Nathan Timmerman and Micah Thompkins
Project Report
CS 396
3/18/20

Introduction and motivation

Our goal at the beginning of the quarter was to create a model that could predict whether or not a business was likely to close. To do so we had a list of questions that we wanted to analyze and test as potential machine learning features for our model. As the project went on our list of questions grew and changed to better fit our project. For example instead of trying to build a model across the entire Yelp dataset and country, we focused only on Arizona, that way we could add features involving location.

Data Cleaning

Our question did not require us to clean any data in the dataset. If any data was erroneous (e.g., wrong stars amount given, wrong times entered, etc.) we would have no way of ascertaining which data is incorrect and thus could not fix it.

EDA

Again, for our EDA, data processing was minimal for the same reason stated above. While no cleaning was done, we reorganized the data into frames containing just the information relevant to our analysis. Analysis was done on a random sample of ten-percent of businesses across the entire dataset. We answered the following questions.

—*Are the distributions of stars for closed and open businesses the same?*
Using a two sample t-test, we found a p-value << 0.05, providing strong evidence that the distributions are not the same.

—*Is there a difference between review length for closed and open between?*
Again using a two-sample t-test, we found a p-value << 0.05, providing strong evidence that the distribution of review lengths are not the same. On average, reviews for closed businesses are longer.

—*Are the number of reviews the same for open versus closed businesses?*
Using a two sample t-test, we found a p-value << 0.05. Although this suggests that the distribution of the number of reviews for closed and open businesses is different, we determined that this is likely coincident on the fact that open businesses have accumulated more reviews by virtue of being open.

—Are the proportion of each number of stars for open versus closed businesses the same?
This analysis was perhaps the most surprising. We found out that for star ratings ranging from 1 to 2.5, open and closed businesses had approximately the same proportion. For 3 to 4 stars, closed businesses had a higher proportion, though only slightly. However, for 4.5 to 5 stars, open businesses had 4% and 7% higher proportions, respectively. This was an interesting insight that we tried to employ in the features of our machine learning models.

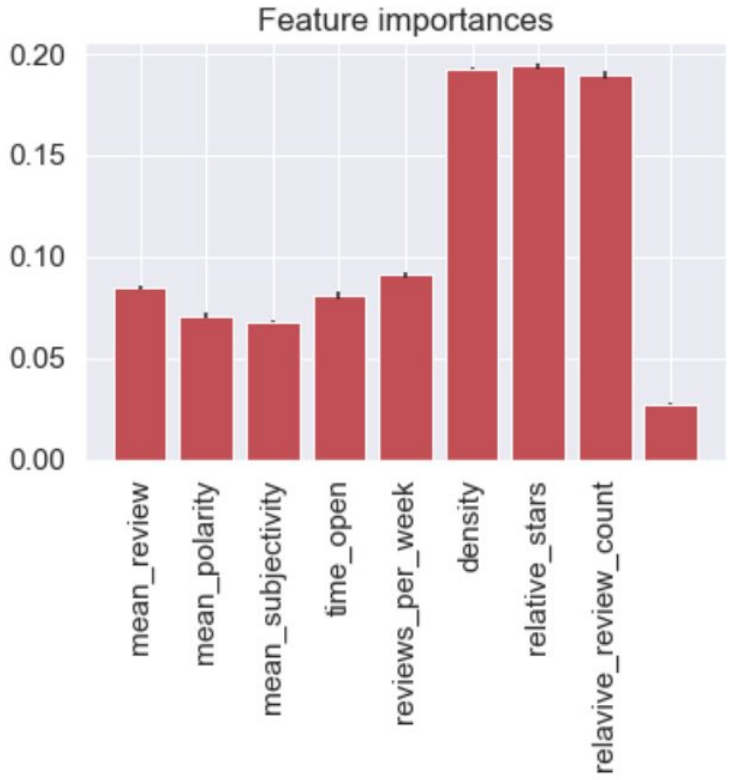
ML, Text Processing, & Social Network Analysis
The first time through the machine learning component of our project our features consisted of the average review length for each business as well as the distribution of the stars given in reviews. We chose these features because during our EDA we saw that these two features had significant differences amongst closed and open businesses. A third feature we added in was sentiment analysis of the actual review text itself. We think this feature is helpful to our model because it takes in one of largest components of data in our set, the review text, and analyzes it. Finally, we added to our features the distribution of the stars for the reviews of each business. Our EDA showed that this was also statistically-significantly different between open and closed businesses. Our targets were whether the business is open or closed (1 and 0, respectively).

In terms of algorithms, we tested multiple classifiers: gradient boosting tree, KNN, decision tree, random forest, and SVC. We chose accuracy because we want our classifier to be useful for business owners in determining whether their business is performing at a level that could lead to closure.

Gradient boosting tree, KNN, and SVC all returned accuracies of about 82%. After analyzing the predictions of each, we realized that each of these models was predicting ‘open’ for every business, performing as if they were dummy classifiers that merely chose the class with largest proportion.

Because we are dealing with imbalanced classes, we turned focus to classifiers that better deal with this case: decision tree and random forest classifiers. Out of the two, random forest performed significantly better, so further model building dealt exclusively with this model of classification. Moreover, employing AdaBoost on top of the random

forest classifier further improved results. The resulting confusion matrix is below, along with a bar graph showing the importance of each feature in the model.

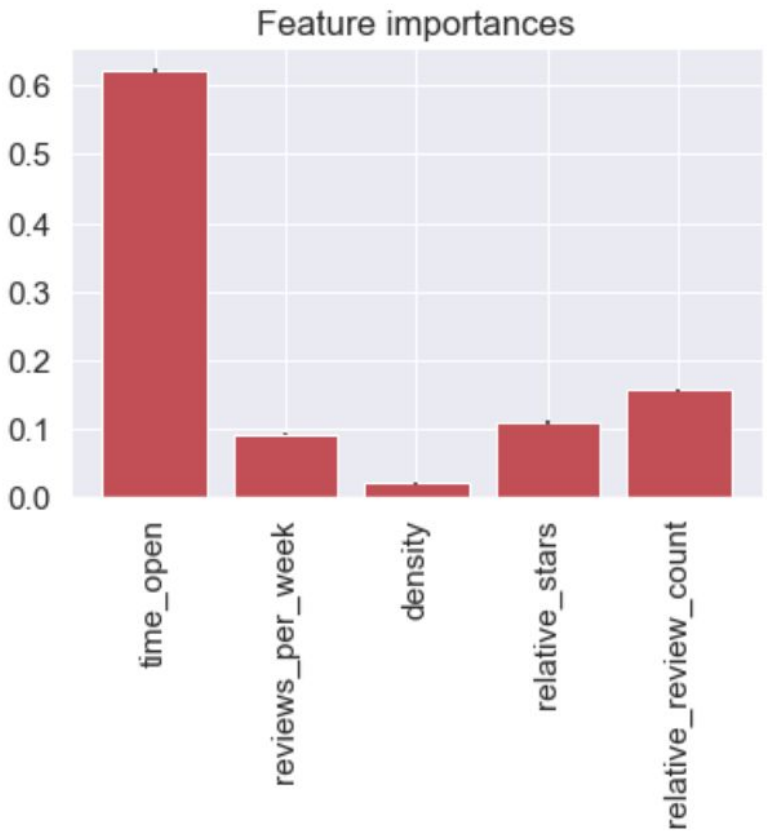


For our second round of machine learning, we delved further into feature engineering to take into account both the location of businesses and their types. This is because we realized that there are likely geographic factors that contribute to a business's success.

Because of this, partnered with the fact that businesses are very sparse in most states (some have as little as one business in the dataset), we focused only on Arizona, the state with the largest number of businesses in the dataset. For each business we collected all other businesses within one-mile that shared at least two categories. We chose two rather than one due to the existence of very general categories that were not good predictors of whether businesses were truly similar.

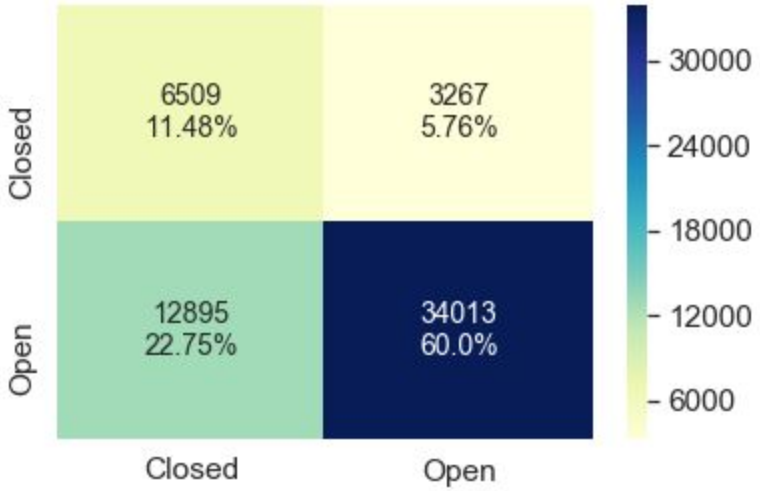
Using these businesses, we calculated the density of the business area and the stars and review counts of a business relative to its neighbors. Furthermore, we took into account whether a business was a chain and the amount of time that the business has (or had) been open.

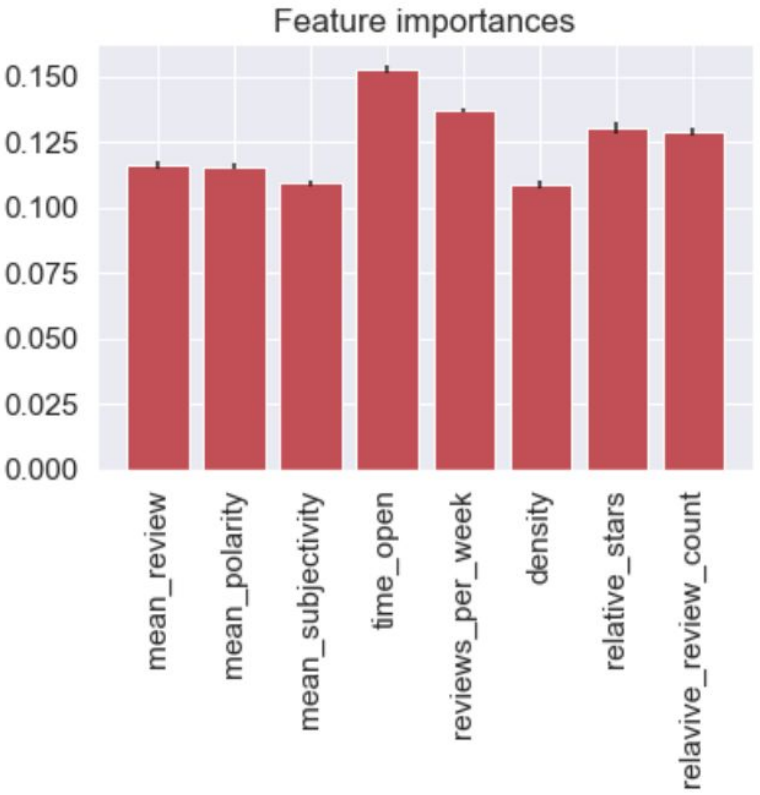
This new set of features, using AdaBoost with a random forest classifier, performed slightly worse in overall accuracy. However, the proportion of correctly classified closed businesses was larger than that of the previously attempted model. As the goal of our project is to be able to predict whether a business will close, we consider this the most important statistic. Therefore, between the two sets of features tested, the second set performed better for our purposes. The resulting confusion matrix is below.



As can be seen through comparing the two confusion matrices, the first model performed with about a 72% overall accuracy with only around 54% of closed classified correctly. In contrast, the second model performed with about 66% overall accuracy with around 66% of closed classified correctly. The second model had a higher portion of open businesses classified as closed. These results could mean one of two things: either neither of our models are strong enough to properly classify open and closed businesses, or a sizeable proportion of open businesses are at risk of closing. Without further information, we cannot determine which of our models is better.

Based on the feature importance graphs, we created one final model using all features from the second model and the three most important from the first. The results are below:





This model combined the best qualities of each of the previous two: it has about a 71.5% overall accuracy and correctly classified 66% of all closed businesses. Still, without further information, we cannot conclude which of our three models most accurately represents reality.

Summary of findings

Potential implications and improvements

One potential improvement that we found as we were looking for ways to further improve our accuracy was adding whether or not the business was claimed on Yelp as a potential additional feature. This is not a part of the Yelp dataset and would require a large amount of web scraping. However, adding this as a feature would likely help the accuracy of our model, as a business that takes the time to claim itself on Yelp is likely in a better position to remain open. Another potential improvement for our model would be if we spent time researching about how business reviews and yelp data actually affects business or talked to people in industry. We have no expertise on running a business or how a businesses Yelp reviews and information help or hurt their success. If we gained more knowledge in this field we could create a more informed and likely better machine learning model. A potential implication of this project is that business owners could use our model and run their yelp data through it and if our model predicts that they are closed or likely to close think about ways in which they can improve their business. If they are getting lower reviews and star ratings compared to other similar businesses near to them they could go check out these other businesses and their reviews in order to see what those businesses are doing differently than them.