David Terpay
Abhishek Pingle
CS 396
03/01/2020

Yelp Project Checkpoint

1. The features that we will be looking are the different types of attributes and amenities that businesses can have (free parking, delivery, etc) as found in tip.json. The features array will be all possible attributes that you can have given that they are booleans (has parking, kid friendly, etc.). Each data point will have a 1 if it contains a specific feature and 0 if it does not. The targets will be whether or not the business has a rating higher than or equal to 4.5. As such, any business with a rating higher than or equal to 4.5 will have a target of 1, anything lower than 4.5 will be marked as 0. This problem becomes a classification problem. The reason why we combine all of the attributes is because similar businesses will have similar attributes. Hence even if we are comparing a restaurant to a repair shop, we can still score and evaluate them relative to the attributes found in similar businesses.

   The ML algorithms that were considered were KNN, Logistic Regression, and Gradient Boosting Tree Classifier. We decided to utilize the Gradient Boosting Tree Classifier because it happens to be an ensemble method, works well with classification problems, and is trainable with many different parameters.

   The metrics we used to access our model's performance were classification accuracy, a confusion matrix, and the area under the ROC curve. We used the area under the ROC curve as a metric because it can provide insight into our model's ability to discriminate between positive and negative classes. Adding on, the confusion matrix is a useful representation of the accuracy of our model. Finally, we used the classification accuracy just to see the number of correct predictions relative to all predictions made.

David Terpay
Abhishek Pingle
CS 396
03/01/2020

2. The Machine Learning framework described above is useful in determining which attributes of a business drive ratings. The reason why is straightforward. Similar businesses that have similar scores will have similar attributes. As such, we can classify (to some accuracy) whether a given business with certain attributes will be rated poorly or not by implicitly comparing the business to other businesses with similar features.

   This is useful to know as the goal of our project is to identify which attributes distinguish a highly reviewed business from a poorly reviewed one. Additionally, by generating a model that is able to determine valuable target attributes to businesses, we are able to more clearly define which attributes matter to different businesses. While our model doesn't explicitly tell us which attributes are important for different categories of businesses, it does gauge how well it will do given the features it provides. This is a small change to the original project proposal that we had. Instead of directly finding attributes that are valuable, we can classify if the business will have a rating higher than or equal to 4.5. We decided to go down this path because there are too many different types of businesses that were difficult to distinguish by only looking at the raw data.

3. We use reviews in tip.json to do text processing. To start, we only used reviews that originated from businesses in Phoenix, Arizona because it is one of the more reviewed areas. We process our reviews by removing reviews that have any non alphanumeric characters, converting all characters to lowercase, checking for spelling errors and lemmatization. Lastly, we remove any reviews that have fewer than 3 words.

David Terpay
Abhishek Pingle
CS 396
03/01/2020

Additionally, we process the string of attributes that a particular business has to offer by creating a feature mapping and filtering out unwanted attributes. To begin, we only care about attributes that can be expressed as booleans ('has parking', 'reservations', etc.). As such, we ignore all other attributes. Next, we created a mapping that maps a feature to a particular location in the feature vector. This is done by simply finding all relevant attributes and mapping them to an index in some fixed length feature vector. We determined the length of the feature vector by finding all relevant attributes. Once we have a feature mapping, we simply iterate through all businesses and map each feature into the corresponding index where a 1 signifies true (it has that attribute) and 0 is false (it does not have that attribute).

Difficulties:

Since each review does not have the rating that the user gave the business, we found it to be difficult to build out a NLP model to score reviews or utilize any of the NLP frameworks we initially wanted to work with. We did not have any targets for each review except for the rating the business got. However, we cannot assume that each review will have the average rating of the business. Because of this, we decided not to incorporate the sentiment score of reviews in the features or targets.