Mitchell Lee

Coursera – Applied Data Science Capstone

08/13/20

**Problem Description and Background**

For this capstone project, I am pretending that I have been approached by representatives of CAVA—a popular chain of Mediterranean restaurants with 105 locations in the north east, California, Colorado, and Texas—to create a ML model that can predict profitable new locations for the chain. They inform me that they have capital to invest in new locations, but want to be confident that their investments will likely succeed so that they can increase their profitability with minimal failures and losses. They then tell me that they are confident that they have already expanded to every viable zip code in states where they have locations (meaning that zip codes without a CAVA in those states are not viable), so they want to expand to new states and are open to any area of the country as long as the viabilities of proposed locations are supported by a strong model. After discussing my uncertainty as to whether ML can provide strong predictions and what the predictors should be, they agree to fund a preliminary investigation of whether the success of a CAVA in a new location can be predicted based on the types of venues that are nearby. I agree to attempt to develop a ML model to predict which of the most populous zip codes in states across the US that do not already have CAVA would be profitable locations for new CAVA restaurants. Hopefully, this model will serve as a fruitful starting point for the development of better models that will enable CAVA to maximize its return on investment during expansion to new areas in the US.

**Data Description**

Creating this model will involve the following development steps that will acquire data from several sources, including calling Foursquare’s ‘Explore’ endpoint:

1. Scrape zip code of every CAVA restaurant in the US from CAVA’s website (<https://cava.com/locations>) to serve as positive training controls.
2. Scrape lists of 30 most populous zip codes from states with CAVA (excluding zip codes that have CAVA) from a demographics website (e.g., <https://www.virginia-demographics.com/zip_codes_by_population>) to serve as negative training controls.
3. Geocode (get coordinates) for every zip code found above using the OpenCage Geocoder API (<https://opencagedata.com>).
4. Use the coordinates obtained above in calls to Foursquare’s ‘explore’ endpoint to find the most popular/recommended venues around each zip code found above. The combination of those zip codes and their corresponding venues will be used as the training dataset.
5. After processing the venues to show the number of each type in each zip code, train and optimize a logistic regression model, a support vector classifier, and a k-nearest neighbor classifier (optimization will include dropping low-frequency and then collinear variables).
6. Repeat steps 2 - 4 to obtain the 10 most populous zip codes in states withoutCAVA and their corresponding coordinates and venues. This combination of information will be used as the prediction dataset.
7. After processing the training dataset as in step 5, use the best performing of the optimized logistic regression, support vector classifier, and k-nearest neighbor classifier to predict whether each zip code in the prediction dataset would be good locations for the CAVA