Summary of Independent Study
Joyce Tian
December 2018

In this independent study, I set out to investigate the possibility of leveraging machine learning models to speed up search fund evaluation process. Data samples were collected through the collaboration with a search fund principle and his research team. The data set contains a target column identifying whether to “accept” or “reject” any particular company as acquisition target, and seventeen attribute columns containing company profile information that are used as input for the machine learning. Nine machine learning classification algorithms were tested in this study under various conditions with prediction performance measured and compared using metrics like accuracy, confusion matrix, precision, recall, and ROC (receiver operating characteristic). Challenges I experienced during this study include small data size problem, target class imbalance, and incomplete and unaccredited information sources. Data pre-processing techniques such as Synthetic Minority Over-sampling Technique (SMOTE) and dimension expansion were applied to overcome class imbalance by generating synthetic samples for the minority class and increasing the data size by adding synthetic features to the data set. Prediction accuracy were compared before and after applying the dimension expansion technique using Logistic Regression and a noticeable increase in accuracy was observed along with the reduction of false negative (type 2 error). On feature engineering side, Recursive Feature Elimination (RFE) was used to rank and select most important attributes, and impact of applying Principal Component Analysis (PCA) on the data set was also tested. A particularly interesting result is that the choice of estimator algorithms used in RFE process has direct impact to downstream classifier in terms of accuracy. Overall, the machine learning models achieved classification accuracy in the range of 75% to 90% depending on specific algorithms used, indicating the idea of creating a predictive model to facilitate search fund process holds potential and should be explored further with more accurate and robust datasets.