For the data we got, we split the data into two sets: training set and test set, which respectively account for 80% and 20%. The training set and test set has the same target distribution as the dataset we got.

Then we select the features in the following different ways:

1. Original set
2. Data set after using ridge to select features
3. Data set after removing features which are highly correlated.
4. Data set after removing constant, Quasi-constant feature, and duplicate features.

For every feature selection, we run the 4 kinds of classifier model, which are Random Forest, Decision Tree, Logistic Regression, and Naïve Bayes. By using the 5-fold cross-validation, we can get rough estimate of accuracy for each model.

According to the performances of the models, we choose the best one, RF, and try to adjust the hyperparameters of the model such that the model can have a better performance.