# 6000B Project 1 Report

Liu Yang (20475562)

1. Preprocessing the Data

   Since the train data attributes have different scales, I standardize the train data by the formula

$$x_{ij'} = \frac{x_{ij} - mean(X_i)}{std(X_i)}$$

   where x-ij is the i-th observation for the j-th attribute, mean(Xi) is the mean of the i-th attribute of all observations and std(Xi) is the standard deviation of the i-th attribute of all observation.

2. Training the Model

   In this project, we use totally seven single classifiers and take the majority votes as the final prediction result. The seven classifiers are:

   (1) K-Nearest Neighbors Classifier (KNN)

   (2) Logistic Regression Classifier (LR)

   (3) Random Forest Classifier (RF)

   (4) Decision Tree Classifier (DT)

   (5) Support Vector Machine Classifier (SVM)

   (6) Gradient Boosting Classifier (GBDT)

   (7) XGBoost Classifier (XGB)

3. Evaluation the Result

   We divide the training data into 4:1 and use the 20% of training data as the validation set to evaluate the performance of the classifiers based on accuracy. The result is shown in the table below.

| classifier | accuracy |
|---|---|
| KNN | 0.89 |
| LR | 0.925 |
| RF | 0.944 |
| DT | 0.899 |
| SVM | 0.947 |
| GBDT | 0.957 |
| XGB | 0.949 |
| Ensemble | 0.96 |