**Predicting Yelp Restaurant Ratings Based on Business Attributes**

**Prepared by: Timothy Robbins**

**Bellevue University**

**Spring 2020**

**Introduction/Executive Summary**

Online review sites play a key role in the purchasing behavior of today’s consumer. Via the internet, individuals can observe millions of reviews on nearly every product and service in the market. Website’s such as Angie’s List, Amazon, Yelp, and Facebook help facilitate this process with online reviews. Since the features of a business have a huge impact on how a business is rated, businesses may be able to encourage higher ratings through targeted alterations to their offerings.

This study created a model to predict future Yelp business ratings (on a scale of 1-5) based on key business attributes. The scope was limited to restaurants in the Phoenix metropolitan area and analyzed data provided by Yelp through their 9th Dataset Challenge. Data analysis was conducted using both R and Python, with modeling done using the caret package in R. A random forest model was created to identify the business attributes that were most closely tied to the star rating. This study found that attributes such as whether a business provides outdoor seating, takes reservations, has waiter service, parking, and a moderate price range, to be factors that contribute to positive reviews.

The remainder of this paper will discuss the methodology, results, and conclusions of this study. The methodology includes the steps taken to collect, clean, and analyze the data provided. Subsequently, the results of the analysis will be presented. Finally, the study concludes with a section that explains the findings and discusses their implications.

**Methods**

***Dataset and Features***

Data was first collected from the .csv file provided by Yelp for their 9th Dataset Challenge. The Yelp data used in this study originally contained reviews for 42,153 businesses in 11 metropolitan areas. The original file can be found at <https://www.kaggle.com/z5025122/yelp-csv#yelp_academic_dataset_business.csv>.

To get an overview of the data, the dfsummary function from the summary tools package was used, which provided a great summary of each variable. The dfSummary function provided the variable name, data type, all possible values, frequency percentages of valid rows at each level, small graph of frequency distributions or small histograms of numerical variables, and the total number of valid and missing rows for each variable, all in a concise table format.

***Preprocessing***

A subset of the data was created to focus on restaurants in the Phoenix metro area. By filtering on the value “AZ” of the state variable I was able to restrict the analysis to Arizona. The categories field in the business dataset lists all categories that relate to each business (i.e. Dentists, Hairstylists, Restaurants, Sporting Good, etc.). A function was applied to the categories field to determine if it contained the string “Restaurants” and a Boolean variable, “is\_restaurant”, was added to indicate that the business was a restaurant. A filter was then applied to restrict the data to values that evaluated to “TRUE” in the “is\_restaurant” field and a revised dataset was created. This narrowed the dataset considerably from 42,153 rows to 7,439 rows and 105 columns. A summary of the initial variables is shown in Table 1, below:

***Table 1***

|  |  |
| --- | --- |
| **Varaiable** | **Description** |
| business\_id | Unique ID to identify business (primary key) |
| categories | Text field containing the type of business (filtered on "restaurants," as well as subcategories (i.e. chinese, italian, mexican) |
| city | City where the business is located |
| full\_address | Full street address for business |
| latitude | Latitide of business location |
| longitude | Longitude of business location |
| name | Business name. Note: some names are duplicated for chains, such as McDonald's, but have separate business\_id's |
| neighborhoods | Region of city where business is located |
| open | Whether business is open or not |
| review\_count | Number of business reviews |
| stars | The business rating on a scale of 1-5, in .5 increments (9 possible values) |
| state | State where business is located |
| type | The type of entity. All records contain "business" in this field |
| attributes.Accepts.Credit.Cards | 70 attributes relating to the business that are mainly binary, such as accepts credit cards, valet parking, happy hour, price range, etc. Only listed one for conciseness. |
| attributes.Hair.Types.Specialized.In.coloring | 8 attributes related to hair salons. This is just one example. |
| hours.Friday.close | 14 variables containing opening and closing times of business for each day of the week. This is just one example. |
| is.restaurant | Boolean variable created to indicate whether business is a restaurant or not. |

Most of the 70 attributes were binary, with a value of True or False. Certain attributes, such as “Alcohol,” were factors with 3 or more levels. Note that not all possible attributes were included for each business – only the ones that the reviewer had listed.

After the filters were applied, the 8 variables relating to hair salons (which had no values in any of the rows) were removed, as well as the 14 variables relating to business hours which were deemed unnecessary for analysis. This left 95 variables remaining: 12 of which were variables relating directly to the business

***Clean-up and Feature Selection***

Next, I cleaned-up the variable names: removing spaces, capitalization, and replacing periods with underscores. I then eliminated duplicate rows and columns, keeping the one’s with the most information for the model. I also dropped columns that didn’t apply to restaurants, such as hair, appointment, and insurance. I then dropped variables using the select and contains functions to eliminate variables that had more than 25% of the data missing, along with a few other variables I deemed less important; double checking that these drops had no impact on the overall dataset. After the drops, I reran the dfsummary function, as well as other summary functions in R, such as head, dim, glimpse, names, etc. I was now down to 7,335 rows and 36 columns.

I then cleaned up the ‘city’ variable, as it contained duplicate city names and misspellings. Originally there were 52 cities which I reduced down to x, again using the stringr package. I used functions from the stringr package (str\_detect, str\_trim, str\_replace), as well as case\_when to combine different string versions of the city name. Cities not in the Phoenix area were assigned to ‘not in area’ and later removed, while also moving cities with low record counts to ‘other.’

***Feature Creation***

Next, I did some string manipulation using the stringr package, along with the case\_with function to create a new variable called “food\_types.” Since the case\_with statement pulls the first field it finds, it was necessary to arrange the food types so that they extracted in a hierarchical fashion: having the more detailed food types prior to the wider categories. For example, having Mexican food prior to Latin American food. I then ended up keeping the food types with high frequency counts (i.e. Italian, Chinese, Mexican) to reduce the number of possible categories.

***Missing and Special Values***

In regard to missing and special values, I first converted all blank values in the fields to NA, eliminated 104 business\_ids which had “#NAME?” in the field. I used the sum(complete.cases()) function to count the number of rows with at least one NA. As for the missing values, I held off on imputing them as I was planning to use a random forest model that would allow for missing values. There appeared to be no outliers, as all nearly all variables contained the responses, true or false.

Next, I performed type conversions: converting all attributes to factors, along with open, city stars, city and food\_type to factors. I then dropped unused levels in the factors using the ‘droplevels’ function. After data cleaning, I saved the condensed file to my laptop so I could reopen in Python to perform exploratory data analysis (EDA).

Table 2 provides summary statistics of the sample after the reduction and clean-up of variables:



***Exploratory Data Analysis***

After data preprocessing, EDA was conducted using Python. An analysis of the histogram for the “stars” target variable indicated an approximate normal distribution with a (left) skew of -0.5366 and kurtosis of 0.2038. Similarly, a boxplot of the target variable indicated a mean of 3.5 with an interquartile range between 3.0 and 4.0. There was a possibility of potential outliers at 1.0 stars, which was determined to be valid as the minimum rating (ratings cannot take on the values 0.0 or 0.5).

Further analysis revealed that a majority of the restaurant reviews were in the cities of Phoenix and Scottsdale, Arizona, with the percentage breakdown of stars indicating that 5-star reviews were rare. The city of Anthem showed the greatest number of 4.5 reviews. In addition, an overwhelming number of reviews related to American food, followed by Italian and Mexican. The highest concentration of 4.5+ star reviews were related to European and “Other” food types.

Positive correlations between the attribute variables and the target variable were noted for divey\_false, classy\_false, touristy\_false, hipster\_false, trendy\_false, intimate\_false, casual\_true, romantic\_false, and upscale\_false. In addition, two of the parking variables, parking\_lot\_true, packing\_validated\_false, also indicated positive correlations as shown in Table 3, below:

***Table 3 – Correlation Matrix***

|  |  |
| --- | --- |
| **Variable** | **Correlation** |
| attributes\_ambience\_casual\_True | 0.204961 |
| attributes\_parking\_lot\_True | 0.200893 |
| attributes\_ambience\_touristy\_False | 0.188147 |
| attributes\_ambience\_upscale\_False | 0.177596 |
| attributes\_ambience\_romantic\_False | 0.174931 |
| attributes\_ambience\_intimate\_False | 0.174845 |
| attributes\_ambience\_classy\_False | 0.171918 |
| attributes\_ambience\_hipster\_False | 0.170722 |
| attributes\_ambience\_divey\_False | 0.170096 |
| attributes\_ambience\_trendy\_False | 0.162632 |
| attributes\_parking\_validated\_False | 0.155653 |

***Modeling***

After cleaning the data in R and conducting EDA in Python, I began the modeling process with the R caret package. The random forest model in caret required some additional processing steps to get the data in the proper form. The levels in stars variable were renamed from numeric (i.e. 1, 1.5, 2) to character, as the models in caret would throw an error if not done so properly. The models in caret allowed for the determination of near zero variances among predictor variables via its preprocessing function. After identification, these features were then dropped from the final data sets. Next, I randomly split the dataset into train and test sets using a 60/40 ratio, as this was a classification problem with a fairly small data set.

Caret automatically splits the variables into dummy variables. However, when it does so, some variables are left with constant and almost constant predictors across samples. Since one categorical variable gets broken into many, some of the dummy variables had zero observations, and thus become a dummy variable full of zeroes. The caret function “nearZeroVar” removes such variables that have both few unique variables relative to the number of samples (zero variance predictors), and a large ratio of the most common value to the frequency of the second most common value (near-zero predictors).

I initially created 3 different kinds of random forest models: randomForest, ranger and rf. The accuracy using the default settings for each model are shown below:

**Table 4 – Random Forest Models Analyzed**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Method** | **Train Accuracy** | **Test Accuracy** |
| mod\_ranger3 | ranger | 0.3706 | 0.3674 |
| mod\_randomForest | randomForest | 0.3564 | 0.3554 |
| mod\_rf | rf | 0.3525 | 0.3429 |

The accuracy between the different models was similar which made choosing the appropriate model more difficult. After numerous iterations, it was decided that the “ranger” model (mod\_ranger3) consistently provided the best accuracy.

**Results**

The ranger model’s hyperparameters were further tuned to optimize accuracy. Custom grid searches and train controls were created to select the optimum hyperparameters for the ranger model. The main tuning parameter in the ranger model is mtry, which is the number of randomly selected variables used at each split. Values of mtry tested ranged from 2 to 37, with the optimum accuracy being achieved using an mtry of 2. The k-nearest neighbors method was used to impute missing values. The results were validated using 5-fold cross validation. The final model using the “ranger” method resulted in a test accuracy of 0.3695, along with training accuracy of 0.3612.

The table below shows the confusion matrix for the 9 possible categories of the “stars” target variable, along with the sensitivity and specificity, for the final model utilized.

***Table 5 – Confusion Matrix***



The two key metrics shown above are sensitivity and specificity. Sensitivity is a metric that evaluates a model’s ability to predict true positives of each available category, whereas specificity is a metric that calculates the model’s ability to predict true negatives of each available category. These metrics are calculated for each possible category as there are more than two possible prediction options.

***Variable Importance***

Based on this study, the features contributing the most to restaurant rating are shown in Table 6 below:

***Table 6 – Variable Importance***



**Conclusion**

As shown in the results, the key business attributes identified in this study that contribute the most to the business ratings are whether the restaurant offers outdoor seating, tv, a casual ambience, moderate price range, average noise level, parking, takes reservations, offers delivery and waiter service. In addition, restaurants in the city of Phoenix showed a stronger relationship with the target variable then other cities in the Phoenix metropolitan area.

Now that a model has been created to predict Yelp restaurant ratingsbusiness owners will be better able to tailor the features of their business to match those that promote positive reviews. Based on insights from the model, business owners can now focus on the features that provide the most influence on business ratings.

***Risks and Limitations***

The observations analyzed in this study were from the Phoenix metropolitan area, in the state of Arizona. Since local preferences and behaviors may differ significantly among geographic regions, the sample taken may not be representative of the entire US population. Inclusion of more regions in the study may have provided for greater insight into the significance of physical location on review behavior and may thus be an opportunity for further studies.

In addition, text and sentiment analysis were not conducted in this study and may provide for further potential. Information such as review text, hours of operation, and a variety of details about the user and other variables should also be considered. Finally, research could be enhanced by the use of additional modeling techniques.

**Acknowledgements**

Since I worked on this project independently, I have no individual team members to thank. I would like to thank Professor Ashley Kern for recommending the caret package for R. I found the caret package to be more transparent, as to which variables were being fed into the model, than the scikit-learn package in Python.

Finally, a previous study was conducted by Peter Schellenberg that I would like to thank for providing insights that were helpful in planning for this project.

**References:**

Abbott, D. (2014). *Applied predictive analytics: Principles and techniques for the professional data analyst*. John Wiley & Sons.

Albon, C. (2018). *Machine learning with python cookbook: Practical solutions from preprocessing to deep learning*. " O'Reilly Media, Inc.".

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, pp. 3-7). New York: springer.

Kazil, J., & Jarmul, K. (2016). *Data wrangling with python: tips and tools to make your life easier*. " O'Reilly Media, Inc.".

Shellenberger, Peter Mark Jr., "Predicting Yelp Food Establishment Ratings Based on Business Attributes" (2017). Honors Theses and Capstones. 374.