Andrew Tang
Oliver Goodman
Daniel Letscher

EECS 349: Predicting if a Business Will Close

Overview

Yelp serves as a great resource to research the quality of many restaurants and businesses. However, given the competitive environment associated with operating a personal business, it is not uncommon for owners to be forced to close their doors, only to be quickly replaced with another business. The goal of our project was to use data from the Yelp Dataset Challenge to predict whether a given business will close down in the near future. This concept was interesting to us given the amount of Evanston restaurants that seem to go out of business on a whim, whereas other businesses may thrive in similar settings. We wanted to explore what types of qualities help determine whether a business will close down or not. The machine learning algorithms we explored were logistic regression, decision tree, and a Naïve Bayes network. We compared the accuracy, precision, and recall of these different approaches. Using these different models, we were able to examine a problem that many business owners deal with personally.

Process

Our dataset consisted of around 77,000 businesses from 10 major cities (mostly in the US). To begin, we first translated our data from a JSON to a csv file using Yelp's provided data-parser. Before processing our data, we cleaned our dataset to only contain information we deemed necessary to predicting whether a given instance would go out of business or not. Although Yelp's dataset provided us with many specific attributes to describe the different types of businesses being reviewed, we decided to focus on attributes that applied to every type of business. The main ones we focused on were:

- ➢ `open` - boolean value indicating if the business is currently open or closed down. This was our main determining attribute
- ➢ `stars` - float value between 0.0 to 5.0 indicating a business's rating
- ➢ `review_count` - integer value indicating how many Yelp reviews that business has
- ➢ `state` - dummy variable representing what state a business is located in
- ➢ `city` - dummy variable representing what city a business is located in
- ➢ `longitude` - float value of the business' longitude
- ➢ `latitude` - float value of the business' latitude

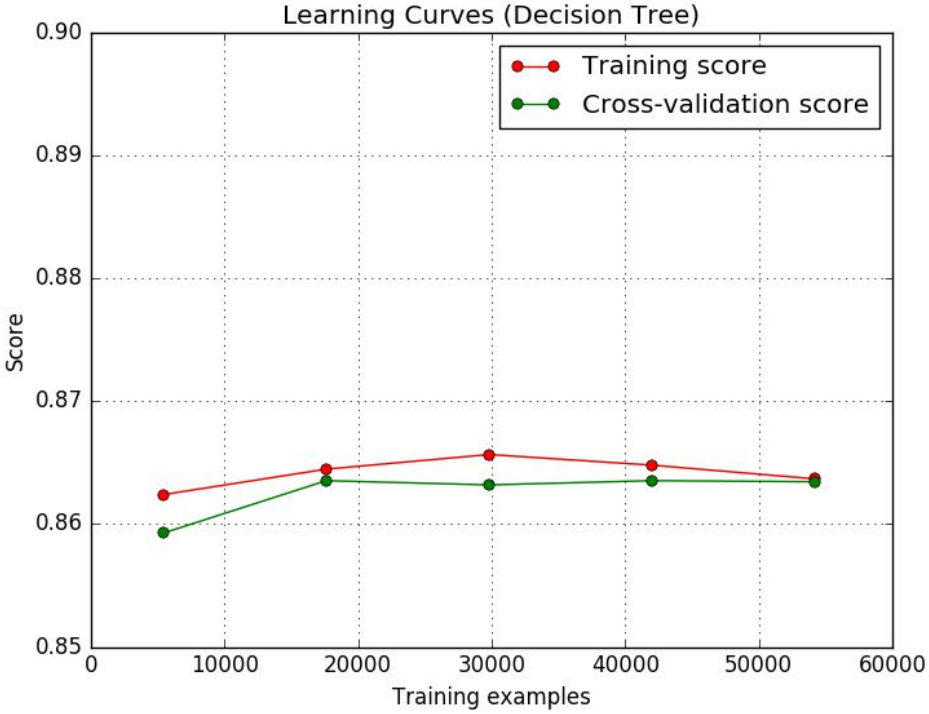
The original dataset included almost 80 attributes in total, many of which were very specific to the type of business (ie. if the business specialized in children's haircuts). The specificity of these types of attributes resulted in many of the businesses in our dataset having no information for the majority of the attributes. As a result, we decided to remove these attributes completely from our dataset because we felt that they would not provide us with a significant amount of information about whether or not a business will close down or not.

The three models we used to test were logistic regression, decision tree, and a Naive Bayes network. To select our models and classify our data, we used the scikit-learn machine learning package for python. This package allowed us to process our csv file using the various models, outputting the training accuracy, test accuracy, test precision and test recall and training time for each model. We experimented with changing different parameters for our models. We found that a depth limit of 5 for the decision tree resulted in better testing scores.

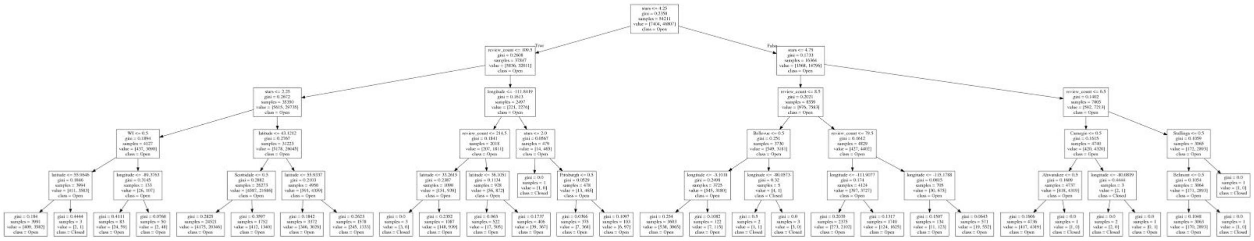
Results

After processing our data, we found that our models yielded the following results:

Model Type	Train Accuracy	Test Accuracy	Test Precision	Test Recall	Training Time
Logistic Regression	0.86368	0.86373	0.86384	0.99985	0:55.413
Decision Tree	0.86366	0.86343	0.86383	0.99945	0:12.507
Naïve Bayes	0.151463	0.14749	0.94000	0.01405	0:05.791



Learning curve for Decision Tree (max depth 5)



Decision Tree Visualization

From looking at the results from the decision tree, we found that the most important feature in determining if a business will close down is its star-rating on Yelp. Naive Bayes had much lower accuracy and recall, but higher precision because it was much more likely to predict a 0 for a dataset that was largely 1s. The scores for Logistic Regression ended up being very similar to the scores for our decision tree. We believe this to be the case because while our models were able often able to successfully classify instances from our dataset, there was not a substantial number of significant attributes that could help influence whether a business would be determined open or closed. Additionally, since the majority of businesses in the dataset were already determined to be open, each model has a higher likelihood to correctly classify an instance as open, which could influence results.

In the testing data set, the percentage of closed business was 84%, so our classifiers only modestly added predictive power. This is likely due to the difficult nature of predicting which and when businesses will go out of business and the impact of omitted variables. Future improvements would be to include more influential features like the sentiment of user reviews and proximity to highly populated areas. This would require joining in not only other Yelp data, but data from external sources, and would not only improve results, but help better distinguish between the different models we tested.