Applied Data Science Capstone Project – Churn

1. Introduction/Business Problem

   Machine learning techniques like K Nearest Neighbors and Support Vector Machines are often used to predict the likelihood of customer loss (or churn) for various businesses such as credit card issuers or cell phone carriers. By identifying customers most likely to switch brands, businesses can potentially develop strategies for retaining those customers. Since the cost of acquiring new customers is typically greater than the cost of retaining existing customers, predictive analytics can be a cost-effective tool.

   These machine learning techniques potentially can be used to analyze various aspects of the public sector (i.e., government) economy as well. Government could conceivably provide services more cost-effectively or to a larger group of constituents if it had a clearer picture of where and when to intervene.

   Many municipalities in the U.S. require businesses operating within their city limits to register for a business license, which is a combination of taxes, fees, and regulatory requirements. The specific regulations and costs vary by city.

   New businesses start up all the time. While some fail, a few relocate to neighboring cities to find cheaper rent or a more lucrative customer base, and others survive for years. Those businesses that fail or move away often leave behind unpaid debts such as taxes or fines to the city. Using predictive analytics, cities could potentially intervene earlier, identifying debtor businesses that are “flight risks” or businesses that could potentially be persuaded to stay in town and support the local economy. Churn analysis is therefore potentially of interest to both finance and economic development officials at any city hall.

2. Data

   For this project, I will use a subset of business license data from the City of San Rafael, a municipality in Marin County, California. More specifically, I will look at restaurants operating in San Rafael in 2019 to predict whether they would disappear in 2020 (regardless of whether they closed their doors, moved out of town, or sold to a new owner).

   Included in the dataset will be business name, address, district name associated with the address, form of ownership (sole proprietor, corporation, etc.), and number of years in business. Some of the restaurants have been “red tagged;” that is, they were temporarily shut down by county officials for health code violations.

   Only non-confidential aspects of the business license data will be used. Regrettably, this may greatly weaken the accuracy of the churn predictions, as financial results are presumably a key factor in a business’s survival.

   Instead of financials, I will use the techniques from this capstone course to retrieve Foursquare ratings for each business. After joining the Foursquare information to the dataset, I will test the hypothesis that low customer ratings are a sign of impending business failure.

3. Methodology

   After importing the dataset in a Jupyter Notebook, I conducted some exploratory data analysis. Specifically, I looked at value counts for the dependent variable 'Business status' and two of the features, 'Geo' (the geographical neighborhood) and 'Ownership type' (business structure). Because most businesses survive in any given year, "Business status' has an imbalanced proportion of classes.

   'Ownership type' exhibited a reasonably even distribution across its classes. And after binning the classes into 10-year intervals, it appears that most business failures occur in the early years, with a secondary bump in failures around the 25/30-year mark. This suggests that it might be useful to square the 'Years' variable at some point in the analysis.

   'Geo,' which represents various districts such as the Downtown, is another feature with an imbalanced proportion of classes. By scrapping 'Geo' altogether and creating a new feature 'Downtown,' with yes or no values, we get an evenly balanced proportion of classes. Binning the 'Downtown' feature into 10-year intervals also indicates that failures are concentrated toward the first 10 years with a secondary bump around the 30-year mark.

   For classifiers, I utilized some of the common ones such as K Nearest Neighbors and Logistic Regression. These have been employed in other courses in this Coursera specialization, and they are often used in churn analysis. I also took a chance on AdaBoost and some other classifiers that are not as well known.

4. Results

   Many of the classifiers yielded 94-95% accuracy in predicting business failure. However, because of the imbalanced classes (only about 5% of businesses fail), one can achieve 95% accuracy just by predicting that all businesses survive. Only 5 classifiers predicted any failures, so I give K Nearest Neighbors credit for accurately predicting 2 of the failures while keeping false positives low.

   Including the square of 'Years' in the model, either with or without the original 'Years' feature, only improved the accuracy slightly. And still no classifier had higher accuracy than simply predicting all businesses survive.

   The 'getFeatureImportance' method of the so-called 'ensemble' style of classifiers allows us to see which features were the most impactful. 'Ratings' was either the first or second most important for Gradient Boosting, AdaBoost, K Nearest Neighbor. 'Years' was also at the top of the list.

5. Discussion

   I originally focused on restaurants because they seem to garner a disproportionate share of the ratings on Foursquare, Yelp, etc. I also did not want to use too large of a dataset because of the limited number of calls that can be made to the Foursquare API

using the sandbox. If expenses were not any issue, expanding the database to include other types of businesses might yield more meaningful results.

Another problem with the dataset is that too many of the restaurants did not have Foursquare ratings. Newer establishments might not yet have attracted the attention of Foursquare users. Some of the failed businesses either could not be found in Foursquare (perhaps listings eventually expire) or they had no ratings. I chose to drop the records with 'N/A,' though I considered assigning them a 0 rating or some other suitably low value. I also considered using Yelp ratings to fill in missing values.

6. Conclusion

It appears that Foursquare ratings act as a somewhat useful predictor of business failure. With the current dataset and methodology, the models can at best tie the accuracy of simply predicting all businesses survive. K Nearest Neighbors, however, has the highest weighted average F1 score.

It would probably be beneficial to expand the dataset to include more types of businesses. And other approaches to handling businesses without Foursquare ratings (for example, assigning a 0 rating or looking up the Yelp rating) might improve the results for all classifiers.