DATS 6103 FINAL PROJECT
EXPLORING COMPANY BANKRUPTCY
05.03.2021
Amar Adusumilli, Kardelen Cicek, Tarush Gupta
Group Report

1. Introduction

The increasing inter-connectivity of the world has placed significant importance on international trade and commerce, with the caveat of creating new sources of risk from international partners. In finance literature, counterparty risk is defined as the risk that one or more of the parties involved in an agreement may not be able to meet their financial obligations. The severity of the consequences associated with this risk are evidenced by the global nature of recessions, most prominently the Great Recession, where a lending crisis in the United States directly contributed to economic decline in Europe. Perhaps the most severe consequence any individual firm can face is bankruptcy. Firm bankruptcies can damage the financial health of other companies and erode the financial ratings of national economies.

Because of such financial uncertainty, publicly traded firms are expected to report their financial health in annual filings to their respective national financial bureaus, such as the Securities and Exchanges Commission in the United States. In these reports, firms must audit and declare their accurate economic standing using financial ratios and transparency statements. These financial ratios have been a touchstone of financial analyses and speculation fodder for investors. In addition to financial ratios, companies also prepare corporate governance indicators that include the mechanisms and processes by which these corporations are controlled and directed (Shailer, 2004). Such controls permit shareholders to exercise their right to oversee the operations of the firm and ensure that the financial and regulatory expectations are met.

The severe consequences of firm bankruptcies make identifying the factors which contribute to bankruptcy, as well as predicting which firms will go bankrupt themselves, important. Modeling the parameters that can determine whether a company is likely to declare bankruptcy can assist financial institutions and investors with making appropriate lending and trading decisions. An array of statistical, and more recently, machine learning methods are being employed in the economics and data science literature to predict these outcomes. Both model types rely on these financial ratios and corporate governance indicators to capture whether a firm is likely to go bankrupt. While earlier studies that lacked richer data only analyze the financial ratios, of late, academia has embraced a combination of these factors in predicting bankruptcy (Altman, 1968; Lee and Yeh, 2004; Lin et al., 2011).

We undertake the task of predicting bankruptcy and identifying its determinants. To that end, we use a Kaggle dataset, collected from the Taiwan Economic Journal (https://www.finasia.biz/) between 1999 and 2009 – to test machine learning and inferential models to predict which firms

will go bankrupt, and to determine the most important indicators of bankruptcy. The Kaggle data were obtained from the University of California, Irvine, Machine Learning Repository (Liang et al., 2016). We delve into the characteristics of this dataset in the next section.

We begin with inspecting the data and conducting exploratory data analysis. Built on these findings, we pre-process through feature reduction and outlier cleaning. We then model the data using an econometric, inferential approach, in order to ascertain the partial effect of the features on the probability of bankruptcy. Next, we run tree-based shallow machine learning models on this data to best predict which firms will file for bankruptcy. We elaborate on these steps in section 3. The findings of these models help us identify the most informative determinants of bankruptcy, which we discuss further in section 4. We conclude this report with our key findings, their impact, and the limitations of our study in section 5.

2. Data

We use the dataset, "Company Bankruptcy Prediction" which can be accessed through the following link (https://www.kaggle.com/fedesoriano/company-bankruptcy-prediction). This dataset was uploaded by Kaggle user fedesoriano in February 2021. It has since received 74,300 views and 8,159 downloads (or an interaction rate of 0.11), with 12 notebook submissions.

This data was collected by the Taiwan Economic Journal spanning the years between 1999 and 2009 and describes the financial health of firms operating in Taiwan during this this period. The target variable is a binary indicator variable that indicates company bankruptcy (1, if firm went bankrupt; 0, otherwise). The definition of bankruptcy in this context is based on the Taiwanese Stock Exchange regulations. The companies represented in this dataset are public firms with at least 3 consecutive years of public filings prior to the financial crisis. The companies included represent a wide variety of industries: manufacturing, service, and retail, among others.

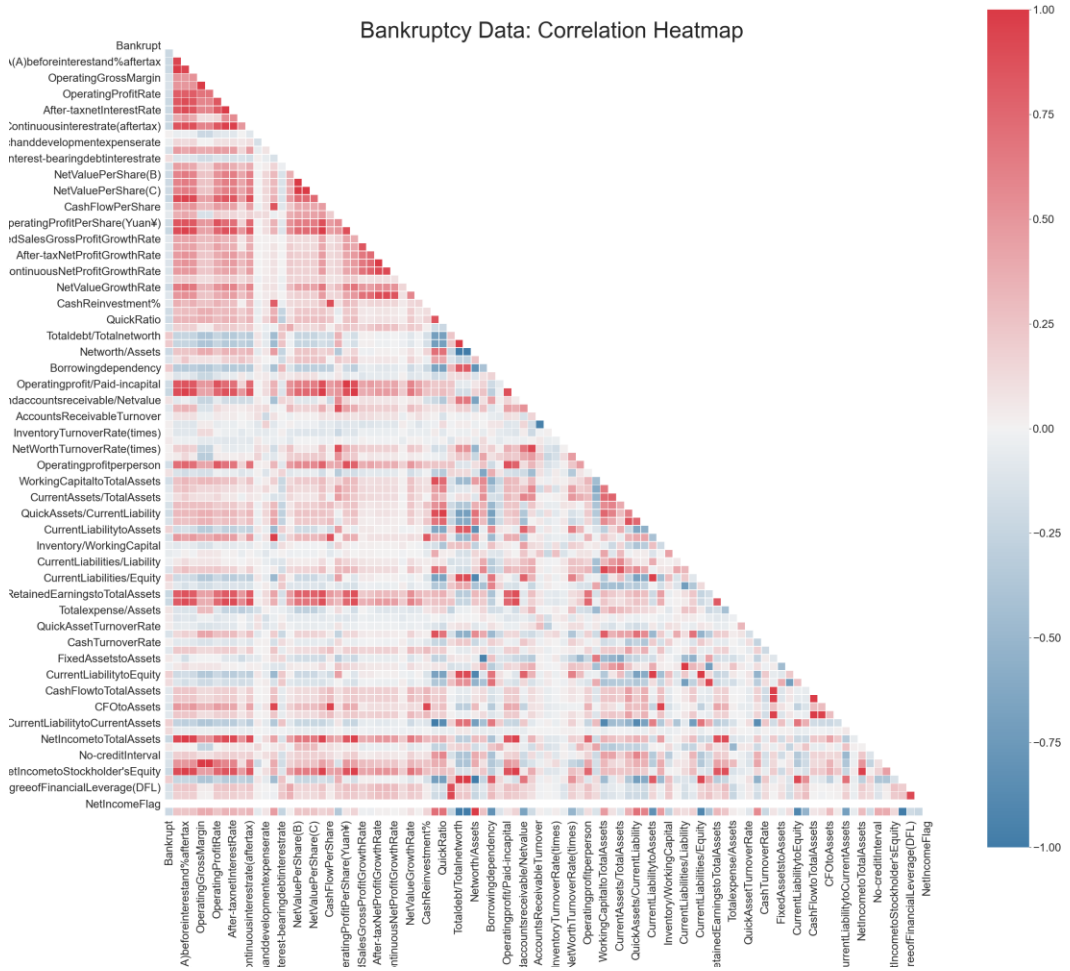
This dataset contains 6,819 observations and 96 columns – one bankruptcy indicator variable and 95 variables that can be mostly characterized as either financial ratios or corporate governance indicators. The data exhibits no missing values. It is well-organized and does not contain duplicate values, and hence does not require extensive cleaning; however, the binary target variable is highly imbalanced, and requires handling. Unattended, the imbalanced class issue can lead to biased machine learning models that will lean towards predicting against bankruptcy outcomes, as the proportion of non-bankruptcy cases far outweighs the proportion of bankruptcies.

We begin by exploring the properties of this dataset using Python 3.9. We import the comma-separated-values (.csv) file that contains our data and obtain the column properties. Each of the 96 columns is coded as an integer or a float. Consequently, we do not need to dummy encode any of the features. The column names contain spaces which are removed for modeling and manipulation purposes, and one blank column is removed ('Net Income Flag'). Exploring the variables, we immediately notice that many are redundant – there are three measures of Return on Assets (ROA), Net Value per Share, among other connected features. This can confound the statistical issue of multicollinearity in our data, wherein the variance of the parameter estimate in

inferential models will be largely inflated, potentially leading to Type 2 errors. We tackle this redundancy issue by reducing the dimensionality of our dataset by eliminating highly correlated features. Specifically, we plot correlation matrices that help us visualize and observe the magnitude of their interconnectedness. We come up with a 95% significance level rule, whereby variables that are at least 95% collinearly related, are removed from our analysis.

To roughly measure the importance of each of these features against the target variable, we plot the Spearman's Rank Correlation matrix that assesses the fit of the monotonic relationship between these variables. The Spearman correlation between any two variables is the Pearson correlation between their ranked values. The Spearman Correlation Matrix is shown as a heatmap in Figure 1, below. This indicates that most of our variables are only weakly correlated with Bankruptcy. Intuitively, we notice that the debt measures tend to be positively correlated with a likelihood of Bankruptcy, and the profitability measures are negatively correlated with a Bankruptcy event. The Spearman's Rank Correlation Coefficients indicate that there are 30 variables that meet our criterion for at least a |95%| correlational relationship.

Figure 1: Spearman Correlation Heatmap



Next, we calculate the Pearson correlation matrix, and apply our multicollinearity rule to find that 17 variables meet our definition. We compare the columns identified by both the correlation matrices, and inspect the 13 additional columns captured by the Spearman's Rank. Of these 13, we find that some variables – including, but not limited to, (Total Debt)/(Total Net Worth), Quick Ratio and Current Ratio – are statistically significant explanatory variables behind Bankruptcy, and hence, drop the 17 columns captured by the Pearson Correlation matrix as opposed to the Spearman matrix. We show the statistical summary for this data in Table 1 and distribution plots in Figure 2.

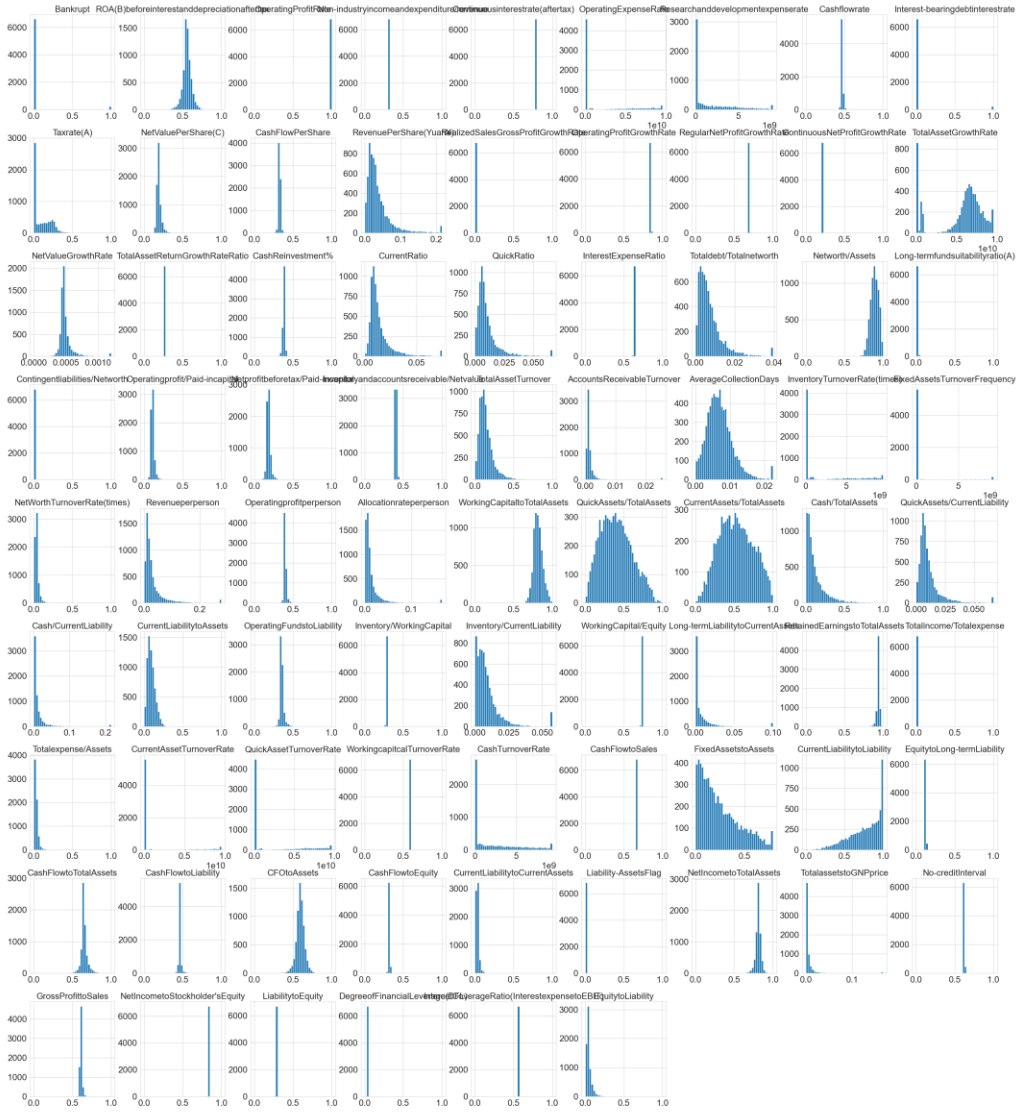
Table 1: Bankruptcy Data Summary Statistics

Variable	Mean	Std. Dev.	Min	25%	50%	75%	Max
Bankrupt	0.032	0.177	0	0.000	0.000	0.000	1
ROA(B)beforeinterestanddepreciationaftertax	0.554	0.062	0	0.527	0.552	0.584	1
OperatingProfitRate	0.999	0.013	0	0.999	0.999	0.999	1
Non-industryincomeandexpenditure/revenue	0.304	0.011	0	0.303	0.304	0.304	1
Continuousinterestrate(aftertax)	0.781	0.013	0	0.782	0.782	0.782	1
OperatingExpenseRate	1995347312.803	3237683890.522	0	0.000	0.000	4145000000.000	999000000
Researchanddevelopmentexpenserate	1950427306.057	2598291553.998	0	0.000	509000000.000	345000000.000	998000000
Cashflowrate	0.467	0.017	0	0.462	0.465	0.471	1
Interest-bearingdebtinterestrate	16448012.906	10827503.533	0	0.000	0.000	0.001	990000000
Taxrate(A)	0.115	0.139	0	0.000	0.073	0.206	1
NetValuePerShare(C)	0.191	0.033	0	0.174	0.184	0.200	1
CashFlowPerShare	0.323	0.018	0	0.318	0.322	0.329	1
RevenuePerShare(Yuan¬•)	1328640.602	51707089.768	0	0.016	0.027	0.046	302000000
RealizedSalesGrossProfitGrowthRate	0.022	0.012	0	0.022	0.022	0.022	1
OperatingProfitGrowthRate	0.848	0.011	0	0.848	0.848	0.848	1
RegularNetProfitGrowthRate	0.689	0.014	0	0.689	0.689	0.690	1
ContinuousNetProfitGrowthRate	0.218	0.010	0	0.218	0.218	0.218	1
TotalAssetGrowthRate	5508096595.249	2897717771.170	0	4860000000.000	6400000000.000	7390000000.000	999000000
NetValueGrowthRate	1566212.055	11415938.9.518	0	0.000	0.000	0.000	933000000
TotalAssetReturnGrowthRateRatio	0.264	0.010	0	0.264	0.264	0.264	1
CashReinvestment%	0.380	0.021	0	0.375	0.380	0.387	1
CurrentRatio	403284.954	33302155.825	0	0.008	0.011	0.016	275000000
QuickRatio	8376594.820	24468474.8.447	0	0.005	0.007	0.012	923000000
InterestExpenseRatio	0.631	0.011	0	0.631	0.631	0.631	1
Totaldebt/Totalnetworth	4416336.714	16840690.5.282	0	0.003	0.006	0.009	994000000

Networth/Assets	0.887	0.054	0	0.851	0.889	0.927	1
Long-termfundsuitabilityratio(A)	0.009	0.028	0	0.005	0.006	0.007	1
Contingentliabilities/Networth	0.006	0.012	0	0.005	0.005	0.006	1
Operatingprofit/Paid-incapital	0.109	0.028	0	0.096	0.104	0.116	1
Netprofitbeforetax/Paid-incapital	0.183	0.031	0	0.169	0.178	0.192	1
Inventoryandaccountsreceivable/Netvalue	0.402	0.013	0	0.397	0.400	0.405	1
TotalAssetTurnover	0.142	0.101	0	0.076	0.118	0.177	1
AccountsReceivableTurnover	12789705.238	278259836.984	0	0.001	0.001	0.001	974000000
AverageCollectionDays	9826220.861	256358895.705	0	0.004	0.007	0.009	973000000
InventoryTurnoverRate(times)	2149106056.608	3247967014.048	0	0.000	0.001	4620000000.000	999000000
FixedAssetsTurnoverFrequency	1008595981.818	2477557316.920	0	0.000	0.001	0.004	999000000
NetWorthTurnoverRate(times)	0.039	0.037	0	0.022	0.030	0.043	1
Revenueperperson	2325854.266	136632654.390	0	0.010	0.019	0.036	881000000
Operatingprofitperperson	0.401	0.033	0	0.392	0.396	0.402	1
Allocationrateperperson	11255785.322	294506294.117	0	0.004	0.008	0.015	957000000
WorkingCapitaltoTotalAssets	0.814	0.059	0	0.774	0.810	0.850	1
QuickAssets/TotalAssets	0.400	0.202	0	0.242	0.386	0.541	1
CurrentAssets/TotalAssets	0.522	0.218	0	0.353	0.515	0.689	1
Cash/TotalAssets	0.124	0.139	0	0.034	0.075	0.161	1
QuickAssets/CurrentLiability	3592902.197	171620908.607	0	0.005	0.008	0.013	882000000
Cash/CurrentLiability	37159994.147	510350903.163	0	0.002	0.005	0.013	965000000
CurrentLiabilitytoAssets	0.091	0.050	0	0.053	0.083	0.120	1
OperatingFundstoLiability	0.354	0.035	0	0.341	0.349	0.361	1
Inventory/WorkingCapital	0.277	0.010	0	0.277	0.277	0.277	1
Inventory/CurrentLiability	55806804.526	582051554.619	0	0.003	0.006	0.011	991000000
WorkingCapital/Equity	0.736	0.012	0	0.734	0.736	0.739	1
Long-termLiabilitytoCurrentAssets	54160038.136	570270621.959	0	0.000	0.002	0.009	954000000
RetainedEarningstoTotalAssets	0.935	0.026	0	0.931	0.938	0.945	1
Totalincome/Totalexpense	0.003	0.012	0	0.002	0.002	0.002	1
Totalexpense/Assets	0.029	0.027	0	0.015	0.023	0.036	1
CurrentAssetTurnoverRate	1195855763.309	2821161238.262	0	0.000	0.000	0.000	1000000000
QuickAssetTurnoverRate	2163735272.034	3374944402.166	0	0.000	0.000	4900000000.000	1000000000
WorkingcapitcalTurnoverRate	0.594	0.009	0	0.594	0.594	0.594	1

CashTurnoverRate	2471976967.444	2938623226.679	0	0.000	1080000000.000	4510000000.000	1000000000
CashFlowtoSales	0.672	0.009	0	0.672	0.672	0.672	1
FixedAssetstoAssets	1220120.502	10075415 8.713	0	0.085	0.197	0.372	8320000000
CurrentLiabilitytoLiability	0.762	0.207	0	0.627	0.807	0.942	1
EquitytoLong-termLiability	0.116	0.020	0	0.111	0.112	0.117	1
CashFlowtoTotalAssets	0.650	0.047	0	0.633	0.645	0.663	1
CashFlowtoLiability	0.462	0.030	0	0.457	0.460	0.464	1
CFOtoAssets	0.593	0.059	0	0.566	0.593	0.625	1
CashFlowtoEquity	0.316	0.013	0	0.313	0.315	0.318	1
CurrentLiabilitytoCurrentAssets	0.032	0.031	0	0.018	0.028	0.038	1
Liability-AssetsFlag	0.001	0.034	0	0.000	0.000	0.000	1
NetIncometoTotalAssets	0.808	0.040	0	0.797	0.811	0.826	1
TotalassetstoGNPprice	18629417.812	37645005 9.746	0	0.001	0.002	0.005	9820000000
No-creditInterval	0.624	0.012	0	0.624	0.624	0.624	1
GrossProfittoSales	0.608	0.017	0	0.600	0.606	0.614	1
NetIncometoStockholder'sEquity	0.840	0.015	0	0.840	0.841	0.842	1
LiabilitytoEquity	0.280	0.014	0	0.277	0.279	0.281	1
DegreeofFinancialLeverage(DFL)	0.028	0.016	0	0.027	0.027	0.027	1
InterestCoverageRatio(InterestexpensetoEBIT)	0.565	0.013	0	0.565	0.565	0.566	1
EquitytoLiability	0.048	0.050	0	0.024	0.034	0.053	1

Figure 2: Distribution of Variables



Most of our variables are within the range [0,1], however some variables have extremely large outliers. Most of these outliers tend to fall in the upper percentiles of the data, which we try capture using Figure 3 below.

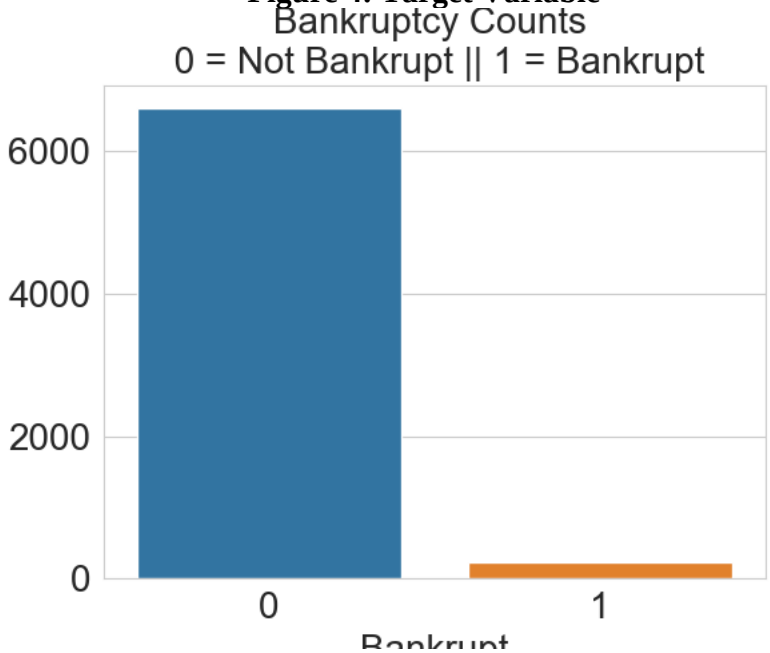
Figure 3: Outlier Variables

The plot shows these 23 variables with their distributions. In most of these variables, the majority of observations are less than 1. For the columns with over 99% of their values less than or equal to 1, we impute the outliers with the largest value <= 1. For the remaining columns, we compare their grouped means for bankrupt and non-bankrupt observations to ensure no systemic bias towards outliers with bankrupt companies. However, most differences appear after the 98th percentile, based on which we Winsorize our data, meaning that we impute values above the 98th percentile with the 98th percentile value. This allows us to avoid trimming the data, which would

exclude potentially important information. We consider Winsorizing the data as a prerequisite to Normalizing as the outliers were in the millions and normalizing them would not return them to the [0,1] range. Additionally, log transformations of the outliers would also yield very large values. Winsorizing the data clips the extreme outliers to a maximum or minimum, based on the prespecified percentile level.

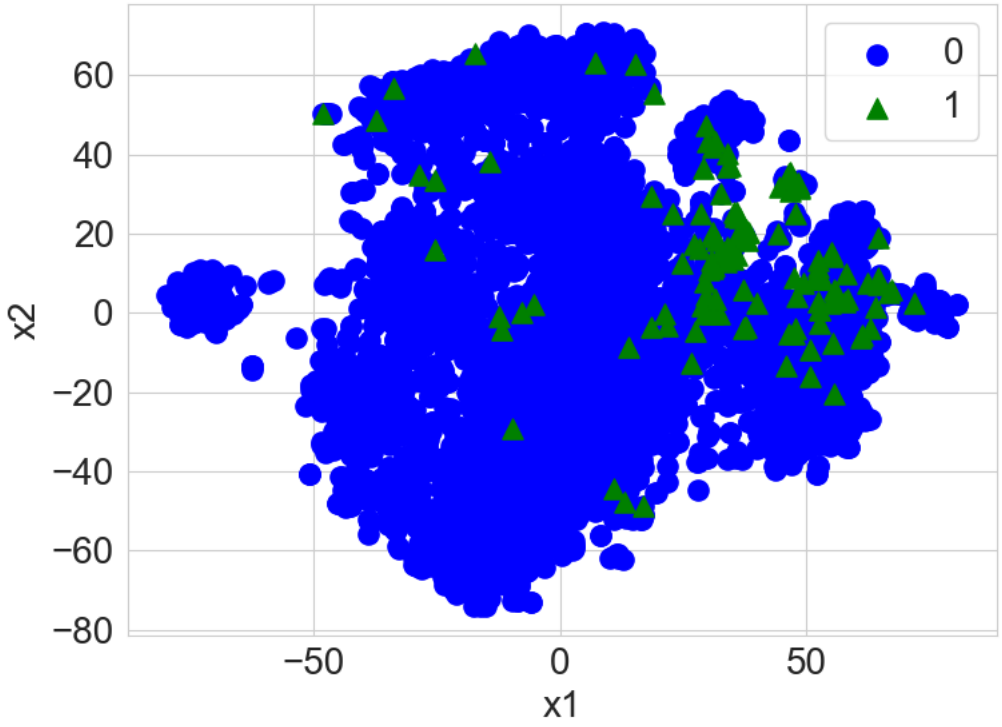
We commence our exploratory data analysis by plotting the distribution of our target variable: Bankruptcy, which is a binary Boolean that indicates whether a firm went bankrupt (1 = yes, 0 = no). Figure 4 shows the highly imbalanced target class, wherein only 3.2% of the target values fall in the minority class. This is a moderate case of imbalance, which we address using statistical techniques.

Figure 4: Target Variable



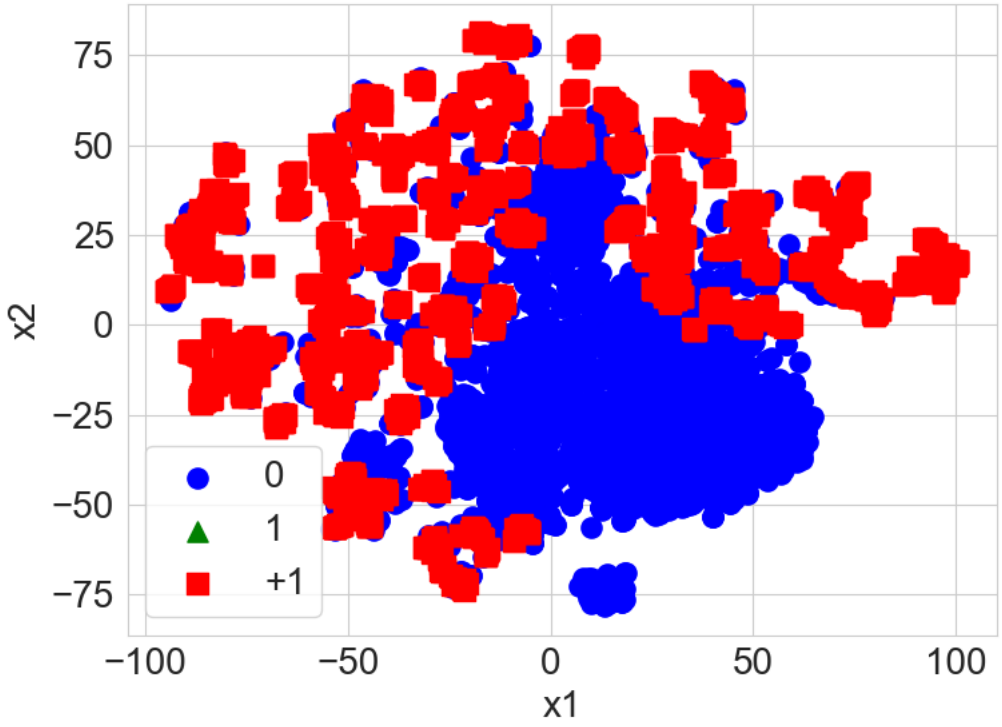
To address this imbalance issue, we apply the Synthetic Minority Oversampling Technique (SMOTE) as developed by Chawla et al. (2002). Whereas in classing oversampling, the minority data is duplicated from the minority data population, SMOTE applies the k-nearest neighbors (KNN) algorithm to generate synthetic data. Here, SMOTE randomly selects data from the minority class and selects the number k for the neighbors. Depending on the amount of over-sampling required, neighbors from the KNN are randomly chosen. SMOTE makes its synthetic data based on these two factors, wherein the procedure is replicated until the two classes balance. Figure 5 below shows the distribution of the imbalanced classes, which we attempt to fix with SMOTE.

Figure 5: Distribution of Imbalanced Target Variable



The plot above shows the imbalance within our target variable – Bankruptcy. From the imblearn library, we use the SMOTE to handle our class imbalance in the target. This augments our data and will hopefully prevent training models that lead to spurious results. Once SMOTE generates our new observations, we appear to have a dataset with a balanced target variable, as seen in Figure 6 below. Here, we can observe the new synthetic data marked with red-squares and a "+1" label.

Figure 6: Distribution of Balanced Target Variable



This plot shows the distribution of the original and synthetic data within our target class variable. We keep our synthetic data separate from our original data to prevent confounding our model results. Over this complete data, we scale the data by standardizing it using sklearn's Standard Scaler technique, which centers the distribution of the data around a mean. We did not find it practical to apply this scaling procedure over high outlier values, which we remove earlier during preprocessing, as it would have skewed our data. For the inferential models, multi-collinearity remained a problem even after the initial removal of highly correlated variables. Multi-collinearity increases coefficient standard errors, potentially leading to Type 2 error through incorrectly classifying a variable as not statistically significant. To further reduce multicollinearity, the Variance Inflation Factor (VIF) was used to identify which variables contributed the most to multi-collinearity (VIF > 10), these variables were dropped.

Now, with our data cleaned and processed, we proceed with separating this data for predictive analysis using our shallow machine learning models. To that end, we split each of the training, validation and testing datasets into subsets for the target vector and the feature matrices. That is, for each of the three datasets, we generate respective target and feature matrices, which we use for our statistical analyses as described in the section below.

3. Methodology

Within our analysis dataset, we contain 6,819 observations over 78 columns – that is, we have 77 feature variables to predict whether a company went bankrupt. Intuitively, not all variables contribute the same influential weight on the target variable, and we must understand which attributes of this data and most meaningful. For prediction, we employ a Decision Tree Classifier, a Multi-layer Perceptron classifier, Histogram Gradient Boosting, and Logistic Regression. For inference, we use a Linear Probability Model and Probit Regression. We explain the basics of each procedure below.

3.1 Decision Tree Classification

Decision Tree is one of the predictive approaches in machine learning which uses decision tree (predictive model) go from observations (decision tree branches) to conclusion about information about the target value. Where the tree models can take discrete set of values are called **classification trees**. In tree structures, leaves represent class labels and branches represent conjunctions of features that lead to class labels. Decision trees can be used to represent decisions and decisions making. The most important decision selection measures are Information gain, Gain Ratio, and Gini Index.
Entropy: is measure of randomness of dataset.
Information gain: it measures decrease in entropy. Information gain estimates the difference between before split and average split of the dataset based on given values.

$$Info\ (D) = -\sum_{i=1}^{m} pi\ \log_2 pi$$

Where pi is the probability that an arbitrary tuple in D belongs to class Ci:

$$Info_a^{(D)} = \sum_{j=1}^{V} \frac{|D_j|}{D}\ X\ Info(D_j)$$

$$Gain\ (A) = Info\ (D) - Info_A^{(D)}$$

Gain Ratio: it manages issue of bias by normalizing the information gain using split info.

$$Gain\ Ratio\ (A) = \frac{Gain\ A}{Split\ Info_A(D)}$$

Gini Index: it works for larger partitions. It uses squared proportion of classes. If Gini index equals to zero it means Gini perfectly classified.

The algorithm of Gini Index: $1 - (P(class1)^2 + P(class2)^2 + \cdots + P(classN)^2)$

$$Gini = 1 - \sum_{i=1}^{C} (p_i^2)$$

3.2 Multi-layer Perceptron Classifier (MLP)

MLP classifier is a type of feedforward artificial neural network (ANN). MLP consist of important tree layer of nodes: an input layer, a hidden layer, and output layer. Each node

uses nonlinear activation function (Rosenblatt, 1961). MLP is supervised learning technique, and it helps to distinguish data which is not linearly separable.
Activation functions of MLP are sigmoid and described by:

$y(v_i) = \tanh(v_i)$ and $y(v_i) = (1 + e^{-v_i})^{-1}$

In first equation, hyperbolic tangent ranges from -1 to 1.
Second equation is logistic function and ranges from 0 to 1.

y_i is output of the i th node and v_i is the weighted sum of the input connections. Alternative activation function can be proposed with including rectifier and softplus functions. More specified functions include radial basis functions.
In MLP there are three or more layers, input and output layer with more hidden layers. MLPS are connected, each node in one layer connects with certain weight to every node in fallowing layer.

3.3 Histogram Gradient Boosting

Gradient boosting is an association of decision trees algorithms. It is a popular technique for tabular classification and regression predictive modelling problems given that it performs across large datasets. Major problem with gradient boosting is that it slows to train the model on large datasets with tens of thousands of rows. Training the trees that are added to the association cab be accelerated by binning the continuous input variables to some unique values. Gradient boosting associates that implement this technique and tailor the training algorithm around input variables under this transform referred to as histogram based gradient boosting ensembles.

Boosting means to a class of ensemble learning algorithms which add tree models to an association sequentially. Each tree model added to association correct the prediction errors made by the tree models. Gradient boosting performs very effective during implementation. Trees should be added and created sequentially unlike other association models like random forest.

3.4 Linear Probability Model (LPM)

The standard linear regression model:

$$y_i = u + \beta_0 + \sum \beta_i x_i$$

With a binary dependent variable is called the Linear Probability model. Each β coefficient can be interpreted as the change in the probability that $y_i = 1$, ceteris paribus. The strength of this model lies in the ease of estimation and interpretability in inferential use-cases; however, it is not useful for prediction as fitted values are not bounded to the $[0, 1]$ interval.

3.5 Logistic Regression and Probit Regression

The primary issue with the LPM is that the conditional probability function is estimated as a linear function, meaning that the model can output probabilities lower than 0 or greater than 1. This flaw can be corrected when we consider a class of binary response models of the form:

$$P(y = 1|x) = G(\beta_0 + \beta_1 X_1 + \cdots + \beta_K X_k) = G(\beta_0 + x\beta)$$

Where G is a function with an image between zero and one for all real number inputs z. In the logistic regression model, G is the logistic function:

$$G(z) = \frac{\exp(z)}{[1 + \exp(z)]} = \gamma(z)$$

Which is between zero and one for all real numbers z. This is cumulative distribution function for standard logistic random variable. In the probit model, G is the standard normal cumulative distribution function (cdf) which is expressed as integral:

$$G(z) = \emptyset(z) = \int_{-\infty}^{z} \emptyset(v)dv,$$

Where $\emptyset(z)$ is the standard normal density:

$$\emptyset(z) = (2\pi)^{-1/2}\exp\left(-\frac{z^2}{2}\right)$$

The choice of G again ensures that binary response model function between zero and one for all values of this parameters and the x_j.

3.6 Principal Component Analysis

Principal component analysis (PCA) is a statistical procedure which uses orthogonal transformation to convert a set of observations of correlated variables into a set of linearly uncorrelated variables (PCA, Pearson 1901). PCA finds directions with maximum variability. PCs are uncorrelated, orthogonal, linear combinations of linear series of equations ($Z_1, \ldots, Z_p$) whose variances are large as large as possible. PC s are form of new coordinate system by rotating the original system constructed by $X_1, \ldots, X_p$.

The principal component analysis (PCA) is concerned with explaining the variance covariance structure of

$$X = (X_1, \ldots, X_p)'$$

Through a few linear combinations of these variables.
Main purposes of PCA are dimension reduction and interpretation. Define the random vector and its mean vector:

$$X = (X_1, \ldots, X_p)'$$

$$\mu = E(X) = (\mu_{1,} \dots, \mu_p)'$$

The variance-covariance matrix of X is the

$$\sum = \ Cov\ (X) = E(X - \mu)(X - \mu)'$$

its ij-th entry $\sigma_{ij} = E(X_i - \mu_i)(X_j - \mu_j)$ for any $1 \le i \le j \le p$.

Sample mean:

$$\bar{X} = \frac{1}{n} X' 1_{n,}$$

X is the design matrix, and 1_n is the vector of $1'$ of length n.
(Unbiased) sample variance- covariance matrix:

$$S_n = \frac{1}{n-1} X_c' X_c = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})\ (X_i - \bar{X})\ '$$

Where X_c the centered design matrix, and $X_i = (X_{i1}, \dots, X_{ip})\ '$ for i=1,…,n.
Linear combination of inputs of PCA can be described as:

$$Z_1 = v_1' X = v_{11} X_1 + v_{12} X_2 + \cdots + v_{1p} X_p,$$
$$Z_2 = v_2' X = v_{21} X_1 + v_{22} X_2 + \cdots + v_{2p} X_p,$$
$$Z_p = v_p' X = v_{p1} X_1 + v_{p2} X_2 + \cdots + v_{pp} X_p$$

Then,

$$Var(Z_j) = v_j' \sum v_j, \ \text{j=1, …, p}$$
$$Cov(Z_j, Z_k) = v_j' \sum V_k,$$

PCA procedure seeks the direction of high variances: the first PCA = linear combination $Z_1 = v_1' X$ that maximizes Var($v_1' X$) subject to $\|v_1\|$= 1.
The second PC = linear combination $Z_2 = v_2' X$ that maximizes Var ($v_2' X$) subject to $\|v_2\|$= 1 and Cov($v_1' X, v_2' X$) = 0
The jth PC satisfies max Var ($v_j' X$) subject to $\|V_j\|$= 1, Cov ($v_1' X, v_j' X v_j$) = $v_1' \sum v_j$ = 0,
for I=1, …, j-1,
where j=2, …, p.
$Z_1 = v_1' X$ has the largest sample variance among all normalized linear combinations of the columns of X.
$Z_2 = v_2' X$ has the highest variance among all normalized linear combinations of the columns of X, satisfying v_2 orthogonal to v_1.
The last PC $Z_p = v_p' X$ has the minimum variance among all normalized linear combinations of the combinations of the columns of X, subject to v_p being normalized orthogonal to the earlier ones.
There are two ways to solve PCs are: eigen-decomposition of cumulative functions.
Singular value decomposition (SVD) of X_c.

4. Results and Discussion

4.1 The LPM results are below, sorted by magnitude:

Table 2: LPM Results

Variable	Beta	Standard Error	T-Stat	P-values
Totaldebt/Totalnetworth	0.056928514	0.006524709	8.725066282	2.66E-18
Constant	0.032262795	0.001935207	16.67149746	2.11E-62
Operatingprofit/Paid-incapital	0.031497086	0.00615543	5.116959648	3.11E-07
TotalAssetTurnover	0.020661977	0.005881382	3.513115917	0.00044 2884
Totalexpense/Assets	0.01997927	0.005956478	3.35420888	0.00079 5923
Cash/CurrentLiability	0.018490886	0.00730577	2.530997459	0.01137 3867
NetValuePerShare(C)	0.017618261	0.003850694	4.575346833	4.75E-06
Revenueperperson	0.014133286	0.004593376	3.076883962	0.00209 1767
Liability-AssetsFlag	0.010022477	0.004268733	2.347881013	0.01888 0552
Contingentliabilities/Networth	0.009314309	0.001136143	8.198180515	2.44E-16
CurrentLiabilitytoLiability	0.00620577	0.002704328	2.294754419	0.02174 7202
Totalincome/Totalexpense	0.001790716	0.000767261	2.333906765	0.01960 0604
TotalAssetGrowthRate	-0.005394018	0.00183532	-2.93900705	0.00329 2656
Taxrate(A)	-0.006388254	0.001961173	-3.257364168	0.00112 4521
NetValueGrowthRate	-0.007907885	0.002270913	-3.482249769	0.00049 722
AccountsReceivableTurnover	-0.010650307	0.002696245	-3.950052333	7.81E-05
Operatingprofitperperson	-0.014840975	0.003995807	-3.714137545	0.00020 3898

Netprofitbeforetax/Paid-incapital	-0.021578202	0.007452767	-2.895327463	0.003787632
ROA(B)beforeinterestanddepreciationaftertax	-0.026042856	0.007405364	-3.516755642	0.000436856
RevenuePerShare(YuanÂ¥)	-0.049013524	0.006823678	-7.182860249	6.83E-13

Note that in table 2, variables that are not statistically significant at the 5% level have been omitted. Since the dataset has been standardized, coefficient magnitudes can be directly compared against one another regarding variable importance. In this case, Totaldebt/Totalnetworth is the most important determinant of bankruptcy. The coefficient of 0.056 means that a one standard deviation increase in Totaldebt/Totalnetworth is estimated to increase the probability of bankruptcy by approximately 5.6%. The other coefficients have the same interpretation. We see that in general, debt and liability ratios increase the probability of bankruptcy, while value/profitability metrics decrease the probability of bankruptcy, which is logical. However, there are some curious results from this estimation – we see that Net Value per Share and Asset Turnover have positive coefficients, implying that as they increase, the probability of bankruptcy increases. This may be the result of omitted variables bias – indeed, it is easy to reason that an intangible omitted variable such as 'Executive Competence' would surely influence the probability of bankruptcy. The model could therefore be improved by incorporating a proxy to mitigate this potential bias.

4.2 The Probit Results are below:

Table 3: Probit Results

Variable	Parameters	T-Stat	P-values
Operatingprofit/Paid-incapital	0.467751671	3.192098992	0.001412429
Totaldebt/Totalnetworth	0.374387096	7.561067932	4.00E-14
Revenueperperson	0.185454735	3.111594243	0.001860801
GrossProfittoSales	0.097805344	2.107706823	0.035056357
AverageCollectionDays	0.087668754	2.009843246	0.044447782
Cash/CurrentLiability	0.079282144	2.459584177	0.013909807
Cashflowrate	-0.16056375	-2.572506752	0.010096499
QuickRatio	-0.212959966	-2.638603834	0.00832482
Cash/TotalAssets	-0.23139262	-2.152194726	0.031382021
AccountsReceivableTurnover	-0.24360273	-4.33921339	1.43E-05
RevenuePerShare(YuanÂ¥)	-0.302684324	-2.179360221	0.029304919

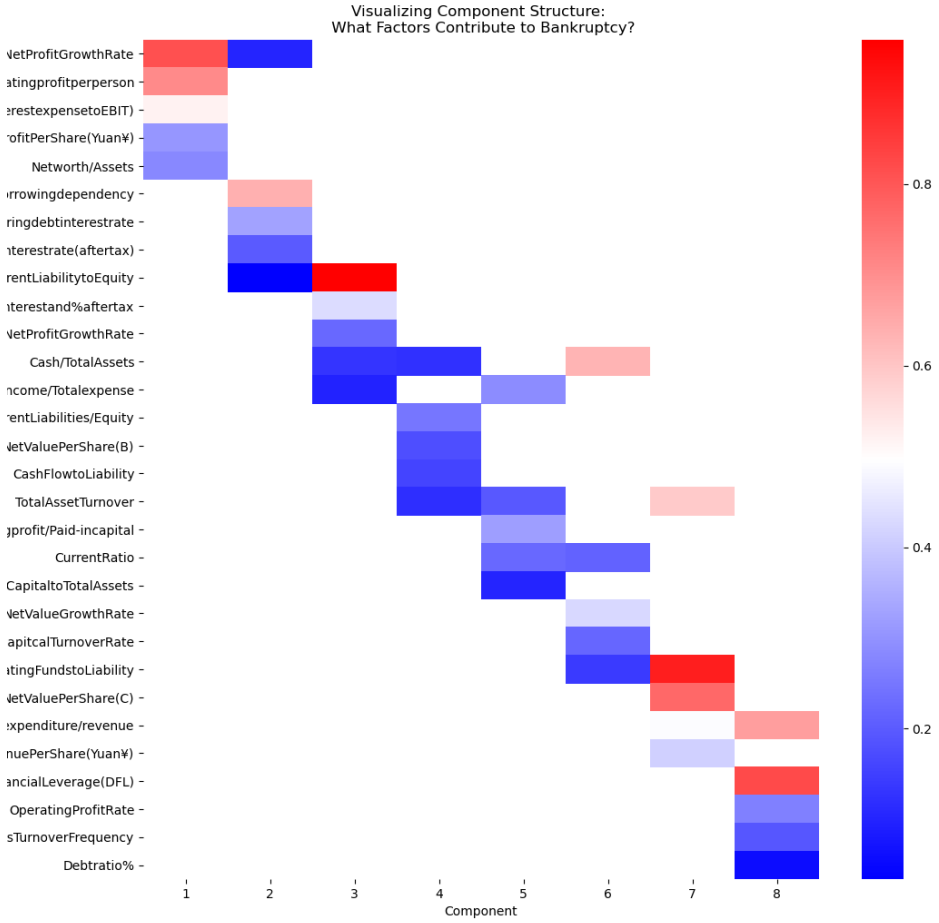
ROA(B)beforeinterestanddepreciationaftertax	-0.323479296	-3.803013532	0.000142947
Netprofitbeforetax/Paid-incapital	-0.529058932	-3.341383889	0.000833619
Constant	-2.797022563	-24.80639267	7.65E-136

Probit estimation produces a different set of important variables. We see that the most important determinant of bankruptcy in the Probit is Operatingprofit/Paid-incapital, with a coefficient of approximately 0.467. This means that a one standard deviation increase in Operatingprofit/Paid-incapital is estimated to increase the Z-score by 0.467. As the Probit is a nonlinear model, the actual partial effect of this increase will change depending on where in the distribution the initial Z-score falls. For instance, $\Phi(0) = 0.5; \quad \phi(0.467) \approx 0.68$. In this case, the probability of bankruptcy increases by some 18%. However, note that $\Phi(-3) \approx 0.001; \Phi(-2.53) \approx 0.005$. As such, the magnitude of the partial effect is dependent on the magnitude of the Z-score. Unfortunately, interpretation is further complicated by omitted variable bias, more severe in the probit than in the LPM. This is evidenced by variables such Revenue per Person having a positive coefficient, implying an increased probability of bankruptcy as Revenue per Person increases, which is nonsensical. The reason for this increased bias is that, unlike in OLS regression with the LPM, omitting features which are correlated with the outcome but not with the other predictors leads to bias in coefficient estimates of the predictors. Therefore, the bias is more severe in the Probit, leading to the confusing results that we see here.

4.3 PCA

Using PCA on the dataset without standardizing, we find that 8 components explain approximately 97% of the total variation. Adapting the procedure outlined by Kaggle user Ella Pham in her submission (PCA with Business View (SVC: 96% acc, 43% f1) | Kaggle), we are able to deconstruct the components into the variables which comprise them.

Figure 7: PCA Variable Composition



We see that each component is mostly comprised of a group of related variables. The first component is comprised of profitability metrics, the fifth component is mostly comprised of liquidity metrics, and so on. While the PCA did not feature heavily into our analysis, it provides an intriguing path for future analysis.

4.5 Predictive Modeling

Table 4: Predictive Model Results

F1 Score	Estimator
0.75	Decision Tree Classifier
0.74	MLP Classifier
0.74	Histogram Gradient Boosting Classifier
0.61	Logistic Regression

Table 4 shows the predictive model results. We find that the decision tree classifier had the best performance, closely followed by the MLP classifier and the Histogram Gradient Boosting Classifier. Logistic Regression performed notably worse than the other models. Decision Tree Classifier predicts company bankruptcy data with 75% accuracy. Based on all possible outcomes and conditions decision tree predicts the bankruptcy with 75% variability. Decision tree has various advantages over other methods. It conducts variable screening and needs less effort for data preparation. Secondly, non-linear relationships do not affect tree's performance. Thirdly, decision tree is beneficial for data exploration and does not make any assumptions on the linearity of data, but number of trees increases accuracy decreases. Drawback of decision tree is outcome based on expectations so it can lead bad decision making. By looking the result of decision tree classifier of this study, 75% accuracy is a good performance. It can be improved by changing the number of trees. Second way prediction accuracy can be improved by boosting algorithms. Decision trees are non-linear so boosting will work with decision trees. Also boosting method have power on bias/ variance tradeoff and it gives opportunity to reduce variance.MLP classifier and Histogram Gradient Boosting Classifier predict the model with 0.74 accuracy. This is the second-best accuracy score. This score can be improved by doing hyperparameter tuning.

5. Conclusion

In our project, we analyzed the Taiwan Bankruptcy dataset and attempted to both predict bankruptcies and determine some of the most important causal factors towards bankruptcy. We preprocessed the data through removing redundant features as well as applying Winsorization to mitigate the effect of outliers. For predictive modeling, we then applied the SMOTE technique to handle the class imbalance present in the target variable. We find that the Decision Tree Classifier performed best of the 4 predictive models tested, suggesting further usage of Tree based methods. In future analysis, we could potentially improve our F1 recall score by employing a random forest or other ensemble methods. Additionally, using the PCA dimensionality reduced dataset may provide similar accuracy with superior computational efficiency.

For inferential modeling, we mitigated the issue of multi-collinearity through removing an additional set of 10 variables with high VIF values, and obtained coefficient estimates through using a Linear Probability Model as well as a Probit model. Unfortunately, the probable presence of omitted variable bias has rendered some of these estimates difficult to trust. The inferential portion of this analysis would be improved by identifying proxy variables for some of these missing factors, or by using an alternative estimation strategy such as Two Stage Least Squares, in the case of the LPM.

6. References

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, *23*(4), 589-609.

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research* 16 (2002): 321-357.

Lee, Tsun-Siou, and Yin-Hua Yeh. "Corporate governance and financial distress: Evidence from Taiwan." (2004): 378-388.

Liang, Deron, Chia-Chi Lu, Chih-Fong Tsai, and Guan-An Shih. "Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study." *European Journal of Operational Research* 252, no. 2 (2016): 561-572.

Lin, Wei-Yang, Ya-Han Hu, and Chih-Fong Tsai. "Machine learning in financial crisis prediction: a survey." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42, no. 4 (2011): 421-436.

Rosenblatt, Frank. x. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington DC, 1961

Shailer, Gregory EP. *An introduction to corporate governance in Australia*. Pearson Education Australia, 2004.

7. Appendix

a. Packages Used in this Project:
    i. Sklearn
    ii. Pandas
    iii. Numpy
    iv. Statsmodels
    v. Matplotlib
    vi. Seaborn
    vii. Plotly
    viii. Imblearn (SMOTE)
    ix. PyQT (GUI)