TRI-AD

Data Scientist Challenge

Ankush Khullar

May 6, 2020

**Part 2: Model Creation**

The model to predict whether a given student will pass the test was created using Python. Python was selected because of its extensive set of intuitive libraries for data analysis (numpy, pandas, scipy), data visualization (matplotlib, seaborn), and machine learning (scikit-learn).

1. **Data Cleaning**

Based on the results from Part 1, it was determined that sex, age, and dojo\_class had a significant relationship to whether a student passes the test. Therefore, the first step was to remove the variables first, last, country, lang, and notes. Next the categorical variable sex and the boolean variables dojo class and pass were mapped to integers (i.e., 1 for Male/True, 0 for Female/False).

1. **Data Imputation**

Hours studied is likely to be related to whether a student passes the test. However, 40% of the observations are missing values for the hours\_studied variable. One possible solution is to remove these observations. However, this would result in the loss of a significant amount of data. Therefore, the missing values were imputed.

One possible imputation method is to replace all missing values with the mean of hours studied. However, this would reduce the variance of hours studied, which would result in a loss of information. Instead, hours studied was imputed using k-Nearest Neighbors (k-NN) with k = 5. The variable pass was excluded in the k-NN analysis to prevent leakage of information from pass into hours\_studied.

1. **Model Selection**

The data was split into training and test sets using 80% and 20% of the data, respectively. Approximately 78% of students in the training set passed the test, while 22% failed. Since the training set was imbalanced, the SMOTE algorithm was used to create synthetic observations of the minority class, which in this case was students that failed the test.

Several machine learning models for classification were tested, including logistic regression, support vector machine, decision tree, random forest, gradient boosted trees, multiple layer perceptron (MLP), k-NN, linear discriminant analysis, AdaBoost, and XGBoost. A mix of linear and non-linear and parametric and non-parametric methods were selected to ensure that that proper relationship between the dependent variable (pass) and the predictors (age, sex, hours\_studied, and dojo\_class) was identified. The best model was selected using 10-fold cross-validation on the training set using accuracy as the model evaluation criteria.

The models with the highest mean accuracy on the training set was random forest (see figure below). It is interesting to note that the models that performed the worst all assume a linear decision boundary. This indicates that the relationship between the predictors and whether a student passes the exam is not linear.

A screenshot of a cell phone

Description automatically generated

The random forest model was selected as the best model. The training set accuracy for the random forest was 92.0%, and the test set accuracy was 90.6%.