Classification Exercise Write Up

Pujan Malavia

The two classification algorithms I used were Random Forest and AdaBoost. The Random Forest is a bagging technique where a certain number of decision trees are grown on different subsets of the training data. It is also known to have low bias and high variance and has an equal amount of say in the final decision in modeling. AdaBoost is a boosting technique that uses stumps instead which would be a decision tree with only one split. These stumps, also known as weak learners, have high bias and low variance). The AdaBoost also has a different amount of say in the final decision.

In the first trial, I performed the following data preparation/data cleansing techniques for both the train/test and validation set.

- Checking missing values
- Doing mean imputation since roughly less than a fraction of a percent of data was missing
- Standardizing/normalizing data based on a scale from 0 to 1.
- Performed PCA (feature extraction) to see which variables show most variance among dataset
- Identified the top ten predictors based on PCA plot to test out on model.
- Visualized data via a histogram, boxplot, PCA plot, heat map(s), and a Scree plot.

The Random Forest performed with an **82.75% accuracy** on the training set and an **81.43% accuracy** on the test set. The AdaBoost performed with an **80.6% accuracy** on the training set and an **80.7% accuracy** the test set.

In the second trial, I performed the following data preparation/cleansing techniques for both the train/test and validation set.

- Treated outliers by winsorizing and capping certain variables
- Bucketed/binning/categorized numerical data
- Performed interactions between predictors to see new patterns
- Tuned the model to help improve the overall model accuracy

However, in the second trial, the accuracy level did improve. The Random Forest performed with an **91.31% accuracy** on the training set and with an **91.36% accuracy** on the test set. The AdaBoost performed with an **93.2% accuracy** on the training set and with an **96.8% accuracy** on the test set. Thus, the highest model accuracy I received was from AdaBoost.

I believe that the AdaBoost would perform the best with what the data I was given. I believe that the choices I made in the context of the exercise would be different in a business context. For example, this dataset only included numerical variables as maybe in a business context, there would categorical variables that could be ordinal in nature (which allows transformation of those variables) to see their impact on the model. Sometimes, real-world data seems to be much more messier in nature where they could have a higher percentage of missing values in the dataset as well as more outliers that could skew the model (obviously depending on the methods used to treat them). In addition, while working a real-world setting, you would know some variables that historically tend to predict well (from a business context), and which ones do not.

To a business partner, I would explain that the Random Forest (all else equal) would work better for complex data (high variance, low bias) that’s a bit more unknown in terms of predictors’ effect on the response variable since it looks at all predictor variables equally in terms of its importance. However, for AdaBoost, although it has a higher accuracy rate, is better for ‘biased’ data vs. data with a lot of variance. However, the caveat is that sometimes the results of AdaBoost has a higher probability of seeing new data and predicting ‘wrong’ if that new set of data has more variance.