Food Inspection Predictor for Chicago's food facilities

Capstone Project with General Assembly's Data Science Immersive course

Data Acquisition

The Chicago Data Portal(https://data.cityofchicago.org/Health-Human-Services/Food-Inspections/4ijn-s7e5/data) logs the inspection results for all the food facilities in the city of Chicago. The dataset for my capstone containes information from inspections of restaurants and other food establishments in Chicago from November 1, 2016 to the present. The dataset has a total of 20635 rows and 17 columns.

Data Cleaning, Transformation and Preprocessing

The Violations column in the dataset was checked for null and was replaced with "None" values and rest of the null values were dropped from the dataset. Since it a classification problem, the Results column was checked for the unique values and the values were classified into two values : Pass or Fail. This value was in turn changed into a binary value , where 1 stands for Pass and 0 stands for Fail. The Risk column was a string column and was transformed into a category.
The project here used the Violation column and predicted the result whether the food facility passed or failed the inspection. The Violation column contains the remarks which were noted during the inspection. To change these remarks into categories Natural Language Processing was used. First the punctuation was removed from the Violations column. After the punctuation removal the stop words were removed. Then the Stemmer normalized the words to a common root. Then train_test_split was used to split up the data set into training and testing data. Then CountVectorizer was used which converted the collection of text documents to a matrix of token counts

Modeling

The modeling techniques explored here include a neural network , bagging and boosting ensemble methods(RandomForestClassifier, GradientBoostingClassifier, AdaBoostClassifier), KNeighborsClassifier,Linear Support Vector Machines Classifier and Logistic Regression which are all part of the Sci-kitLearn package of programs. The train set was fit and transformed to the given models and then the test set was run to check for the accuracy in predicting the target.
The receiver operating characteristic (ROC) curve was also plotted, which is defined as a plot of test sensitivity as the y coordinate versus its 1-specificity or false positive rate (FPR) as the x coordinate. It is an effective method of evaluating the performance of diagnostic tests.

Results

The base score for the model was calculated to be .673 . All the models achieved a good accuracy score but the one model that outshines in the accuracy was Logistic Regression. The test set achieved the accuracy of 0.900 which makes this model a success. The precision and recall score were also considered just to make sure that a lot of credence was not put on raw accuracy alone. The precision and recall score came out to be 0.91 and 0.90 respectively, which is a pretty good estimate of the model's accuracy in predicting the target on the test set.

Next Steps

Look more in-depth at the relation of the Results based on the ZIP code and build a way to make the output attractive and interactive