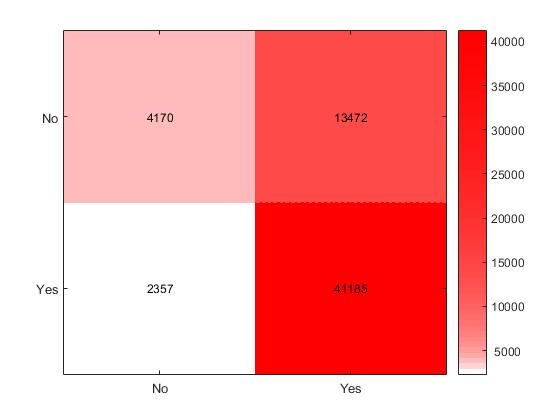
**Most Valued Business Attributes By Location**

Team: Andrew Toscano 1000808071

The objective of my project is to identify trends of business attributes and categories, to determine which are the most important for a store to be rating highly. I was inspired to use MATLAB for this project after attending a seminar in which they showed Machine Learning techniques that I thought would be useful for my project. The deliverables I will have are the most influential categories for all businesses as well as a confusion matrix of a classifier created to determine the rating of the business just by the categories.

The biggest difficultly I have faced is dealing with the data set itself. The json file with the business information was very inconsistent in several fields, and a normal json parser could not make sense of the data. I had to parse data into separate inconsistent structures line by line, then at the end bring it all together to create a data set that was usable. The json file is not included in this repository for memory reasons, but the data read is.

I created classification trees using a bagged tree model. First, I read in the business data from the yelp\_academic\_dataset\_business.json, then found every unique category for every business. I only got the first word of each category, which combined about 200 of the categories given. I then created a table with a column for each category type and a column for which day of the week they are open. These columns were used as predictors. The final column was the response column, and was either 1 if it was rated 3 stars or higher and 0 if not. I then used MATLAB’s Classification Learner to create a classifier, which was a decision tree with a maximum of 500 leaves and cross partitioning/validation, and applied the classifier to the data set to get a Confusion Matrix. The total accuracy was 73%.



Because there is a high true positive and false positive rate, I believe it implies that it is inconclusive to evaluate the businesses by category alone. I have not included the business attributes in this decision tree, so maybe by only including those I can create better classifiers.

Even with this result, I wanted to see the most important features with a decision tree as a classifier. I used a BaggedTree creator in MATLAB to determine the feature importance. The feature importance is a built in feature that figures out what changes to features give the classifier a higher misclassification rate. These would imply that they are the most important factors in determining whether a business was rated badly or not. I also use the stat of which features result in a high classification rate. This implies that those are the most important factors in determining whether a business was rated well or not. This is the output of those statements.

Top Error Features

'French' 'Vape' 'Thai' 'Active' 'Arts' 'Mediterranean' 'Car'

'Latin' 'Cafes' 'Buffets'

Top Success Features

'Vape' 'Arts' 'French' 'Thai' 'Monday' 'Chiropractors'

'Sunday' 'Active' 'Musical' 'Specialty'

This by itself isn’t particularly interesting, but when I separate the data by city, it may reveal more.

I still need to add business attributes and separate the data by cities. I may also look into other methods of evaluating feature importance. My final report will have more information, graphs, and results of these calculations.