Final Project Report

Las Vegas Strip Hotel Reviews Analysis

Ivan Shapirov

4/27/2020

Business Understanding

Business Problem

My project topic is on maximizing the reviews left on hotels in the Las Vegas Strip Area. The business problem is that a hotel corporation has opened a new hotel in the Las Vegas Strip, but they have received more negative reviews than their competitors, causing the number of tourists who select their hotel to decrease. The hotel is looking for some sort of data analytics solution so that they are able to increase the reviews left by customers on their hotels, increasing the number of customers who choose their hotel.

Dataset

The dataset has 504 instances, with each instance representing a review left by a customer on a hotel in the Las Vegas Strip area on TripAdvisor. There are 20 features for each review.

5 of the features give information about the reviewer: Country of origin of the reviewer, continent of the reviewer, total number of reviews left by the reviewer, total number of hotel reviews left by the reviewer, and the number of years they were a member of TripAdvisor.

6 of the features were about the review: The number of "Helpful votes" (People who found the review helpful), the period of stay, the traveler type (solo, business, friends, etc.), the month the review was posted, the weekday the review was posted, and finally the score left by the reviewer.

3 of the features were about the hotel: The stars the hotel has, the number of rooms in the hotel, and the name of the Hotel.

The next 6 features give whether or not a specific amenity was provided: Pool, gym, tennis court, spa, casino, and free internet.

This dataset can be used in a couple of ways. One way is to focus more on which kind of customers are more likely to leave a good review. For example, it could be analyzed whether families are more likely to give good reviews compared to people traveling solo. This could give the company their target audience which they could advertise to. Another possibility is to focus more on the services provided by each hotel and seeing which services contribute more to a positive review and which services result in a more negative review.

Proposed Analytics Solution

A solution to this problem is to analyze the reviews left by consumers at the hotels in the area and see which factors most contribute to the customer leaving a favorable review. The target variable will be the rating score left by the customer. The ratings on the dataset range from 1 to 5, so the goal will be to find what makes reviewers leave a higher score.

Data Exploration and Preprocessing

Analytics Base Table

Descriptive Features

User country	Nr. Reviews	Nr. Hotel Reviews	Helpful Votes	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	Score

Score is the target feature.

Data Quality Report

Numerical features

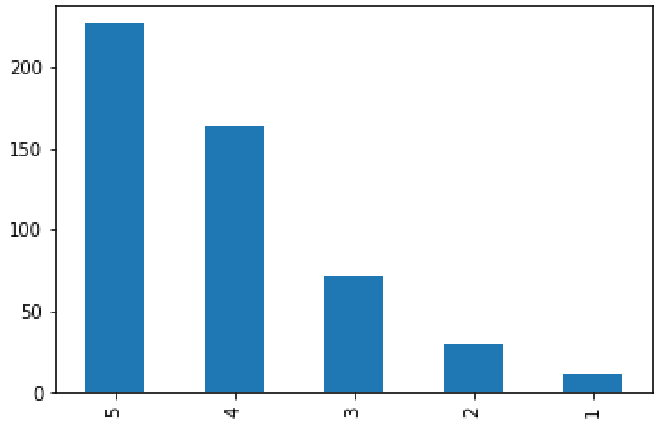
	Nr. Reviews	Nr. Hotel Reviews	Helpful Votes	Nr. Rooms	Member Years
Count	504	504	504	504	504
% Missing	0	0	0	0	0
Mean	48.131	16.024	31.752	2196.381	0.768
Standard Dev.	74.996	23.958	48.521	1285.477	80.693
Min	1	0	0	188	-1806
25%	12	5	8	826	2
50%	23.5	9	16	2700	4
75%	54.25	18	35	3025	6
Max	775	263	365	4027	13
Cardinality	139	64	109	21	15

Categorical features

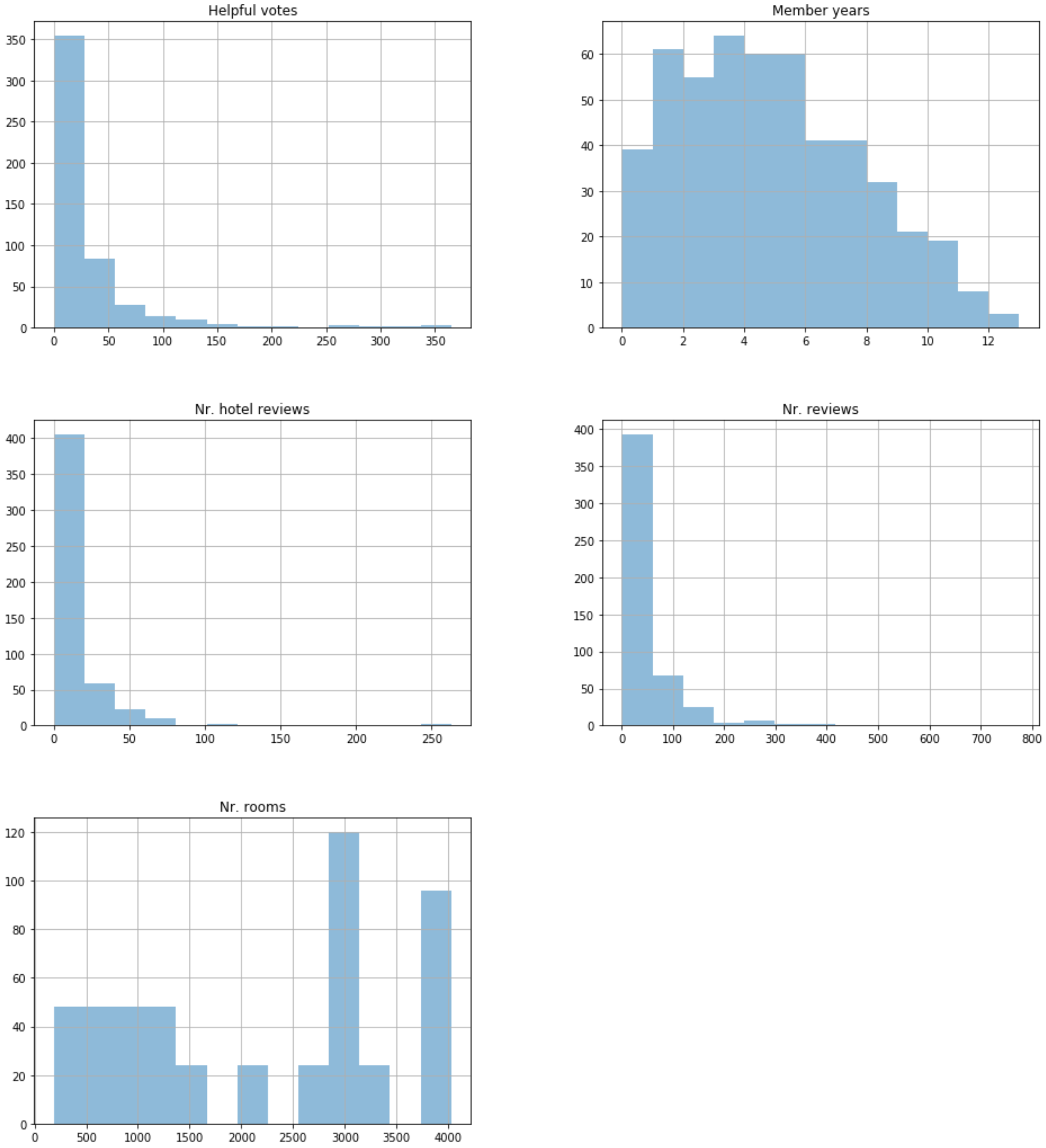
	User Country	Period of Stay	Traveler Type	Pool	Gym	Tennis Court	Spa	Casino
Count	504	504	504	504	504	504	504	504
% Missing	0	0	0	0	0	0	0	0
Mode	USA	Mar-May	Couples	Yes	Yes	No	Yes	Yes
Mode Freq.	217	128	214	480	480	384	384	456
Mode %	43.056 %	25.397%	42.460%	95.238%	95.238%	76.190%	76.190%	90.476%
2nd Mode	UK	Sep-Nov	Families	No	No	Yes	No	No
2nd Mode Freq.	72	126	110	24	24	120	120	48
2nd Mode %	14.286 %	25%	21.825%	4.762%	4.762%	23.810%	23.810%	9.524%
Cardinality	48	4	5	2	2	2	2	2

	Free Internet	Hotel Name	Hotel Stars	User Continent	Review Month	Review Weekday	Score
Count	504	504	504	504	504	504	504
% Missing	0	0	0	0	0	0	0
Mode	Yes	Monte Carlo Resort&Casino	5	North America	July	Wednesday	5
Mode Freq.	480	24	192	295	42	85	227
Mode %	95.238%	4.762%	38.095%	58.532%	8.333%	16.865%	45.034%
2nd Mode	No	The Venetian Las Vegas Hotel	3	Europe	May	Tuesday	4
2nd Mode Freq.	24	24	168	118	42	80	168
2nd Mode %	4.762%	4.762%	33.333%	23.413%	8.333%	15.873%	33.333%
Cardinality	2	21	3	6	12	7	5

Bar plot for the Target feature (Score)

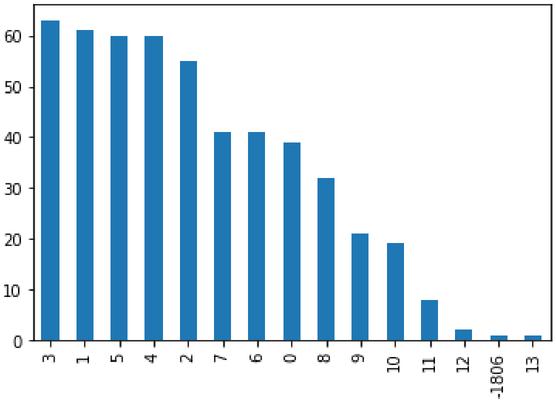


Histograms of all the continuous data



Missing Values and Outliers

I noticed that the minimum value for member years was -1806, which clearly cannot be right, so I decided to look into the column further to see if this was a common problem with this feature.



Interestingly, there was only the one incorrect feature, and I simply decided to replace it with the mode.

For outlier removal, at first, I found the interquartile range and then removed all records which fall outside of the Q1 - IQR * 1.5 to 1.5 * IQR + Q3 range. However, I realized that this caused me to remove nearly 20% of my total records. I found that this was because the helpful votes, number of reviews and number of hotel reviews columns had an exponential distribution, and I was simply removing records that were far to the right of the data. I still decided to use this method, but I excluded the columns relating to the number of reviews and helpful votes, because it doesn't seem like those features' outliers are irregularities which will harm my data analysis. I think this may be especially true for helpful votes, as the outliers in that feature would be the reviews people found most helpful.

Normalization

I used a min-max of [0,1] to normalize my data. I did not want to introduce negatives into the data as that would make the data harder to intuitively understand so I kept the range within positive numbers.

Feature Selection and Transformations

20 features seemed too many to create an effective prediction model, especially when intuitively it seemed that many of the features would likely have no correlation with the target feature, such as the user country, user continent, how many reviews were made by the user, and how long the reviewer was a member of TripAdvisor. I simply decided not to use some of the features as it was difficult to see their use, or their predictive ability would not be a lot of help for the hotel. I removed User Country, User Continent, Hotel Name, Review Month, Review Weekday, Nr. Reviews, Nr.Hotel Reviews, Member Years, and Helpful votes (helpful votes would be useful in filtering through reviews, but not really helpful in predicting the quality of review a customer would leave).

Model Selection and Evaluation

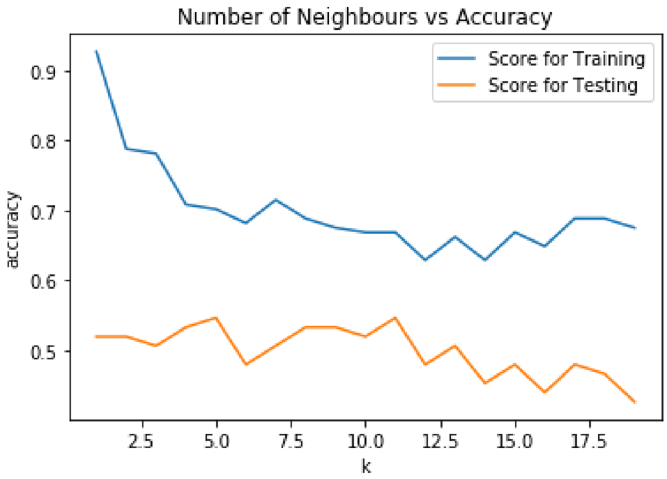
Evaluation Metrics

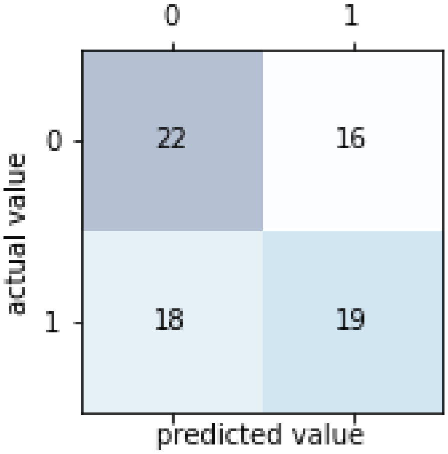
The models will be evaluated based on its accuracy and how easy it is to understand how it came up with its results. The speed of the prediction is not critical here, as the goal is to try to understand the criteria which make a review be higher from a customer, so there will not be many, frequent uses of the model to predict other reviews in the future. However, accuracy will be critical as the goal is to maximize the reviews left by customers, and an error in the model could result in lower reviews for the hotel, harming its reputation. The model should also be easy to understand as this will make the model believable and valuable in extracting conclusions about which features contribute to the better reviews.

One of the most important evaluation metrics for this problem would be the f1 score for the “Low” label. This is because the hotel is trying to maximize the average review score it gets, and this will happen by minimizing the number of lower reviews that they will receive. Since this dataset shows “High” reviews seem to happen much more frequently than “Low” reviews, being able to get an accurate prediction of what causes a customer to leave a “Low” review is more important that accurately guessing why reviews are high. We need to know the true negatives, because we want to maximize our accuracy of predicting it, false negatives because then our model’s predictions would think certain features which people rated highly are actually bad, and false positives so we accurately get the number of low reviews.

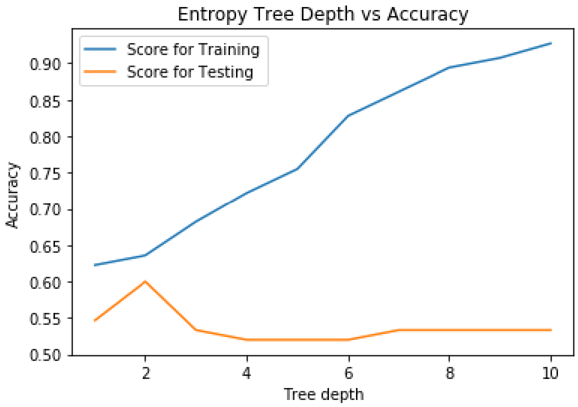
Models

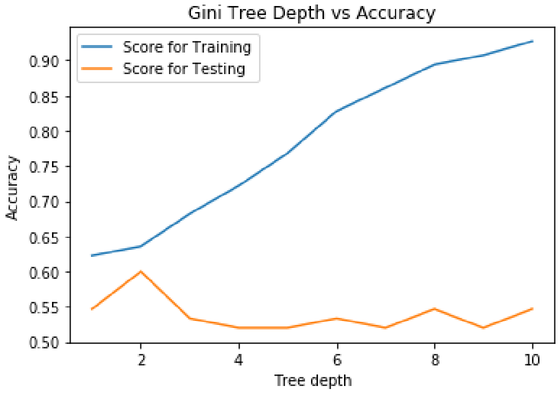
I first used KNN to create my similarity-based learning model. Unfortunately, the model was not able to perform really well, performing at its best at 54.67% accuracy when k=5.

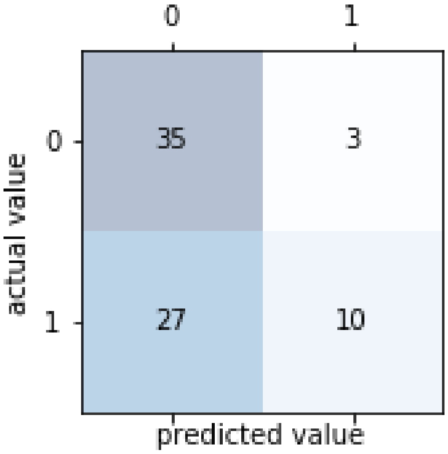


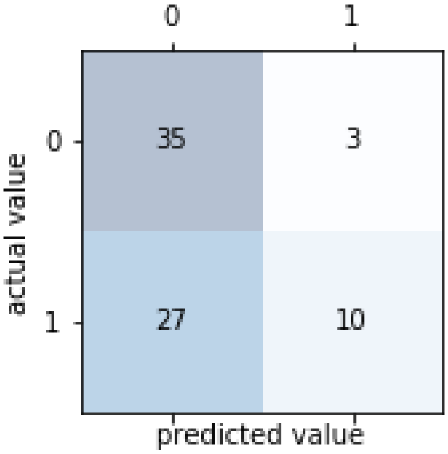


I then created a decision tree for my information-based-learning model. Using entropy, the accuracy maxed out at 60% when the depth of the tree was 2. For Gini, the max was also 60% when the depth was 2.

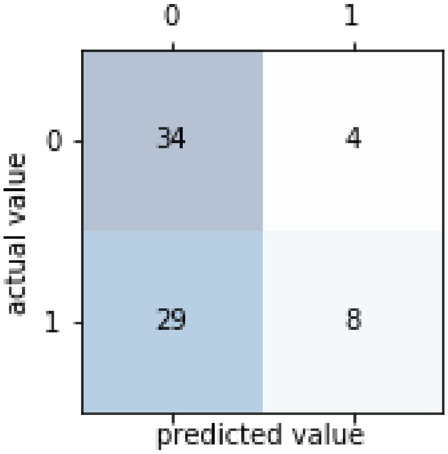




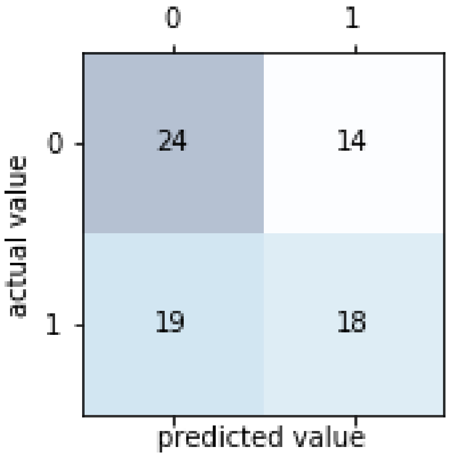




For my probability-based learning model, I used a Gaussian Naïve-Bayes classifier. The accuracy was at 56% and this was the confusion matrix.



Finally, for my error-based learning model I used logistic regression because my target feature Score was categorical. My accuracy was 56% and this was the confusion matrix.



Sampling and Evaluation Settings

For all of the models, I split the dataset using 2/3 for training and 1/3 for testing. When I first created the models, the accuracy was very low so I decided to split score into “High”(4-5) and “Low” (1-3) in hopes that this would make the data more separable so that a useful model can be created. There were also many more “Low” scores than “High” scores, so I used under sampling to make them be equal.

Hyper-parameter Optimization

I did some hyper-parameter optimization when I was doing the information-based model and the instance-based model. For the KNN classifier, I looked at the accuracy of the classifier for different values of K to select the best one. For the decision trees, I used both Gini and entropy for the different models and would prune the tree at different depths.

Evaluation

F1 score for “Low” for each model:

Instance-based Model (KNN)	0.53
Information-based Model (Decision Tree Entropy)	0.40
Information-based Model (Decision Tree Gini)	0.40

Probability-based Model (Naïve Bayes)	0.33
Error-based Model (Logistic Regression)	0.52

For this business problem, understanding the model is also very important, because if the model cannot be understood then the hotel will not be able to exactly know which steps to take and how taking those steps will result in the reviews increasing. I therefore think that the similarity based KNN model is the best, as it does predictions on the outcome of the review based on the similarity of the conditions for this reviewer compared to another reviewer. This makes it easy to understand why the model outputs the predictions it does, and it also had the highest F1 score for “Low”, making it the best model out of the four models.

Results and Conclusion

In conclusion, while the similarity-based model seems to be performing best based on the metrics given, no model really gave very convincing accuracy results. I think this means one of two things. First, either the features in the dataset were as good as it gets in separating this kind of problem, and that customer reviews are naturally going to have a lot of noise. This seems very possible, because 2 different groups of people going to the same hotel could leave with completely different impressions of it. This could be simply because of their differences or maybe one unfortunate event happened to one of the groups which caused a member to leave a poor review.

Another possibility is that more features are needed to predict the features more accurately. Many more features such as price, quality of room service, the location of the hotel, friendliness of the employees, etc. could play a very important role in determining whether or not a person left a favorable review. I believe that it is likely a combination of both of these factors that the accuracy is not very high.

My recommendation for the business is to start collecting data on its customers on its own. Perhaps by giving an optional survey for each customer to fill out, as this will allow future datasets to be more descriptive with the features being more relevant to the business problem, and the entire dataset to be customers which have been to the hotel. The best model from this dataset does not seem to be convincing enough to make business changing decisions, and I believe that this model should not be deployed. However, once the more detailed data is obtained and new models are created, the new model should be evaluated again. If again, its very hard for any model to separate the data convincingly, then maybe this is not the kind of business problem that has an easy data analytics solution.

On the other hand, if one of the models seems to be performing really well in separating the data, then that model should be analyzed in detail to see which features are most contributing to its predictions, and the business could simulate the reviews they would receive if they made certain changes, allowing them to evaluate the importance of each of their services in affecting the reviews left by customers.

Source for dataset

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.