Fatima Warren

Final Assignment

KSB-621

10/15/2023

**Technical Report**

**Project Question and Area of Opportunity for Client**

Question: Can a service or product’s value be determined by its ratings and/or its reviews?

**Some Areas of Client Opportunity:**

* Cybersecurity Ratings
* Ratings from Customer feedback on pilot programs
* Ratings from Real time and smart sensor data from products
* Services ratings for various solutions and volunteer work
* Ratings for Finance Planning and Investment

**Data Description**

The Yelp dataset includes detailed information about businesses, user reviews, ratings, and attributes among other variables. This dataset is suitable for addressing questions related to business performance and user behavior.

**Pre-Processing**

SQL Query (BigQuery) was used to extract the initial dataset (converted .json to CSV) by business\_ID based on the categories "Food" and "Restaurants" along with all the 33 total columns. Missing values were converted to N/A. Also, all ‘No’ and ‘Untrue’ cells were converted to False. Also, converted the datset from array to dataframe.

The data was then filtered by city. Four cities were chosen: Tucson, Nashville, Philadelphia, and Tampa Florida.

**Key Variables**

There are two dependent variables that were analyzed: Stars or ratings of the restaurants and Review Count. These two are important metrics as they show what the customer, consumer, or business owner values. They also helps to gauge interest and can narrow down results where comparison metrics and their evaluation is important.

The 10 Independent variables included:

State: The sates that the four cities belong to.

Restaurant Attributes: Reservations, Takeout, Delivery, Catering, Has TV (or not), Good for kids, Good for Groups, Outdoor Seating, and Bike Parking.

These variables were chosen based on various preliminary data analysis results and after preprocessing.

**Joining Dataset**

An external dataset of median and mean income based on the US Census Bureau (American Community Survey) 2019 Household Income by zip code was joined to this dataset.

**Dataset Reshaping and Data Validation**

Columns with a high level of NAs and problematic values were removed. The data had columns with punctuation that seemed to be hard coded. These included goodformeal, ambience, drivethru, restaurantsattire, wheelchairaccessible, happyhour, businessparking, restaurantstableservice, coatcheck, noiselevel, wifi. After removing these columns and NAs, the data had around 125,000 rows and 26 columns.

The data was checked to make sure that all variables fit the right class for analysis.

**Strategy Resolution of Missing Values and Outliers**

Outliers below a 2.5% quantile (lower bound) and above a 97.5% quantile (upper bound) were isolated and removed. A new dataset was created with these removed outliers which produced slightly better results in the linear regression models. All observations below 26 and above 5721 are considered outliers according to the quantile limits set, which is 3,448 entries.

**Rescaling**

The continuous variable review\_count was rescaled. A histogram of review\_count showed skewness. After logging the variable, the data seemed a bit more normal.

A graph of blue squares

Description automatically generated with medium confidence A graph of a graph

Description automatically generated

**Preliminary Exploratory Data Analysis Results**

Stars: A histogram of stars showed that the highest number of stars are in the 4 star category followed by 3.5 stars and 4.5 stars.

Histogram of Stars

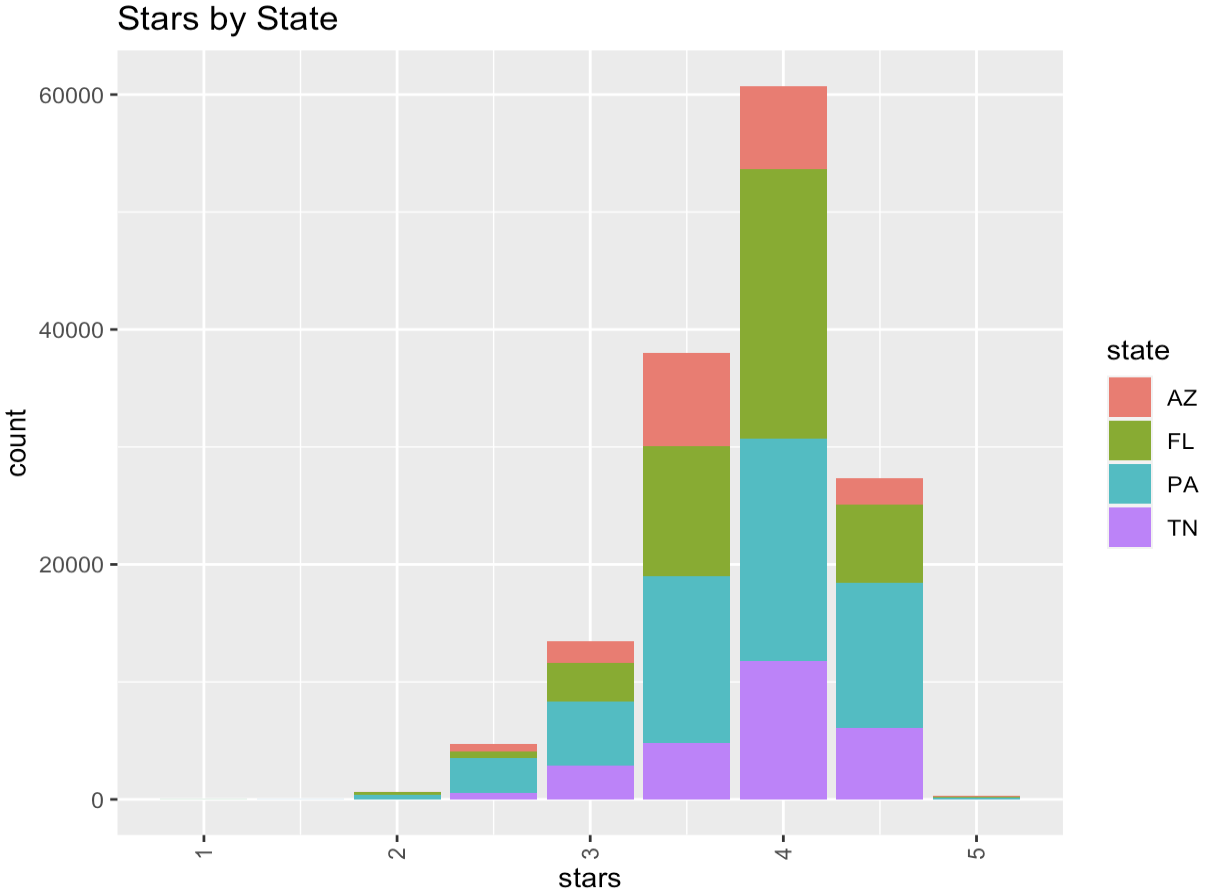
A graph with blue bars

Description automatically generated

Review (review\_count): As mentioned, this variable’s histogram showed that it was skewed to the right. After logging review\_count the qqplot showed improvement as well as the histogram and boxplot.

State (A variable that proved more important based on modeling than cities):

Distribution of Stars by State



Other Key Variables:

Relationships between other key variables were explored to see how they might be incorporated into the modeling plans and interpretation of results.

A graph with a bar and text

Description automatically generated with medium confidence A graph of a bar chart

Description automatically generated with medium confidence

**Modeling Analysis Results**

Two different methods of modeling were used to evaluate the best fit for the Yelp dataset. Linear Regression did not produce the best results based on fit statistics and explanatory analysis. Bootstrap Aggregation and Randm Forest gave better results.

**What Models were Chosen or not Chosen, Modeling Outcomes, Evaluation Metrics**

Three linear regression (LR) models used and 4 Bootstrap Aggregation and Randm Forest models were used.

**LR (dependent variable: review\_count):** Initial results from the LR model showed a very low Adjusted R Square of 0.2883 which means that 28% of the variance of the dependent variable is explained by the independent variable. After data transformation such as removing NAs and outliers and cleaning columns stepwise LR was used to find best fit and it showed a very slight improvement of the Adjusted R Square which jumped to 0.2956. An interaction model also showed a slight improvement (0.2986) but it did not seem significant.

**LR (dependent variable: stars):** Linear regression was not performed for this categorical variable as LR is not fit for use for the prediction of such variables. Also, since LR was not performing well overall with the dataset, alternative modeling methods were searched and the package randomForest was used for modeling.

**Random Forest (dependent continuous variable: review\_count):** The variation explained by Random Forest was 96%.

The variable importance plot shows that the most important variables according to %IncMSE for Random Forest are stars,restaurantreservations, restaurantdelivery, caters, hastv, state, outdoorseating, goodforkids, bikeparking, restaurantstakeout, restaurantstakeout, restaurantsgoodforgroups.

**Random Forest (dependent categorical variable: stars):** The variation explained by Random Forest was 94%.

The variable importance plot shows that the most important variables according to %IncMSE for bagging are review\_count, state, restaurantsdelivery, outdoorseating, hastv, goodforkids, bikeparking, caters, restuarantsreservations, restaurantsgoodforgroups, restaurantstakeout.

Variable Importance Plots

A graph of a restaurant review

Description automatically generated with medium confidence A graph with text overlay

Description automatically generated

**Bootstrap Aggregation:** The variation explained by this method was 100% for the dependent continuous variable revew\_count and the variation explained for the dependent categorical variable stars was 96%. This method seemed to be overfitting the dataset.

**Limitations**

**Part 1: Dataset**

* The dataset had a few hardcoded scripts that provided difficulty in analysis.
* The dataset itself is limited in scope due to the type (restaurant under the hospitality business) the areas chosen, and the period from when the reviews were generated.
* The categories section was not entirely sectioned and grouped properly. For instance, categories such as restaurants, food, and beverage all have similarities, but they do not specify the type of business so that the results can be narrowed down accurately.

**Part 2: Modeling Techniques**

* Random Forest is not good a good method used for extrapolation. It is also prone to bias due to this reason.
* Bootstrap Aggregation can overfit the dataset as it reduces variance.
* Both of these techniques can be computationally expensive.

**Conclusion and Reflection**

RandomForest provided some of the best results as it is able to handle imbalanced or skewed data. In the case of the Yelp dataset, not all the normality assumptions could be met. This method also effectively handles outliers. The Yelp dataset did not fit any of the linear models well even after removal of outliers and handling any missing values. Additionally, categorical variables can be used for prediction with randomForest without transformation.

**Key Takeaways and Client Recommendations**

Ratings provide key metrics of evaluation for products, services, and new as well as existing projects. They can be used in technologically enhanced projects such as ratings from sensor data or in evaluation of volunteer work to provide better work environments for case handlers and volunteers alike.

The various attributes used here to predict the ratings and reviews can be replaced with important indicators that are specific to the service or product. Where nonlinear data and complex systems are concerned, complex relationships between variables are handled well with randomForest compared to linear regression.

Appendix

Exploratory Data Analysis

Initial analysis of some of the key variables showed outliers and NAs along with densities and counts among other stats.

A screenshot of a graph

Description automatically generated

Cities: Distribution of cities showed that Philadelphia and Tampa had the highest number of businesses. Also a distribution of stars by cities.

A graph of different colored squares

Description automatically generated with medium confidenceA graph of different colored squares

Description automatically generated

Decision Tree

Additionally, decision trees were made using the “tree” package before randomForest was chosen as the second modeling technique as the fit statistics for the randomForest based bagging models were more clearly interpretable while I was not able to retrieve fit statistics for the decision trees without errors/glitches that were breaking up the code so those fit statistic retrieval methods had to be removed. However, I was able to get trees for the attached dataset’s mean\_income variable.