Nathan Timmerman and Micah Thompkins

Machine Learning Checkpoint

1.

For our machine learning component of our project, our features first 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, 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. The following table displays the accuracy results of each.

Classifier	**Accuracy**
Gradient Boosting Tree	0.8197394028038192
KNN	0.8197395779852465
Decision Tree	0.7203676830778033
SVC	0.8197396184117297

2.

The machine learning above is a building block to the majority of our final project. We still want to hone our model and are looking into potentially adding further features to our project. ML helps with our project because the main component of our project is going to be using a machine learning model to see if we can predict whether or not a business is likely to be closed as a result of it's yelp data.

3.

We process text by doing sentiment analysis on the review text of the businesses. We used the TextBlob package and the sentiment command to analyze the review itself and then used the polarity value it returned to add to our machine learning model.

Difficulties.

We were unable to get our model above 82% accuracy. Adding in the distribution of the stars of the reviews did not actually make a difference to the accuracy, although it is statistically-significant between open and closed businesses.