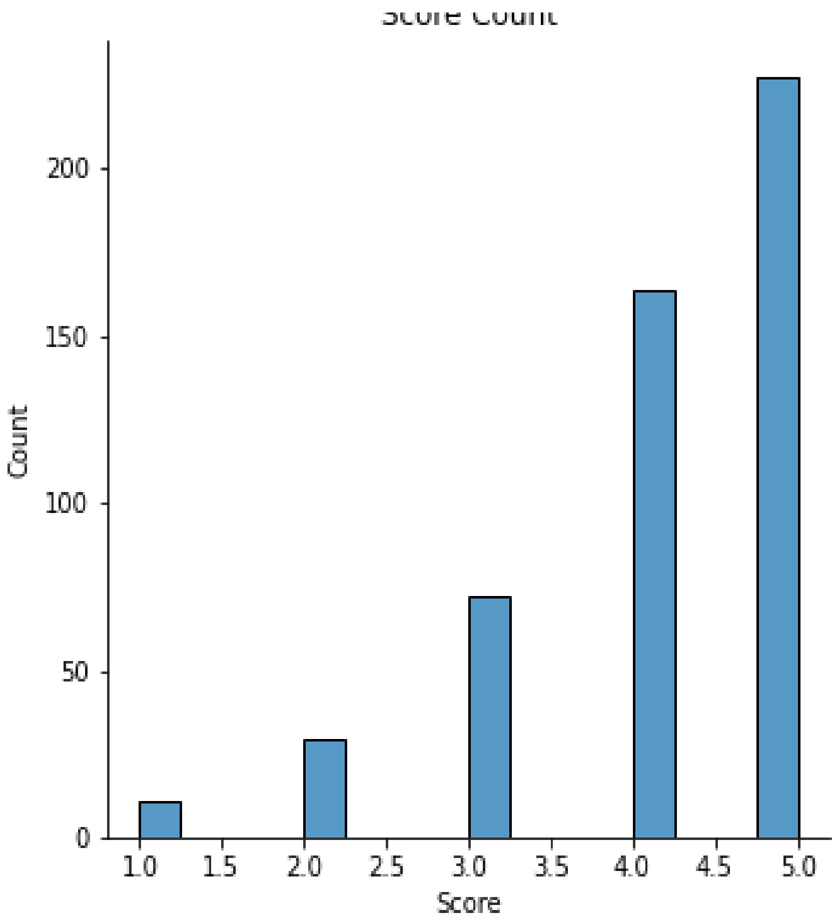


Figure 2. A histogram bar chart shows the overall distribution and count of hotel scores within the dataset. Counts are recorded in increments of fifty on the left side of the chart, with scores reported on the x-axis in 0.5 increments from 1.0 to 5.0. As scores were only reported in the dataset as whole numbers, counts from the n whole number to the n+0.5 tick represent the count of scores within the bin from whole number to whole number, for example the count from 1.0-1.5 represents the count for the bin from 1.0-2.0.

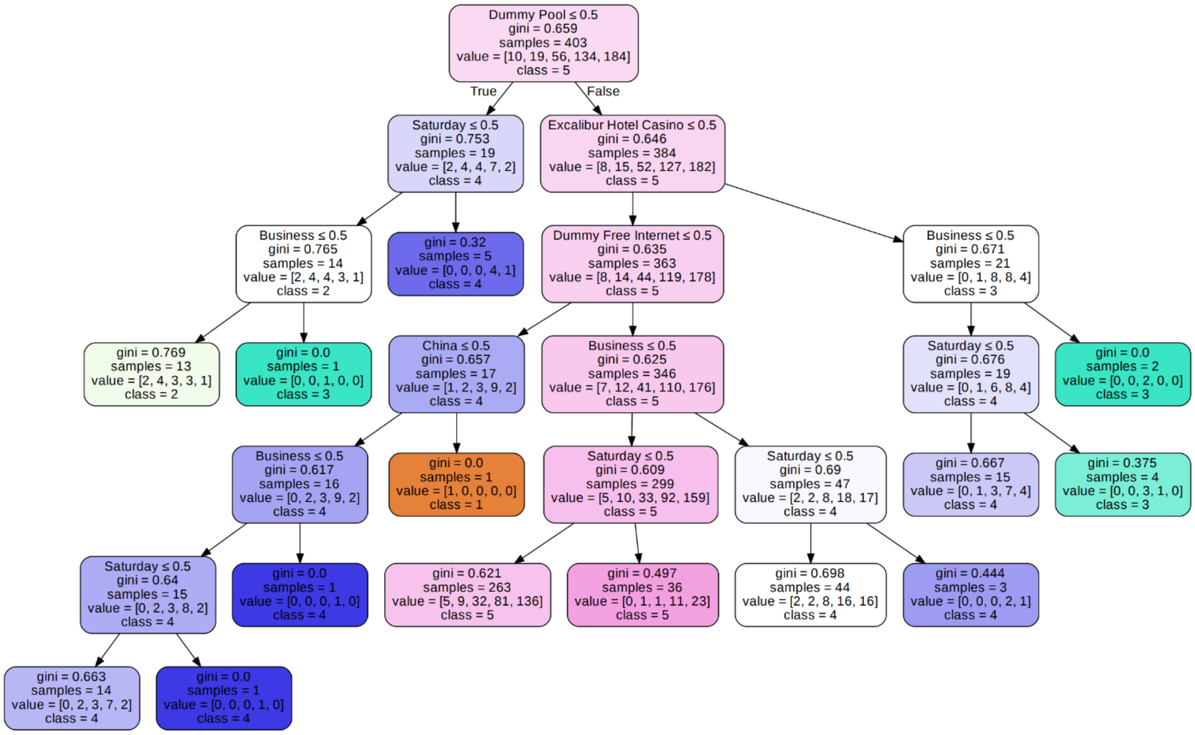


Figure 3. A decision tree using the features ‘Dummy Pool’, ‘Saturday’, ‘Excalibur Hotel and Casino’, ‘Business’, ‘Dummy Free Internet’, and ‘China’ attempts to predict the TripAdvisor score given by users. It is to be read as a flow chart, following a true or false pattern that leads to a prediction about the TripAdvisor score, following the arrows that lead to a true statement for each data point.

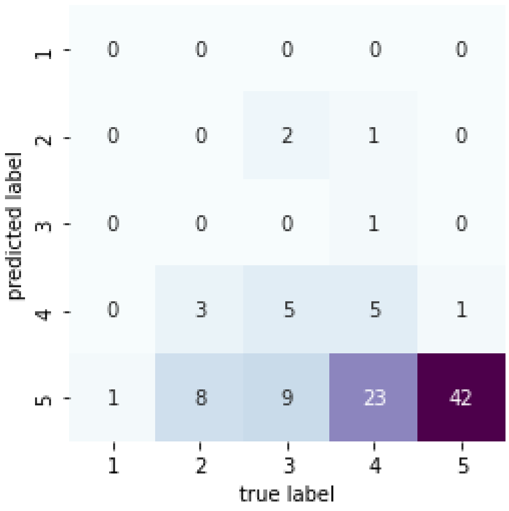


Figure 4. A confusion matrix for the random forest model of the imbalanced data set shows there are very few scores lower than four being predicted. The bottom axis shows the true value of the score being predicted, whereas the left side axis shows the predicted value of the score. A highly accurate model has highlighted colors along the left diagonal, in boxes where predicted and true scores meet. A confusion matrix can also show how many values are being predicted for each score value, which can be valuable when ensuring the model is running correctly.

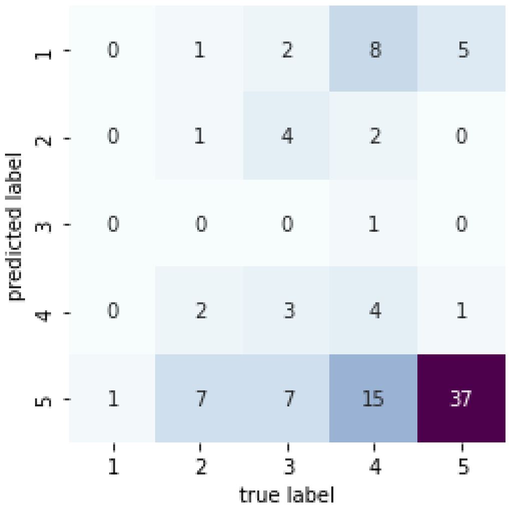


Figure 5. Same as Figure 3, but this confusion matrix is for a weighted random forest model.

Tables

	Original Data	Oversampled Data	Weighted Data
Accuracy	0.503	0.397	0.481
F1 Score	0.25	0.399	0.409
Mean Absolute Error	0.729	1.517	0.729

Table 1. The accuracy, F1 score, and mean absolute error are shown for the decision tree classification model for the original dataset, the oversampled dataset, and the weighted dataset.

	Original Data	Oversampled Data	Weighted Data
Accuracy	0.491	0.404	0.481
F1 Score	0.399	0.335	0.412
Mean Absolute Error	0.746	1.472	0.893

Table 2. The accuracy, F1 score, and mean absolute error are shown for the random forest classification model for the original dataset, the oversampled dataset, and the weighted dataset.

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.

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.

UCI Machine Learning Repository: Las Vegas Strip Data Set. UCI Machine Learning Repository: Las Vegas Strip data set. (n.d.). Retrieved December 1, 2021, from https://archive.ics.uci.edu/ml/datasets/Las+Vegas+Strip.

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.