DATAROBOT EXERCISE

Yong Phelan Nkfum

06/05/2015

Introduction

The purpose of this exercise is to use the Google Prediction API to derive a predictive model on a given data set, run predictions against the data and answer the question *"Do you think these predictions are good".*

Before we jump to the conclusions of this exercise, let's explore the steps that brought about these conclusions.

Data selection

The data was acquired from a Kaggle competition at https://www.kaggle.com/c/facebook-recruiting-iv-human-or-bot/data. The competition organized by Facebook required competitors to build a predictive model from the given data that can predict whether the bid on an auction site was generated by a human (*the negative class represented by 0*) or a robot (*the positive class represented by 1*).

The training set from this competition had about 2200 instances, with about 2100 representing the negative class and about 100 representing the positive class (which we are more interested in). These numbers tell us that the data has *high class imbalance* which degrades the quality of our predictive model.

Data preprocessing

Before training the model with the Google Prediction API, a few steps were taken to preprocess the data. Using the program Weka, I created a 10% random subsample of the training set (from Kaggle) with no replacement of instances. I also used the option of keeping the classes balanced when creating the subsample to account for the class imbalance problem. I chose 10 percent to keep the distribution of classes fairly balanced. This resulting file had about 200 instances with a balanced distribution of both classes.

Next I checked to make sure there were no missing values for the different instances, then I randomized the instances to maintain the class imbalance when I split the subsample into a training set and a test set. The final step for preprocessing was to split my subsample into a training set which contained about 2/3 of the subsample population and a test set which contained the remaining 1/3

Data mining

The training set was uploaded to the Google Prediction API and the prediction model was built.

Conclusion – Pattern Evaluation

Using the python script and the test file included with this report, the predictive model was used to make predictions. The predicted outcomes were compared to the actual outcomes to evaluate the accuracy of the predictive model. We would focus on the *Sensitivity and Specificity evaluations*.

The Sensitivity, calculated by (TruePositives/ActualPositives) was 0 for our model. This means that the model did not identify any robot (our positive class).

The Specificity, calculated by (TrueNegatives/ActualNegatives) was 1. This means that the model identified every human (our negative class).

So the answer to the question asked at the beginning is *"No" these are not good predictions because the model has failed to correctly identify any member of the class of interest*. In my opinion, the poor results are because there is not enough data for the model to accurately identify a pattern.