**ADMN5006 Analytics for Financial Markets**

**Assignment #1**

Your assignment is to design and develop machine learning models for corporate bankruptcy prediction 2 years into the future using the data provided. You can use either R or Python. The column “BK” in the data provided denotes whether the company goes bankrupt (indicated by 1) 2 years in the future, or not (indicated by 0). Some of the input variables included are *“Assets Growth”, “Sales Growth”, “Earnings-per-share (EPS)”, “Return-on-equity (ROE)”*, etc. In addition, the paper entitled: “Machine learning models and bankruptcy prediction” by Barboza et. al. is posted to the portal and is an excellent reference for you to use. Note however that the dataset you are given is not the same as the one used in the aforementioned paper. Therefore, don’t be alarmed if your results don’t agree with the Barboza et. al. paper. You will need to partition your data into testing and training sets and explain your experimental set-up.

**You need to try and implement at least 2 machine learning models from the following list:**

*Random Forest, Support Vector Machines, Logistic Regression, Discriminant Analysis (LDA & QDA), Naïve Bayes Classification.*

You are free to try others as well. You should think about the business application that you plan to use your algorithm and clearly explain the criteria you will use when comparing different algorithms.

You should clearly explain data preprocessing (if any), and outline the methodology that you use for each of the models. You should clearly mention a step by step process of how you train the machine, test and validate the models with hyper parameter tuning. Note that bankruptcy problems are highly imbalanced with far more observations being non-bankrupt than bankrupt. You will have to deal with this problem. In the end, you need to explain your results and provide a comparative recommendation on model selection for the bankruptcy problem. Your report should also include a recommendation of how your models may be used in a practical setting in the financial world. All explanations are to be written in the Python Notebook

The input and output variables are described below:

EPS – Earnings Per Share

Liquidity – Working Capital/Total Assets

Profitability – Retained Earnings/Total Assets

Productivity – EBIT/Total Assets

Leverage Ratio – (Total Long-term debt + Debt in Current liabilities)/Stockholders Equity

Asset Turnover – Sales/ Total Assets

Operational Margin – EBIT/Sales

Market Book Ratio – (Price Close Annual Fiscal \* Common Shares Outstanding)/Book Value Per Share

Asset Growth – Change in assets from previous year

Sales Growth – Change in sales from previous year

Employee Growth – Change in employees from previous year

Tobin’s Q – (Total market value of company + liabilities)/ (Total asset or book value + liabilities)

BK – Company bankrupt or not

**Note**: *While model parameters are learned during training — such as the slope and intercept in a linear regression — hyperparameters must be set by the data scientist before training. Some examples of hyper parameters are:*

***Random Forest*** *– No. of trees, No. of features considered by each tree when splitting, Stratified sampling, Sampling with/without replacement etc.*

***SVM****: Regularization Parameter, C, Class Weighting, etc.*

***Naïve Bayes****: Smoothing, Subset validation, etc.*