**Executive Summary:**

The business target is to generate a credit default forecasting model by using the financial and accounting measurements from 10000 U.S. public companies. The dataset has 9 different class such as EBITA and ranges from 2008 to 2017 for each variable. These classes are all important and can be used to analyze a company’s financial condition. Two machine learning models are used in building the model. The first one is ordinary least square(OLS) which is a linear regression model, and the second one is decision tree model. For results, since the dataset contains too many missing values, the OLS is not desired with a close to zero model explained variances. The decision tree model perform quite well with predetermined tuning parameters with a 45% R^2. In order to improve the model, we could run other methods such as Lasso which is good at dealing with sparse dataset. A more strict filling for missing points might also help improve the model.

**Technical Methodology:**

*Data Management & Workflow* : The dependent variable(y) is whether a company’s S&P credit rating being downgraded during 2008-2017. If so, value marked as 0, otherwise marked as 1. Independent variables(X) is 90 measurements. I will clean the data first and then train two models and test performance.

*Training & Out-of-Sample Datasets* : Missing values are randomly assigned 0/1 if entire class is missing, forward/backward fill with nearest values or filled with average of previous and next values. The dataset is split in 80%-20% for test/holdout set. No need for data standardization since unit is the same for X.

*Model Parameters* : Nothing sets for OLS. As for Decision tree, tree depth is 8, minimum units for node is 20, and minimum improvement for split is 1% for R^2.

*Fit & Performance* ：OLS model does not fit due to large number of self-assigned missing values. Test R^2 is almost zero and test MSE is around 0.03. Decision Tree model has relatively good performance with a test R^2 of 0.45% and test MSE of 0.017. In general, we could use other models(e.g. Lasso) to improve performance.

the company is marked as default and assigned a dummy variable of 0. Conversely, 1 is assigned to a company which never been downgraded before. For a company with no available data points, we will delete it from the data set since it provides nothing to the company