**MSc in Data Sciences and Business Analytics**

**Ensemble Learning**

**Taiwanese Companies Bankruptcy Prediction Model**

**Members:**

Syed Ahsan Abbas Naqvi

Raghuwanch Raj

Basma Wehbe

Rishikeshwan Rangarajan

Shivaditya Meduri

**Outline**

Be it policy makers, risk managers or investors, having the ability to signal bankruptcy predictions of companies is of utmost importance to them. However, as much as exogenous variables have a bearing on the going concern of an enterprise, such as demand, backward supply or economic environment, an enterprise’s own financial indicators inform a lot about sustainability of an entity. In this regard, we attain a decade-long dataset from the Taiwan Economic Journal for the years 1999 to 2009 containing financial statement data on companies. The dataset contains 6,819 observations containing 95 explanatory variables. It is a classification problem which predicts whether a company is set to bankruptcy or not. We believe that this exercise would help stakeholders better inform their investment, regulatory or risk management decision making when presented with financial statement data of companies. More specifically, all these variables may be categorized into financial ratios (FRs) and corporate governance indicators (CGIs). We propose to employ ensemble methods such as decision trees, random forests for implementation of our predictive algorithms.

**Proposed Methodology**

We would first explore if there are any missing values in the dataset. If they are very few, we could drop them out. However, if that is not the case, then we would impute them. In addition, we would be conducting exploratory data analysis first in order to determine distributional characteristics of the datapoints, and identify any outliers. One thing we would have to be careful about is the nature of the attributes. Since these are financial information, outliers could actually carry predictive power in informing bankruptcy classification. So, domain knowledge of corporate finance would be required to decide whether to retain some outlier variable or not.

Furthermore, as these are a lot of variables, we will try to implement dimensionality reduction first using PCA. Once we have information as to which variables explain the most in predicting bankruptcy, we will extract them out. Subsequently, we will break the dataset into train and test sets. Ensemble methods would be utilized on train set for the models to learn and then used to generate predictions on test set. The exercise will be restricted to methods taught in this course only. Cross validation will be used to generate suitable parameters as well. Lastly, models will be evaluated based on F-1 scores and confusion matrix. The best model would then be proposed along with prediction results.