# Technical Methodology

*Jiaxi Wu*

## 1. Data management and overall workflow

For the credit default , we estimate the performance of linear regression and regression tree. We follow the following workflow:

(1). Data preprocessing: denote default as 1 and no default as 0. Delete variables and samples with entire data class missing. Ignore the two metrics with less than 5% data available because it's not reasonable to assign missing values for them. Assign remaining single missing value with average of previous known and next known values. Then restore the whole data set for usage of regression tree and pick the last values of each financial metric as predictors for OLS.

(2). Divide the data into a training sample and a hold-out sample.

(3). Run each prediction algorithm on training sample. For linear regression, just run the algorithm on the full training sample and store the result. For regression tree, we tune empirically on the training sample by cross-validation.

(4). After storing all the prediction functions, we run each algorithm on the hold-out sample and produce the statistics we are interested in such as MSE and $R^2$.

## 2. Training & Out-of-Sample Datasets

It's necessary to separate a hold-out(testing) sample to assess the model performance since good in-sample performance may overstate the overall performance and overfit the model. In this assignment, we separate 20% of the original data as a hold-out sample to test the model performance.

## 3. Model Parameters

For OLS, we can't use too many variables, so we only use the last value of each financial metric as predictors. Then we run the algorithm on the full training sample and store the result.

For regression tree, we choose complexity parameters like tree depth by cross-validation within the training sample. We randomly divide the training data into 5 folds. For each fold, we fit the model for every tuning parameter value on all other folds and predict on the given fold. Then we obtain one prediction for every tuning parameter on each fold and average the prediction loss over the full training sample for each tuning parameter. Based on the loss, we choose the tuning parameter and then fit the model with the chosen tuning parameter on the full training sample. Finally evaluate the fitted function on the hold-out sample to test the performance of model.

When considering multiple tuning parameters, such as the depth of the tree and the minimal size of a non-terminal node in this assignment, we choose these parameters jointly; that is, we run the above procedure over a grid of multi-dimensional parameter. As a result, minsplit = 20 and depth = 8 has the best hold-out performance.

## 4. Fit & Performance

As a result, the OLS model has in-sample and hold-out $R^2$ around 0.1 and MSE around 0.09. While after tuning, the Regression Tree has optimal hold-out $R^2$ of 0.2 and MSE 0.079, with tree depth 8 and minimum

units in each non-terminal node 20. Under such setting, in-sample $R^2$ for Regression Tree is 0.62 and MSE is 0.037. Hence Regression Tree has a better performance than OLS on this data set.