**Bankruptcy Prediction**

**Benjamin Caine | Jiating Li | Qiukai Chen | Xing Ying | Yiwei Wang**

***Original Work Statement***

We the undersigned certify that the actual composition of this proposal was done by us and is original work

**Executive summary**

At the start of this process, we set out to build a project and model that could accurately depict and predict the likelihood of a business to face bankruptcy. Using a multitude of data mining techniques, we gathered an understanding of models that improved our prediction capabilities along with models that offered less suitable results. Through data processing, application of machine learning techniques and comparative analysis amongst the models results, we were able to arrive at findings that shed light on the variables and factors that may lead businesses to face bankruptcy more often than others

After running our models, it is evident and clear which model should be utilized when approaching our business questions. The DNN model, which has the highest average sensitivity rating amongst our seven models, is the model of choice. It is far and above the best model based on average sensitivity as its value is more than 0.07 than the second highest model’s average sensitivity. From our findings and ranking by average sensitivity, it is interesting to see that of the seven models we tested, five of them record average sensitivities in the .60 to .75 range with the other two models, DNN and ParPF being high and low outliers respectively. It is also worthwhile to note that the models with the highest average sensitivity typically have the lowest average accuracy as seen in our results and findings section. It can be concluded that there seems to be trade off between sensitivity and accuracy.

## **Data Description**

**Overview**

The company bankruptcy prediction data set was constructed by Taiwan Economic Journal and covers 1999 to 2009. When defining bankruptcy, the dataset utilizes the definition given from the business regulations of the Taiwan Stock Exchange. The dataset consists of 95 variables, with two independent variables and 93 dependent variables, of which we narrowed down in our model that cover various corporate financial indicators, such as *Operating Profit Rate*, *Operating Expense Rate* and so on.

**Data Source**

**Source:** Deron Liang and Chih-Fong Tsai, deronliang '@' gmail.com; cftsai '@' mgt.ncu.edu.tw, National Central University, Taiwan

**Website:** <https://www.kaggle.com/datasets/fedesoriano/company-bankruptcy-prediction>

**Data Description**

**Input:**

* 6,819 rows of 96 total variables
* There are 93 numeric independent variables and 2 categorical independent variables. The independent variables are comprised of several financial indicators, where the numeric variables are measured in numeric counts such as 0.50 and the categorical variables are measured as 0 or 1

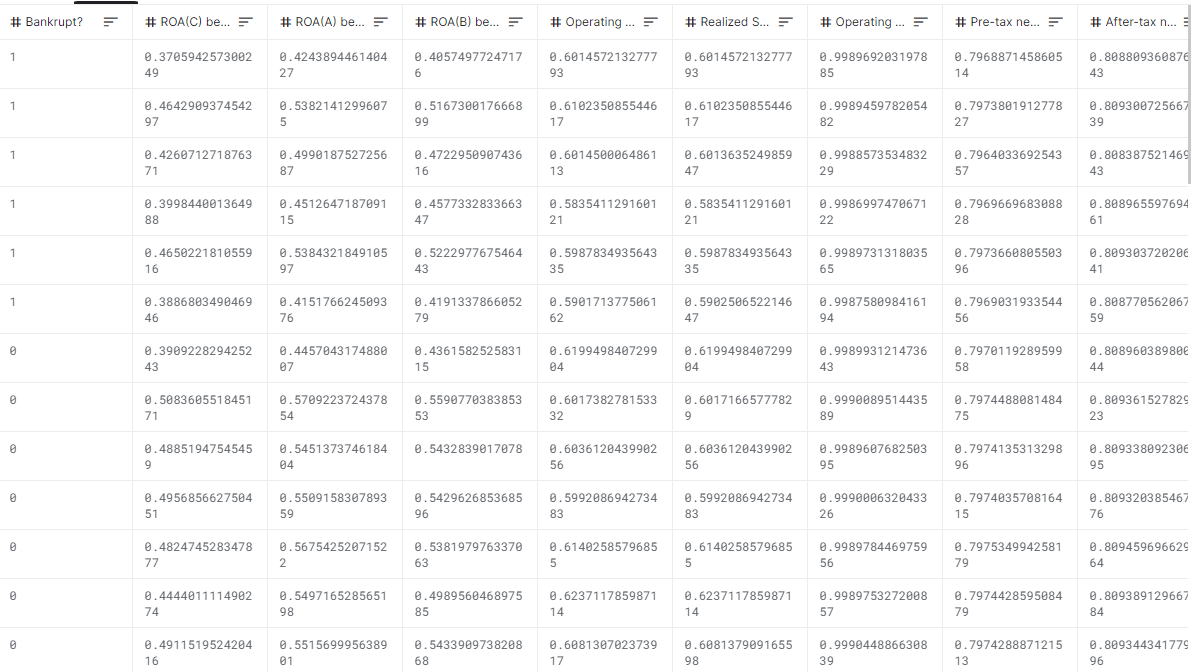
**Output:**

* “Bankrupt?”
  + Our dependent variable will be measured as a categorical variable
    - 1 if bankrupt
    - 0 if not bankrupt.

**Sample Size:**

* **(n)** 6,819 observations
* **(k):** 96 variables

**Sample of Observations:**

****

**Interest:**

We are excited to dive into the ‘Company Bankruptcy Prediction’ dataset because of the real world application and insights into company performance that we can derive from our findings. Utilizing several data mining techniques from class, it will be interesting to see which variables heavily impact the bankruptcy measure of certain companies as well as which ones don’t. Following this project, it will be insightful to take a look at companies that have gone bankrupt in years past and see if any of the impactful variables in our model are evident in these companies leading up to declaring bankruptcy. Additionally, companies that face bankruptcy are minority among all companies. Are there any specific features those companies have or how to use those features to make predictions if a company will be likely to get into bankruptcy, this is our kernel conception we want to dive into.

## **Research Questions**

The financial status of the global economy is ever changing, where businesses, large and small alike, face year over year success and more often than not, failures. According to JonesDay.com, “from June 30th, 2008 to June 30th 2009, bankruptcy filings increased 35%” year over year (Douglas). With the U.S. financial crisis of 2008 mentioned above and the coronavirus global pandemic that impacted billions across the world in 2020, businesses are facing volatility in their year to year processes and must do what they can to be prepared.

Preparing business for the unexpected in an ever changing economy grabbed our attention as a group. To be able to apply practical data mining models and the processes we learned in class, we were more than excited to apply our knowledge to analyze and offer some insight to this real-world problem.

Our initial step was to formulate research questions that both allow us to dive deep into the data and each variable's meaning as well as come away with conclusive and informative findings.

**Research Question 1:** How to make the appropriate lending decision according to a companies performance

Our first question reflects our initial purpose for choosing this dataset: How can we offer businesses facing bankruptcy or those seeking to lend to a potential business, a recommendation with the supportive data on whether or not they should expect the company to go bankrupt, allowing them to make an informed decision. Utilizing predictive models and classification of our variables, our goal is to provide a high accuracy predictive model that can assign a value of 1 or 0 to a potential company on the prediction of them going bankrupt

**Research Question 2:** Extracting statistical significant variables that can affect our models decision

Our second model dives into the supporting data, mentioned in our first research question. Here, we seek to find variables in our dataset that are statistically significant and have a greater weight in determining our models bankruptcy prediction than other variables.

## **Methodology**

We implemented a variety of data mining techniques in our research process that were required in achieving the results we initially were seeking from our established research questions.

Our first data mining technique was dimensionality reduction. Here, we incorporated a low variance filter, high correlation filter and PCA to reduce dimensionality. Our low variance filter removed variables that had repeated values up to 97% while our high correlation filter picked out and discarded variables whose correlation coefficient was higher than 0.8. The principal component analysis or PCA for short, utilized the prcomp() function to derive low dimensional features from the variables in our model. Together, these three processes were an integral first step in ensuring our data was fit for modeling.

The next process, to evaluate generalized error of each algorithm, we used was K-Fold Cross-Validation. Here, we split the data into four separate parts and repeated this process ten times. Using the K-Fold Cross-Validation was a resampling procedure we included in order to ensure that all of the data points in the dataset were accounted for in our models.

Our third technique was to normalize our data through a normalization function. Here we scaled each of our test using the mean and standard deviations of training data set to prevent data leakge. Given the nature and format of our data, it was effective to normalize our variables to reduce the scale of their values as well as the distribution of the dataset as a whole. And because we will use SMOTE in the next step, which is an oversampling technique based on KNN. The oversampling performance will be much better if we scale the data first.

Next, we used SMOTE, which stands for Synthetic Minority Over-Sampling Technique, was our way to approach the issue of our dataset being unbalanced. In R, we initially put our dataset into a chart that breaks down the number of observations that are 0 and 1 for our dependent variable. It was immediately clear that we had a balance issue as only 3% of observations were 1. Why not build a model upon this? Because classic machine learners are more likely to learn the distribution of majority, which leads to high prediction but low sensitivity (or to say specificity). For example, we predict all the labels using the label of majority, the accuracy will be 97%, which is amazing high, but the sensitivity would be 0. Therefore, to correct this, we implemented SMOTE to synthesize new samples to give balance to our dependent variable.

Our final data mining technique was the implementation of nested cross-validation. Here we use 2-repeats 5-folds cross-validation to tune our models to find the optimal hyperparameters for training data set. Thus, we got two cross-validation, one for outer loop and one for inner loop. The aim of inner loop is to tune hyperparameters and outer loop is to evaluate our model’s performance on different test data set to evaluate the generalized error of different algorithms.

## **Results and Findings**

We began our research process with a simple research question, how can we assist those with making the appropriate lending decision according to a company's performance. The formulation and selection of our methods on the other hand were not as simple. To test our goal of this study, we ran a multitude of models against our cleaned dataset in order to analyze their predictive capabilities through their performance metrics. What we found was that each model differed when it came to their average accuracy and average sensitivity ratings. How our models computed these performance metrics are unique to the processes of each model. While the processes within each model differ, the test remained the same throughout: How does the model predict on the ‘Bankruptcy?’ variable. Below, find our models and their respective performance metrics with the complete performance metrics and the models themselves placed in the appendix.

|  |  |  |
| --- | --- | --- |
|  | **Avg. Sensitivity Rank** | **Avg. Accuracy Rank** |
| **Deep Neural Network** | 0.8273 1 | 0.8044 7 |
| **Rpart(Cost-Sensitive CART )** | 0.7545 2 | 0.8897 6 |
| **Boosted Generalized Linear Mode** | 0.7318 3 | 0.8928 5 |
| **Penalized Discriminant Analysis** | 0.7227 4 | 0.8978 4 |
| **Logistic Regression** | 0.6227 5 | 0.9259 2 |
| **KNN** | 0.6045 6 | 0.9130 3 |
| **ParRandom Forest** | 0.2955 7 | 0.9666 1 |

## **Conclusion**

From the tables above, we can see models’ performance varies a lot. So how to choose the best one model? According to average sensitivity, deep neural network is the best solution, but on the other hand, it has lowest accuracy. For conversative investors, we recommend them to choose DNN to make predictions to assist them in investment. According to accuracy, the random forest has the best performance, it is suitable for aggressive investors. Additionally, it is obvious that there is trade off between sensitivity and accuracy. So, when put it into practice, we might consider to give a concession. For example, we can set a threshold for accuracy, like 0.8, and then we choose the best algorithms with the highest sensitivity.

## **Future Work**

There are still many things we want to finish but limited by time and computing redundancy. First, from the perspective of algorithms, for all the models above, we find some of them have large variance in sensitivity, which means they are overfitting to some extent. But for those models which have relatively stable performance, it does not mean it will not overfit in another new test data set. Thus more cross-validation (repeated k-folds) should be done to evaluate their performance more generally. Additionally, ensemble methods can also be implemented to increase both accuracy and sensitivity. We can combine our seven models or even more together to predict classification labels based on majority votes or average prediction probability.

Second, from the perspective of data, we can see that the imbalanced data set is still problematic even after oversampling. After all, oversampling is just to artificially synthesize new samples, which could cause overfitting no matter which technique we use, and undersampling will cause information missing. Therefore, obtaining more real samples can be best. But what if that is all we could have. Different from oversampling and undersampling, there are also some techniques to deal with this kind of problem without changing the original data. For example, one class SVM, which is designed for abnormal detection, in our problem, bankruptcy can be seen as an abnormal label. And cost sensitive learning, which takes the misclassification costs into consideration, can be optimized towards trading off sensitivity and accuracy. After all, in this project, we do care more about sensitivity than accuracy, but the influence of accuracy should also not be ignored.

**Appendix (Anything additional)**

Bibliography

Douglas, Mark. “The Year in Bankruptcy: 2009: Insights.” *Jones Day*, 2010, https://www.jonesday.com/en/insights/2010/02/the-year-in-bankruptcy-2009#:~:text=The%20Administrative%20Office%20of%20the,up%20from%20934%2C009%20for%202008.

**Complete Performance Metrics**

