ST 952 Assignment 2 - Group B

Student IDs: u2026624, u2056942, u2119251, u2116856

December 9, 2021

Contents

1 Introduction

This report aims to investigate the subset of a financial dataset, `PolishBR`, introduced from the UCI Repository by Tomczak et al. (2016). With the help of a logistic regression model, we try to predict the probability of company bankruptcy in terms of various financial indicators.

2 Exploratory Data Analysis

2.1 Data Preprocessing

`PolishBR` contains 291 missing values. After closer inspection, we discovered that cost of products sold (*CostPrS*) and current liabilities (*CLiabil*) are missing together in 266 observations. The remaining missing values also seem to be missing together with other variables. That indicates that these are not missing at random, and we cannot use any imputation method on them. However, noting the size of our data, we can safely ignore these values altogether. The proportion of companies that went bankrupt in the data made of observations containing some missing information is about 0.127, while in the complete data it is only about 0.061.

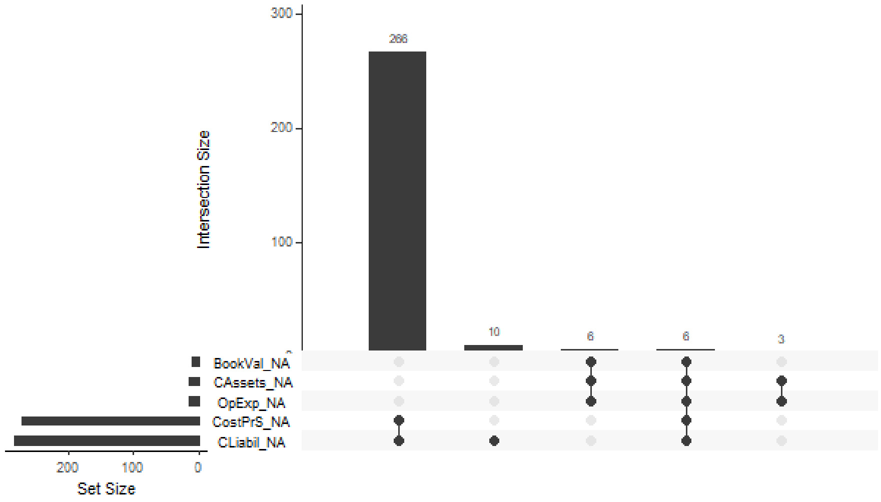


Figure 1: Missing values.

After removing missing values, we end up with 5218 observations of 11 variables, two of which are categorical (factor type).

2.2 Preliminaries

The distributions of numerical predictors are highly skewed to the right and contain multiple extreme outliers. Naturally, we expect some unusually high values in the tail of distributions. However, there is one suspicious observation that looks more like a mistake considering other values of this variable:

Table 1: Suspicious observation.

bust	TAssets	TLiabil	WkCap	TSales	StLiabil	CAssets	OpExp	BookVal	CostPr	CLiabil
0	5.961368	-2568.575	Low	391.1075	69.77781	31.27302	382.3126	2025.989	382.311	78.91632

Since we cannot determine the nature of such a high negative value of total liabilities (*TLiabil*), we decided to discard this observation. After the literature review, we found out that the book value of equity (*BookVal*) can be calculated by subtracting total liabilities from total assets, i.e.

$$\text{Book value of equity} = \text{Total assets} - \text{Total libilities.}$$

Indeed, these values look alike (accounting for the rounding error and few unusual observations). Table2 displays the difference between *BookVal* and a variable calculated according to the formula above.

Table 2: Book value of equity difference.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1911.97	-1.67	0	-3.04	0	0.14

Furthermore, the working capital (*WkCap*) can be calculated by subtracting current liabilities (*CLiabil*) from current assets (*CAssets*). However, since we are given categorical labels for this variable, we decided to retain it in its original form.

A word on current liabilities

The data does not provide any explanation of the difference between current and short-term liabilities. Many sources use these terms interchangeably to quantify financial accountability to be settled within the fiscal year of the operating cycle. However, plotting these variables against one another shows almost perfect alignment up to some point where the pattern breaks. Furthermore, one would expect the current liabilities to be part of total liabilities - similarly, current assets to be included in total assets. Even though the data supports the latter, it does not justify the former. Nonetheless, short-term liabilities do behave as expected concerning total liabilities, and therefore, we believe they are more reliable in our analysis. From now on-wards, unless otherwise stated, we shall treat short-term liabilities (*StLiabil*) as current liabilities and use these terms interchangeably.

2.3 Variables Relationships

The percentage of companies that went bankrupt split by their working capital is presented in Table3. 16% of companies with low working capital became insolvent. This number drastically decreases for medium, high and very high working capital, suggesting that this variable could be a valuable predictor.

Table 3: Working capital and bust (%).

	Low	Medium	High	Very High
0	84	95.1	97.1	98.94
1	16	4.9	2.9	1.06

2.3.1 Correlation

It is always a good practice to examine the correlation between explanatory variables. Figure2 shows the correlation plots for raw and transformed variables. To produce the second plot, we used log-transformation on all numerical predictors, but *BookVal*, which has been discarded. We can see how the extreme outliers impact the correlation (especially *CostPrS* and *CLiabil*) and how a simple transformation can reduce their effect.

Not surprisingly, the correlation between liabilities and assets (be it current, short-term or total) is high. We note an almost perfect correlation between operating expenses (*OpExp*) and total sales (*TSales*). Plotting these variables does not reveal much. However, we may reason that as the company grows, it produces and sells more goods (high *TSales*), and its operating expenses tend to increase accordingly. Therefore, in this case, the company's size might be a lurking variable responsible for such a strong correlation between these variables. Perhaps unexpectedly, operating expenses and the cost of products sold are highly correlated. It turns out that in about 60% of observations, operating expenses observations are almost identical to the cost of products sold (correlation of 0.999). Data does not give any indication of why such a perfect relationship exists.

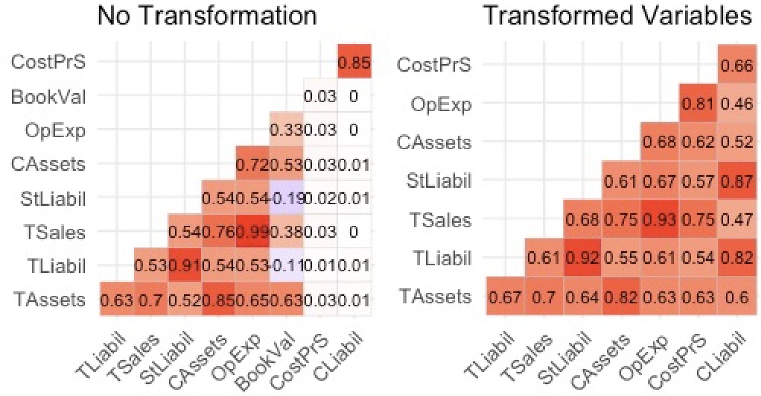


Figure 2: Correlation plots.

2.3.2 Short and Long-Term Obligations

Let introduce a measure helping us quantifying companies liquidity (ability to pay debts as when they fall due):

$$\text{Current ratio} = \frac{\text{Current Assets}}{\text{Current liabilities}}.$$

A ratio lower than 1 suggests that short-term (within one fiscal year) liabilities are greater than accessible assets. Next, we investigate the relationship between going bankrupt and the current ratio. Table4 displays the percentage of companies that went bankrupt and had the current ratio below or above 1. We can see that almost 17% of companies, having a ratio lower than 1, became insolvent.

Table 4: Current Ratio and bust.

Ratio Below 1	Bust	Percentage of Bust
No	No	96.46%
No	Yes	3.54%
Yes	No	83.07%
Yes	Yes	16.93%

To measure companies capabilities to meet their long-term financial obligations, consider the following metric:

$$\text{Shareholder Equity Ratio} = \frac{\text{Book value of equity}}{\text{Total assets}} = \frac{\text{Total assets} - \text{Total liabilities}}{\text{Total assets}}.$$

High Equity Ratio indicates that the company effectively manages its finances, and therefore, it will be able to meet its long-term financial liabilities Companies with Shareholder Equity Ratios bigger than 0.5 are considered conservative because they rely mainly on the stakeholders' equity rather than on debt. On the other hand, a company with a ratio smaller than 0.5 has a substantial amount of debt in its capital structure - it is called a leveraged company. If the ratio is considerably small, the company's financial liabilities might outweigh its income.
We can see that companies that went bankrupt tend to have smaller Shareholder Equity Ratios and low-to-medium working capital. This observation indicates that the ratio could help determine the likelihood of becoming bankrupt. The negative Equity Ratio is an indication of financially dangerous struggles. It is, therefore, no surprise that among 246 companies with a negative ratio, over 28% became insolvent.

2.3.3 Remaining Predictors

So far, we have investigated how assets, liabilities and their combination, such as equity ratio, can be used to explain a company's finances and impact the probability of a company going bankrupt.

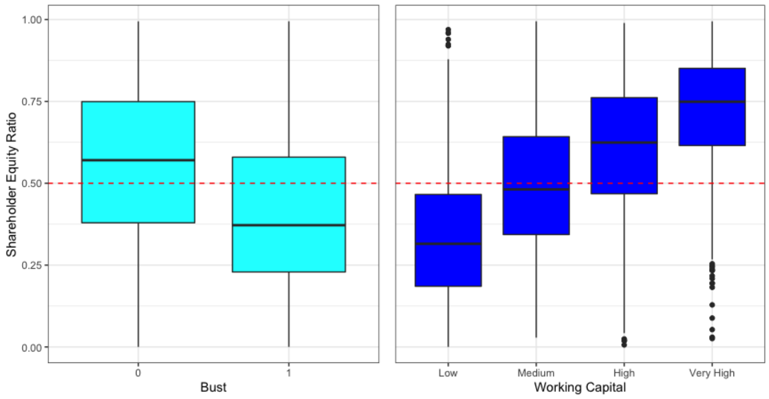


Figure 3: Equity Ratio Relationship.

We are left with three predictors to examine. Consider Figure4 showing log-transformed variables plotted against log odds ratio. We can see that for all cases, the variable's values increase results in a (more or less apparent) decrease in log odds ratio, that is, the lower the odds of a company going bankrupt. This decrease is the most evident for total sales. Intuitively, more sales imply more income and more financial stability. It is less obvious why operating expenses and the cost of goods sold can decrease the odds ratio. As mentioned earlier, we suspect that these measures indicate the company's size and growth, generally speaking. Therefore, it may explain the present trends. Plotting log odds ratio against current or total liabilities, however, presents the opposite relation - the bigger the debt and financial obligations, the bigger the chance of going bankrupt.

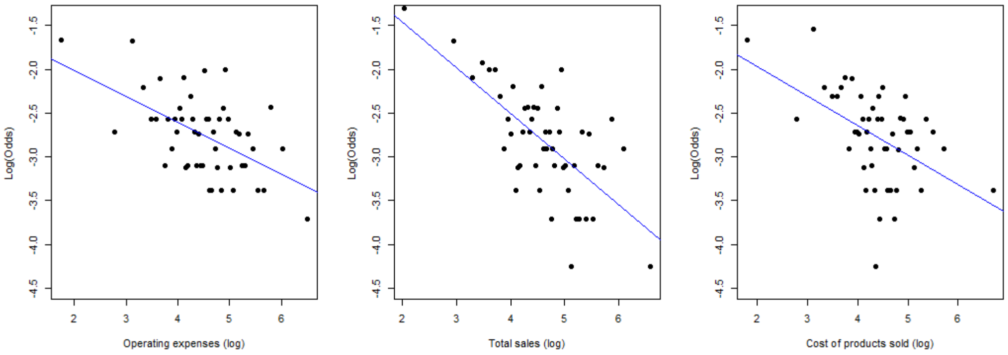


Figure 4: Remaining predictors.

3 Modelling

3.1 Class Imbalance

Before we begin the modelling, it is crucial to address a critical issue our data exhibits, namely the class imbalance. As noted in the second section, companies that went bankrupt constitute only about 6.1% of the dataset. That means this category is rare, and the logistic regression model will struggle to predict this particular outcome. If we do not find a solution taking into account this fact, predicting non-bankruptcy will always be better, in terms of accuracy, than predicting insolvency. The context is of vital importance here. Failing to predict bankruptcy is far more dangerous than anticipating it where it will not actually occur. The model's predictions could be used as a precaution, and the company could take appropriate measures regardless of the outcome. We can translate this scenario into the model's sensitivity and specificity. Here, we are willing to sacrifice some specificity (forecasting a company not going bankrupt when, in fact, it will) in order to increase sensitivity (predicting a company going bankrupt when in fact it will not). As logistic regression returns predicting probabilities vector, a question now arises: how do we decide on the optimal probability threshold for assigning an observation to either class?

3.2 The Need of Validation Set

We use the following steps to find the best logistic model. First, we created dummy variables from the *WkCap* variable. Then, the data was split into three subsets - training, validating and testing set (in proportions 0.7, 0.15 and 0.15, respectively). We ensure that each set contains the same proportion of companies that went bankrupt ($Bust = 1$) by utilising the `caret` package and `createDataPartition` function.

All proposed models *learn* on the training set. Then, using the validation set, we find the optimal probability cutoff for the binary classification. We will explain how to find the optimal cutoff in more depth with an example in Section 3.4.

3.3 Model Fitting

3.3.1 Data Preparation

Following our discussion in the explanatory data analysis section, we discard *CLiabil*, with the belief that *StLiabil* variable is more reliable from an accounting point of view. Moreover, we replace *BookVal* with its definition but retain total assets and total liabilities, which may contain some important information for our models. To address the right skewness for some numerical predictors we apply log transformation.

On top of the ratios explained in the analysis section, we define one additional ratio, the return on capital employed (ROCE), as

$$\text{ROCE} = \frac{\text{TSales} - \text{CostPrS} - \text{OpExp}}{\text{TAssets} - \text{StLiabil}}.$$

This ratio can be used to assess a company's profitability and capital efficiency. Further, for one of the models, we include an interaction between current assets and working capital because the behaviour of *CAssets* may differ for each level of *WkCap*. It is a part of experimentation on variables.

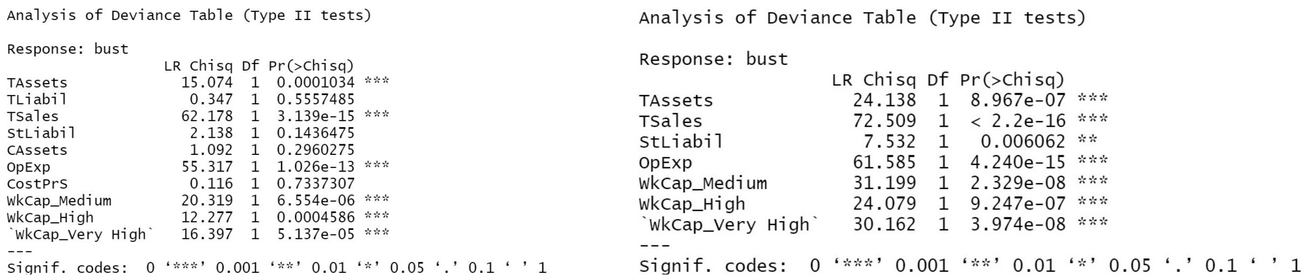
3.3.2 Proposed Models

In this section, we fit numerous logistic models using different variable combinations and selection method techniques to find the best performing model.

- `ModelInitial` is the baseline model fitted on the full data without any transformations and interactions.
- `Model1Base` is the full model with transformed data without any interactions.
- `Model1B` is a reduced version of `Model1Base` using manual backward selection method.
- `Model1L` is a reduced version of `Model1Base` using LASSO.
- `Model2Base` contains transformed variables with added ratios variables and interactions.
- `Model2B` is a reduced version of `Model2Base` using automated backward selection method.
- `Model2L` is a reduced version of `Model2Base` using LASSO.

Deviance Methodology
Here, we explain how to reduce `Model1Base` using manual backward selection method to obtain `Model1B`.

From Figure5(a), we can see that not all p-value of chi-squared test of `Model1Base` are significant, which means that some of the variables make little contribution to predicting bankruptcy. We use



```
Analysis of Deviance Table (Type II tests)

Response: bust
                  LR Chisq Df Pr(>Chisq)
TAssets             15.074  1  0.0001034 ***
TLiabil              0.347  1  0.5557485
TSales              62.178  1  3.139e-15 ***
StLiabil             2.138  1  0.1436475
CAssets              1.092  1  0.2960275
OpExp               55.317  1  1.026e-13 ***
CostPrS              0.116  1  0.7337307
WkCap_Medium        20.319  1  6.554e-06 ***
WkCap_High          12.277  1  0.0004586 ***
`WkCap_Very High`   16.397  1  5.137e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(a) Table of Model1Base

```
Analysis of Deviance Table (Type II tests)

Response: bust
                  LR Chisq Df Pr(>Chisq)
TAssets             24.138  1  8.967e-07 ***
TSales              72.509  1  < 2.2e-16 ***
StLiabil             7.532  1   0.006062 **
OpExp               61.585  1  4.240e-15 ***
WkCap_Medium        31.199  1  2.329e-08 ***
WkCap_High          24.079  1  9.247e-07 ***
`WkCap_Very High`   30.162  1  3.974e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(b) Table of Model1B

Figure 5: Analysis of deviance tables of Model1Base and Model1B

analysis of variance to check if a variable should be retained in the final model. The basic logic of this method is to remove variables that have the highest p-value (the lowest test statistic) step by step. For example, the cost of products sold (*CostPrS*) has a p-value of 0.7337307, i.e. there is no evidence suggests that this variable is useful in the model, consequently, it will be removed in the next step. Carrying this procedure leaves us with the final analysis of deviance table of `Model1B`.

Stepwise Method

This method is similar to deviance methodology. However, the criterium for removing variables for this method is Akaike's information (AIC), it is equivalent to significance level of about 0.15.

LASSO

Another way to reduce the full model is the shrinkage approach. Using the LASSO method we aim to minimise

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})^2 + \lambda\sum_{j}^{p}|\hat{\beta}_j|$$

Commonly, we use cross-validation to choose the optimal λ which can be realised by function `cv.glmnet` in package `glmnet`. However, we get different results for the final model if we repeat this process several times, which means instability of the final result by this method. To address this issue, we use bootstrap, repeating the procedure 100 times to find the average of optimal lambdas, which should stabilise the model. Then we plot the trace to ensure the reasonability of this choice.

3.3.3 Model1 Family - Illustrative Example

Using methods briefly explained above, we can find the models `Model1B` and `Model1L` with coefficients listed in Table5. To use LASSO, we generate bootstrap lambda distribution (one standard error of the minimum of the cross-validation). We find a value of approximately -5.2 as the optimal $log(\lambda)$. Plotting the trace below reassures us of this choice.

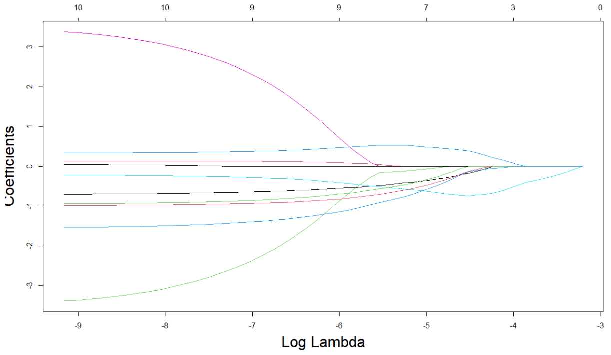


Figure 6: LASSO trace against Log(lambda)

Table 5: Model1 Family coefficients

Model1Base

(Intercept)	log(TAssets)	log(TLiabil)	log(TSales)	log(StLiabil)	log(CAssets)	log(OpExp)	log(CostPr)
-0.13	-0.71	0.14	-3.57	0.33	-0.21	3.58	0.05

WkCap_M	WkCap_H	WkCap_VH
-0.99	-0.95	-1.56

Model1B

(Intercept)	log(TAssets)	log(TSales)	log(StLiabil)	log(OpExp)	WkCap_M	WkCap_H	WkCap_VH
-0.01	-0.71	-3.71	0.36	3.69	-1.10	-1.12	-1.78

Model1L

(Intercept)	log(TAssets)	log(TSales)	log(StLiabil)	log(CAssets)	WkCap_M	WkCap_H	WkCap_VH
0.15	-0.40	-0.11	0.52	-0.58	-0.58	-0.44	-0.73

3.4 Model Choice

We exploit the validation set to find the optimal probability thresholds for each of the models. By using the same validation set for all models we ensure a fair decision on the best performing model, i.e. choosing an optimal cutoff values will not interfere with predictions on the testing set. Let us assess each model on the validation set. Figure7 presents the sensitivity and specificity values of `Model2B` for different probability thresholds. Note that if we used default classification probability of 0.5, model's accuracy would increase to almost 0.95 with extremely high specificity - 0.99, but disappointingly poor sensitivity - only 0.15. Accuracy, in such a case, would be artificially inflated and the model would be unreliable. This is because accuracy-wise, it is better to always predict that a company will not go bankrupt. We find the optimal probability thresholds for each considered model.

	$Bust_A = 0$	$Bust_A = 1$
$Bust_P = 0$	545	7
$Bust_P = 1$	189	40

Table 6: Confusion matrix.

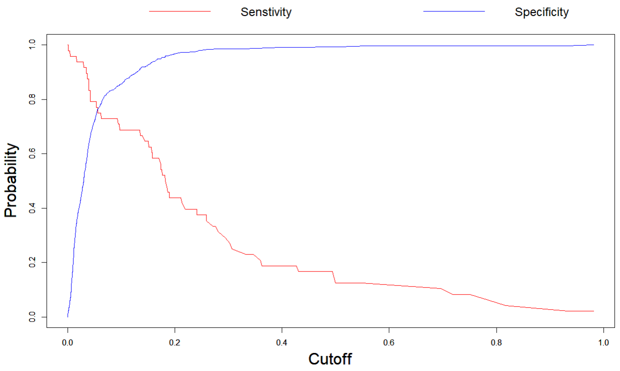


Figure 7: Plot of sensitivity and specificity.

For example, Figure7 helps to choose optimal cutoff for `Model2B`. We take the value of x-axis of the intersection point, which is about 0.0608. Table6 is the confusion matrix for this model with *threshold = 0.0608*, where $Bust_A$ stands for the actual value of *bust* while $Bust_p$ indicates predicted value of *bust*. We can calculate accuracy, sensitivity and specificity from confusion matrix directly.

$$sensitivity = \frac{40}{7+40} = 0.8511$$
$$specificity = \frac{545}{545+189} = 0.7425.$$

Similarly, we can summarise all models performance in the Table7. We can see that the probability cutoff method worked, and the models' sensitivity is reasonable. Comparing models across different metrics suggests choosing `Model2B` as it has reasonably high sensitivity, not too low specificity and gives fairly high accuracy.

Table 7: Comparison for different models

	ModelInitial	Model1B	Model1L	Model2B	Model2L
Accuracy	71.96%	71.57%	67.99%	79.68%	78.56%
Sensitivity	85.11%	85.11%	85.11%	85.11%	87.23%
Specificity	71.12%	70.71%	66.89%	74.25%	69.89%

3.5 Parameters Interpretation

The formula of `Model2B` can be written as:

$$\begin{aligned}\log\left(\frac{\hat{y}}{1-\hat{y}}\right) = &- 0.48 - 0.64\log(\text{TAssets}) - 3.47\log(\text{TSales}) + 0.66\log(\text{StLiabil}) + 3.52\log(\text{OpExp}) \\ &+ 1.61\text{WkCap\_Medium} - 2.24\text{WkCap\_High} - 3.09\text{WkCap\_Very High} \\ &+ 0.02\text{CRatio} - 1.19\log(\text{CAssets})\text{\_Medium} - 0.39\log(\text{CAssets})\text{\_Low}\end{aligned} \tag{1}$$

The optimal threshold for the final model is 0.0608. We can interpret the coefficients in the following manner. For instance, -0.643 coefficient of log(TAssets) means that a unit increase of log(TAssets), while keeping other variables constant, results in the odds ratio of $\exp(-0.643) = 0.5257$, i.e. the odds ratio of going bankrupt decreases by 47.43%. Similarly, for the categorical predictor, e.g. for *WkCap_Medium*, $\exp(1.608) = 5$ - the odds of going bankrupt for companies with medium working capital increases by 400% according to this model.

Total assets are the highest amount that a company can use to repay its liabilities, thus, the larger they are, the less likely the company liquidation is. We can see significant negative coefficients for *WkCap_High* and *WkCap_Very High* as well as their interaction with current assets. The model confirms observations from EDA - more working capital decreases the probability of bankruptcy. On top of that, the combination of it and current assets seems to have some predictive power. On the other hand, short-term liability and operation expenses increase the odds ratio of bankruptcy. The more money a company owes, the more likely it will not be able to pay it back. Operating expenses also cause a decrease in probability.

There is one unexpected observation, namely, the current ratio's positive impact on the probability. That seems counter-intuitive as a high current ratio means higher liquidity of a company, i.e. its current assets can cover more urgent liabilities, which means less repayment pressure. We suspect that extreme values of current assets for companies that went bankrupt impacted this coefficient significantly.

3.6 Prediction of the Given Observation

Table8 gives parameters of the observation that need to be predicted. Using our model we can find the desired probability:

$$\begin{aligned}\log\left(\frac{\hat{y}}{1-\hat{y}}\right) &= -0.48 - 0.64\log(80) - 3.47\log(150) + 0.66\log(23) + 3.52\log(130) \\ &- 2.24 \cdot 1 + 0.02 \cdot \frac{50}{12} = -3.6249.\end{aligned}$$

It follows that

$$\hat{y} = \frac{1}{1+e^{3.6249}} = 0.02596 < 0.0608 = \textit{threshold}.$$

Therefore, our model predicts that the company will not go bankrupt, which is not surprising given the high working capital and total sales.

Table 8: The observation for prediction.

TAssets	TLiabil	WkCap	TSales	StLiabil	CAssets	OpExp	BookVal	CostPr	CLiabil
80	30	High	150	23	50	130	55	70	12

4 Assessing Model Fit

4.1 Added-Variable Plots

Added-variable plot for an explanatory variable X_i is a scatter plot for the regression residual of X_i and other explanatory variables against the regression residual of dependent variable and other explanatory variables. We can find a fitted line of this plot, and the sign of parameter of this line should be consistent with that of the parameter of X_i in the model.

Figure8 shows a series of added-variable plots for `Model2B`. From each plot, it is clear that there are many observations whose value of bust|other is not around 0. *CRatio* added plot's line is distorted by influential observations 2414 and 5082. Interestingly, if we focus on the plots for *Tsales* and *OpExp*, we can easily find that the scatters arrange at a right angle. This may indicate that there exists perfect collinearity between *Tsales* and *OpEXP* for part of the observations. This agrees with the high correlation observed in Figure 2. To sell more goods, a company usually pays more for its operating, managerial and advertising expenses - this can perhaps, partially explain the phenomenon. However, some further investigation would be necessary.

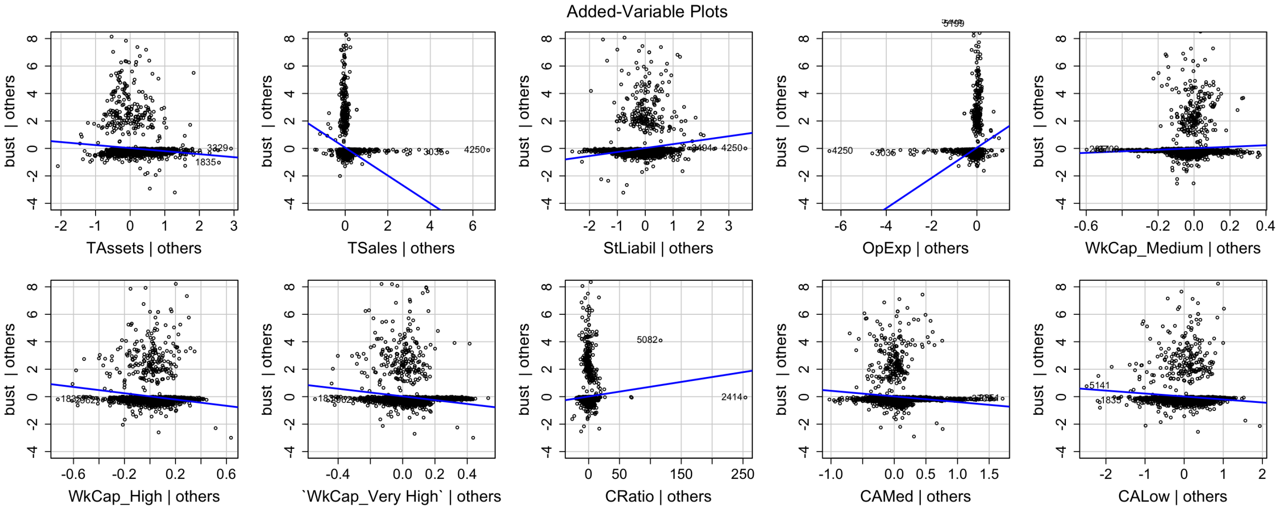


Figure 8: Added-Variable Plots.

4.2 Dfbeta Plots

Figure9 shows the standardised change of each variable when removing each observation. It is worth noting that almost all the plots change significantly by removing the last approximately 250 observations, which are the bankrupt companies. It shows that the observations with *Bust* equal to one are incredibly important for a proper model fitting because they alter the coefficients estimates the most. In other words, the more such observations we have, the better our model fits the data - which again agrees with our previous discussions on the class imbalance issue. We can also see some influential points representing companies that did not go bankrupt. Further analysis of outliers would be useful, leading to perhaps better model fit. Only the Dfbeta plot of *CRatio* variable seems no dramatic fluctuation (apart from one outlier), which is probably because the ratio's coefficient is smaller than for other variables.

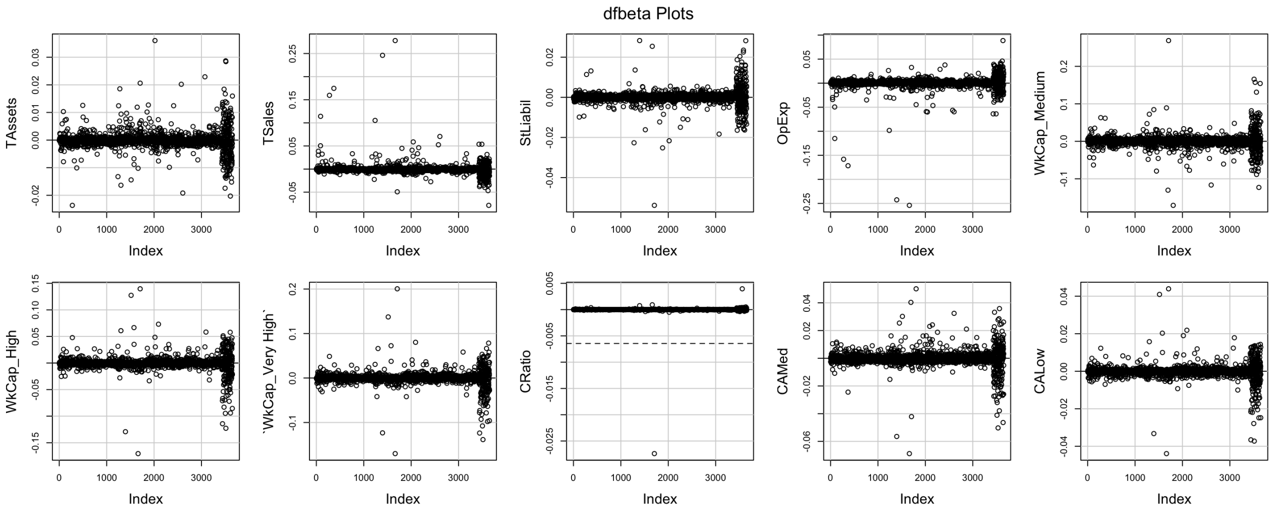


Figure 9: Dfbeta Plots.

5 Further Analysis

This report is by no means exhaustive. Even though we aimed to produce a thorough analysis, time constraints prevented us from delving deeper into more subtle relations within the data. In particular, a proper investigation should be conducted for data with missing values and the disproportion of companies that went bankrupt there. Other methods for dealing with class imbalance, such as SMOTE algorithm, could be implemented and compared with our models. Perhaps more care could be taken for extreme outliers. Some further data explanation could be helpful, for example, data collector's definition of variables and the difference between them. Although we briefly discussed variables' meaning from an accounting perspective, maybe consulting with a specialist would shed light on some patterns we might have missed in our analysis. Another idea would be to use ridge regression to address the multicollinear structure of the data. Finally, a more thorough analysis of the final model fit using various graphs such as binned residual plots and Cook's distance would be desirable.

A Appendix. R Code

```
load("/PolishBR.Rdata")
library(caret)
library(fastDummies)
library(glmnet)
library(car)
library(ROCR)
library(pROC)
library(dplyr)
library(ggplot2)
library(gridExtra)
# Missing values investigation
naniar::gg_miss_upset(PolishBR)
PolishBR_clean <- PolishBR %>% na.omit()
# Check the proportion in bust for missing/non-missing data
Missing <- PolishBR[rowSums(is.na(PolishBR)) > 0, ]
Not_missing <- PolishBR %>% na.omit()
Missing %>% count(bust) %>% summarise(bust, Prop = n/sum(n))
Not_missing %>% count(bust) %>% summarise(bust, Prop = n/sum(n))
knitr::kable((PolishBR_clean %>% filter(TLiabil < 0)),
             caption = "Suspicious observation", booktabs = T)
PolishBR_clean <- PolishBR_clean %>% filter(TLiabil > 0)
# Check that BookVal can be found using TAssets and TLiabil
check <- PolishBR_clean %>%
  mutate(BookVal_2 = TAssets - TLiabil,
         Diff = BookVal - BookVal_2)
knitr::kable(t(round(summary(check$Diff), 2)),
             caption = "Book value of equity difference", booktabs = T)
# Proportion table for WkCap and bust
basic_table <- table(PolishBR_clean$WkCap, PolishBR_clean$bust)
prop_table <- round(prop.table(basic_table, margin = 1) * 100, 2)
knitr::kable(t(prop_table), caption = "Working capital and bust (%)")
# Left-hand side plot
cor_1 <- ggcorrplot::ggcorrplot(round(cor(PolishBR_clean[,-c(1,4)]), 2), lab = T, type =
    "lower", show.legend = FALSE, lab_size = 2, tl.cex = 8, title = "No Transformation")
PolishBR_clean_logs <- PolishBR_clean %>%
  filter(TLiabil > 0) %>%
  mutate(TAssets = log(TAssets),
         TLiabil = log(TLiabil),
         TSales = log(TSales),
         StLiabil = log(StLiabil),
         CAssets = log(CAssets),
         OpExp = log(OpExp),
         CostPrS = log(CostPrS),
         CLiabil = log(CLiabil)) %>%
  select(-BookVal)
# Right-hand side plot
cor_2 <- ggcorrplot::ggcorrplot(round(cor(PolishBR_clean_logs[,-c(1,4)]), 2), lab = T, type = "lower", show.legend = FALSE, lab_size = 2, tl.cex = 8,
                                 title = "Transformed Variables")
grid.arrange(cor_1, cor_2, nrow = 1, ncol = 2)
# Proportion of about 60%
the_same <- PolishBR_clean %>%
  mutate(Diff = OpExp - CostPrS,
         is_the_same = case_when(
           abs(Diff) < 1 ~ "almost identical",
           TRUE ~ "Different"
         ))
the_same %>% count(is_the_same) %>% summarise(prop = n/sum(n))
for_cor <- the_same %>% filter(is_the_same == "almost identical")
# Correlation of 0.9999998
cor(for_cor$OpExp, for_cor$CostPrS)
table_CRatio <- PolishBR %>%
  na.omit() %>%
  mutate(CRatio = CAssets/StLiabil,
         Ratio_Below_1 = ifelse(CRatio < 1, "yes", "no")) %>%
  group_by(Ratio_Below_1) %>%
  count(bust) %>%
  summarise(Bust = ifelse(bust == 1, "yes", "no"),
            Percent_bust = round((n / sum(n) * 100), 2))
knitr::kable(table_CRatio, caption = "Current ratio and bust")
plot_bust <- PolishBR %>%
  na.omit() %>%
  mutate(Equity_Ratio = (TAssets - TLiabil) / TAssets)  %>%
  filter(between(Equity_Ratio,0,1)) %>%
  ggplot(aes(bust, Equity_Ratio)) +
  geom_boxplot(fill = "cyan") +
  geom_hline(yintercept = 0.5, col = "red", lty = 2) +
  theme_bw() +
  xlab("Bust") + ylab("Shareholder Equity Ratio")
plot_WkCap <- PolishBR %>%
  na.omit() %>%
  mutate(Equity_Ratio = (TAssets - TLiabil) / TAssets)  %>%
  filter(between(Equity_Ratio,0,1)) %>%
  ggplot(aes(WkCap, Equity_Ratio)) +
  geom_boxplot(fill = "blue") +
  geom_hline(yintercept = 0.5, col = "red", lty = 2) +
  theme_bw() +
  xlab("Working Capital") + theme(axis.title.y = element_blank(),
                                   axis.ticks.y = element_blank(),
                                axis.text.y = element_blank())
grid.arrange(plot_bust, plot_WkCap, ncol = 2, nrow = 1)
# 70/246 (28.46%) companies with negative Equity ratio went bankrupt
PolishBR %>%
  na.omit() %>%
  mutate(Equity_Ratio = (TAssets - TLiabil) / TAssets)  %>%
  filter(Equity_Ratio < 0) %>%
  count(bust) %>%
  summarise(Percent = n/sum(n) * 100)
library(Stat2Data)
# There were always the same number of observations in c.tab - using clustered observations "bubbles" wasn't helpful - that's why I deleted appropriate parts
myemplogit <- function(yvar = y,xvar = x, maxbins = 10, line=TRUE, ...){
  breaks  <<- unique(quantile(xvar, probs = 0:maxbins/maxbins))
  levs  <<- (cut(xvar, breaks, include.lowest = FALSE))
  num <<- as.numeric(levs)
  emplogitplot1(yvar ~ xvar, breaks = breaks, showline = line, ...)}
par(mfrow=c(1,3))
myemplogit(PolishBR_clean$bust,
           PolishBR_clean_logs$OpExp,
           50, xlab="Operating expenses (log)", ylim = c(-4.5, -1.4))
myemplogit(PolishBR_clean$bust,
           PolishBR_clean_logs$TSales,
           50, xlab="Total sales (log)", ylim = c(-4.5, -1.4))
myemplogit(PolishBR_clean$bust,
           PolishBR_clean_logs$CostPrS,
           50, xlab="Cost of products sold (log)", ylim = c(-4.5, -1.4))
#            Preparation              #
# Remove negative TLiabil value
PolishBR_clean<-PolishBR %>% na.omit()
PolishBR_clean_ready <- PolishBR %>% na.omit() %>% filter(TLiabil > 0)
# Create dummy variables for WkCap (with Low as a base category)
PolishBR_clean_ready<-dummy_cols(PolishBR_clean_ready, select_columns = "WkCap",remove_first_dummy = TRUE) %>%select(-WkCap)#
# Make sure the proportions of bust are the same in all sets
Data_Transformed_inter <- PolishBR_clean_ready %>%
  mutate(BookVal = TAssets - TLiabil,
         CRatio = CAssets/StLiabil,
         ROCE = (TSales-CostPrS-OpExp)/(TAssets-StLiabil),
         ERatio = BookVal/TAssets,
         TAssets = log(TAssets),
         TLiabil = log(TLiabil),
         TSales = log(TSales),
         StLiabil = log(StLiabil),
         CAssets = log(CAssets),
         OpExp = log(OpExp),
          CAMed = WkCap_Medium*CAssets,
          CAHigh = WkCap_High*CAssets,
          CAVHigh = `WkCap_Very High`*CAssets,
          CALow = CAssets - CAMed - CAHigh - CAVHigh,
          CostPrS = log(CostPrS)) %>%
  select(-c(CLiabil,BookVal,CAssets))
set.seed(3456)
split <- createDataPartition(Data_Transformed_inter$bust, p = 0.7, list = FALSE)
train <- Data_Transformed_inter[split, ]
validate_and_test <- Data_Transformed_inter[-split, ]
validate_and_test_split <- createDataPartition(validate_and_test$bust, p = 0.5, list = FALSE)
validate <- validate_and_test[validate_and_test_split, ]
test <- validate_and_test[-validate_and_test_split, ]
## full model
model_3 <- glm(bust ~ ., family = binomial, data = train)
summary(model_3)
# Reduce the model using backwards deviance methodology – manual approach:
Anova(model_3)
model_3_2 <- update(model_3, ~ . - TLiabil)
Anova(model_3_2)
model_3_3 <- update(model_3_2, ~ . - CAVHigh)
Anova(model_3_3)
model_3_4 <- update(model_3_3, ~ . - CAHigh)
Anova(model_3_4)
model_3_5 <- update(model_3_4, ~ . - ERatio)
Anova(model_3_5)
model_3_6 <- update(model_3_5, ~ . - CostPrS)
Anova(model_3_6)
model_3_7 <- update(model_3_6, ~ . - ROCE)
Anova(model_3_7)
model_3_8 <- update(model_3_7, ~ . - WkCap_Medium)
Anova(model_3_8)
summary(model_3_8)

model_3_backward <- step(model_3, direction = "backward",test = "Chi",k=3.841)
summary(model_3_backward)
###### AIC
model_3_backward_AIC <- step(model_3, direction = "backward",test = "F",k=2)
summary(model_3_backward_AIC)
## LASSO MODEL
model_3_lasso <- glmnet(train[, -1], train[, 1], family = "binomial",alpha=1)
plot(model_3_lasso,'lambda',cex.lab=2)
cv_model_3<-list()
lambda<-rep(0,100)
for (i in 1:100) {
  cv_model_3[[i]] <- cv.glmnet(as.matrix(train[, -1]), train[, 1], family = "binomial", alpha = 1)
  lambda[i]<-cv_model_3[[i]]$lambda.1se
}
log(mean(lambda))##around -4.92 -4.86
coef(model_3_lasso,s=exp(-4.86))
### Choose the threshold
## back
pred_3_b_validate <- predict(model_3_backward_AIC, validate[-1], type = "response")
predictions <- prediction(pred_3_b_validate, validate[1])
sens <- data.frame(x=unlist(performance(predictions, "sens")@x.values),
                   y=unlist(performance(predictions, "sens")@y.values))
spec <- data.frame(x=unlist(performance(predictions, "spec")@x.values),
                   y=unlist(performance(predictions, "spec")@y.values))
par(mai=c(1,1,1,.5))
plot(sens,type='l',col='red',ylab='Probability',xlab='Cutoff',cex.lab=2)
par(new=TRUE)
plot(spec,type='l',col='blue',ylab='Probability',xlab='Cutoff',cex.lab=2)
legend(x=par('usr')[2]+xinch(0.5),y=par('usr')[3]+yinch(2),
       legend = c('Senstivity','Specificity'),col=c('red','blue'),
       lty=1,xpd=TRUE,xjust=1,yjust = 0,ncol = 2,bty='n',cex = 1.5)
optimal <- sens[which.min(apply(sens, 1, function(x) min(colSums(abs(t(spec) - x))))), 1]
optimal
pred_3_b_test <- predict(model_3_backward, test[-1], type = "response")
pred_3_b_final <- as.data.frame(pred_3_b_test) %>%
  mutate(Prediction = ifelse(pred_3_b_test > optimal, 1, 0)) %>%
  select(Prediction)
cMatrix <- confusionMatrix(table(pred_3_b_final$Prediction, test$bust),positive='1')
cMatrix
Transformed_Backward <- round(c(cMatrix$overall[c("Accuracy", "Kappa")],cMatrix$byClass[c("Sensitivity", "Specificity")]), 3)
roc(test$bust ~ pred_2_b_test)$auc
## lasso
pred_3_lasso_validate <- predict(model_3_lasso, s = exp(-4.86), newx = as.matrix(validate[-1]))
predictionsl <- prediction(pred_3_lasso_validate, validate[1])
sensl <- data.frame(x=unlist(performance(predictionsl, "sens")@x.values),
                    y=unlist(performance(predictionsl, "sens")@y.values))
specl <- data.frame(x=unlist(performance(predictionsl, "spec")@x.values),
                    y=unlist(performance(predictionsl, "spec")@y.values))
par(mai=c(1,1,1,.5))
plot(sensl,type='l',col='red',ylab='Probability',xlab='Cutoff',cex.lab=2)
par(new=TRUE)
plot(specl,type='l',col='blue',ylab='Probability',xlab='Cutoff',cex.lab=2)
legend(x=par('usr')[2]+xinch(0.5),y=par('usr')[3]+yinch(2),
       legend = c('senstivity','Specificity'),col=c('red','blue'),
       lty=1,xpd=TRUE,xjust=1,yjust = 0,ncol = 2,bty='n',cex=1.5)
```

```

optimall <- sensl[which.min(apply(sensl, 1, function(x) min(colSums(abs(t(specl) - x))))), 1]
pred_3_lasso_test <- predict(model_3_lasso, s = exp(-4.86), newx = as.matrix(test[-1]))
pred_3_lasso_final <- as.data.frame(pred_3_lasso_test) %>%
  mutate(Prediction = ifelse(pred_3_lasso_test > optimall, 1, 0)) %>%
  select(Prediction)
cMatrixl <- confusionMatrix(table(pred_3_lasso_final$Prediction, test$bust),positive='1')
cMatrixl
Transformed_inter_LASSO <- round(c(cMatrixl$overall[c("Accuracy", "Kappa")],
                             cMatrixl$byClass[c("Sensitivity", "Specificity")]), 3)
#          Prediction and diag          #
##prediction
predtest <- Data_Transformed_inter[1,]
predtest[1,] <-as.numeric(c(0,log(80),log(30), log(150), log(23), log(130), log(70),0,1,0,50/23,0,0,0,55,0,0))
# optimal 0.06080186
prob <- predict(model_3_backward_AIC, predtest[-1], type = "response")
##plot the addedvariabelplot and dfbeta plot
dfbetaPlots(model_3_backward_AIC,layout = c(2,5),cex.lab=1.5)
avPlots(model_2_backward_AIC,cex.lab=1.7,layout=c(2,5),cex.axis = 1.5,ylim=c(-4,8))
```