Data Mining Yelp Reviews

Task 6: Hygiene Prediction

1 Data Preparation Methodology

1) There were three files provided for this assignment. The first was the labels indicating if a restaurant had passed the public health inspection test. The second was Yelp reviews for each restaurant that had a label. The third was additional information containing the restaurants cuisine category and review score.

2) The text reviews required significant data cleaning to configure the data in a way that was useful for modeling. All words were converted to lower case, striped of whitespace characters, Unicode symbols were removed, any non alphanumeric characters were removed, and doubles spaces were removed.

3) The three files were then joined together to create one flat file.

2 Feature Creation

1) A term frequency matrix containing (for unigrams, bigrams, and trigrams) with stop words removed, term frequency–inverse document frequency matrix (for unigrams, bigrams, and trigrams) with stop words removed, part of speech term frequency matrix, character length in all reviews for each restaurant, the number of words in all reviews for each restaurant, sentiment and polarity features were created from the text review data.

2) For the additional data provided, dummy variables were created from the cuisine categories for each restaurant to quantify the impact of each cuisine category on passing the health inspection.

3) New interaction features were created from the text reviews and additional data features. These feature were average character length of reviews, average number of words per review, average character length per rating, average number of words per rating, average word length, number of reviews times average rating, character length times subjectivity, character length times polarity, number of words time subjectivity, number of reviews times polarity, average rating times subjectivity, and average rating times polarity.

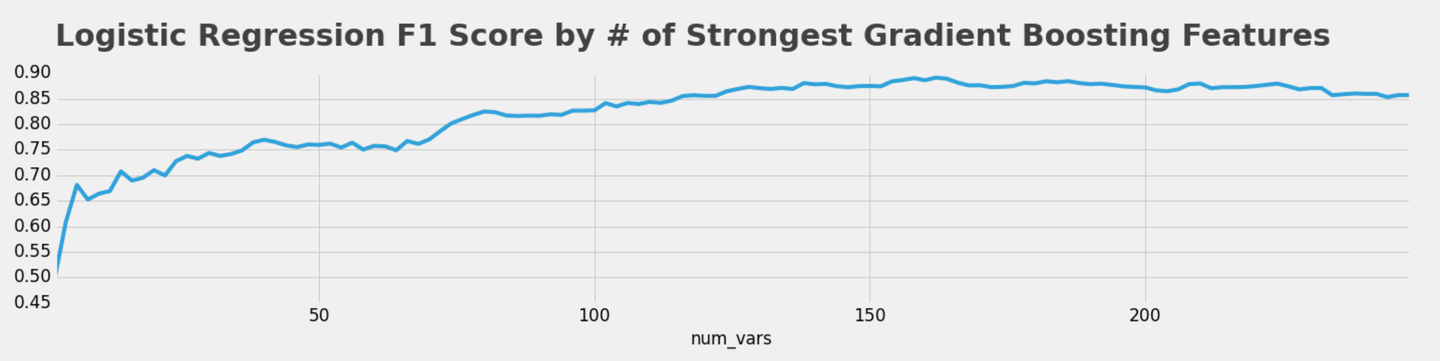
4) The final flat file used for modeling contained over 61,000 features. 546 records were for training and 12,753 were for testing the models performance.

3 Decisions, Analysis and Modeling

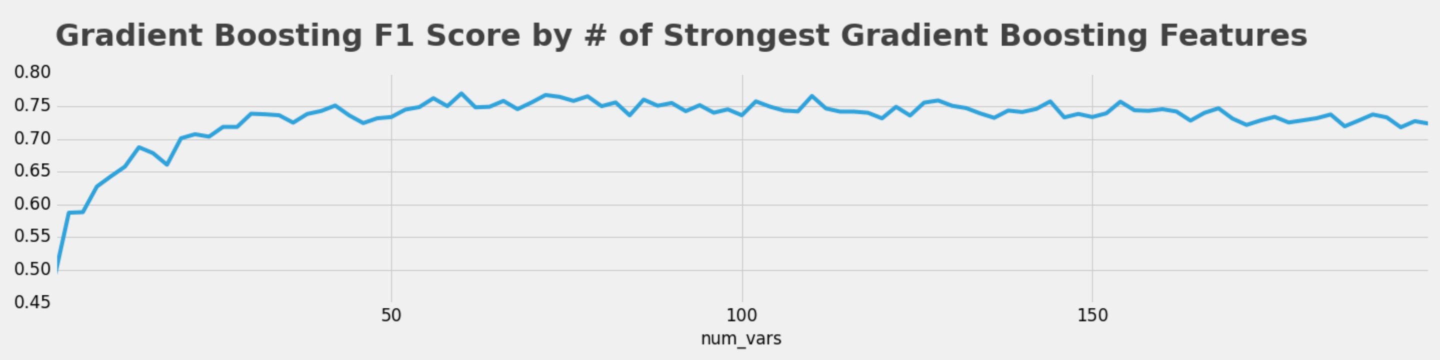
1) The model was built in Python leveraging Pandas for data analysis, scikit-learn for predictive modeling and Text Blob for natural language processing.

2) Making a decision on using a term frequency matrix or a term frequency–inverse document frequency matrix for the text review representation was the first choice that needed to be made to move forward with the model. Each of the matrices were joined with the target label variable. Predictive accuracy was assessed by a random forest model for each matric using 3-fold cross validation. The term frequency matrix scored 0.005 PTTs higher at an F1 score of .611 while the term frequency–inverse document frequency matrix scored at .606. The term frequency matrix was chosen as the primary text representation due to simplicity of the calculation and slightly higher F1 score.

3) Three algorithms were used to compare models: logistic regression, gradient boosting, and random forest.
    - The logistic regression model was built using the feature importance from gradient boosting. The chart below displays the training data mean 3-fold cross validation F1 score by the number of strongest gradient boosting features.
    - Overall, the model fit the training data very well but didn't generalize to the testing. The testing F1 score for logistic regression was 0.51.

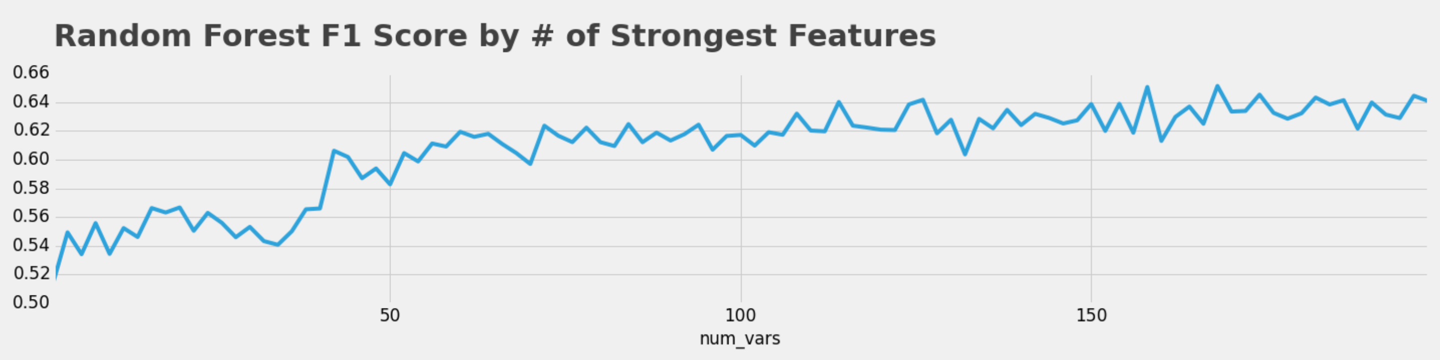


4) Gradient boosting was the best performing algorithm at first; securing 11th place at my first submission but I was not able to improve the model after submission. The chart below shows the performance pf the F1 score capping after the first 125 variables. It was difficult to build a model with gradient boosting that didn't over fit the 546 records contained in the training data.



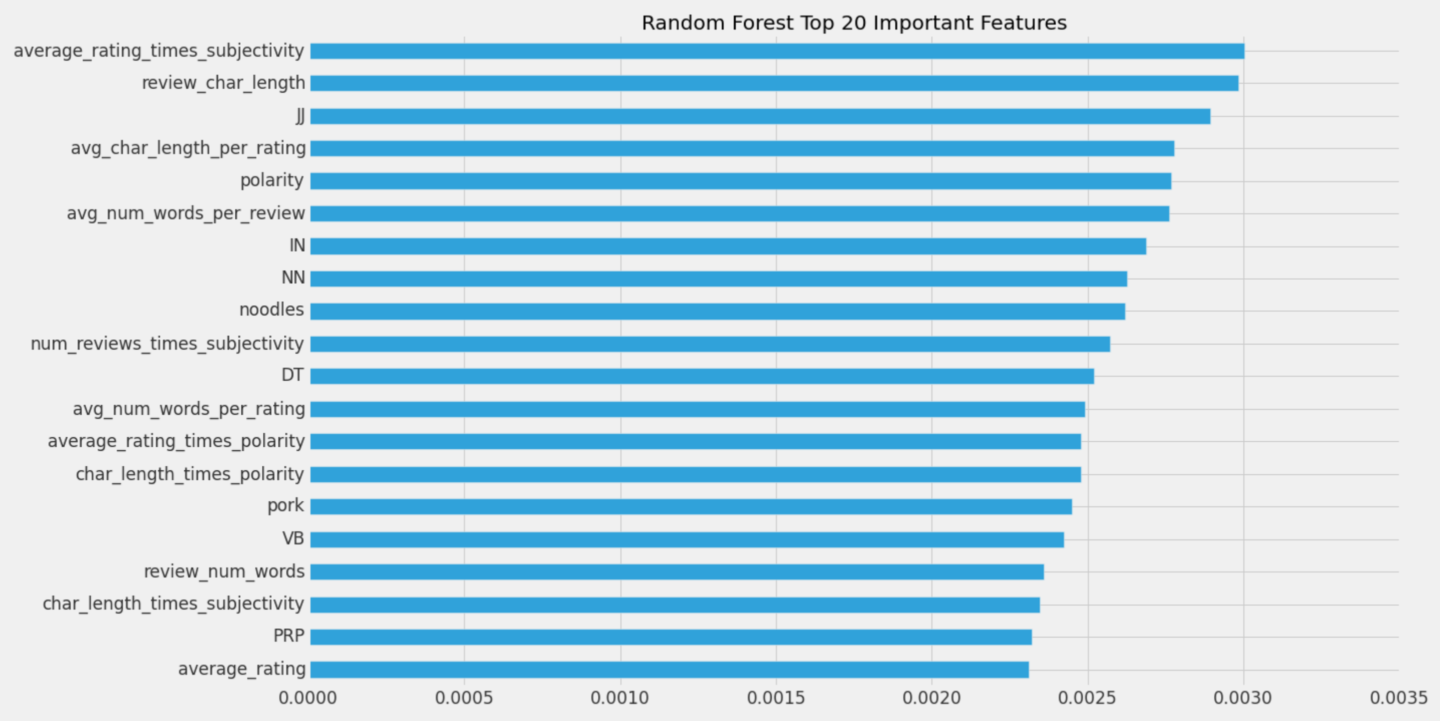
3 Decisions, Analysis and Modeling (continued)

5) Random forest generated the best results and was able to secure **3rd place on the capstone leaderboard** with a testing F1 score of 0.5735.
   - Random forest was able to continually improve its F1 score on the training was well beyond 125 variables making it the clear winner in terms of predictive performance.



4 Additional Insights

1) The image below shows the top predictors of passing a restaurant health inspection.



2) The following final random forest model parameters were used:

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
        max_depth=None, max_features='auto', max_leaf_nodes=None,
        min_samples_leaf=2, min_samples_split=2,
        min_weight_fraction_leaf=0.0, n_estimators=420, n_jobs=1,
        oob_score=False, random_state=None, verbose=0,
        warm_start=False)
```