Amar Adusumilli

Data Mining Project – Individual Report

Introduction:

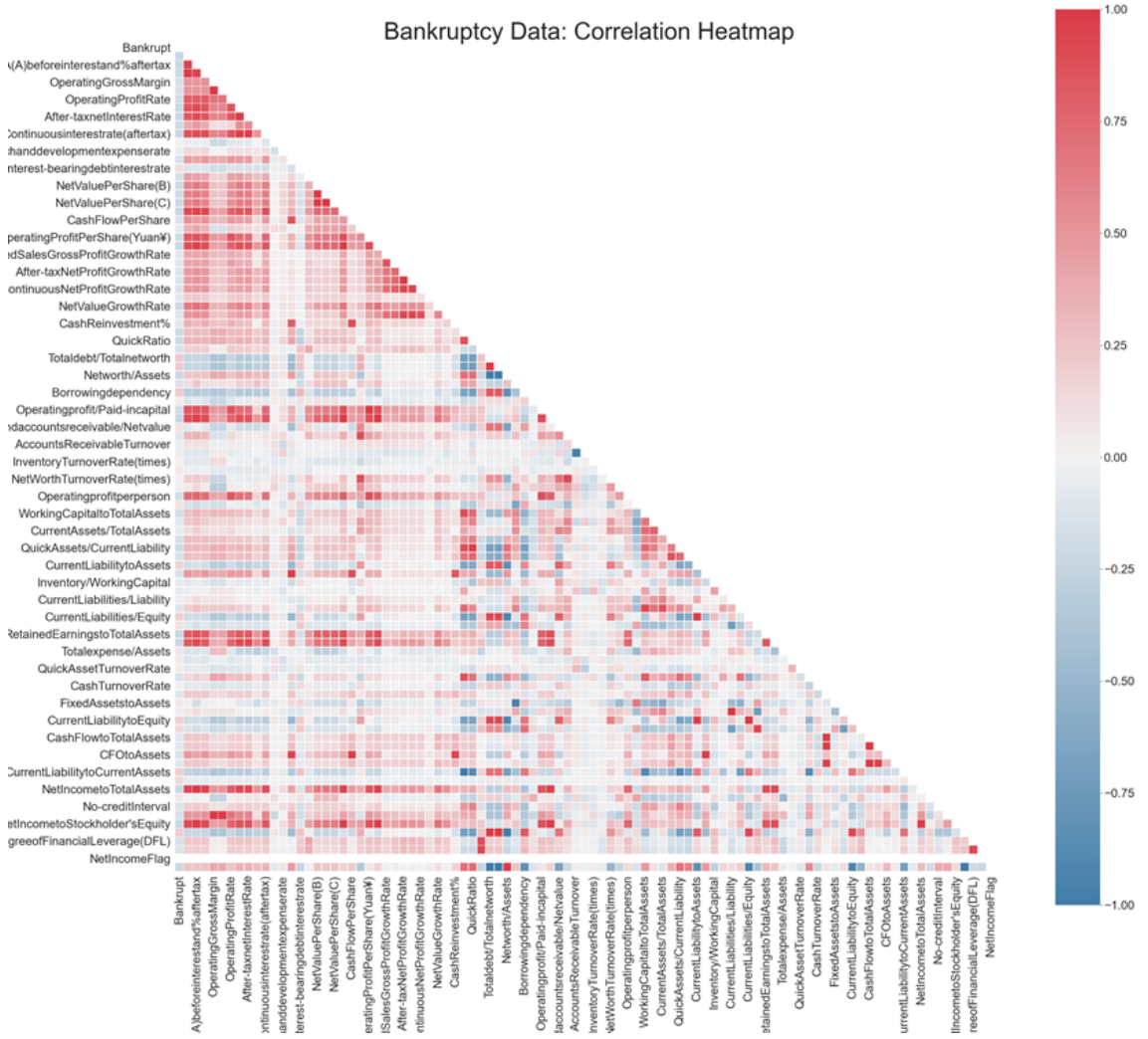
In our project, we analyze bankruptcy data describing corporations in Taiwan over the period spanning from 1999-2009. The dataset consists of 6819 observations of 96 columns. Bankruptcy itself was codified as a binary indicator variable, while the majority of the features are financial ratios bounded on the interval $[0, 1]$, with some exceptions. We carried out our analysis with two main objectives – to identify the determinants of bankruptcy, and to predict bankruptcy. We began with preprocessing the data, performing basic cleaning, outlier handling, standardizing, and plotting. We then estimated econometric regression models to calculate the partial effect of the features on the target and built predicative models to predict which companies went bankrupt. The primary challenge associated with this dataset was contained within the class imbalance in the target, where only some 3% of companies went bankrupt. The secondary challenge was in identifying and removing the redundant features, of which there were many. In the following sections, I will expand upon the project steps and my contribution with more detail.

Description of Individual Work

EDA:

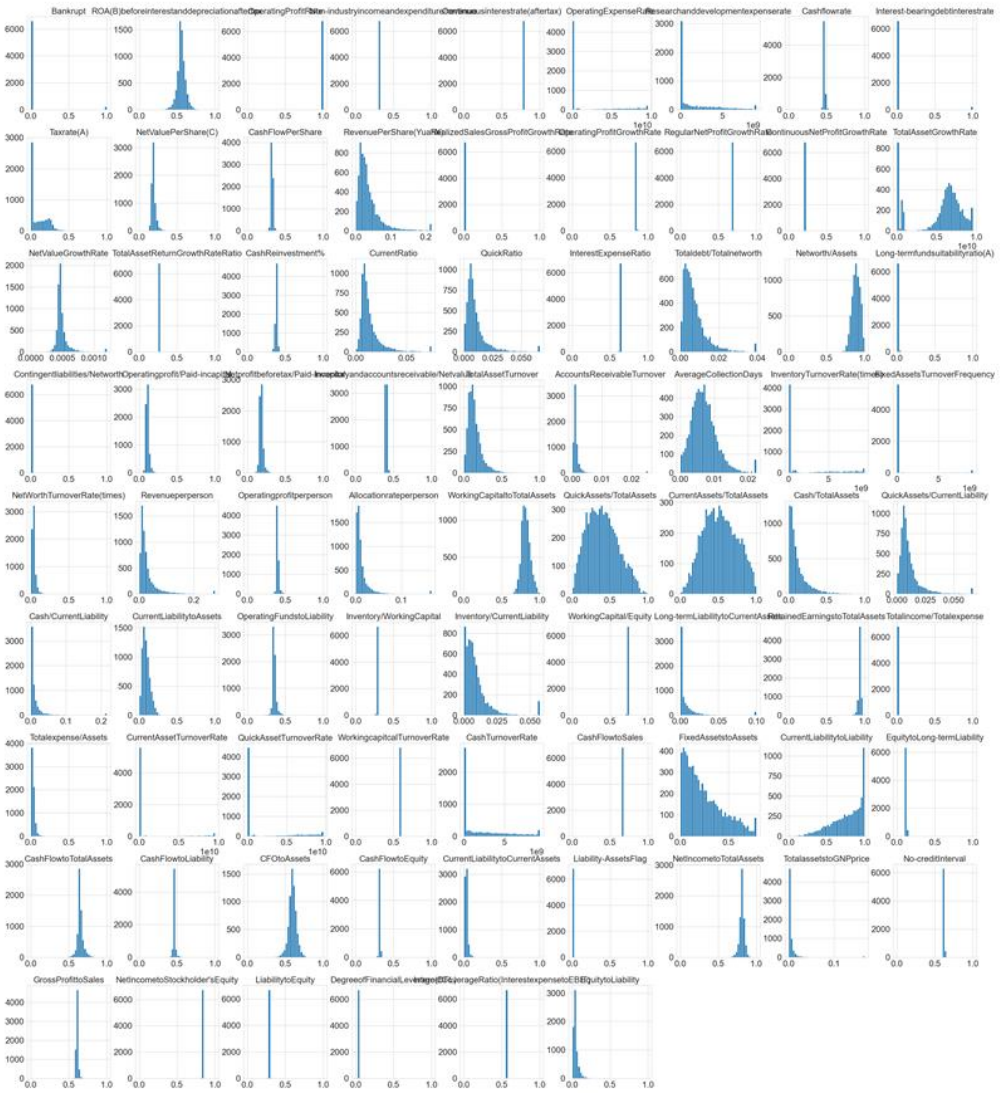
- Removed whitespace from column names and renamed the target from ‘Bankrupt?’ to ‘Bankrupt’
- Created the Spearman correlation matrix, show below:

Figure 1: Correlation Matrix



- Wrote several helper functions to detect the presence of outliers and whether they were biased towards bankrupt companies.
    - IE, do outliers disproportionately correspond to bankrupt companies?
        - We find that they did not disproportionately correspond to bankrupt companies, pointing to data quality issues being the main source of the outliers.
- Created the feature distribution plot, below:

Figure 2: Variable Distributions



- As part of the outlier identification process, I also identified the features possessing the outliers and plotted their distributions separate from the others:

Figure 3: Outlier Variables

- I discovered the pattern in the outliers: that in almost all cases, vast majority of observations were bounded within $[0, 1]$, with a few values exploding into the millions.
    - To counteract this, I wrote a function to Winsorize our data.
    - In the set of 23 outlier columns I identified, I determined that for 21 of them, the 99th percentile observation was within $[0, 1]$, but the small set of values above that trended towards the millions. For these, the function replaces all values above the 99th percentile value with the 99th percentile value.
    - Of the remaining two columns, one had its 98th percentile value within the $[0, 1]$, with values above trending towards the millions. This column was handled by extending the Winsorization function to perform a second sweep.
    - The final column 'Total Asset Growth Rate' is the only column to break the pattern – the data from the minimum to the 50th percentile is within $[0, 1]$, but all values over that trend towards the millions. This column was standardized without further adjustment,

as for new companies or rapidly expanding companies, it makes sense that there can be exceptionally large values, as this is just the percentage growth in Total Assets.

- Manually examined the mean values of the features conditional on bankruptcy. This was our first indication of the data quality problems we attempted to rectify with Winsorization and column removal.
- I created most of the figures shown in the GUI. The figures associated with the EDA are:

Figure 4: Class Imbalance

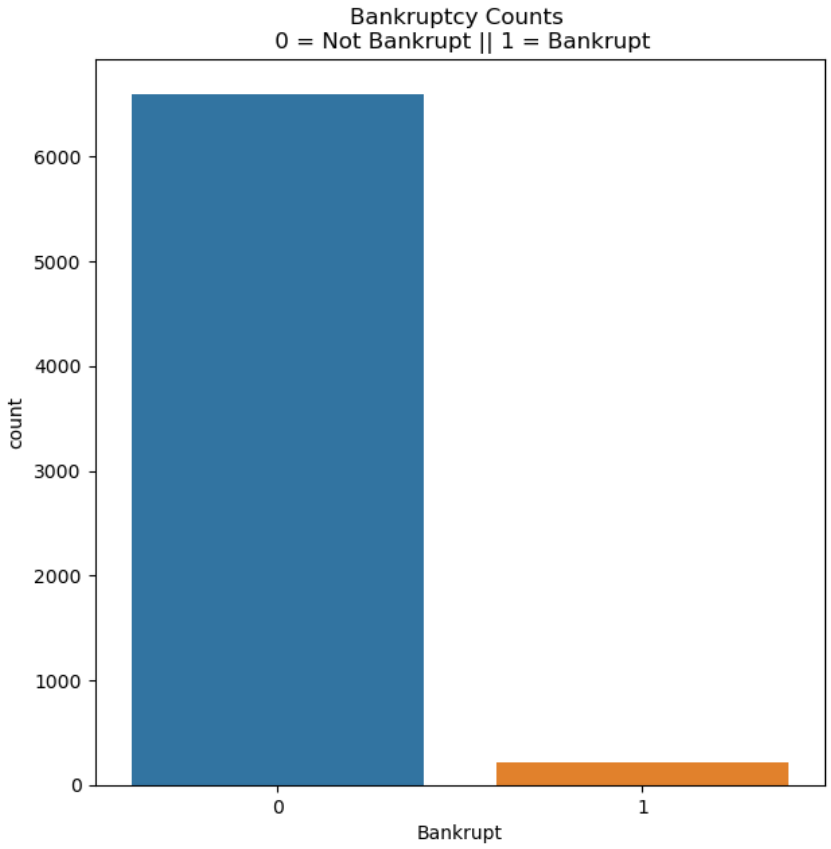


Figure 5: Operating Profit vs Debt Ratio

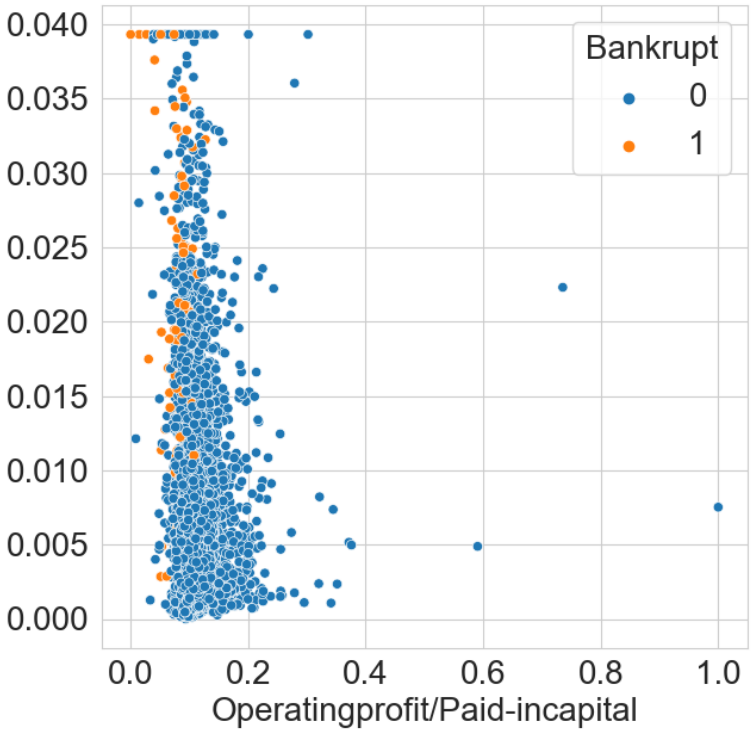


Figure 6: Asset Turnover vs Debt Ratio

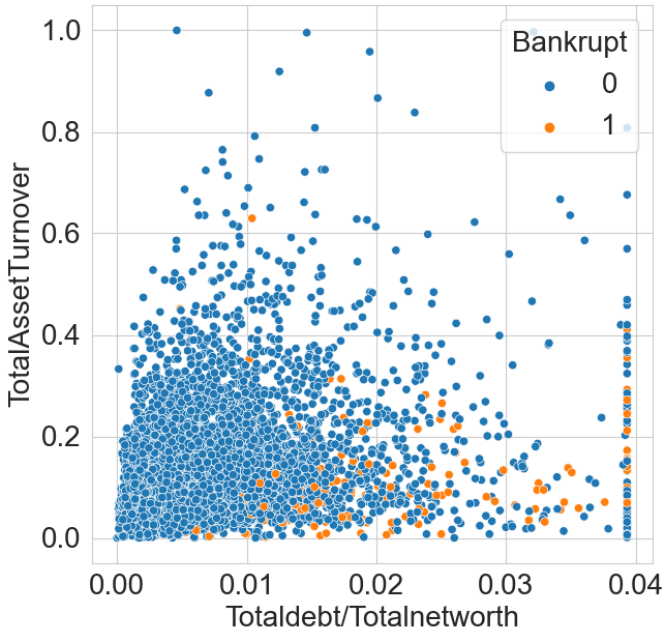
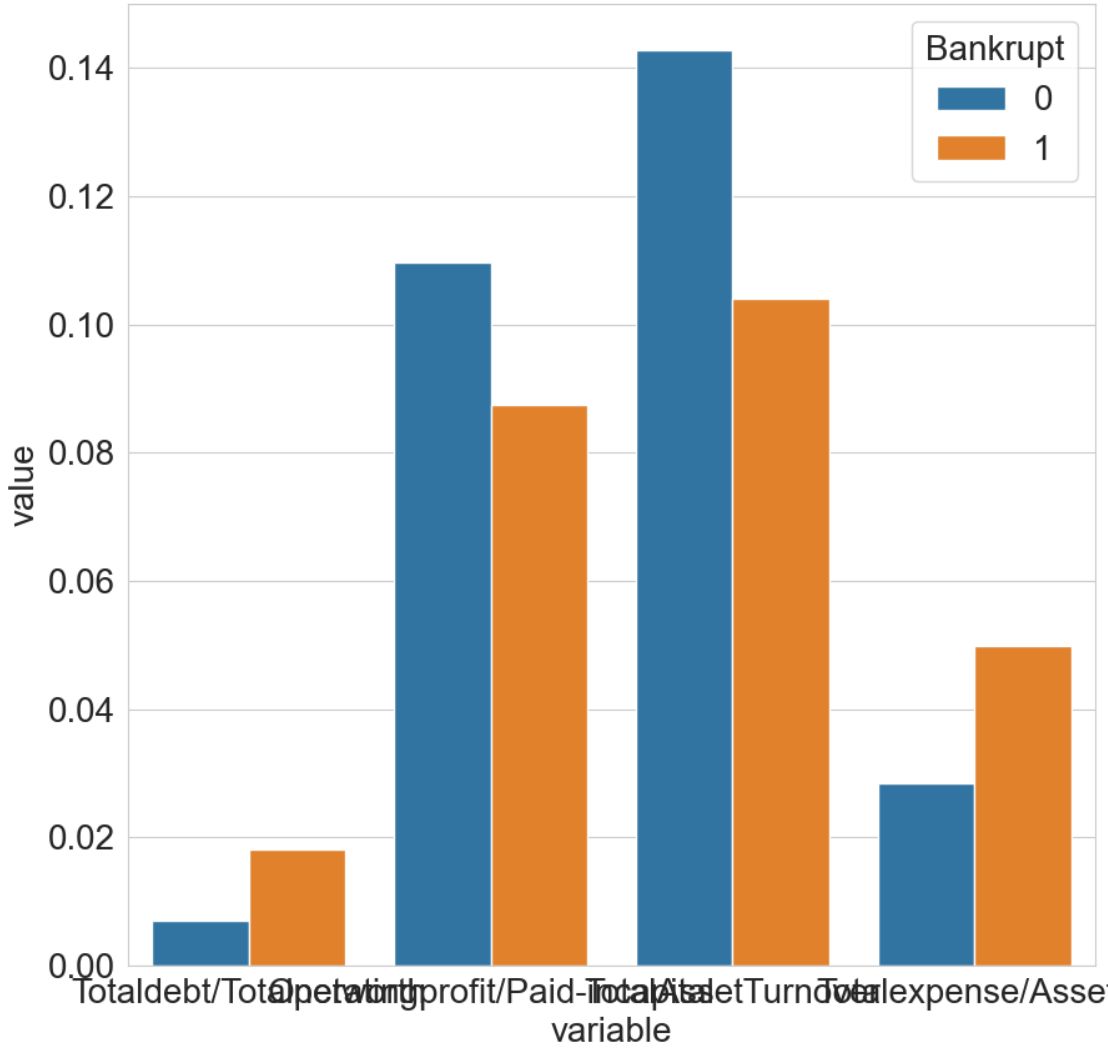


Figure 7: Grouped Means of Important Variables



Inference:

- Removed an additional set of collinear variables. I calculated a new correlation matrix from our pruned data (after the initial removal in the EDA script), isolated the upper triangular portion of the matrix, and identified the columns with Spearman correlation > 0.9. Those that matched the filter were dropped.
- Using statsmodels, I estimated a Linear Probability Model (LPM). As the design matrix was nearly singular and standard errors were exceedingly high, multi-collinearity was still an issue.

    - As such, I found the Variance inflation Factor (VIF) for each remaining feature.
- VIF is estimated using an auxiliary regression to identify R_i^2, the R^2 found by regressing one feature on the set of all others. Symbolically, this is estimating:

$$\boldsymbol{x_i = \delta_0 + u + \sum \delta_j x_j}$$

    - In the above, $j \neq i$, u is the error, δ_0 is the model intercept, and the other δ terms are the coefficients of the features $j \neq i$.
    - After estimating this regression, the VIF for feature i is given by:

$$\boldsymbol{VIF = \frac{1}{1 - R_i^2}}$$

- All features with VIF > 15 were dropped. After this, multi-collinearity was no longer an issue.

- The linear probability model is simply OLS regression applied towards a binary dependent variable. This is the equation:

$$P(Y = 1|X_1, X_2, X_3, \dots) = u + \beta_0 + \sum \beta_i x_i$$

    - Each β coefficient can be interpreted as the change in the probability that $Y = 1$, ceteris paribus. The strength of this model lies in the ease of estimation and interpretability in inferential use-cases.
    - This model is not useful for prediction as fitted values are not restricted to the $[0, 1]$ interval – the model can fit values below 0 and/or above 1.
- As an alternative to the LPM, I also estimated a probit model.
- 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:

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

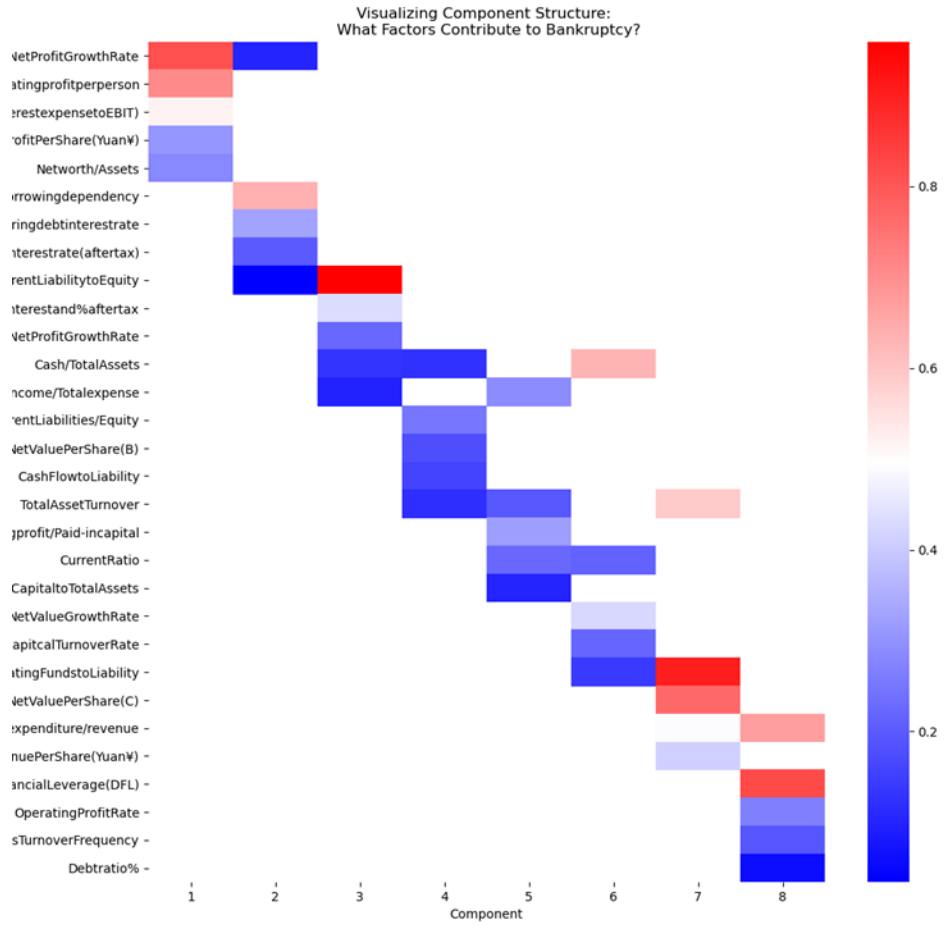
- Where G is a function with an image between zero and one for all real number inputs z. In the probit model, G is the standard normal cumulative distribution function (cdf) which is expressed as integral:

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

    - That is, in the Probit, $G(z)$ is the standard normal distribution function.
- In effect, the linear model within G outputs a Z-score, which is then converted into a probability through the normal distribution function.
    - This ensures that the model has an image bounded by $[0, 1]$, correcting the biggest flaw in the LPM.

- I also conducted Principal Components Analysis on the data in this section. While we did not use the PCA dimensionality reduced dataset elsewhere in the analysis, PCA presents an intriguing avenue for further analysis.
    - As part of the PCA work, I adapted the procedure outlined by Kaggle user Ella Pham in her submission (PCA with Business View (SVC: 96% acc, 43% f1) | Kaggle) in order to break each component into its constituent variables
    - As part of this process, I created the following plot:

Figure 8: PCA Variable Composition



Report/Presentation/Other:

- I wrote roughly half of our group report myself and edited the other portions written by others.
- I was able to use shell scripting to convert the ui file into a py file for the GUI
- I was responsible for and mostly wrote slides 4-6 and 9-14 in the presentation.

Results

1. The LPM results are below, sorted by magnitude:

Table 1: 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	-2.939000705	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	0.000203898
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 1, 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 bias in a coefficient β_1 is the difference between the population parameter and the sample estimate – an unbiased estimate has expectation of the sample estimate as equal to the population parameter. In the case of an OLS coefficient, this can be informally represented as:

$$Bias(\beta_1) = \beta_2\delta_1$$

In the above, an omitted variable x_2 is excluded. β_2 measures the effect of x_2 on the dependent variable, while δ_1 measures the covariance between x_2 and the included x_1. If either of these are 0, there is no bias. The model could therefore be improved by incorporating a proxy to mitigate this potential bias.

2. The Probit Results are below:

Table 2: Probit Results

	Parameters	T-Stat	P-values
Operatingprofit/Paid-incapital	0.467752	3.192099	0.001412
Totaldebt/Totalnetworth	0.374387	7.561068	4.00E-14
Revenueperperson	0.185455	3.111594	0.001861

GrossProfittoSales	0.097805	2.107707	0.035056
AverageCollectionDays	0.087669	2.009843	0.044448
Cash/CurrentLiability	0.079282	2.459584	0.01391
Cashflowrate	-0.16056	-2.57251	0.010096
QuickRatio	-0.21296	-2.6386	0.008325
Cash/TotalAssets	-0.23139	-2.15219	0.031382
AccountsReceivableTurnover	-0.2436	-4.33921	1.43E-05
RevenuePerShare(YuanÂ¥)	-0.30268	-2.17936	0.029305
ROA(B)beforeinterestanddepreciationaftertax	-0.32348	-3.80301	0.000143
Netprofitbeforetax/Paid-incapital	-0.52906	-3.34138	0.000834
Constant	-2.79702	-24.8064	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 $\Phi(-3) \approx \quad 0.001; \Phi(-2.53) \approx 0.005$. 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.

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. The plot of the deconstructed components is shown above in Figure 8. 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.

Conclusions

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. The bulk of my involvement with the analysis was centered on the Exploratory and Inferential analysis. I preprocessed the data through removing redundant features as well as applying Winsorization to mitigate the effect of outliers. I also was able to reduce the dimensionality of the dataset through PCA, discovering that 8 components are sufficient to explain 97% of the variation in the data.

For inferential modeling, I 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. The LPM identified Total Debt/Total Net Worth, Operating Profit / Paid in Capital, Asset Turnover, and Total Expense / Assets as the most important features. These results generally make sense, as I would expect debt metrics to be positively associated with the probability of bankruptcy. In the Probit, we have Operating Profit / Paid in Capital, Total Debt/Total Net Worth, Revenue per Person, and Gross Profit to Sales as the most important determinants of bankruptcy. While the debt ratio's presence in this group makes sense, the other variables listed do not make sense – it is difficult to believe that higher values of Profit to Sales correspond with higher chances of bankruptcy. Unfortunately, the probable presence of omitted variable bias has rendered some of these estimates difficult to trust, especially in the probit model. 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.

In general, I have learned that increased measures of debt are positive determinants of bankruptcy, and higher values in profitability metrics are associated with lower chances of bankruptcy. While this is a bit obvious, what was more relevant are the magnitudes of the effects – namely, the 5.7% increase in bankruptcy likelihood for a standard deviation increase in the debt ratio. The specific estimates can potentially be especially useful for regulators and financial professionals in determining the financial health of a company, and the relative risk associated with said company. Therefore, I believe there is considerable value in attempting to refine and expand upon this inferential analysis. The challenge for the future will be in identifying proxies for the important missing factors, and perhaps refining the variable removal process to retain more useful information.

Percentage of Code

The only code copied from the internet for my share of the project was the PCA variable deconstruction. Between the EDA, Inference, and GUI Plots, the ratio is approximately:

$$\frac{53-14}{437} = \frac{39}{437} \approx 8.9\%$$

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.

Kaggle Link (PCA):

PCA with Business View (SVC: 96% acc, 43% f1) | Kaggle

1. 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)