Assignment

Uni no. [redacted]

Contents

1 Abstract **3**

2 Exploratory data analysis **4**

2.1 Missing data 4

2.2 Distribution of bust, the response variable 4

2.3 Explanatory variables with skewed distributions 4

2.4 Bankrupt vs Not Bankrupt 5

2.5 Linearity assumption 6

2.6 Strategy for training and validation split 7

3 Model screening **7**

3.1 Initial model 7

3.1.1 Transformations 7

3.1.2 Additional predictors and interaction 8

3.1.3 Initial model summary 8

3.1.4 Multicollinearity 9

3.2 Model Reduction 9

3.2.1 Stepwise regression 9

3.2.2 Shrinkage through LASSO regression 10

4 Model assessment **11**

4.1 ROC chart 11

4.2 Performance metrics 11

4.3 Comparison of predictors 12

5 Final Model **13**

5.1 Model diagnostics 13

6 Limitations of the dataset and model **15**

7 Bibliography **16**

8 Appendix **17**

1 Abstract

The purpose of this report is to analyse, explain and predict the financial position of Polish Companies with regards to their susceptibility to bankruptcy. This is based on data collected from the first decade of the 21st century and the early 2010s.

We started with an initial exploratory data analysis identifying missing data, skewed distributions of predictors and the statistical differences between companies that survived bankruptcy and those which did not. We then fitted three different logistic regression models – one with all predictors, a stepwise regression model and a LASSO shrinkage model. These models were then evaluated through ROC charts and various performance metrics. We eventually settled on a subset of the available predictors for a final model.

While model assessment revealed positive signs including the predictors contributing effectively to the explanatory power of models, we also discovered issues like multicollinearity, high leverage from unusual observations and high predictive accuracies conflicting with models biased towards predicting companies remaining solvent.

This report emphasises the importance of financial data in understanding the vulnerability of companies to bankruptcy while also highlighting the need for non-financial predictors and for a dataset that is less class imbalanced.

2 Exploratory data analysis

2.1 Missing data

We observe that the **CostPrS** and **CLiabil** columns have the most missing values - 272 and 282 respectively. Other columns have at most 15 missing values each. Out of the 5509 observations, 291 contain missing values, which is around 5.3% of the dataset. This is not necessarily detrimental since "Missing data under 10% for an individual case or observation can generally be ignored" (Hair 2019).

Hence, we drop the missing values, leaving 5218 observations.

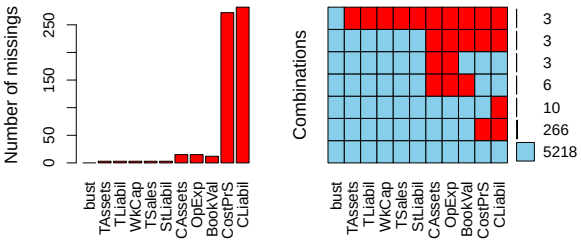


Figure 1: Aggregation plot of missing values

2.2 Distribution of bust, the response variable

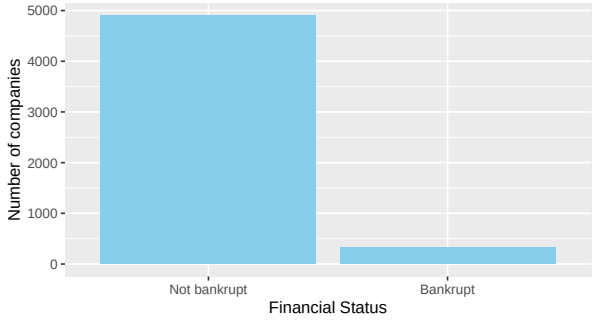


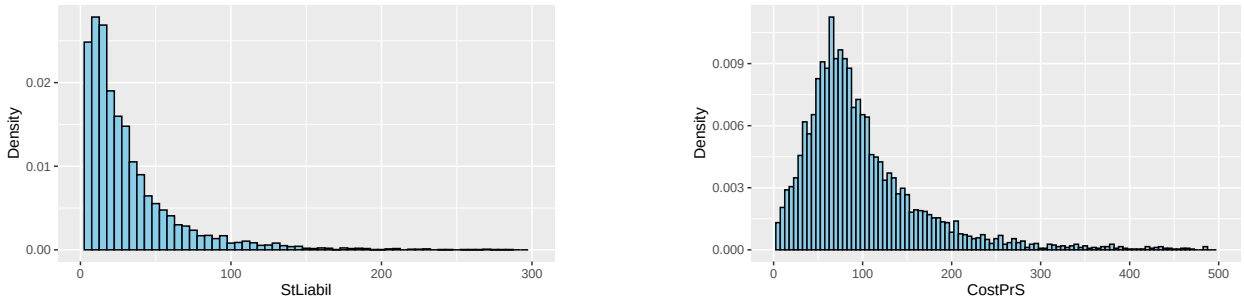
Figure 2: Barplot of factor column "bust"

There are exceedingly more companies that remained solvent after a year in the dataset compared to those which went bankrupt - 4899 observations for the former and 319 for the latter. This might pose a problem when training our model as it may be biased towards predicting "Not bankrupt".

This is something we will need to address when splitting our dataset into training and validation sets later on.

2.3 Explanatory variables with skewed distributions

7 out of the 9 real explanatory variables display a positive skew. These include **TAssets**, **TSales**, **StLiabil**, **CAssets**, **OpExp**, **CostPrS** and **CLiabil**. This is clear from the histograms in Figure 3 below. This is because each of these variables have many abnormally large values acting as outliers. This is expected given the varying sizes of companies and the difference between the financial positions of small and large companies is quite substantial.



(a) Histogram of StLiabil

(b) Histogram of CostPrS

Figure 3: Histogram of explanatory variables with skewed distributions

2.4 Bankrupt vs Not Bankrupt

We have observed the positive skew in the distributions of the variables discussed above. It would be helful to determine whether there are any changes in the distributions depending on whether a company survived bankruptcy or not. We do this using conditional density plots as shown below.

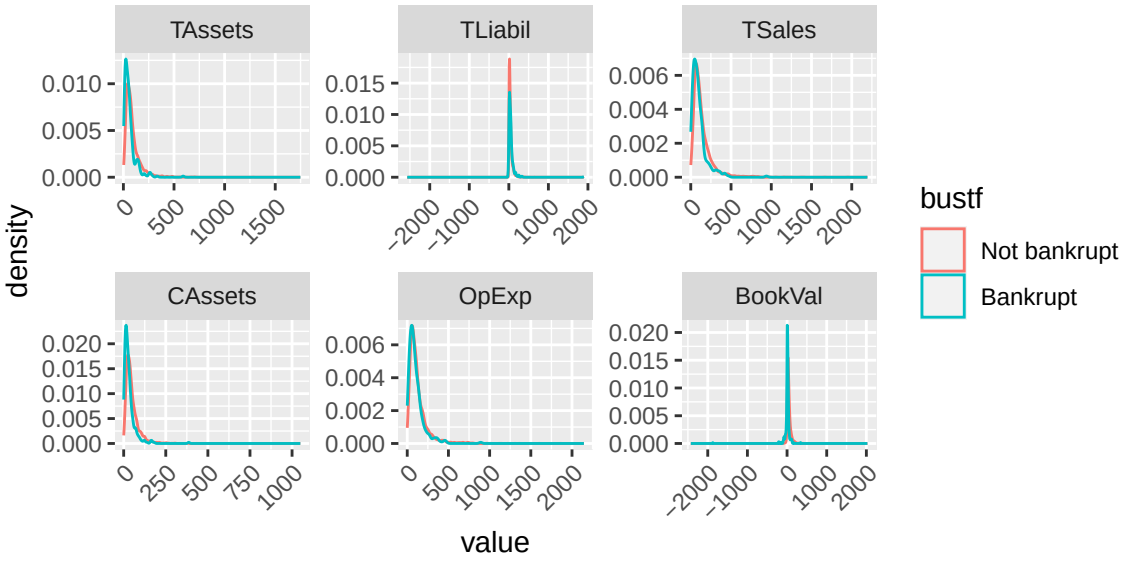


Figure 4: Conditional density plots of selected explanatory variables

From the selection of variables with either skewed distributions or fairly normal distributions - **TLiabil** and **BookVal** - we observe that there are no visible differences in the conditional densities whether companies were bankrupt or not. It would be naive though to use this as a sign of the variables having low explanatory or predictive power with regards to the response variable, bust.

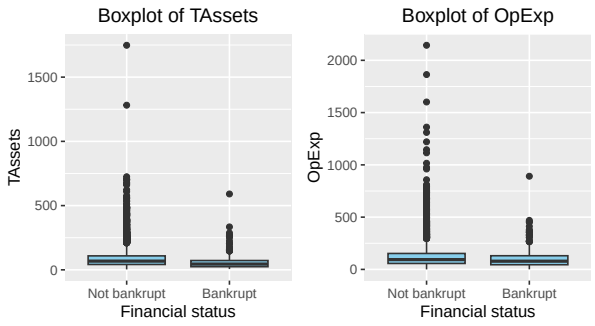


Figure 5: "bankrupt" vs "not bankrupt" through outliers

In fact, there is a discernible pattern displayed by the outliers particularly in the case of variables mentioned in **Section 2.3**.

We first notice that our deduction that there is no consequential differences in the conditional distributions of the variables with regards to bankruptcy is substantiated by the boxplots in **Figure 5**.

Then, we observe that in the case of **TAssets** and **OpExp**, there are far more outliers for companies that did not go bankrupt. This is understandable given that larger companies with greater financial resources are less susceptible to bankruptcy hence indicating the power of the variables in explaining and predicting bankruptcy.

2.5 Linearity assumption

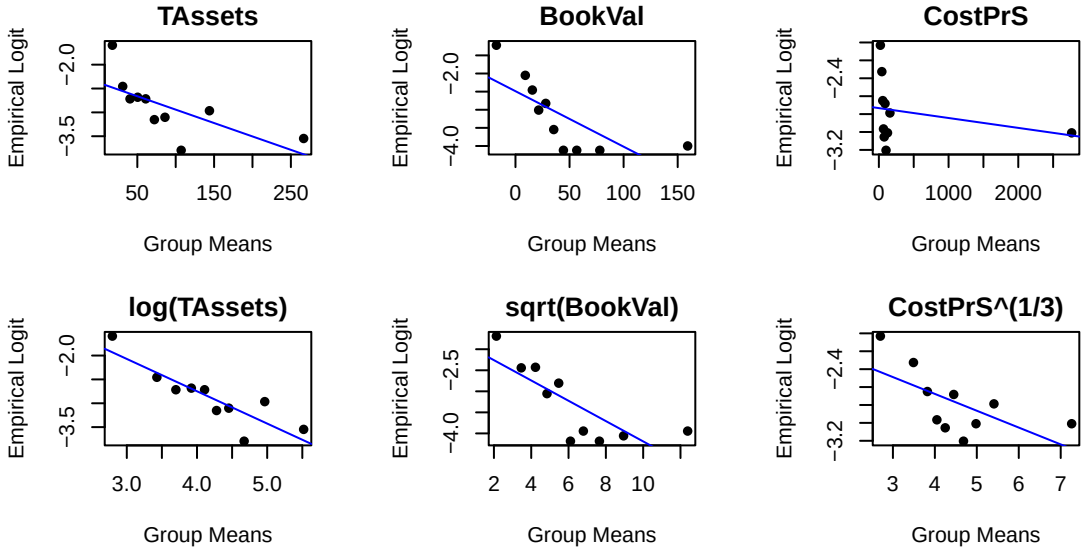


Figure 6: Empirical Logit Plots and Effects of Transformations

In **Figure 6**, we use the empirical logit plot to determine to what extent the explanatory variables satisfy the linearity assumption and whether we can detect clear relationships between the response, bust, and the variables.

In the first row of plots, we observe that the observed logit values are not evenly spread around the line of best fit, indicating that the linearity assumption is not porperly satisfied. This is expected as a result of the skewness of distributions as we previously noted.

To ensure these explanatory variables contribute positively to our model, we need to address this issue. So, in the second row of plots, we investigate several potential transformations which appear to have the desired effect of strengthening the linearity assumption. Clearly, the observed logit values display a more linear pattern. This is particularly evident for $\sqrt[3]{\textbf{CostPrS}}$.

2.6 Strategy for training and validation split

We opt for a 70/30 training and validation split. Since there are 4899 observations for "not bankrupt" and 319 for "bankrupt", this seems to be appropriate enough to avoid having a model biased towards "not bankrupt". We also aim to achieve similar proportions for the two categories across both the training and validation sets - 6% "bust" and 94% "no bust" - resulting in the following split:

	Training	**Validation**
"Bankrupt" observations	223	96
"Not Bankrupt" observations	3430	1469

Table 1: Training and Validation split

3 Model screening

3.1 Initial model

We start with a model that includes all available explanatory variables - **bustf** ~ **TAssets** + **TLiabil** + **WkCap** + **TSales** + **StLiabil** + **CAssets** + **OpExp** + **BookVal** + **CostPrS** + **CLiabil**.

From the summary of the model, four explanatory variables - **TAssets**, **CAssets**, **CostPrS** and **CLiabil**- are not statistically significant. Strikingly, all four of these variables have skewed distributions as identified in **Section 2.3**, which might be the reason behind their statistical insignificance. We therefore resort to transformations to reduce the skewness, as explored in **Figure 6**, and investigate whether the model improves.

3.1.1 Transformations

To decide on the most appropriate transformations, we adapt the Box-Cox method which assesses the effectiveness of power and logarithmic transformations through log-likelihood maximisation. This is illustrated in **Figure 7** where the log-likelihood function is plotted and the 95% confidence interval for the optimal λ values is highlighted.

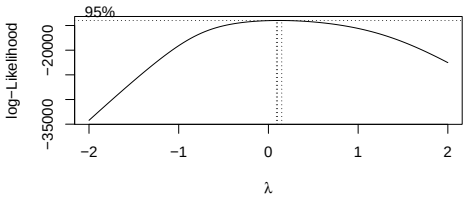


Figure 7: Box-Cox method applied to CAssets

Thus, for **CAssets**, $\lambda = 0$, corresponding to a natural logarithmic transformation. The latter proves to be the best transformation for all the variables with skewed distributions identified in **Section 2.3**.

The updated model is then **bustf** ~ **log(TAssets) + TLiabil + log(TSales) + WkCap + log(StLiabil) + log(CAssets) + log(OpExp) + BookVal + log(CostPrS) + log(CLiabil)**. Consequently, only 2 explanatory variables - **log(CAssets)** and **TLiabil** - are now statistically insignificant judging from the model summary.

3.1.2 Additional predictors and interaction

We consider the following possible alterations to the model to seek an improvement in its explanatory and predictive power:

1. **Debt to Equity Ratio:** This is equal to $\frac{TLiabil}{BookVal}$. This is a financial metric used to gauge a company's financial leverage and risk profile and hence would be a good predictor for bankruptcy. In addition, it might make up for the statistical insignificance of **TLiabil**.
2. **Interaction between log(TSales) and log(CostPrS)** This is because **TSales** and **CostPrS** are involved in the calculation of gross profit. Aside from the natural relationship between profitability and solvency, this would also act as an indicator of the cash flow position and market competitiveness of a company.

3.1.3 Initial model summary

The formula for the initial model is **bust ~ log(TAssets) + log(TSales) + log(CostPrS) + WkCap + log(StLiabil) + log(OpExp) + DebtToEquity + log(CAssets) + log(CLiabil) + log(TSales):log(CostPrS)**.

Three predictor variables are not statistically significant based on their p-values. These are **DebtToEquity**, **log(CAssets)** and **log(TSales):log(CostPrS)**.

The residual deviance of the model is 1237.8 which is an improvement from the model at the start of **Section 3.1** which had a residual deviance of 1312.2.

We could further reduce the deviance and ensure we only include predictors for their contribution to model performance by aiming for a more parsimonious model. This will be explored further in **Section 3.2**.

3.1.4 Multicollinearity

Variable	VIF
log(TSales)	17.12
log(CostPrS)	31.24
log(StLiabil)	40.20
log(OpExp)	34.09
log(CLiabil)	37.36
log(TSales):log(CostPrS)	32.83

Table 2: Variables with high VIFs

It is clear from **Table 2** that multicollinearity is a serious issue with the initial model given that 6 variables have high VIFs, much greater than the threshold of 10. Part of this can be explained through underlying connections between certain variables, for instance, **CLiabil** and **StLiabil** which are quite similar in financial terms while **log(TSales)** and **log(CostPrS)** are potentially impacted by the introduction of the interaction effect. This is yet another argument for reducing the model as proposed in **Section 3.1.3**.

3.2 Model Reduction

3.2.1 Stepwise regression

Having implemented a bidirectional stepwise regression procedure, the resulting model is **bustf ~ log(TAssets) + log(TSales) + log(CostPrS) + WkCap + log(StLiabil) + log(OpExp) + log(CAssets) + log(CLiabil)**.

This is achieved through an AIC criterion, illustrated in **Figure 8**. As the number of variables is reduced, AIC falls indicating progress towards a better model fit.

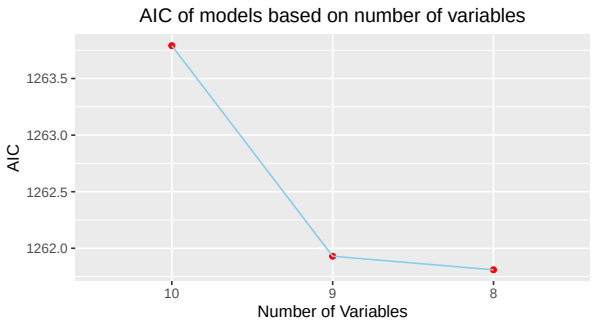


Figure 8: Stepwise regression tracked via AIC

Interestingly, out of all the variables with high VIFs, only the interaction term is excluded through this method which might be a sign multicollinearity still exists. Furthermore, excluding the number of variables, this model is very similar to the initial model in terms of residual deviance, coefficient estimates and p-values. In short, this process merely excluded the variables from the initial model with high p-values, namely **DebtToEquity** and **log(TSales):log(CostPrS)**.

3.2.2 Shrinkage through LASSO regression

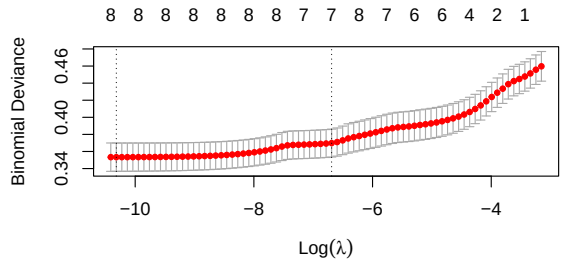


Figure 9: LASSO regularisation path

LASSO shrinkage helps in achieving a parsimonious model while also addressing multicollinearity. It shrinks the coefficients of correlated predictors towards zero, stabilising coefficient estimates and reducing the sensitivity of the model to small changes in the data. It also faciliates variable selection as it may set the coefficient of certain variables to zero.

Since we implemented a cross-validated LASSO regression, **Figure 9** shows the binomial deviance of the model for different λ values, where λ is a regularisation parameter which allows the LASSO to find a balance between model simplicity and predictive accuracy. The optimal λ value is calculated to be 0.00124.

Table 3 lists the coefficients of the predictors having implemented a LASSO model with $\lambda = 0.00124$.

Variable	**Coefficient**
log(TAssets)	-0.774
log(TSales)	-1.760
log(StLiabil)	-
log(CAssets)	-0.886
log(OpExp)	2.650
log(CostPrS)	-0.882
log(CLiabil)	1.117
DebtToEquity	0.001

Table 3: LASSO coefficients

The coefficient of **log(StLiabil)** is set to zero, which is possibly as a result of its contribution to multicollinearity as explored in **Section 3.1.4**. Similarly, the coefficient of **DebtToEquity** is also very low which echoes its exclusion from the stepwise regression model.

log(TAssets), **log(CAssets)** and **log(CostPrS)** also have relatively lower coefficients. A high VIF may be the reason for the latter variable while **log(CAssets)** has been associated with high p-values in previous models we fitted hinting towards its potentially limited contribution to model fit.

Following these observations, we drop **log(StLiabil)**, **DebtToEquity**, **log(CAssets)** and **log(CostPrS)** settling on the model **bustf** ~ **log(TAssets)** + **log(TSales)** + **WkCap** + **log(OpExp)** + **log(CLiabil)**. We include **WkCap** since it has proven to be statistically significant in both previous models we examined.

4 Model assessment

4.1 ROC chart

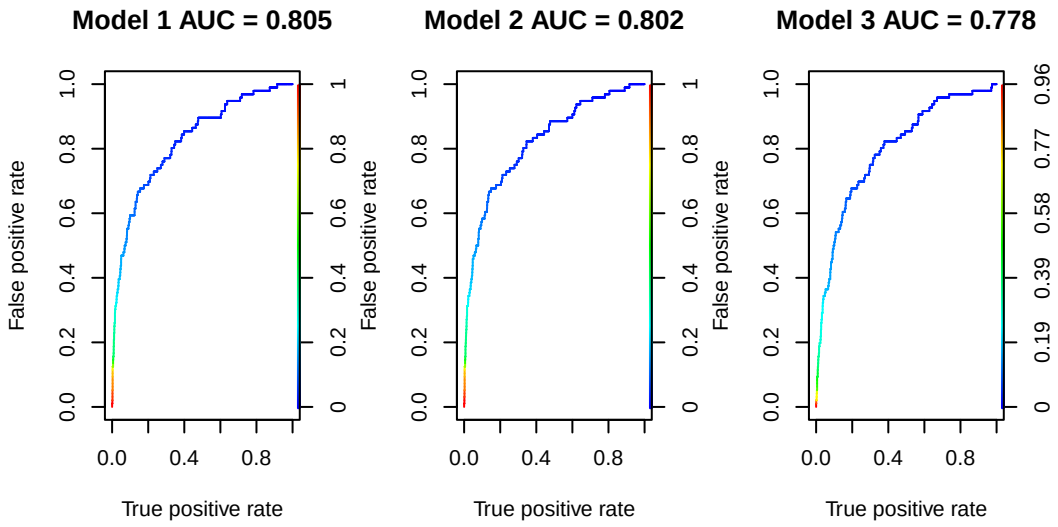


Figure 10: ROC Charts for three models

We aim to assess the performance of the three models in **Sections 3.1.3**, **3.2.1** and **3.2.2** through ROC charts displayed in **Figure 10**. The closer the curve is to the top-left corner, the better the model performance.

It is noteworthy that we used the validation set for this analysis and any further model assessment to expose the models to previously unseen data.

Since the shapes of the curves for the three models are very similar, we resort to the Area Under Curve (AUC) as a measure of model performance. We would ideally desire an AUC above 0.9 although 0.8 is also a desirable threshold, which is met by the first two models. Nevertheless, our previous analysis revealed multicollinearity in these two models. This explains the lower AUC of the third model since we excluded some variables for a less multicorrelated model, hence sacrificing some predictive accuracy.

4.2 Performance metrics

All three models have high accuracy but perform poorly in terms of precision and f1 score.

One striking observation is that the first two models have equal metrics. Since model 2 only differs from model 1 through the exclusion of **DebtToEquity** and the interaction effect between **log(TSales)** and **log(CostPrS)**, this tells us that these two predictors did not contribute to model

fit. Moreover, since **DebtToEquity** is calculated from **TLiabil** and **BookVal**, this shows these variables have a poor relationship with bankruptcy.

	Model 1	**Model 2**	**Model 3**
Accuracy	0.942	0.942	0.938
Precision	0.636	0.636	0.467
F1 Score	0.237	0.237	0.126

Table 4: Performance metrics

In addition, we cannot entirely rely on the metrics due to the class imbalance in the dataset, with solvency data far outnumbering bankruptcy. Thus, the models may be biased towards predicting solvency hence the high accuracy while the poor F1 scores are a result of many false negatives - 89 out of 96 banruptcy observations.

4.3 Comparison of predictors

Variable	**Model 1**	**Model 2**	**Model 3**
log(TAssets)	✓	✓	✓
log(TSales)	✓	✓	✓
log(CostPrS)	✓	✓	-
WkCap	✓	✓	✓
log(StLiabil)	✓	✓	-
log(OpExp)	✓	✓	✓
DebtToEquity	✓	-	-
log(CAssets)	✓	✓	-
log(CLiabil)	✓	✓	✓
log(TSales):log(CostPrS)	✓	-	-

Table 5: Model breakdown

One definitive conclusion is that **DebtToEquity** and the interaction term do not contribute to model fit given their exclusion following both stepwise and LASSO regression.

log(CostPrS), **log(StLiabil)** and **log(CAssets)** are the three variables which stepwise and LASSO regression treat differently. In the case of **log(StLiabil)**, it would be better to side with the result of LASSO because of its high VIF and it is the only variable with a coefficient of zero. As for **log(CAssets)**, it is correlated to both **WkCap** - due to the its involvement in the calculation of the latter - and **log(CLiabil)**. Furthermore, **log(CAssets)** is not statistically significant in model 2. Finally, we opt to include **log(CostPrS)** in the final model due to its importance in net profit calculation alongside **log(TSales)** and **log(OpExp)**. This would allow the model to reflect to some extent the impact of profitability on solvency.

Hence, we settle on the final model: **bustf** $\sim$ **log(TAssets)** + **log(TSales)** + **WkCap** + **log(OpExp)** + **log(CLiabil)** + **log(CostPrS)**.

5 Final Model

Variable	Coefficient Estimate
Intercept	0.3730
log(TAssets)	-1.0498
log(TSales)	-3.2473
log(CostPrS)	-0.5312
WkCapMedium	-1.0671
WkCapHigh	-1.1371
WkCapVery High	-1.5126
log(CLiabil)	0.6641
log(OpExp)	3.7426

Table 6: Coefficient estimates for final model

Table 6 shows the coefficient estimates for the predictors in the final model which explain their relationship with the response.

For instance, all other things remaining unchanged, an increase in the total assets of a company by a factor of exp(1) decreases the log odds of a company being bankrupt to it surviving by 1.0498. Likewise, ceteris paribus, having a medium working capital, high working capital or very high working capital reduces the log odds of a company being bankrupt to it surviving by 1.0671, 1.1371 or 1.5126 respectively in contrast to a company with a low working capital.

The formula for the final model is thus: $\eta_{\mathbf{i}} = \mathbf{0.3730} - \mathbf{1.0498}\log(\mathbf{TAssets_i}) - \mathbf{3.2473}\log(\mathbf{TSales_i}) - \mathbf{1.0671WkCapMedium_i} - \mathbf{1.1371WkCapHigh_i} - \mathbf{1.5126WkCapVeryHigh_i} + \mathbf{3.7426}\log(\mathbf{OpExp_i}) + \mathbf{0.6641}\log(\mathbf{CLiabil_i}) - \mathbf{0.5312}\log(\mathbf{CostPrS_i})$.

Hence, for a company A with a high working capital, TAssets = 80, TSales = 150, OpExp = 130, CostPrS = 70, CLiabil = 12, $\eta_A = -4.025$. Therefore $\mathbb{P}(A \textit{ goes bankrupt}) = \frac{\exp(-4.025)}{1+\exp(-4.025)} = 0.0175$, which is less than 0.5, so we deduce that A remained solvent after one year.

5.1 Model diagnostics

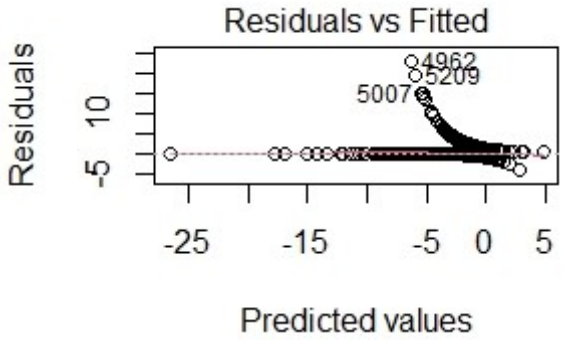


Figure 11: Residuals vs Fitted plot

We observe that the smoother in **Figure 11** is fairly horizontal and most of the points are scattered around zero. Some discrepancy is observed for larger values with three observations higlighted but this is expected given the presence of outliers which is not completely erased by transformations. We explore this issue further by considering the Cook's distance in **Figure 12**.

It is clear that most observations have Cook's distance values lower than the threshold of $\frac{4}{n} = 0.001$. There are however some observations with Cook's distance higher than that rule of thumb as indicated in the plot.

Further investigations detailed in **Figure 13** show that we also identify irregular observations with leverage values higher than the threshold of 0.005. These discrepancies in Cook's distance and leverage suggest that these unusual observations might be influencing the model negatively. In fact, Pearson residuals plot shows observations with high positive values greater than 2 which indicates that the model might be predicting "not bankrupt" when the actual observation is "bankrupt" - something we elaborated in previous sections.

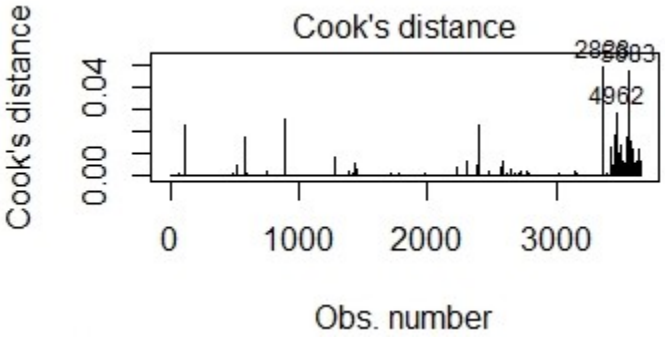


Figure 12: Cook's Distance Plot

However, it is noteworthy that despite these issues, the residual and influence analysis for the first three models we considered revealed a more pronounced leverage problem.

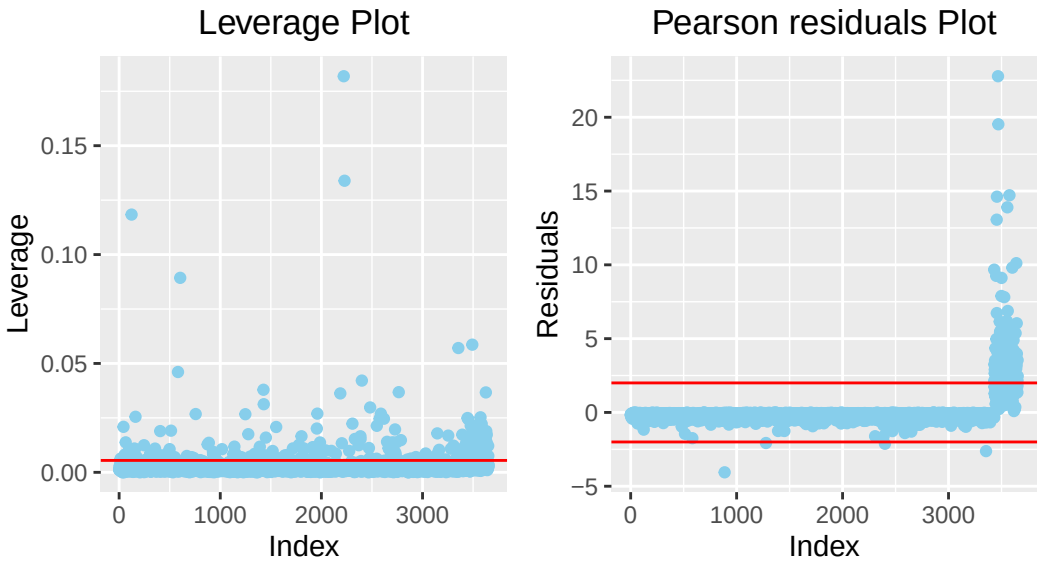


Figure 13: Final Model leverage and Pearson residuals

It would be incredibly time-consuming and complex to identify and remove the causal observations especially given the class imbalance and outlier prevalence in the dataset. So we leave this issue untouched.

6 Limitations of the dataset and model

The most significant limitation of the dataset is its class imbalance. Observations of solvent companies comprise 94% of the dataset. This means that any model trained on the dataset is biased towards predicting "not bankrupt". This problem is somehow veiled by the high predictive accuracy of models but the leverage and residual analysis reveals these flaws. Without time to properly analyse and filter the data for outliers and imbalance, any model will inevitably face these problems. Had there been more time, the dataset could have been scrutinised for big corporations with unusual financial positions acting as outliers, we could have collected data about more bankrupt companies and could have gleaned non-financial predictors like industry the companies operate in or whether the companies have a conglomerate structure since many financial predictors in the dataset are correlated. In terms of evaluating model fit, we could have explored robust regression which specifically deals with data contaminated with outliers and influential observations and we could have performed more in-depth validation and regularisation diagnostics.

Word Count: $\sim$ 2700 words

7 Bibliography

Hair, Black J.F. 2019. *Multivariate Data Analysis (8th Edition).*

8 Appendix

```
# Required packages
library(reshape2)
library(ggplot2)
library(GGally)
library(gridExtra)
library(VIM)
library(dplyr)
library(sm)
library(Stat2Data)
library(MASS)
library(olsrr)
library(glmnet)
library(ROCR)

# Summary of data
load("PolishBR.Rdata")
summary(PolishBR)

# Aggregation plot (Figure 1)
aggr(PolishBR, prop = FALSE, combined = FALSE, numbers = TRUE, sortVars = FALSE,
↪  sortCombs = TRUE)

# Drop missing values
PolishBR <- na.omit(PolishBR)
# Creating a copy of bust but as a factor
PolishBR$bustf <- factor(PolishBR$bust,labels=c("Not bankrupt","Bankrupt"))
attach(PolishBR)

# Barplot of bustf (Figure 2)
ggplot(PolishBR, aes(x = bustf)) + geom_bar(fill = "skyblue") + labs(x =
↪  "Financial Status", y = "Number of companies")

# Plotting histogram for real explanatory variables (code repeated with x
↪  changed)
ggplot(PolishBR, aes(x = StLiabil)) + geom_histogram(aes(y=..density..),binwidth
↪  = 5, fill = "skyblue", color = "black") + labs(x = "StLiabil", y = "Density")
↪  + xlim(0, 300)
```

```r
# Plotting conditional plots for a selection of real explanatory variables
PolishBR.long <- melt(PolishBR[,c(2:3,5,7,8,9,12)], id="bustf")
ggplot(aes(x=value, group=bustf, col=bustf), data=PolishBR.long) + geom_density()
↪  + facet_wrap(~ variable, scales="free") + theme(axis.text.x =
↪  element_text(angle = 45, hjust = 1))

# Conditional boxplot of TAssets
ggplot(PolishBR, aes(x = bustf, y = TAssets, fill = bustf)) +
↪  geom_boxplot(fill="skyblue") + labs(x = "Financial status",y =
↪  "TAssets",title = "Boxplot of TAssets") + theme(plot.title =
↪  element_text(hjust = 0.5))

# Empirical logit plot for BookVal (code repeated for other plots by changing the
↪  predictor variable)
emplogitplot1(bust ~ BookVal, ngroups = 10, xlab="Group Means", ylab="Empirical
↪  Logit", main="BookVal")

# Splitting dataset into train and validation sets
df <- na.omit(PolishBR)
## Reset index
rownames(df) <- NULL
## Getting row indices of "bust" and "no bust" observations
bust_row <- as.numeric(row.names(df[df$bust == 1,]))
no_bust_row <- as.numeric(row.names(df[df$bust == 0,]))
## Set seed for reproducibility
set.seed(123)
## Randomly sampling observations for the training set
bust_index <- sample(bust_row, 223)
no_bust_index <- sample(no_bust_row, 3430)
train_index <- c(no_bust_index, bust_index)
## Train set
df_train <- df[train_index,]
## Validation set
df_val <- df[-train_index,]

# Fitting model with all explanatory variables untransformed and its summary
model1 <- glm(bustf ~ TAssets + TLiabil + WkCap + TSales + StLiabil + CAssets +
↪  OpExp + BookVal + CostPrS + CLiabil, data = df_train, family = binomial)
```

```
summary(model1)

# Box Cox method for CAssets
boxcox(lm(CAssets ~ 1))

# Initial model with transformation variables and its summary
model1 <- glm(bustf ~ log(TAssets) + TLiabil + WkCap + log(TSales) +
↪  log(StLiabil) + log(CAssets) + log(OpExp) + BookVal + log(CostPrS) +
↪  log(CLiabil), data = df_train, family = binomial)
summary(model1)

detach(PolishBR)
attach(df_train)
# Adding Debt to Equity ratio column to train set
df_train$DebtToEquity <- TLiabil / BookVal
# Initial model with transformed variables, addded predictor and interaction and
↪  its summary
model1 <- glm(bustf ~ log(TAssets) + log(TSales) + log(CostPrS) + WkCap +
↪  log(StLiabil) + log(OpExp) + DebtToEquity + log(CAssets) + log(CLiabil) +
↪  log(TSales):log(CostPrS), data=df_train, family=binomial)
summary(model1)

# Calculating VIF for model1 using olsrr package
lm1 <- lm(bustf ~ log(TAssets) + log(TSales) + log(CostPrS) + WkCap +
↪  log(StLiabil) + log(OpExp) + DebtToEquity + log(CAssets) + log(CLiabil) +
↪  log(TSales):log(CostPrS), data=df_train)
ols_coll_diag(lm1)

# Using stepwise regression
step(model1, direction = "both")

# Plot of AIC for stepwise regression
data <- data.frame(num_variables = c(10, 9, 8),
                   AIC = c(1263.79, 1261.93, 1261.81))
data$num_variables <- factor(data$num_variables, levels = c(10, 9, 8))
ggplot(data, aes(x = num_variables, y = AIC, group=1)) +
  geom_point(color="red") +
  geom_line(color="skyblue") +
  labs(x = "Number of Variables", y = "AIC") +
```

```
  ggtitle("AIC of models based on number of variables") + theme(plot.title =
  ↪  element_text(hjust = 0.5))

# Fitting model derived from stepwise regression
model2 <- glm(bustf ~ log(TAssets) + log(TSales) + log(CostPrS) +
    WkCap + log(StLiabil) + log(OpExp) + log(CAssets) + log(CLiabil),
    family = binomial, data = df_train)

# Performing LASSO regrssion
## Creating a copy of df_train and transforming relevant variables
new_df <- df_train
log_index <- c(2,5:8,10:11)
for(i in log_index){
  new_df[, i] <- log(new_df[, i])
}
## Creating target vector and matrix of explanatory variables
y <- new_df$bust
X <- as.matrix(new_df[,c(2,5:8,10:11,13)])
## Performing cross-validated LASSO regression to determine optimal lambda and
↪  plotting results
set.seed(123)
modCV <- cv.glmnet(y = y, x = X, family = "binomial", alpha = 1, nfolds = 10)
plot(modCV)
## Fitting LASSO model with optimal lambda and printing coefficients
model3 <- glmnet(y = y, x = X, family = "binomial", alpha = 1, lambda =
↪  modCV$lambda.1se, maxit = 1000000, intercept=FALSE)
coef(model3)

# Fitting model derived from LASSO regression
model3 <- glm(bustf ~ log(TAssets) + log(TSales) + WkCap + log(OpExp) +
↪  log(CLiabil), data=df_train, family = binomial)

# Producing ROC chart
## Adding DebtToEquity ratio column to validation set
df_val$DebtToEquity <- df_val$TLiabil / df_val$BookVal
## Creating a list of the three models
models <- list(model1, model2, model3)
## Plotting ROC charts
par(mfrow=c(1,3))
```

```r
for(i in 1:3){
  predictions <- predict(models[[i]], newdata=df_val, type="response")
  pred <- prediction(predictions, df_val$bustf)
  perf <- performance(pred,"fpr", "tpr")
  auc_value <- round(mean(perf@y.values[[1]]),3)
  plot(perf,colorize=TRUE, main=paste("Model", i, "AUC =", auc_value))
}

# Performance metrics for models
## Creating vectors for storing metrics
accuracy <- c()
precision <- c()
f1_score <- c()
FN <- c()
## Calculating performance metrics
for (i in 1:3){
  # Accuracy
  pred <- predict(models[[i]], newdata = df_val, type = "response")
  pred <- ifelse(pred > 0.5, 1, 0)
  accuracy[i] <- mean(pred == df_val$bust)
  # Precision
  TP <- sum(pred == 1 & df_val$bust == 1)
  FP <- sum(pred == 1 & df_val$bust == 0)
  precision[i] <- TP / (TP + FP)
  # F1-Score
  FN <- sum(pred == 0 & df_val$bust == 1)
  FN[i] <- FN
  recall <- TP / (TP + FN)
  f1_score[i] <- 2 * (precision[i] * recall) / (precision[i] + recall)
}

# Fitting final model
model4 <- glm(bustf ~ log(TAssets) + log(TSales) + WkCap + log(OpExp) +
↪  log(CLiabil) + log(CostPrS), family = binomial, data = df_train)

# Predicting bankruptcy or solvency for a given company
pred_data <- data.frame(TAssets=c(80),TSales=c(150),OpExp=c(130),CostPrS=c(70),
CLiabil=c(12),WkCap=c("High"),TLiabil=c(30),StLiabil=c(23),CAssets=c(50),
BookVal=c(55))
```

```r
pred_data$WkCap <- as.factor(pred_data$WkCap)
pred <- predict(model4, newdata = pred_data, type = "response")

# Residual vs Fitted plot and Cook's distance plot for final model
plot(model4, 1)
plot(model4, 4)

# Calculating and plotting leverage values of final model
hat_values <- hatvalues(model4)
data <- data.frame(Idx = 1:length(hat_values), Hat_Values = hat_values)
leverage <- ggplot(data, aes(x = Idx, y = Hat_Values)) +
  geom_point(color="skyblue") +
  geom_hline(yintercept = 20/3653, color = "red") +
  labs(title = "Leverage Plot", x = "Index", y = "Leverage") + theme(plot.title =
  ↪  element_text(hjust = 0.5))

# Calculating and plotting Pearson residuals for final model
resid <- residuals(model4, type = "pearson")
resid_data <- data.frame(Idx = 1:length(resid), Residuals = resid)
residuals <- ggplot(resid_data, aes(x = Idx, y = Residuals)) +
  geom_point(color="skyblue") +
  geom_hline(yintercept = 2, color = "red") +
  geom_hline(yintercept = -2, color="red") +
  labs(title = "Pearson residuals Plot", x = "Index", y = "Residuals") +
  ↪  theme(plot.title = element_text(hjust = 0.5))
```