**BANKRUPTCY BETS:**

**HELPING BANKS FIND BETTER PARTNERS**

**Submitted to:**

Dr. Lin Hao

Data Mining for Business

Fordham University

**Prepared By:**

Chloe B. Platek, Dovid Zev Rosenbaum,

Jonkheer David G. Graafland, Leonardo G. Luchetti,

Owakhela Kankhwende, Sruthi Mudunuri

**December 6, 2020**

**Table of Contents Page**

[**Ⅰ. Executive Summary**](#_y05t4t6pmsyv) **3**

[**Ⅱ. Problem Statement**](#_drijfhfpxeuh) **3**

[**Ⅲ.Planning the Prediction Model**](#_1zvrgtovht42) **4**

[Purpose](#_93052bohyox) 4

[Task](#_bpioudop2p2d) 4

[Data Acquisition](#_l0ce5zyrjfus) 4

[Data Preprocessing](#_mbzjpgj1yc5x) 4

[**Ⅳ.Data Exploration:**](#_hujp4r39zap8) **4**

[Predictors Selection:](#_ogjsp8mm7c0) 4

[**Ⅴ. Oversampling:**](#_pq5iw84su5u) **6**

[Initial Implementation:](#_e6t5xvyeb5u1) 6

[Final implementation of over-sampling:](#_fttznf6d4n2d) 6

[**Ⅵ. Analysis and Results**](#_ynra9wq85n6e) **7**

[Model Execution:](#_7evwrryvqvk2) 7

[Model Selection:](#_rtcy63z9pe05) 9

[**Ⅶ.Plans For Future Research**](#_90h6qmyuylul) **11**

[Breaking down the current data further](#_741da2s2ndxg) 11

[Using Macroeconomic or “Environmental Data”](#_rdb2hxroxq3m) 11

[Global](#_9fn67p44h0m5) 11

[**References**](#_e4ka9741r3u3) **12**

# Ⅰ. Executive Summary

Evaluating the risk of corporate bankruptcies is of paramount importance to creditors and investors. Our assignment utilized our client PKO Bank Polski's corporate debtor dataset to develop a bankruptcy prediction model. PKO Bank Polski is the largest bank in Poland and one of the largest in central/eastern Europe. It was founded in 1919 and is in the commercial banking sector. In 2019, their profit was 4,03 billion złoty, or 950 million Euros, and assets valued at over 348 billion złoty, or 81 billion Euros. This model would assist PKO Bank Polski in deciding which corporations to supply credit, minimizing risk, and maximize interest income. The dataset was fully complete with standardized numerical predictors; our preprocessing involved finding the relevant predictors and dealing with the data's skewed aspects. For predictor selection, we utilized two concurrent methods. We narrowed down the 64 attributes to 38 via average analysis, correlation, and box plot visualization for the Mean Analysis method. The other method used Principal Component Analysis for dimensionality reduction, the results being a dataset of seven principal components. These two datasets were used concurrently, later on, to find the best results during model testing. To counteract possible adverse effects from the imbalance dataset, an Oversampling treatment was introduced within the model algorithms. This alteration would allow us to assess whether oversampling and henceforth accounting for our data's imbalanced aspects would yield better results as the models are tested. A variety of models were built and tested, their performance was ranked on their accuracy, AUC\_ROC score, and total time it takes to run such a model. After developing a ranking algorithm, five models were chosen as the most effective while being time-efficient. Utilizing the confusion matrix results we recommend the Neural Networks Model and Decision Tree model would be the most effective. This report aims to bring new knowledge to PKO Bank Polski by outlining the development of efficient and accurate models for investing decisions.

# Ⅱ. Problem Statement

On average, in Poland, in 2019, corporate loans made up 29.2% of bank loans and 19.5% of bank assets in the year. The industry boasts a robust lending system to corporate debtors and is stringent on healthy returns. Our firm, Team 4, was approached by the foremost Polish Bank, PKO Bank Polski, to develop a lending strategy that better assesses their institutional client's creditworthiness for future loans. In the advent of modern data analysis techniques, PKO wanted us to develop a way to predict their clients' potential financial distress. The jeopardy of those assets due to corporate bankruptcy can put 383,523 złoty at risk for banks. A strategy that incorporates the predicted probability of bankruptcy will alleviate PKO's lending risk and protect profit via its projected interest income. The dataset used for this details the bank's polish client corporations' bankruptcy status over five years. The file contains the financial information of 10,000 companies spread out over 64 numerical attributes representing relevant financial ratios for credit analysis. Some companies filed for bankruptcy, and the data makes a distinction between them and the successfully operating ones. Armed with a viable prediction model and subsequent profit analysis, We aimed to provide PKO with thorough means of minimizing the cost associated with the risk of bankruptcy and maximize their interest income profits.

# Ⅲ. Planning the Prediction Model

## **Purpose**

* **DIDA Framework**

**Data:** 10,000 rows of 64 financial ratios (detailed below) of Polish corporations, including the class-dependent variable that details whether the corporation went bankrupt (1) or not (0).

**Insights:** Probability of a corporation going bankrupt and henceforth the riskiest debtors amongst PKO’s corporate clients.

**Decision:** Which corporations to provide loans to, the observations that are not predicted to appear in the bankruptcy class.

**Advantage:** Minimizing the risk of default of loans underwritten by PKO,

maximize interest income.

## **Task**

* Predicting the probability of bankruptcy amongst 10,000 Polish companies.

## **Data Acquisition**

* Standardized values of 64 attributes and the class depicting bankruptcy (1) or not (0)
* Before having eliminated any attributes, for probability as per the formula n = 6 x m x (u + 1) we required at minimum 6 x 2 x (64 + 1) = 780. As we had 10,000, we opted to assess observations that must be eliminated in data exploration and preprocessing.

## 

## **Data Preprocessing**

* There was no need for missing value imputation as there are no missing values in the whole data set
* No need for Categorical variable coding as they are all numerical variables.
* Numerical variables are in Standardized form.

# Ⅳ. Data Exploration:

## Predictors Selection:

We utilized two predictor selection methods to prepare two datasets for model building and deployment. Mean Analysis would hinge on the predictor’s averages, while Principal Component Analysis would rely on choosing predictors via covariance. The datasets from both would be used concurrently in later steps as we searched for optimum results.

Case 1: Mean Analysis

Firstly, we divided the data into two Data sets as bankrupt and non-bankrupt. We then ranked the attributes based on the difference of their means in the two data sets. A more considerable difference between the two means shows that the attribute has significantly different bankrupt and non-bankrupt values.

We moved on to looking at the correlation between each pair of attributes. We removed the attribute with the lower importance for the pairs with a correlation more significant than +0.80 (importance ranked based on the above mean analysis explained).

After the above elimination, we were left with 38 attributes out of 64. The attributes left were placed into our numerical variable list, and our class attribute was coded as the sole categorical variable with the 'class\_0' column dropped. Below are the columns for the dataset made via this predictor selection method.

'Attr2' 'Attr3' 'Attr5' 'Attr6' 'Attr9' 'Attr11' 'Attr12' 'Attr13'

'Attr15' 'Attr16' 'Attr17' 'Attr20' 'Attr21' 'Attr24' 'Attr25' 'Attr27'

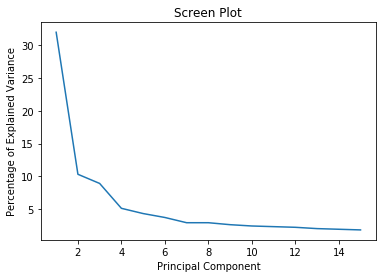
'Attr29' 'Attr30' 'Attr34' 'Attr36' 'Attr37' 'Attr38' 'Attr39' 'Attr41'

'Attr45' 'Attr46' 'Attr47' 'Attr50' 'Attr51' 'Attr53' 'Attr54' 'Attr55'

'Attr57' 'Attr59' 'Attr60' 'Attr61' 'Attr63' 'Attr64' 'class\_1'

Case 2: Principal Component Analysis (PCA)

Principal Component Analysis is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets by transforming a more extensive set of variables into a smaller one that still retains most of the information from the more extensive set.

The PCA method was initialized to find and keep the 15 principal components. The model achieves this by first building a covariance matrix with all possible pairs of the initial variables. The first principal component accounts for the most considerable possible variance in the dataset. The second principal component is calculated in the same way with the condition that it is uncorrelated with the first principal component and so accounts for the next highest variance. This process continues up to indicated principal component 15. Eigenvectors of the covariance matrix are the directions of the axes where there is the most variance. The associated eigenvalues give the amount of variance carried in each Principal Component. The principal components are ordered via ranking their eigenvectors in order of their eigenvalues from highest to lowest. The percentage of variance tied to each component is found by dividing that component's eigenvalue by the sum of all eigenvalues. We then plotted a PCA line graph detailing the Percentage of Explained Variance on the Y-axis and the fifteen Principal components in descending order of explained variance on the X-axis. 

The figure featured here detailed Principal components 1 through 7 representing the complete dataset. After that, Principal components 8 through 15 were dropped. With the seven top Principal components found, the class attribute was added to the dataframe, and all variables were coded as needed. This dataset and the dataset from the Mean Analysis method were brought forth for the next steps.

# Ⅴ. Oversampling:

One issue that arose during our exploration phase is that there is a small target that we are aiming to hit. Less than 3% of the dataset's companies go bankrupt. To counteract possible adverse effects on results from an imbalanced dataset, we introduced oversampling into our methodology. Creating a balanced dataset would benefit machine-learning techniques like neural networks, making more reliable predictions when trained with balanced data. However, some methods like regression do not. The chosen method was SMOTE or Synthetic Minority Over-sampling Technique.

## Initial Implementation:

We oversampled on the training set as a whole. This step caused the train and validation set to have exact copies, leading to a highly overfitted model. We rethought the implementation as an alternate treatment of the training data within the models for comparison later on.

## Final implementation of over-sampling:

Using the imblearn (library for imbalanced datasets) pipeline function, we implemented a way to perform oversampling on the training set and not on the validation set while performing cross-validation. When the Oversample treatment route within the model is initialized, The parameter grid is explicitly made for the oversampling instance. SMOTE is initialized, and a pipeline is made to be loaded with the model instance and SMOTE instance. After that, the GridSearch cross-validation method is utilized to find the optimal model candidate.

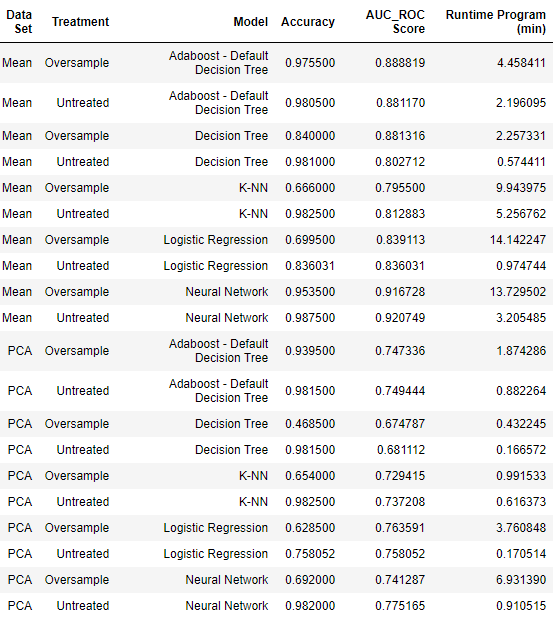
The "treated" or oversampled route and the "untreated" route within each model would be carried onward into model selection when the models are tested. These would be assessed alongside each other for the same Accuracy, AUC score, and Runtime metrics. In order to pick the best composition for final implementation and recommendation.

# 

# Ⅵ. Analysis and Results

## Model Execution:

Below we have generated a table that describes the performance of each model based on their accuracy, AUC\_ROC score, and total time it takes to run such a model in minutes. The models, as previously mentioned, were applied in two different datasets (mean analysis and PCA analysis) with two different treatments (untreated data and oversampled data).

****

*Neural Network:*

With Principal Component Analysis completed and no oversampling done to the dataset, the neural network did have a relatively high AUC of 0.75. However, a major flaw in that result is that the neural network predicted zero companies that would go bankrupt when the test partition has 35 bankrupted companies. Delving further into the classification tree that came from the Principal Component Analysis, we can see a similar issue. The classification tree has a higher AUC, but only correctly predicts 2 out of 35 bankrupted companies.

That being said, when we folded in mean analysis not only did our AUC rise, but the correctly predicted amount of bankrupt companies rose as well. The AUC for this model is 0.91 and the model correctly predicted 16 out of 35 bankrupted companies.

As for the runtime when it came to executing these models, the results were satisfactory. While the treated mean dataset runs a bit longer than others, we deemed it manageable at 13 minutes. As for the untreated mean dataset, the runtime was cut significantly down to a runtime of four minutes. Runtime was significantly shorter when we used the PCA datasets

*Logistic Regression:*

When we ran logistic regression for our datasets, we saw that in both datasets, the untreated datasets performed better than the oversampled datasets during logistic regression. Additionally, the runtime was significantly faster when executing over untreated datasets, with the runtime being, on average, 12 times faster than the oversampled datasets. Ultimately, we concluded that the untreated mean dataset was the most effective out of the four.

*Decision Tree Classifier:*

Impressive in both its AUC\_ROC score and runtime, our decision trees performed in the top tier of our different models. Both Mean Analysis datasets performed well with an AUC\_ROC score of at least .80. As for runtime, the two mean analysis datasets were among the fastest models, with the treated dataset taking a little bit longer to run at almost two minutes. Those two minutes did add about .08 to the AUC\_ROC score, ultimately making the treated mean analysis dataset the third highest ranking method for predicting which potential clients would go bankrupt.

Our top three indicators for predicting bankruptcy are listed below. One thing to take note of is the financial ratio operating expenses / total liabilities which appears in all three rules.

1. If operating expenses / total liabilities is greater than -0.52 and (gross profit + depreciation) / total liabilities is greater than -0.34 but less than or equal to 0.041 THEN predict 1

2. If operating expenses / total liabilities is less than or equal to -0.52 and sales / total assets is less than or equal to -0.62 and equity / fixed assets is greater than -0.039 but less than or equal to -0.025 and (total liabilities \* 365) / (gross profit + depreciation) is less than or equal to -0.55 THEN predict 1.

3. If operating expenses / total liabilities is greater than -0.52 and (gross profit + depreciation) / total liabilities is greater than 0.04 and retained earnings / total assets is less than or equal to -0.016 and total assets / total liabilities is greater than 1.78 and sales / inventory is greater than -0.028 THEN predict 1.

*Nearest Neighbor Classifier:*

Our nearest neighbor model performed in the middle of the pack. Starting with the AUC\_ROC score, both mean datasets, treated and untreated, performed noticeably better than their PCA counterpoints. The AUC\_ROC score rose by at least .07 points when we moved from PCA datasets to the Mean Analysis datasets.

However, a big advantage the PCA datasets had over the Mean Analysis datasets was runtime. Both PCA datasets took less than a minute to run whereas their Mean Analysis counterparts took on average six minutes to run.

*AdaBoost:*

The implementation of the AdaBoost for Decision Tree Classifier scored fairly high on AUC\_ROC of the Mean Analysis dataset both for the untreated and oversampling method. The runtime of the model on the Mean Analysis dataset was slightly above average.

When applying the model to the dataset derived using PCA we can see that the performance is significantly worse for both the untreated dataset and the oversampled dataset. The runtime on the other hand is significantly better than on the Mean Analysis.

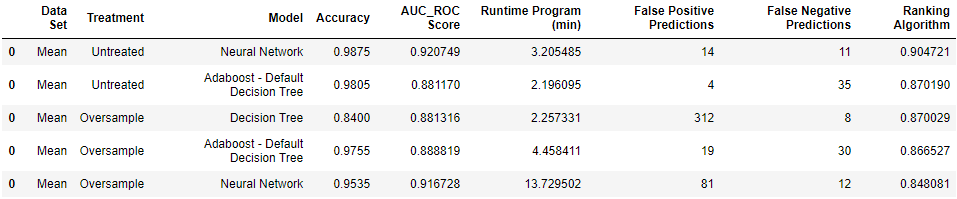
*Random Forest Classification:*

The Random Forest Classification model was thought to be very promising since it combines several different decision tree classifiers into one optimized model. However, the model took an extremely long time (over an hour) to train and was therefore excluded from the project.

## Model Selection:

After analyzing the results described above, our team felt the need to compile the displayed information and attempt to find the best performing model on the combined factors of AUC and Runtime, but also add more information to our selection.

To initiate the selection process, we created a ranking algorithm, which would combine the AUC score with the Runtime with predetermined weights and return a ranking value. The algorithm takes the AUC score of the model and penalizes it for higher runtimes, so the best models would not only have a great AUC score but would also be efficient.

After generating the Ranking Algorithm score for each model, our team selected the five more effective and efficient models to conduct some further analysis to choose the model to recommend to our potential client. One way to conduct this analysis was to observe the values they predicted on a confusion matrix and take into account the goal of our DIDA framework. The table below shows the results of our analysis:

The false-positive predictions and false-negative predictions were taken from the confusion matrix of each model. We found that such information was crucial to choose our recommended model since the goal of the model goes beyond a sole AUC score. By looking at the confusion matrix results, there are two recommendations our team would make for a potential client:

The first model we would like to recommend is the Neural Network model conducted on the untreated database where we conducted the mean analysis. Beyond its distinctive AUC\_ROC score and efficient running time, we found that its false-negative predictions were also comparably low to its competitor models. This means that beyond making good predictions, it is able to minimize false-negative predictions. This is important to our DIDA framework because this model would most likely be used to support investment decisions. This means that when the model shows that it has predicted a company not going bankrupt when it actually would, this would be a lost investment for our client organization. So minimizing those predictions is extremely important for our model selection, which the Neural Network does well on the untreated sample.

Our second prediction goes along the same thought process that we just described. Our previous model maximizes performance while at the same time lowering the risk of supporting an investment on a company that will go bankrupt. However, that risk is still not the lowest in our best predictive models. In a more conservative investment approach, we would recommend the Decision Tree model on the oversampled database where we conducted the mean analysis. The focus of this recommendation is the absolute minimization of false-negative predictions. Even though it presents a lower AUC\_ROC score than our previous recommendation, it has a reduced number of false-negative predictions, which depending on the business model of the organization, might be the best model to fit their business practices.

It is important to emphasize that both models are good performers as we have noted on our first table in this section. As to our recommendations, the Neural Network with the untreated database is the best performer on a more liberal investment approach, where performance is maximized over a slightly greater risk. On the other hand, the Decision Tree model over the oversampled database sacrifices performance to have the lowest risk among our top predictor models.

# Ⅶ. Plans For Future Research

With a broader, international dataset, it would be interesting to see the differences play out as we move from country to country. The ability to localize our data, as we did here focusing on Polish companies, is a strength, and with more data properly standardized, we would be able to draw larger conclusions about the global economy and how they might influence a company's probability of bankruptcy. To further push that idea, if we can fold in global economic metrics into our dataset, our accuracy of predicting should increase as we can better predict which companies are flexible enough to withstand difficult periods.

## Breaking down the current data further

Firstly, with the financial ratios' findings, we can break these down further to identify the singular financial metrics that go the furthest towards predicting bankruptcy when tested against datasets from other counties. This route could serve to expand the capabilities of the user's ability to predict bankruptcy amongst companies.

Also, the predictors for bankruptcy may change based on what industry the companies may belong within. Our prediction model proved useful for PKO's client dataset across the board. However, the accuracy for prediction could be improved when accounting for how a predictor's value can change per industry.

## Using Macroeconomic or “Environmental Data”

Another further step would be to find the macroeconomic indicators that would play a part in predicting bankruptcy. Corporate performance is always affected by macroeconomic conditions. Macroeconomic variables could include GDP, Personal Income Index, Consumer Price Index, and M2 Index, which reflect the amount of money supply in the economy. These would be employed as year-to-year ratios concurrent with the years of the initial data.

## Global

Globally, the implications of the development of new bankruptcy prediction models are currently increasing. Many quoted companies operate in several countries, which means differences between them are reduced, regardless of their location or the factors particular to the country of origin. A comprehensive analysis that considers different countries, predictors significance within those countries, industries, and banking composition would yield precious insights into bankruptcy prediction in the modern globalized world.

# **References:**

Jaadi, Zakaria . (September, 24 2019) “A Step-by-Step Guide of Principal Component Analysis.” Builtin.com. <https://builtin.com/data-science/step-step-explanation-principal-component-analysis>

“Polsih Banking Group Books Record Profit.” thefirstnews.com. (February, 12 2020).

<https://www.thefirstnews.com/article/polish-banking-group-books-record-profit-10447>

“STATUT POWSZECHNA KASA OSZCZĘDNOŚCI BANK POLSKI SPÓŁKA AKCYJNA.” Bank Polski. <https://www.pkobp.pl/media_files/65ceb636-27eb-4dd8-b892-4e48c0d84352.pdf>