**A Comparative Analysis of ML Techniques for Student Retention Management**

**Maggie Tang**

**The College of Saint Rose**

**Fall 2019**

Purpose of this study is to use different machine learning models making prediction of college students’ retention and identify the most important variables that effect the retention.

Dataset used for this study came from a U.S. institution, entailed to 16066 students enrolled as freshmen between 2004 and 2008. Seven machine learning models are applied to this study: 4 individual models (Artificial Neural Network, Decision Tree, Logistic Regression, support vector machine); 3 ensemble models (Bagging, Busting, Information fusion). Each model run twice on both unbalanced data and balanced data with a 10-fold cross valiadition. For the unbalanced data, SVM produced the best results with an overall prediction rate of 87.23%. But accuracy for “No” class got from unbalanced data were under 50%, it’s not acceptable. The next round of experiment, the study used a well-balanced data, SVM generated the best overall prediction accuracy again. The prediction models did significantly better on predicting the “No” class with well-balanced data.

Although SVM predicted the best, but decision tree might be a better choice, because decision tree is more interpretable, and show the reasoning process of different outcomes.

It is also show ensemble models predicted better than individual models.

The aggregated sensitivity analysis report indicated the most important factors for student retention are related to the past and present academic success, and financial help. Administers and decision-maker can craft targeted programs to retain students.