Nice link: https://www.mdpi.com/2813-2203/4/3/22

Intro/Motivation/Context (Ian)

Welcome to our project presentation. Our project is about trying to predict corporate bankruptcy based on various ratios and other accounting factors. For this we relied on several publicly available data sets with information on companies and whether they went bankrupt or not. But Max will shortly tell you more about the datasets.

Our primary goal is to identify the most important predictors of bankruptcy. We want to do this because being able to predict bankruptcy in financial sectors is hugely important. Bankruptcy causes large financial and societal losses. Being able to accurately predict bankruptcy would for example reduce the risk to investors on the stock market or make it

easier for banks to decide whether to give out a loan or not. Also, classical models like Altman Z-score and Ohlsen O-score may be outdated.

Hypotheses (Ian)

We formulated a handful of research questions at the start of the project. Our first research question is: “are there models that which can predict bankruptcy better than classical models?”. Here our goal is to find through various tests which types of tests and models are best able to predict bankruptcy. We think that some form of machine learning model will outperform classical ways of predicting bankruptcies - like the Altman Z-score.

Our second question is “Which financial ratios predict bankruptcy best?”. We suspect that financial ratios will have little predictive power of bankruptcies, because if there were clear indicators like that there would have long been a more solid way to predict bankruptcies.

The third research question we had is “Are smaller firms more predictable than larger firms?”. We think that smaller firms will be more predictable because they have more visible financial patterns, meaning they have a simpler structure than larger firms.

Our final research question is “Do failing firms have a higher instability than healthy firms?”. Yes we think we will see more variability in failing firms than in healthy firms.

Data(Max)

For this project we used publicly available data sets on corporate bankruptcy from three countries: Poland, the US and Taiwan. Our objective was to explore the predictive power of financial ratios and other accounting indicators for bankruptcy prediction

Exploratory Data Analysis

We conducted exploratory data analysis on the Polish, U.S., and Taiwanese bankruptcy datasets to understand their properties and prepare them for modeling. We looked at:

- Class Imbalance
- Data Cleaning
- Model results
- Feature Distributions

Focussing on Taiwan

After looking at the results and data from the different datasets, we decided to focus on the Taiwanese dataset. This dataset had the best results during early testing and supported classical models: Altman Z-score, Ohlson O-score, as well as modern machine learning approaches: Logistic Regression, Random Forest, XGBoost, KNN, and LDA.

Datacleaning process

We cleaned the data by winsorizing extreme values, removing low-variance and highly correlated features, and creating a composite logsize metric to approximate firm size.

The data was then split into training and test sets using an 80/20 stratified split, ensuring that the proportion of bankrupt and non-bankrupt firms was preserved in both subsets. This separation allows model performance to be evaluated on unseen data and reduces the risk of overfitting. All input features were subsequently standardized using z-score scaling, giving each variable a mean of zero and a standard deviation of one. This step is particularly important for distance-based and regularized models, as it prevents features with larger numeric ranges from dominating the learning process. Finally, SMOTE was used to create extra training cases for the bankruptcy class (this class was under represented in the dataset).

Multicollinearity (Ian)

Since we are dealing with a lot of variables per row in our datasets. One of the first things we did was create a correlation graph between all the variables in the dataset. Here we show the correlation graph for the Taiwan dataset. Our aim with this was to gain an intuition on how the various variables in our data are related to each other. As you expect to see that on the diagonal, there is a perfect correlation between the variables. Apart from that we can see that most variables have little to no correlation apart from a few blocks where similar variables are grouped together. We also notice how a few variables seem to have much more correlation with other variables than most others. Visible by the more lightly colored bands. Because we don't want to skew the results by using variables that are closely related, we dropped columns with a similarity >0.9. This would be similar to counting things twice.

Classic Models (Max)

For our classical models we used the Almant Z-score and the Ohlson O-score. The Z-score was developed in 1968 and uses 5 key financial ratios. The O-score was created in 1980 and already is somewhat more sophisticated than the Z-score. It uses a combination of 9 financial ratios, and similar to the Z-score it is a linear combination. Both use a limited amount of financial calculations and assume linear relationships. Despite them being used for multiple decades, we want to find out if there are models which can predict bankruptcy better.

ML Models (Wenyi)

To predict corporate bankruptcy in the Taiwanese dataset, we implemented a range of machine learning models, each with different strengths. We started with Logistic Regression (LR). This is a linear model that estimates the probability of bankruptcy based on the financial ratios. Because of the class imbalance (only a few companies went bankrupt), we applied SMOTE. Synthetic Minority Oversampling Technique, by generating synthetic examples of the minority class you create more data to work with. This created LR + SMOTE, which improves the model's ability to detect bankrupt firms.

We also tested Logistic Regression with Principal Component Analysis (PCA), which reduces feature dimensionality while retaining variance, resulting in LR PCA and LR PCA + SMOTE.

Next, we used Random Forests (RF). This uses a group of decision trees that captures non-linear relationships and interactions between features. We also trained RF + SMOTE to further handle the imbalance. To reduce dimensionality and focus on the most informative features, we applied Linear Discriminant Analysis (LDA) before also applying Random Forest, resulting in RF LDA and RF LDA + SMOTE.

For non-linear, gradient-based methods, we used XGBoost. This is a powerful boosting algorithm that builds trees in sequence to correct errors from previous trees, along with XGBoost + SMOTE.

We also explored K-Nearest Neighbors (KNN) with k values of 3 and 99, which predicts bankruptcy based on the closest firms in feature space. SMOTE was applied for both k

values, creating four KNN models: KNN k=3, KNN k=3 + SMOTE, KNN k=99, and KNN k=99 + SMOTE

This selection of models allowed us to compare linear, non-linear, ensemble, and distance-based approaches.

Evaluation Metrics(Vincent)

We decided to measure the performance of our models using the following metrics:

1) AUC
    a) This is a way to check how good a binary classification model works. It shows how good the model is at telling the difference between True Positive Rate (TPR) and False Positive Rate (FPR). The plot represents a trade off between sensitivity and specificity of a classifier. AUC has a range between [0,1] with a value of 0.5 meaning random guessing. However, with datasets with a high imbalance, which is the case, the AUC prediction may be overly optimistic, because the FPR is dominated by the majority class.

2) Recall
    a) This measures how often a model correctly identifies true positives from all the actual positive samples. In simpler terms: it measures the percentage of actual bankruptcies that the model correctly identifies.It is calculated by dividing the True Positives by True Positives + False Negatives. Downside it does not account for the false positives.
3) Precision
    a) In simple terms: how often is the positive prediction correct? (How many predicted bankruptcies are actual bankruptcies. It is calculus by dividing the True Positives by True Positives + False Positives. This metric works well with imbalanced classes, however it does not take into account false negatives.

4) F1
    a) This is a somewhat simpler metric, it is the harmonic mean of recall and precision. It can show if a model optimizes recall or precision, it reflects are more balanced classification.

We did not use accuracy since there is a class imbalance, bankruptcy only occurs around 3% of the time. And if the model would return safe 100% of the time, it would still be 97% accurate.

Our main metric will be recall, since missing a bankruptcy (False Negative) is both socially and economically costly, while a “wrong” investigation into a healthy firm does little to no harm. However, the classical models can only be scored using the AUC score, since they produce a continuous value, instead of a binary one.

AUC: fair for every model.
Recall: best for ML models.

Model Comparison (Max)

Overall, most models achieve high AUC values, indicating strong discriminatory power between bankrupt and non-bankrupt firms. However large differences emerge once recall and precision are considered, which is critical given the severe class imbalance and the high cost of missing bankruptcies.

The Altman Z-score and Ohlson O-score are traditional bankruptcy prediction models that calculate a single numeric score from financial ratios. In our dataset:

Altman Z-score achieved an AUC of 0.884, indicating strong discriminatory power between bankrupt and non-bankrupt firms.
Ohlson O-score achieved a slightly lower AUC of 0.847, also demonstrating reasonable predictive ability.

For these classical models, only AUC can be meaningfully calculated, as they just give a float score instead of a class. Metrics such as recall, precision, or F1 are not applicable unless a threshold is manually chosen to classify companies as bankrupt or not. The high AUC values suggest that both classical scores rank firms correctly in terms of bankruptcy risk, though they do not directly provide a probability threshold for actionable decisions.

Spray and Pray: Logistic regression performs consistently well across configurations. Both the baseline and SMOTE-enhanced versions achieve high AUC scores (≈0.94–0.95) and strong recall (0.86), indicating that the model successfully identifies most bankrupt firms. Precision remains low, which reflects a relatively high number of false positives.

Picky: The standard Random Forest achieves the second-highest AUC (0.95) but performs poorly in terms of recall (0.16), indicating that it misses most bankruptcies despite being very precise. This behavior reflects a strong bias toward the majority class. After applying SMOTE, recall improves substantially to 0.55 while maintaining a reasonable precision of 0.51, resulting in one of the highest F1-scores among all models. This demonstrates that Random Forest benefits significantly from class rebalancing.

XGBoost shows strong overall performance, particularly after SMOTE is applied. The SMOTE-enhanced model achieves a balanced trade-off with a recall of 0.55, precision of 0.56, and the highest F1-score in the table (0.55). Combined with the highest AUC score (0.95), this makes XGBoost with SMOTE one of the strongest overall performers.

KNN models show highly unstable behavior depending on the choice of *k* and the use of SMOTE. With a small *k* (k=3), performance is weak without SMOTE but improves notably with oversampling, especially in recall (0.70). However, precision remains low. For large *k* (k=99), the non-SMOTE model completely fails to identify bankruptcies, while the SMOTE version achieves very high recall (0.95) at the cost of extremely low precision. This indicates that KNN is highly sensitive to class imbalance and parameter selection, making it less reliable for practical bankruptcy prediction.

Applying PCA to logistic regression results in a noticeable performance loss, especially in recall, suggesting that compressing the feature space into a single component removes important bankruptcy-related information. While SMOTE restores recall to very high levels (0.98), this comes at the cost of extremely low precision, producing many false alarms.

Random Forest with LDA performs moderately without SMOTE but deteriorates sharply once SMOTE is applied. The reduced AUC and low F1-score suggest that combining aggressive dimensionality reduction with synthetic oversampling leads to overfitting and loss of discriminatory structure.

Across all models, SMOTE generally increases recall but often reduces precision, highlighting the inherent trade-off between identifying bankruptcies and limiting false positives. Logistic Regression (with or without SMOTE) and XGBoost with SMOTE emerge as the most suitable models. Logistic regression offers stable, interpretable, and high-recall predictions, while XGBoost provides the best overall balance between recall, precision, and AUC.

Financial Ratios (Wenyi)

In this section, we analysed which financial ratios were most important for predicting bankruptcy across three models: Logistic Regression, Random Forest and XGBoost.

Several ratios consistently appeared among the top features. In particular, Persistent EPS in the last four seasons, Borrowing dependency and Total debt / Total net worth ranked highly across models. These features reflect profitability stability and leverage, which makes sense economically – firms with unstable earnings or high reliance on debts are more likely to face financial distress.

We also observed model-specific patterns. Logistic Regression emphasized profitability-related ratios, while Random Forest highlighted leverage and liquidity indicators. XGBoost gave the strongest weight to Net Value Growth Rate and Borrowing dependency, suggesting that changes in firm value and reliance on external financing are particularly informative for this dataset.

Overall, the results show that both profitability stability and leverage-related ratios play the largest role in distinguishing bankrupt firms and healthy ones.

Firm Size Analysis(Vincent)

To examine if firm size affects bankruptcy predictability we divided the firms into 3 groups, small, medium, and large based on the log-transformed proxy “Total assets to GNP price”. After which we test each ML model on the size groups.

For almost all models, the general trend for the AUC curves is decreasing when firms size increases, so this would suggest an inverse relationship, even though the changes are only

marginal. This indicates that the financial ratios separate bankrupt and healthy firms more clearly for smaller firms.

Unlike the AUC plot, the recall does not follow a single trend across firm sizes. Instead it depends on the model, and its sensitivity to imbalances. For some models recall is higher for smaller firms, while for others it is higher for the larger ones. The story is the same with SMOTE, it improves some models, but deteriorates others.

So overall, there is some evidence to suggest that firm size does influence the performance of the models, but the effect is not uniform across the metrics. There is a consistent decline with firm size with the AUC metric, implying that smaller firms are more predictable. For recall however, the trend varies between models, which indicates that different models pick up different distress signals across firm sizes. This highlights that the relationship between size and bankruptcy predictability is complex and model-dependent.

Variance Analysis(Vincent)

To test the hypothesis that failing firms have a higher instability than healthy firms, we perform a variance analysis and a coefficient of variation analysis. The latter shows the relative variability and is computed by dividing the standard deviation by the mean. It shows the spread relative to the mean. A high CV means that the data points are more spread, and a low CV means they are clustered more together. For our analysis we decided to look at 4 ratios:

1) Working Capital to Total assets (Liquidity ratio)
2) Cash Flow to Total Assets (Liquidity/Performance ratio)
3) Total Debt to Total Net worth (Leverage ratio)
4) Total Assets Turnover (Efficiency ratio)

Our results show that instability is not greater for failing firms. The variance plot even shows that although marginal, variance is greater for the healthy firms. The CV table shows similar results, with healthy firms having a higher number for the first 3 ratios. This can potentially be explained by the fact that failing firms perform consistently poor.

Total Assets Turnover does have a higher spread for failing firms, which could indicate operating inconsistency as bankruptcy approaches.

Overall we do not find enough support to accept that falling firms have a higher instability than healthy ones. It appears that the instability is ratio-specific and not firm specific.

Limitations (Ian)

One of the limitations we encountered is that the variables in the data sets didn't match the variables used to calculate the Altman Z-score and the Ohlson O-score. Some values we where able to use directly from the datasets, but for some others we had to find substitutes or omit them, which reduces performance.
Furthermore because of the number of entries in the datasets it was impossible for us to manually verify all data for accuracy and outliers, so we had to make compromises by Winsorizing the data.

And lastly because there is an inherit class imbalance in the datasets, we had to apply techniques like SMOTE to reduce the imbalance in the classes. Otherwise, any model could always predict that a firm wouldn't go bankrupt and be correct 97% of the time

Conclusion (Wenyi)

In this project we compared classical models with a range of machine learning models on three bankruptcy datasets and then focused on the Taiwanese firms. Overall, the machine learning models clearly outperformed the Altman Z-score and Ohlson O–score. Especially Logistic Regression and XGBoost with SMOTE achieved high AUC and recall, meaning that they can separate bankrupt firms from healthy ones well and pick up many of the failing companies.

Looking at the feature importances, we found that both profitability stability and leverage-related ratios are key. Persistent EPS in the last four seasons, Borrowing dependency and Total debt to Total net worth appear among the top predictors in several models. This fits our economic intuition that firms with unstable earnings or a high reliance on debt are more likely to experience financial distress.

For the other research questions the results are more mixed. There is some evidence that smaller firms are slightly easier to predict in terms of AUC, but the patterns in recall depend a lot on the model and on whether we use SMOTE. We also do not find strong support for the idea that failing firms are always more unstable; higher variability seems to be ratio-specific rather than firm-wide.

Taken together, our findings suggest that ML models combined with careful data cleaning and class-imbalance handling, provide a useful tool for bankruptcy prediction. At the same time, model performance still relies on design choices such as SMOTE and dimensionality reduction. What's more, predictions should be interpreted together with economic reasoning about profitability, leverage and firm characteristics.