Sant'Anna
Scuola Universitaria Superiore Pisa

Statistical Learning and Large Data
Exam - Module Two

Quality over Quantity

Analysing Models and Features Predicting Bankruptcy

Davide Bacigalupi and Pietro Pianini

May 2024

Abstract

This work explores methods for predicting a firm's risk of bankruptcy through the analysis of a comprehensive financial dataset. With a focus on feature selection and logistic regression, our study addresses the challenges posed by imbalanced datasets and high dimensionality. By employing undersampling and oversampling techniques, we ensure a balanced representation of instances, thereby enhancing the model's predictive performance. Our investigation identifies the ISIS+LASSO model, applied to an undersampled dataset, as the most effective feature selection method. From an initial pool of 94 features, we pinpoint four key indicators (Net Income to Total Assets, Current Liability to Assets, Equity to Long Term Liability, and Cash to Total Assets) that significantly influence bankruptcy risk assessment. While our findings offer valuable insights, further research is required to validate their generalizability across different contexts.

Contents

1 Introduction

Bankruptcy or business failure can significantly impact both individual enterprises and the global economy. Business professionals, investors, governments, and academic researchers have long explored ways to identify potential risks of business failure to mitigate economic losses caused by bankruptcy (Balleisen (2001))

Predicting bankruptcy is crucial for many financial institutions, as they need to foresee the likelihood of a firm going bankrupt to make informed lending decisions. Various techniques have been used in the literature to develop bankruptcy prediction models, including both statistical methods and machine learning approaches (Balcaen and Ooghe (2006); Kumar and Ravi (2007); Lin et al. (2011)). Machine learning techniques, in particular, have shown better performance compared to statistical methods.

While numerous studies have focused on creating new machine learning or statistical methods to improve prediction accuracy, few have examined the impact of input variables (or features).

Our objective is to study which variables have the greatest impact on a firm's risk of bankruptcy, starting with a dataset containing numerous financial explanatory variables and a target variable representing bankruptcy.
We use features screening and features selection algorithms, along with logistic regression (which is the predictive model that provides the most information on the explanatory variables) to achieve this goal.
More specifically, we aim to answer three questions:

1. Regarding methods: we seek to effectively address the issue of a dataset with many highly correlated features and significant imbalance (with only 3% of instances being bankrupt);

2. Regarding models: we want to determine the best feature selection model. By the best feature selection model, we mean the one that has the best performance in terms of the area under the Precision-Recall curve;

3. Regarding features: we aim to select only the most significant features for predicting bankruptcy risk and study their impact.

Our work is organized as follows: the next section describes the dataset and the data pre-

processing, necessary to address the issue of highly correlated features mentioned in question 1. Section 3 primarily addresses question 2 by analyzing the various feature selection models we decided to use. Additionally, in subsection 3.2, it tackles the other issue mentioned in question 1: the imbalanced dataset problem. Section 4 addresses question 3. Finally, Section 5 concludes the paper. An appendix and the bibliography follow for interested readers.

2 Dataset: Description and Pre-processing

2.1 Dataset Description

This work utilizes data sourced from the Taiwan Economic Journal spanning the years 1999 to 2009 (UCI (2020)). Two specific criteria guided the collection of data samples:

- The selected companies were required to possess a minimum of three years of comprehensive public information preceding the financial crisis event.
- There needed to be an adequate number of comparable companies within the same industry and of similar size, facilitating a comparison between bankrupt and non-bankrupt cases.

Consequently, the sample encompassed companies primarily from the manufacturing sector, consisting of industrial and electronics firms (346 companies), the service sector encompassing shipping, tourism, and retail enterprises (39 companies), and miscellaneous industries (93 companies), excluding financial institutions. The total number of instances is 6819.
Company bankruptcy is the target binary variable. The definition of company bankruptcy adhered to the business regulations outlined by the Taiwan Stock Exchange.
The explanatory variables are 94 financial features. Financial ratios (FRs) are widely recognized as critical factors in bankruptcy prediction models (Altman (1968); Beaver (1966); Ohlson (1980)). These ratios can be divided into seven categories: solvency (28 features in the dataset), profitability (18 features), cash flow ratios (5 features), capital structure ratios (9 features), turnover ratios (13 features), growth (8 features), and others (13 features).
It is important to recognize that a substantial disparity in the number of bankrupt versus non-bankrupt cases creates a class imbalance issue. This imbalance can adversely affect the accuracy and effectiveness of the final prediction performance.

2.2 Dataset Manipulation

There are many highly correlated features in the dataset, as can be seen from the corrplot below. The names of several of these variables are often very similar to each other (like, for example, "Current Liabilities Liability" and "Current Liability to Liability"). This results in a clear multicollinearity problem and violates the irrepresentability condition of methods like Ridge and LASSO.

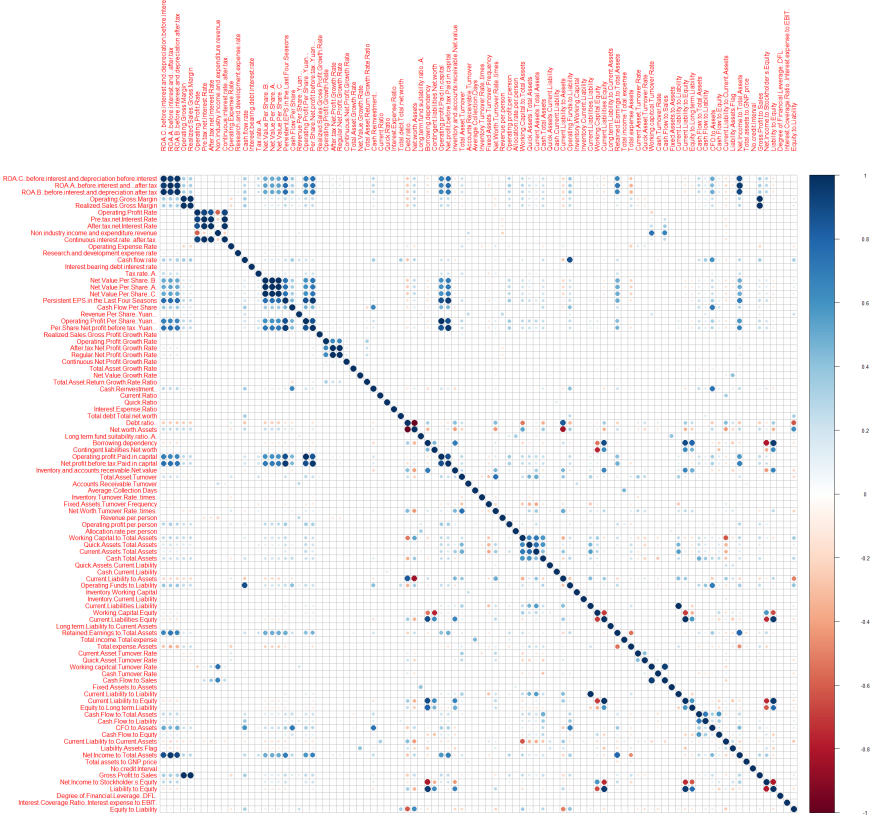

Figure 1: Corrplot before the data preprocessing

To address this issue, we group the variables using a dendrogram. We then cut the dendrogram at height 80, thereby creating 60 clusters. Most of the clusters with more than one variable are made of variables which share the same economic meaning and the same denomination, thus for each cluster we choose one variable at random as representative and eliminate the others, obtaining in the end 60 variables. These are listed in a table in section 6.6.

The corrplot obtained with the remaining variables displays correlations between features that are much weaker and challenging to eliminate using only data pre-processing techniques.

After the preprocessing, we scale the dataset and we divide it into a training set and a test set (70% - 30%). The division is performed using a stratified splitting technique in order to

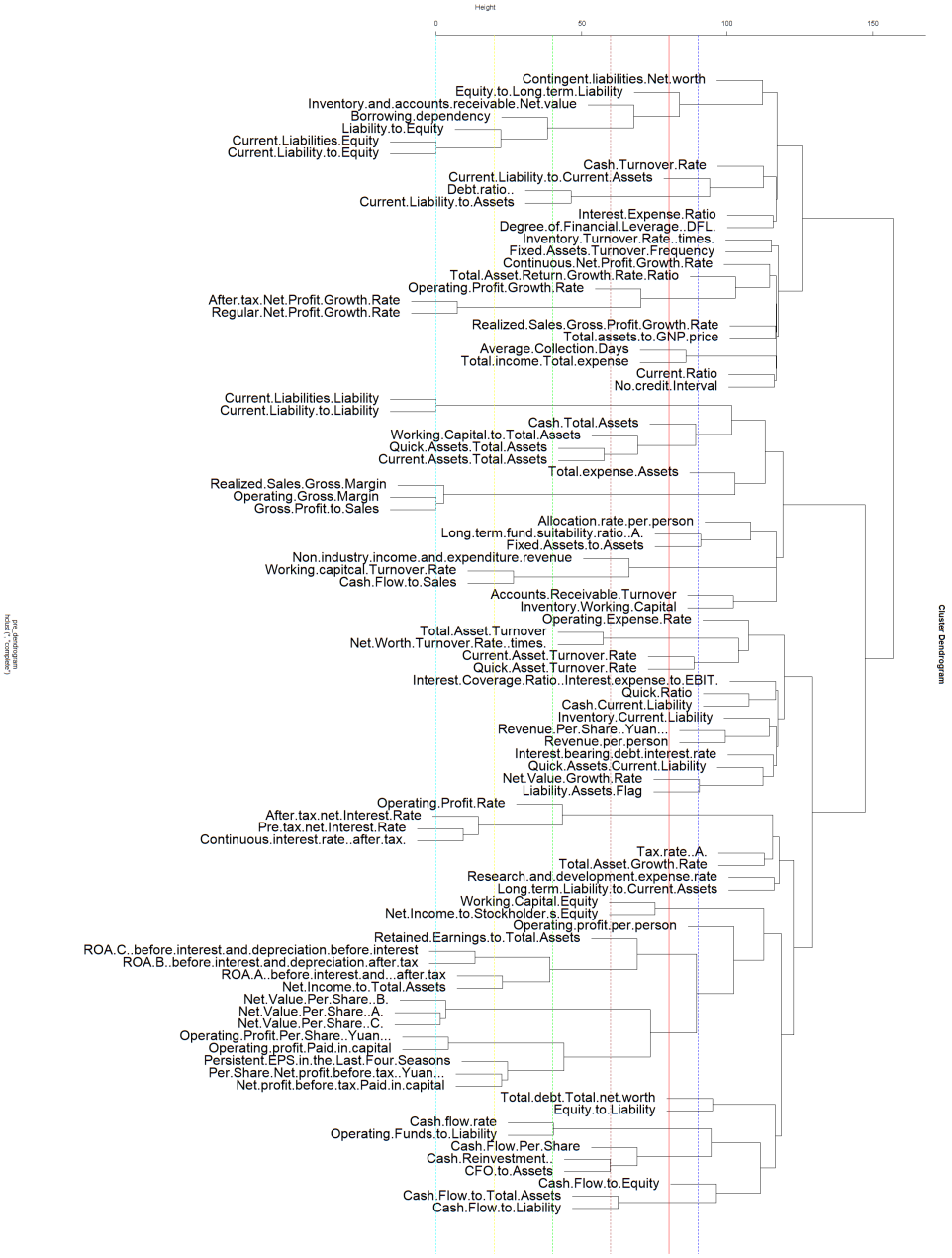

Figure 2: Dendrogram of variables

preserve the same proportion of the two classes in both subsets. This is needed since the dataset is imbalanced and there is a risk of neglecting the less abundant class in the splitting process if done simply at random.

We have thus that the proportion of bankrupt firms in the training set is 0.03226482 and on the test set is 0.03225806, that are almost equal values (while on the original dataset it was 0.0322628).

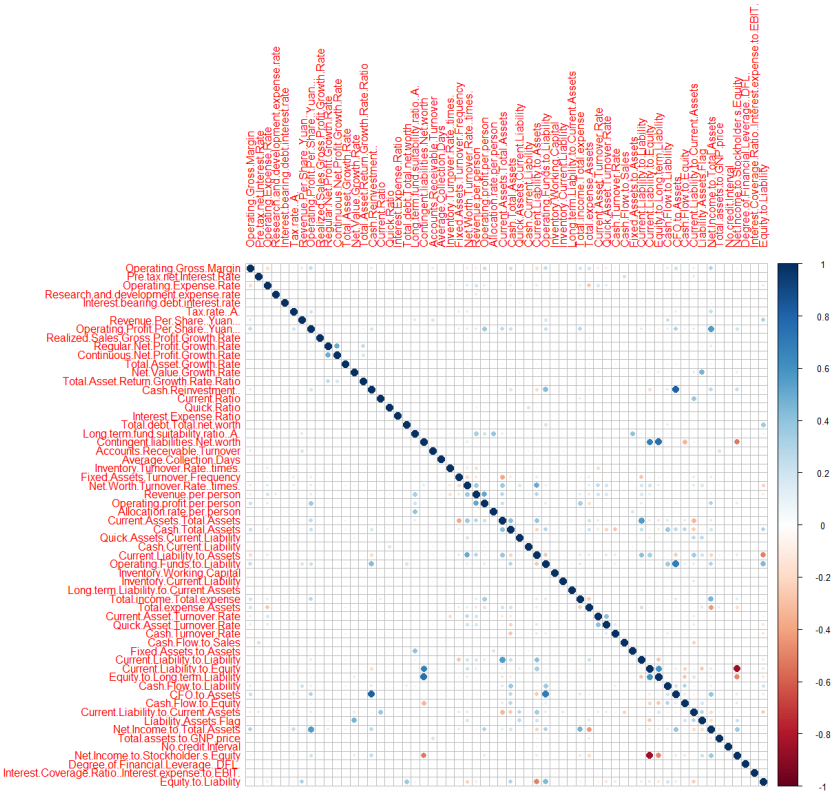

Figure 3: Corrplot after the data preprocessing

3 Models Performances Analysis

3.1 Models on Imbalanced Dataset

In order to solve our classification problem, we perform a series of different models on the training set, evaluating then their performances on the test set.

For each model, since the dataset is severely imbalanced, we use as measure of performance the Precision-Recall (PR) curve, and also the Area under the Precision Recall Curve (AUPRC) both on the training and on the test set is indicated for each model (Saito and Rehmsmeier (2015)).

The Precision-Recall curve is the curve of the values of precision and recall for every value of the threshold used to choose the predicted class of the observation.

This curve is particularly indicated for our problem also since it concentrates itself only on the "bankrupt" class. We are interested in predicting bankruptcy of firms, and since the quantity of nonbankrupt firms in our dataset is much higher than the one of bankrupt ones, using other measures of performance like Accuracy and the ROC curve tends to select models based almost only on the performance on the majority class (which for us is of no interest at all).

We use then also the Area-under-the-PR-curve since we want an overall measure of performance which is independent from the threshold used to compute the precision and recall.

To begin to tackle the problem, we perform a baseline logistic regression with all 60 features. This model is not only huge and difficult to be intepreted, but also displays significant overfitting since the Area under the Precision-Recall curve on training set is 0.4454847, while on test set AUPRC = 0.3351596 (an appalling result).
Thus, selection of features is probably needed in order to have results not affected by overfitting.

3.1.1 LASSO and Elastic Net

To select the relevant features, we firstly perform the LASSO regression and an Elastic Net regression with $\alpha = 0.5$ (James et al. (2023)).
The Elastic Net can help us only if we have residual multicollinearity issues, while if we have no issue of this kind it is only a problem since it biases the results towards zero.
We tune the parameter λ of both approaches with (10-fold) cross-validation (using the binomial deviance as tuning parameter). In another run, we use as tuning parameter directly the AUPRC, but this does not affect sensibly our results, thus it is not displayed here (see the Appendix 6.5 for further details).

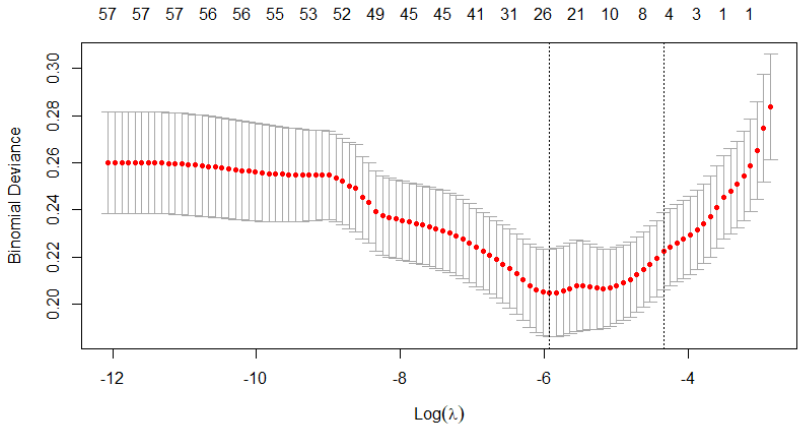

Figure 4: CV fit - LASSO

As we can see in the figures, the λ chosen for both LASSO and Elastic Net as the one which minimizes the tuning parameter involves the selection of 26 features. This is probably too much, since we can see that the tuning parameter does not change very much in a large neighbourhood of the minimum measure.

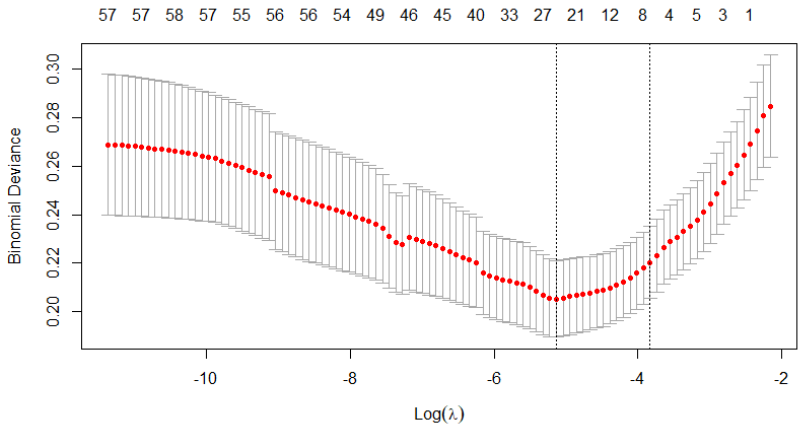

Figure 5: CV fit - elastic net

We decide then to use, instead of the minimum λ, the λ at 1-standard error distance from the minimum, thus selecting only 7 features out of 60.

The features selected here are the same for both LASSO and Elastic Net and are: Current Liability to Assets, Fixed Assets to Assets, Equity to Long-term Liability, Current Liability to Current Assets, Liability-Assets Flag, Net Income to Total Assets, Net Income to Stockholder's Equity.

We can see below also the plot of the dynamic of the coefficients of each feature by varying λ, and we notice that they are almost the same for both LASSO and Elastic Net.

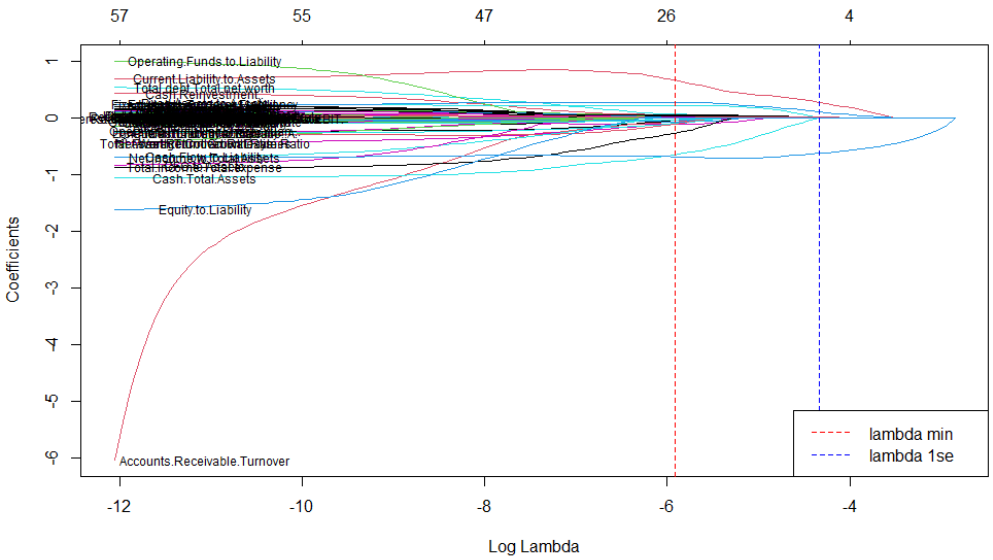

Figure 6: Lambda Plot - LASSO

Afterwards, we performed a logistic regression using only the features selected with the LASSO in order to eliminate the bias resulting from the direct use of this model and we calculate the

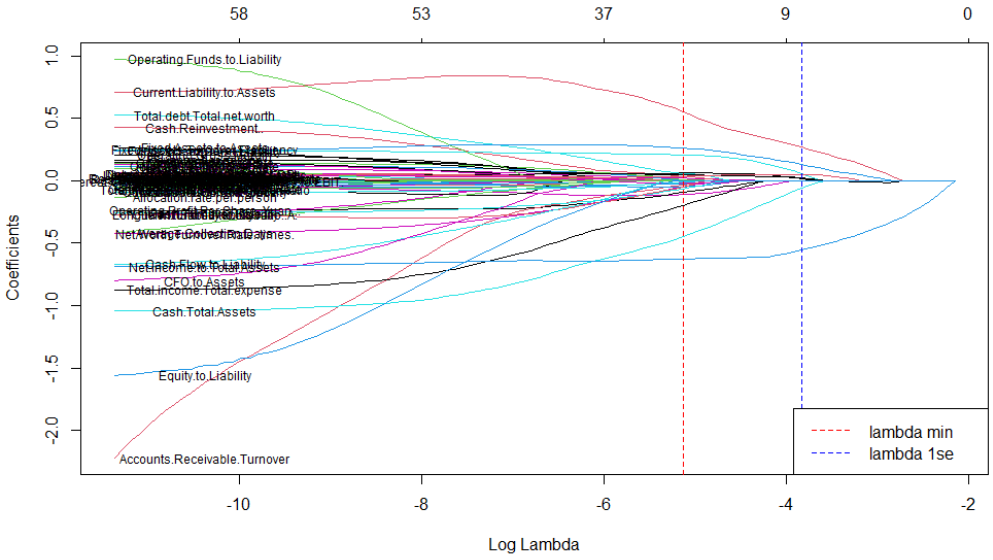

Figure 7: Lambda Plot - Elastic Net

performance of this classification method both on the training and on the test set.

The Elastic Net approach reports performances worse than the logistic after LASSO in both training set (AUPRC of Elastic Net = 0.3413642, AUPRC of LASSO = 0.3526428) and test set (here AUPRC of Elastic Net = 0.3061739, AUPRC of LASSO = 0.3179982). Thus, we can safely conclude that there is no residual multicollinearity issue, and discard the Elastic Net results as essentially useless.

In any case, the results obtained are still not good, even compared with the baseline logistic performed before.

3.1.2 Feature Selection and Feature Screening

After the LASSO, we try to do feature selection also using more traditional methods (James et al. (2023)). We perform the stepwise regression algorithm since the best-subset selection is too computationally expensive, and we tune it via (5-fold) cross-validation using as tuning parameter the AUPRC on the training set.

However, since the features are probably still too much for the algorithm to perform well, and also in order to save computational time, we try to apply to the training set some screening algorithms before the stepwise selection one. In particular, we use the Iterative Sure Independence Screening (ISIS) algorithm (Fan and Lv (2008)).

After the ISIS, it is common in the literature also to apply other selection algorithms to further restrict the field, like the SCAD and the LASSO already used before.

Thus, we perform only ISIS, ISIS+SCAD and ISIS+LASSO on the training set and select the relevant features, then using the features selected by each of these results we perform a preliminary logistic regression and afterwards the stepwise selection algorithm to further reduce the set of features selected. We fit also a stepwise selection on all the features without screening, but the features selected are much more and the performance is remarkably worse with respect to the models performed after the screening algorithms.

The performances of the models compared to the LASSO and the elastic net (and the baseline logistic) in the test set are shown in the figure below.

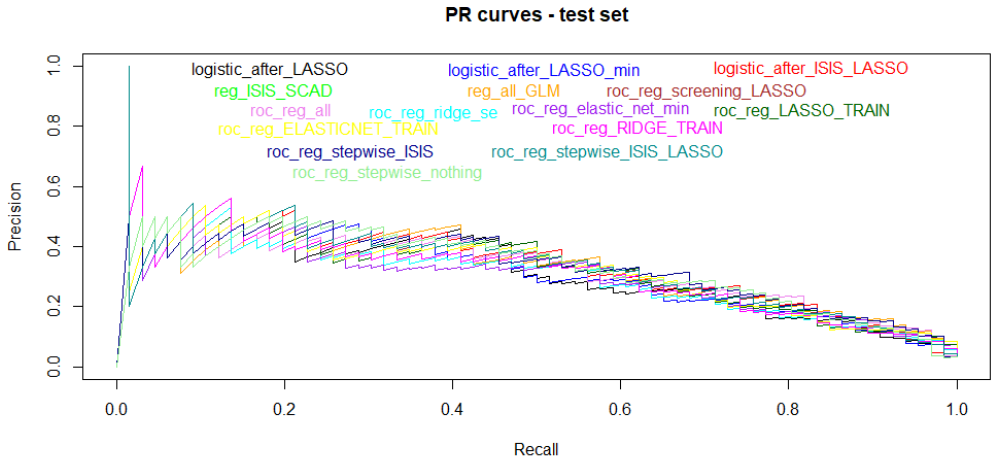

Figure 8: PR curves on test set - all models fitted on original training set

We cannot really see it from the figure, but the best model from the ones performed on the original data set is the stepwise selection after ISIS (AUPRC = 0.3450373 in the test set, while in training set it is 0.4202862), thus proving the usefulness of this selection technique. The features selected by this model are still 7, but different from the ones chosen by the LASSO (we will see this later in section 4).

The best model on the training set is instead still the baseline logistic, but as said before it displays a more significant overfitting and does worse in the test set than the one chosen here (and with much more features used...), thus we can safely discard it.

In any case, the results shown here are particularly low for every model, and we do not get rid of overfitting, even in the best model. We suspect this is because the dataset is imbalanced and the models trained on it fail to give enough importance to the minority class (i.e., bankrupt firms).

3.2 Solving the Imbalanced Dataset Problem

Since the performance on the test set of the models chosen by using the original dataset is particularly low, we use some techniques of oversampling and undersampling on the training set in order to bring balance to the dataset and train models on them. The hope is that, with these new balanced datasets, the models trained can predict much better the (interesting) minority class and thus gain in performance on the test set.

We use as undersampling technique the simple random undersampling with replacement of the more abundant class (i.e., the nonbankrupt firms) and then joining the set obtained from this subsampling with all bankrupt firms, to form a perfectly balanced undersample (Prusa et al. (2015)). We then repeated the sampling to form 19 balanced subsamples on which we trained all the models already used before on the original training set.

Each subsample consists of 308 observations (154 bankrupt firms and 154 nonbankrupt).

As oversampling technique instead we use the SMOTE method (Chawla et al. (2002)), and in this case we create a completely new synthetic dataset by artificially increasing the number of observations of the less abundant class (i.e., bankrupt firms) in order to obtain an (almost) balanced dataset.

We managed to test the models only on one oversample (mainly for shortage of computational time) of 8777 observations, in which the proportion of bankrupt firms is 0.4737382 (thus, very similar to a perfectly balanced dataset).

3.2.1 Undersampling Results

We show here the results of the undersampling approach.

In the graph below, we can see the Area under the Precision-Recall curve for each model performed on each of the 19 subsamples, both for the (original) training set and for the test set. It is clear from the graph that some models, like the baseline logistic and the logistic after the screening, perform very badly both on training and on test set.

Instead, some models like the Logistic after LASSO tuned using cross-validation and the stepwise selection after ISIS seems from the graph to perform particularly well both on training and on test set.

We then calculated the mean AUPRC for each model on every subsample and the maximum AUPRC for each model, on both the training and the test set.

In the table below we show the results for mean AUPRC and maximum AUPRC on the test set and we find that the best model is the logistic after screening and LASSO CV (with mean

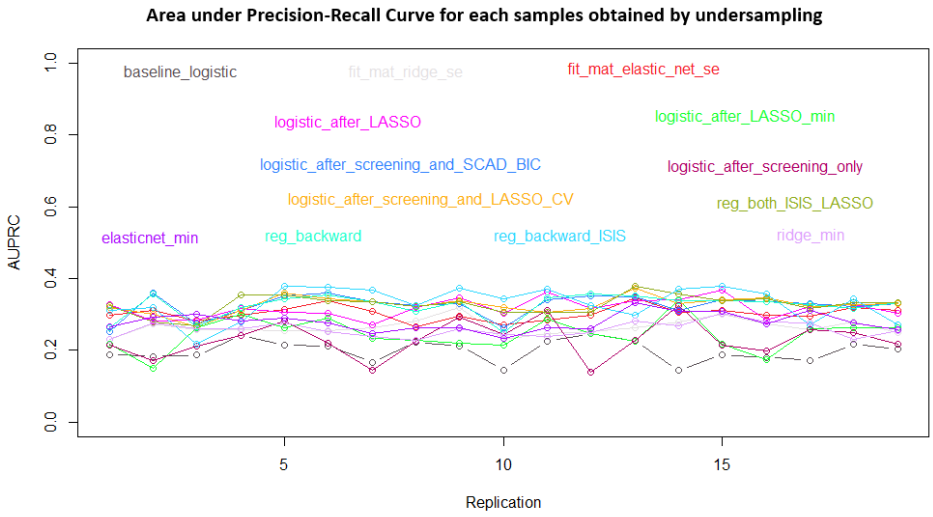

Figure 9: AUPRC of undersampling models on the training set

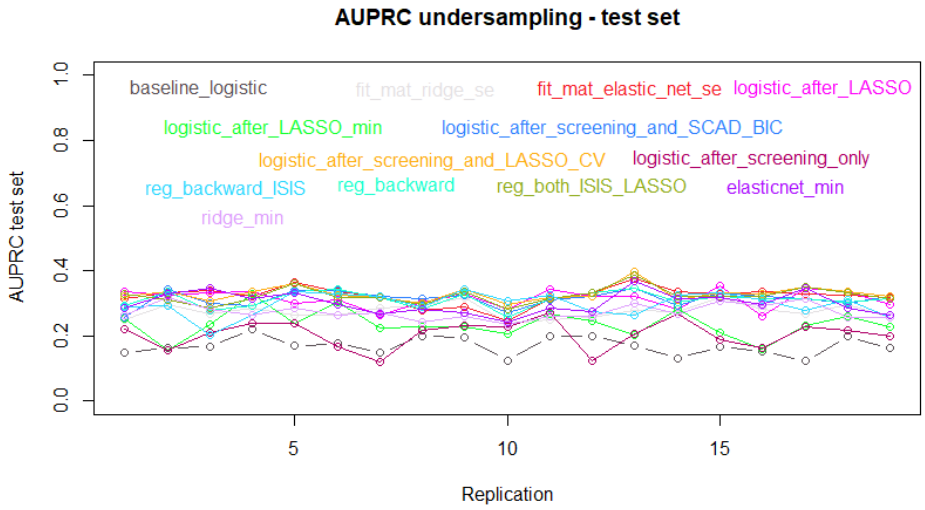

Figure 10: AUPRC of undersampling models on the test set

AUPRC of 0.329 and max AUPRC of 0.396).

Thus we can choose as our best model given by the undersampling technique the logistic after screening and LASSO tuned via CV replication that gives us the max AUPRC on the test set (0.39604953), which is the best replication also on the training set (AUPRC on training set = 0.3745879).

The results are not so good overall, but surely better than the ones obtained by using the original dataset (there max AUPRC on test set was 0.3450373) and without overfitting.

Mean AUPRC	Max AUPRC	Model performed
0.16949885	0.22119095	baseline logistic
0.27413786	0.31903822	ridge λ-min
0.30118422	0.36797344	elastic net λ-min
0.27587816	0.31970551	ridge λ-1se
0.32286157	0.37539557	elastic net λ-1se
0.31506921	0.35317675	logistic after LASSO λ-1se
0.23670158	0.32938533	logistic after LASSO λ-min
0.31380733	0.3454652	logistic after screening and SCAD BIC
0.32979597	0.39604953	logistic after screening and LASSO CV
0.20489695	0.27090144	logistic after screening only
0.31223534	0.3484965	stepwise selection after nothing
0.32336864	0.3854025	stepwise selection after ISIS + LASSO
0.29765063	0.34368236	stepwise selection after only ISIS

Table 1: Mean and Max AUPRC on test set for each model

3.2.2 Oversampling Results

Afterwards, we perform the same models on the oversampled dataset, hoping that this will give us some interesting results.

We then plot the Precision-Recall curves for each model on both the (original) training and test sets and calculate the area under them.

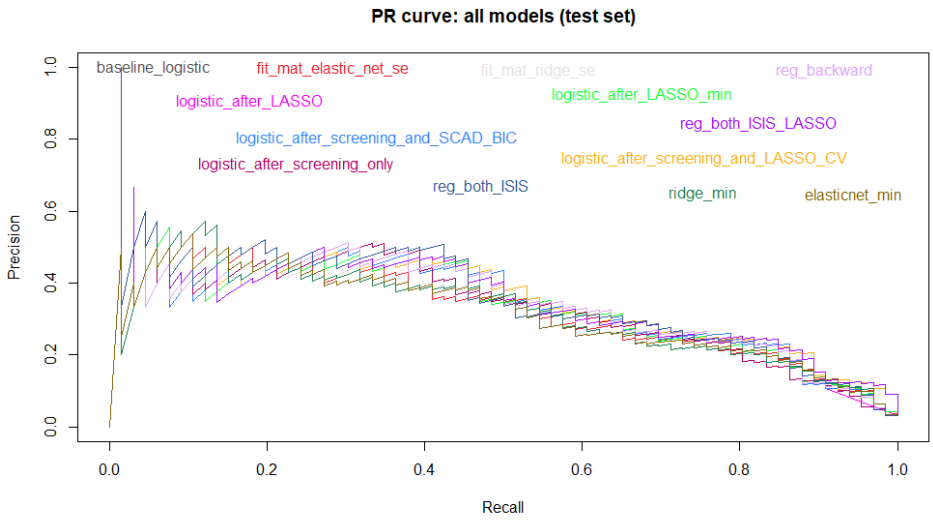

Figure 11: PR curves on the test set of the oversampling models

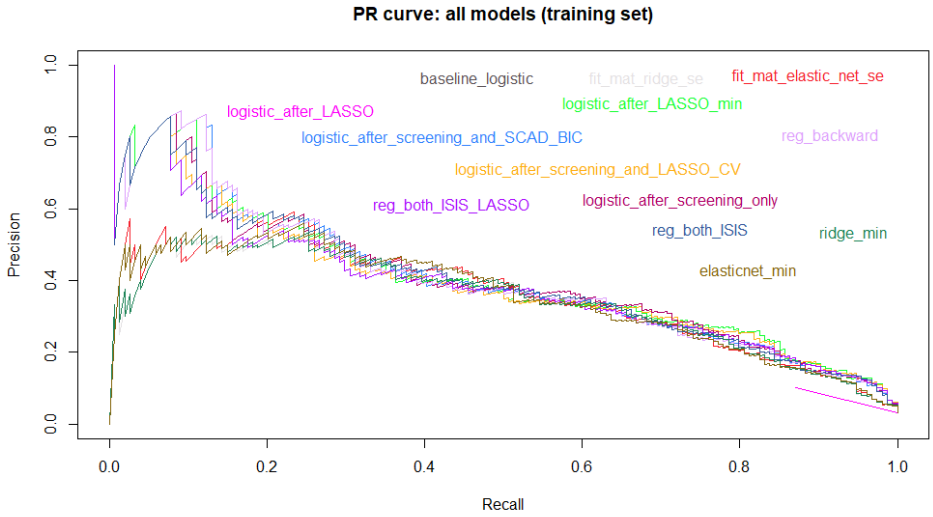

Figure 12: PR curve on the training set of the oversampling models

From the results on the training set, we can see that the best model is the logistic after LASSO with λ 1-se (AUPRC = 0.4111132), but its performance is far worse on the test set (AUPRC = 0.1281449), displaying an enormous amount of overfitting.

Incidentally, we can see from the figure that the part of the curve which accounts for most of the difference in performance between the models on the training set is in the region with very low recall, i.e., the least interesting one for our point of view (we are more interested to catch all the bankrupt firms than to catch only the bankrupt firms, thus we are more interested in the area with high recall).

In the test set, instead, there is not this problem. It is thus better to use directly the test set to evaluate the performance of the model.

The best model on the test set is the stepwise selection after only ISIS, that instead do not display overfitting at all and has good performances also on the training set (AUPRC on the test set = 0.36931, while it is 0.4041888 on the training set).

We choose thus as the best model chosen by the oversampling method the stepwise selection after only the screening algorithm ISIS.

Even here the results obtained are better than the ones obtained by using the original data (and without overfitting), confirming the validity of the oversampling method here used.

However, since even using the oversampling and undersampling techniques we continue to obtain some results that are not so good overall, we should conclude that the problem of "not good results" is to be ascribed to the low predictive power of our variables and models themselves.

We can see this also by fitting for comparison purposes a theoretically more flexible and more

performant (but less interpretable) classifier: the Random Forest (James et al. (2023)) (see Appendix).

3.2.3 Supervised Dimension Reduction

Another approach possible to ameliorate the fitting of our models is to take the best linear combination of all our variables in explaining the bankruptcy of firms. This is the so called "Supervised Dimension Reduction" approach.

This approach is in principle similar to Principal Component Analysis (PCA) since also PCA finds the best linear combinations of all variables (it is too a Dimension Reduction algorithm) but in an unsupervised setting.

We can use in principle two algorithms to perform this analysis: the SIR and the LDA. In both cases we have to do an "inverse regression" of the Xs on the Y but the rescaling matrix used to find the best linear combination is different: the SIR treats the problem as a continuous regression and tries to find the best slicing by using the overall sample covariance matrix, the LDA instead takes into account the class structure by using as rescaling matrix the between sample v/cov matrix (Sachin et al. (2015), Ma and Zhu (2013)).

We use LDA since it succeeds better in dividing the two classes (see Appendix for SIR results) and since it is by construction better tailored to classification problems.

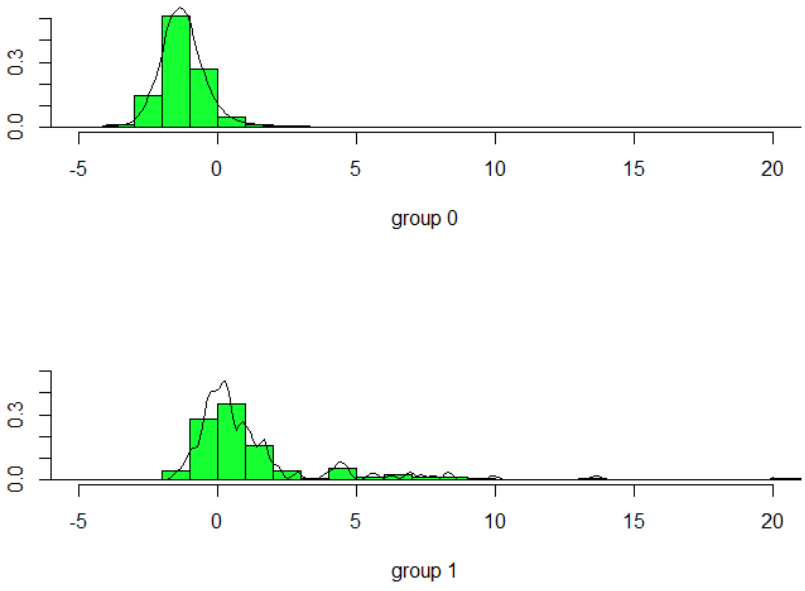

Figure 13: LDA on original dataset

We performed the LDA both on the original training set and on the oversampled training set.

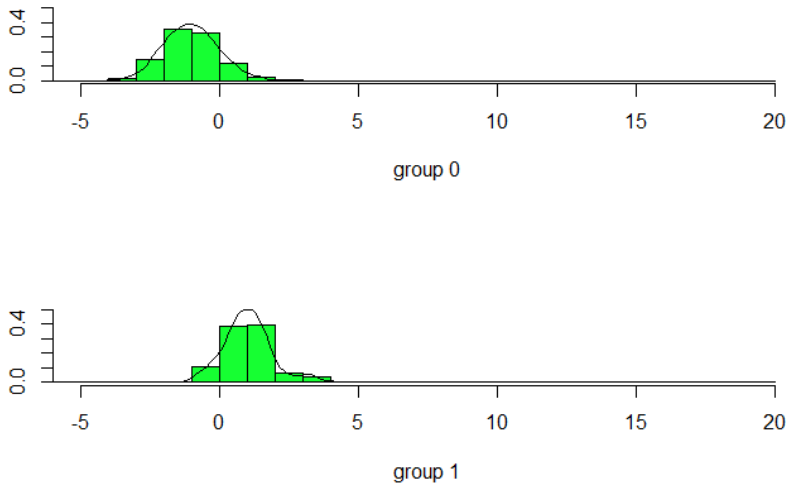

Figure 14: LDA on oversample dataset

Since the LDA is an approach based on a geometric distance, it is expected to give much better results in balanced datasets (like the oversampled one).
Of course, since the problem is a binary classification, the LDA gives us only one linear combination of variables. Thus, we cannot plot the first two best vectors on each other; we can only see a histogram of the distribution of each class on the same single vector.
From the figure, there seems to be no visible difference between the two LDA specifications.
A big difference instead arises when we see the results and the performances of both LDA specifications in classifying the test set (see figure below).
In fact, the logistic model performed using LDA on the original training set gives us an AUPRC on the test set of 0.2990436, while the one performed on the oversample training set has an AUPRC on the same test set of 0.3764772.

As expected, here the use of oversampled dataset gives us much better performances than the original one. The LDA approach on this dataset, furthermore, gives us results that are better than those obtained from standard models performed both on the original and on the oversampled datasets.
An issue still open for further research is whether the LDA performances could be even better if we perform it also using the undersampling approach.

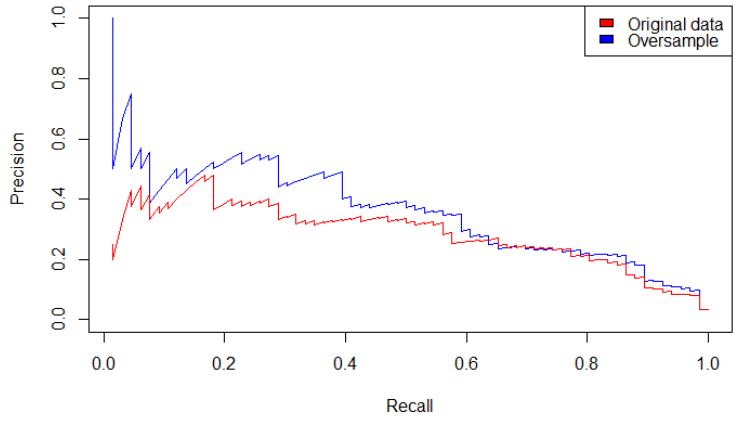

Figure 15: Precision-Recall curve on test set of LDA models - comparison

4 Features Analysis

In this section, we leverage the informative power of logistic regression on the explanatory variables to examine their significance and impact.
For each of the following seven models, we analyze the selected features, their signs, and their significance in the regression:

- Model A: logistic regression after LASSO with 1 standard error λ on imbalanced dataset
- Model B: Elastic Net ($\alpha = 0.5$) with 1 standard error λ on imbalanced dataset
- Model C: stepwise selection after ISIS+SCAD on imbalanced dataset
- Model D: stepwise selection after ISIS+LASSO on imbalanced dataset
- Model E: stepwise selection after only ISIS on imbalanced dataset
- Model F: logistic selection after ISIS+LASSO on undersampled dataset (the best selection model on undersampled dataset)
- Model G: stepwise selection after only ISIS on oversampled dataset (the best selection model on oversampled dataset)

We notice that only 6 features are selected in at least 4 models. The following table presents all six of them. For each one, it indicates the number of models from which it was selected, the sign it appears with, and whether it is significant at the 10% level or not. In section 6.6, the reader can find a comprehensive summary table containing this information for all variables,

disaggregated by model.

Feature	Number of models	Sign	Significance at 10%
Net.Income.to.Total.Assets	7	-	always
Current.Liability.to.Assets	7	+	always
Equity.to.Long.term.Liability	6	+	always
Cash.Total.Assets	5	-	always
Fixed.Assets.to.Assets	5	+	never
Fixed.Assets.Turnover.Frequency	4	+	always

Table 2: Features selected in at leat 4 out 7 models

We therefore decided to further analyze these six variables. We conducted three logistic regressions (referred to as "parsimonious", as they utilize very few explanatory features compared to the initial 94): the first was conducted with only the two features selected by all models, the second adds the two variables selected by at least 5 models and consistently significant, and finally the last one is conducted with all six variables.

Table 3 below presents the coefficients of the features and their significance in the three parsimonious regressions. Additionally, it shows the predictive performance of the regression on the test set in terms of the area under the Precision-Recall curve.

Features | Number of variables:	**Two**	**Four**	**Six**
Net.Income.to.Total.Assets	-0.8335*	-0.9154*	-0.8800*
Current.Liability.to.Assets	0.6344*	0.5080*	0.5921*
Equity.to.Long.term.Liability		0.4128*	0.4022*
Cash.Total.Assets		-1.7365*	-1.5382*
Fixed.Assets.to.Assets			0.2244
Fixed.Assets.Turnover.Frequency			0.3289*
Performance on test set (AUPRC):	*0.2873*	*0,3468*	*0,3447*

*significance at 1%

Table 3: Parsimonious Regressions

Moving from the two-variable regression to the four-variable regression, a considerable increase in predictive power is observed. However, moving from the four-variable regression to the six-variable regression results in a slight loss thereof. This parsimonious four-variable regression

performs well even when compared to all other models presented in Section 3, as evidenced by the precision-recall curves summarized in the graph below.

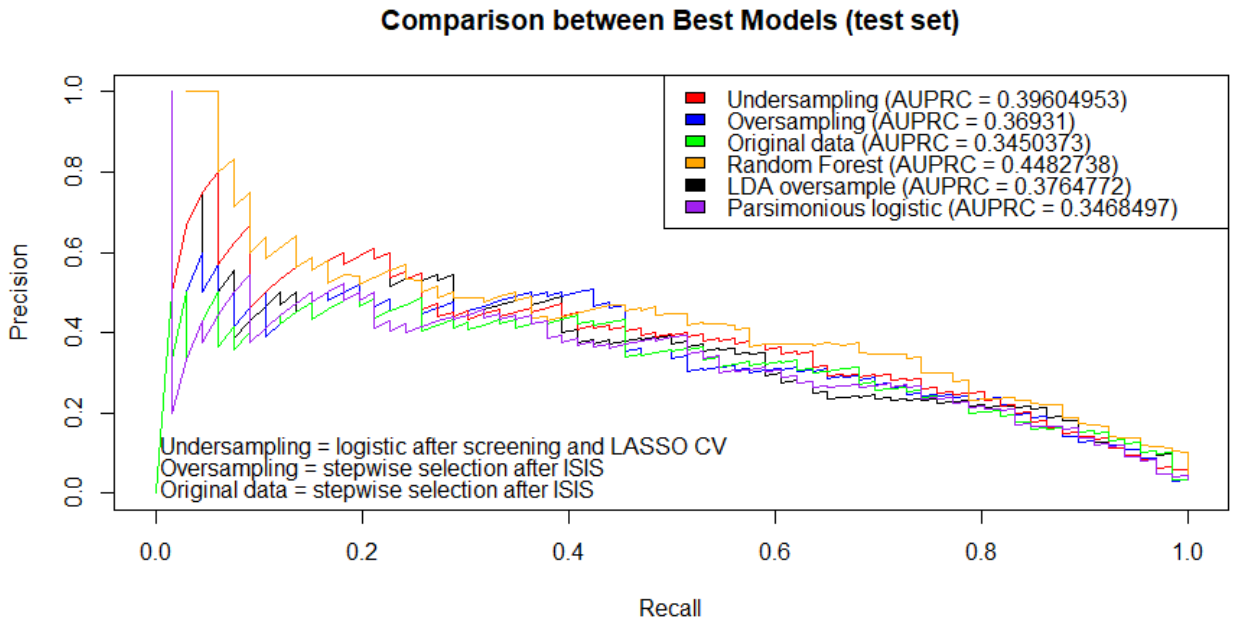

Figure 16: Comparison between Best Models (PR curve on test set)

We can consider the four features used as regressors as four valid indicators (positive or negative) of a company's bankruptcy risk. In particular, we note that focusing solely on them results in minimal loss of predictiveness while significantly enhancing interpretability.

5 Conclusions

Our study aims to identify the most impactful variables for predicting a firm's risk of bankruptcy, utilizing a comprehensive dataset of financial explanatory variables and a binary target variable representing bankruptcy. Our approach incorporates feature screening and selection algorithms in conjunction with logistic regression to provide insights into the explanatory variables.

5.1 Addressing Imbalanced Datasets

To effectively manage the challenges posed by an imbalanced dataset, where only 3% of the instances represent bankrupt companies, we employ undersampling and oversampling techniques. These methods help balance the dataset, ensuring that the model is trained on a more representative sample and thus improving its predictive performance. We have observed this

improvement in performance not only, or not so much, in the feature selection models, but primarily in the results of supervised dimension reduction.

5.2 Best Feature Selection Model

Our analysis reveals that the ISIS+LASSO model applied to an undersampled dataset emerged as the best feature selection model. This combination of ISIS and LASSO is particularly effective in handling the high dimensionality and multicollinearity in our dataset, leading to a model with improved interpretability and predictive performance.

5.3 Key Predictive Features

From the original set of 94 features, our analysis highlights four key indicators that provide a more concise and informative assessment of a company's bankruptcy risk:

- *Net Income to Total Assets*: higher values indicate lower risk, reflecting a firm's profitability. This ratio measures how efficiently a company uses its assets to generate profit, with higher profitability suggesting better financial health and lower bankruptcy risk.

- *Current Liability to Assets*: higher values indicate higher risk, signifying solvency issues. This ratio assesses the proportion of a company's assets that are financed by short-term liabilities, with higher values indicating potential liquidity problems and a greater risk of financial distress.

- *Equity to Long Term Liability*: higher values indicate higher risk, reflecting the capital structure's influence on financial stability. A higher ratio suggests that a company relies more on equity financing relative to long-term debt, which could be a sign of cautious financial management, but in our case, given regressions results, it's an indication of insufficient leverage to optimize growth.

- *Cash to Total Assets*: higher values indicate lower risk, emphasizing the importance of liquidity. This ratio demonstrates the proportion of a company's assets that are held in cash or cash equivalents, with higher values suggesting better liquidity management and a reduced likelihood of facing cash flow problems that could lead to bankruptcy.

5.4 Generalizability and Model Performance

While our results provide valuable insights, their generalizability to other contexts (such as different types of companies, countries, or time periods) remains to be validated. Future research should explore these aspects to confirm the applicability of our findings across various settings. This could involve testing the model on datasets from different industries, regions, or historical periods to assess its robustness and adaptability.

Regarding the performance of the predictive models, feature selection algorithms that we did not utilize can be explored. Specifically, Group LASSO (Meier et al. (2008)) and Exclusive LASSO (Zhou et al. (2010)) could be considered, as features can be pre-classified into groups. Additionally, techniques such as gradient boosting and neural networks could potentially offer enhanced predictive capabilities on this dataset.

These models should be investigated to compare their performance against logistic regression and the feature selection methods used in this study, providing a comprehensive evaluation of various approaches to bankruptcy prediction.

6 Appendix

6.1 Features List

Solvency (1-28 | 28):

X1 Cost of Interest-bearing Debt

X2 Cash Reinvestment Ratio

X3 Current Ratio

X4 Acid Test

X5 Interest Expenses/Total Revenue

X6 Total Liability/Equity Ratio

X7 Liability/Total Assets

X8 Interest-bearing Debt/Equity

X9 Contingent Liability/Equity

X10 Operating Income/Capital

X11 Pretax Income/Capital

X12 Working Capital to Total Assets

X13 Quick Assets/Total assets

X14 Current Assets/Total Assets

X15 Cash/Total Assets

X16 Quick Assets/Current Liability

X17 Cash/Current Liability

X18 Current Liability to Assets

X19 Operating Funds to Liability

X20 Inventory/Working Capital

X21 Inventory/Current Liability

X22 Current Liabilities/Liability

X23 Working Capital/Equity

X24 Current Liabilities/Equity

X25 Long-term Liability to Current Assets

X26 Current Liability to Current Assets

X27 One if Total Liability exceeds Total Assets

X28 Equity to Liability

Capital structure ratios (29-37 | 9):

X29 Equity/Total Assets

X30 (Long-term Liability+Equity)/Fixed Assets

X31 Fixed Assets to Assets

X32 Current Liability to Liability

X33 Current Liability to Equity

X34 Equity to Long-term Liability

X35 Liability to Equity

X36 Degree of Financial Leverage

X37 Interest Coverage Ratio

Others (38-50 | 13):

X38 Operating Expenses/Net Sales

X39 (Research and Development Expenses)/Net Sales

X40 Effective Tax Rate

X41 Book Value Per Share(B)

X42 Book Value Per Share(A)

X43 Book Value Per Share(C)

X44 Cash Flow Per Share

X45 Sales Per Share

X46 Operating Income Per Share

X47 Sales Per Employee

X48 Operation Income Per Employee

X49 Fixed Assets Per Employee

X50 total assets to GNP price

Profitability (51-68 | 18):

X51 Return On Total Assets(C)

X52 Return On Total Assets(A)

X53 Return On Total Assets(B)

X54 Gross Profit /Net Sales

X55 Realized Gross Profit/Net Sales

X56 Operating Income /Net Sales

X57 Pre-Tax Income/Net Sales

X58 Net Income/Net Sales

X59 Net Non-operating Income Ratio

X60 Net Income-Exclude Disposal Gain or Loss/Net Sales

X61 EPS-Net Income

X62 Pretax Income Per Share

X63 Retained Earnings to Total Assets

X64 Total Income to Total Expenses

X65 Total Expenses to Assets

X66 Net Income to Total Assets

X67 Gross Profit to Sales

X68 Net Income to Stockholder's Equity

Turnover ratios (69-81 | 13):

X69 (Inventory +Accounts Receivables) /Equity

X70 Total Asset Turnover

X71 Accounts Receivable Turnover

X72 Days Receivable Outstanding

X73 Inventory Turnover

X74 Fixed Asset Turnover

X75 Equity Turnover

X76 Current Assets to Sales

X77 Quick Assets to Sales

X78 Working Capital to Sales

X79 Cash to Sales

X80 Cash Flow to Sales

X81 No-credit Interval

Cash flow ratios (82-86 | 5):

X82 Cash Flow from Operating/Current Liabilities

X83 Cash Flow to Total Assets

X84 Cash Flow to Liability

X85 CFO to Assets

X86 Cash Flow to Equity

Growth (87-94 / 8): X87 Realized Gross Profit Growth Rate

X88 Operating Income Growth

X89 Net Income Growth

X90 Continuing Operating Income after Tax Growth

X91 Net Income-Excluding Disposal Gain or Loss Growth

X92 Total Asset Growth

X93 Total Equity Growth

X94 Return on Total Asset Growth

6.2 Ridge Regression

Here we present the results for the Ridge regression, including the CV fit and the plot of the features at the varying of λ.

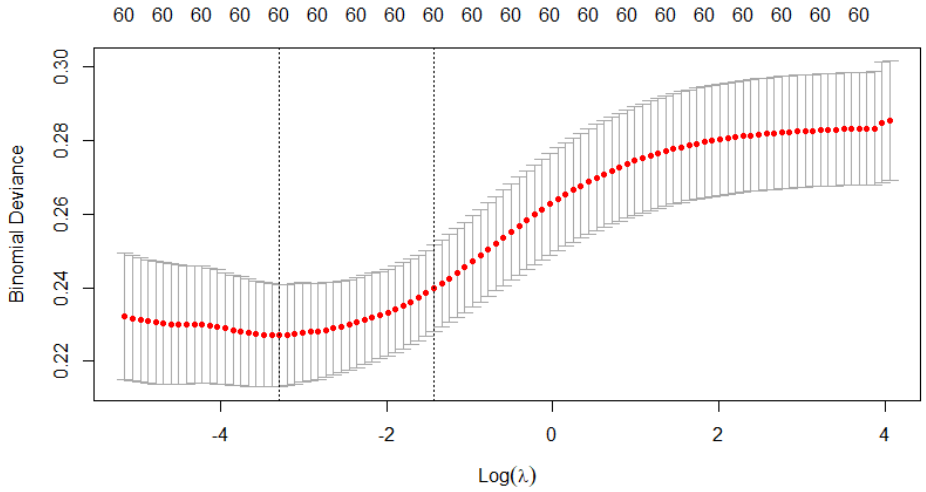

Figure 17: CV fit - Ridge

If we choose the λ min ($= 0.0372$) model we obtain an AUPRC of 0.393363 on the training set and of 0.3076649 on the test set, while for the λ 1-se ($= 0.2392$) the AUPRC is 0.371659 on the training set and 0.3120704 on the test set. Both are substantially less than the ones obtained using the baseline logistic (AUPRC on training set = 0.4454847, on test set = 0.3351596). Thus, we can conclude that residual multicollinearity issues are not present here and Ridge regression is basically useless for our purposes.

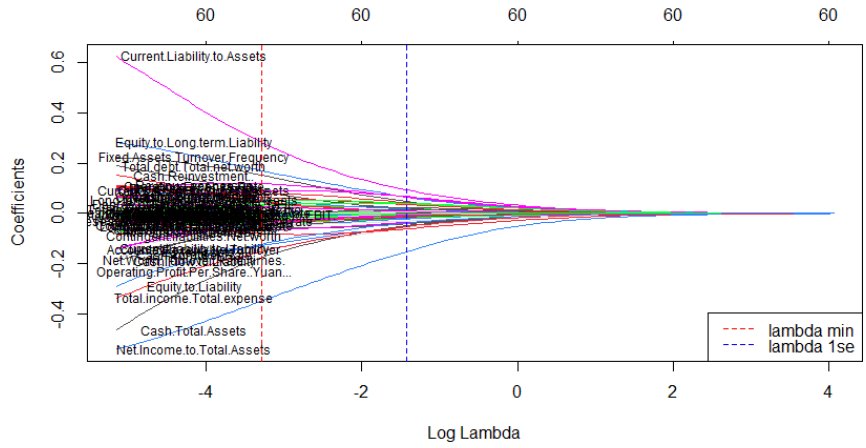

Figure 18: Lambda Plot - Ridge

6.3 Random Forest

We fit a Random Forest tuned with cross-validation using as tuning parameter the Area under the Precision Recall curve. As hyperparameter we tune only mtry = number of variables randomly sampled as candidates at each split and not the number of trees themselves (which is set = 500 by default) in order not to spend too much computational time in tuning both hyperparameters.

In order to find the best hyperparameter region, we proceed at first to perform a Random-Search model on the training set, then a Grid-Search in the interval in which we have the maximum performance of the hyperparameters.

The interval chosen by the Random Search model is with a mtry between 6 and 16. The Grid-Search model furtherly refines the result by setting mtry = 7 as the best performing one.

The performance of the Random Forest model with mtry = 7 on the test set is clearly the best one among all models considered until now (the AUPRC on the test set is 0.4482738) but it displays very significant overfitting since on the training set the fit is almost perfect (AUPRC ≈ 1).

Since the Random Forest is robust to multicollinearity issues, we tried to perform it also on the original dataset including all variables discarded in the preselection process (except the perfectly multicollinear variable "Net worth Assets" and the "Net income Flag" column of all 1). The performances here, however, are slightly worse than the ones obtained on the preprocessed dataset (AUPRC on test set = 0.4469252), thus we can safely discard it and consider only the model performed on the preprocessed dataset.

We can see below also the Feature Importance plot in order to give some interpretation to the

results obtained by this classifier.

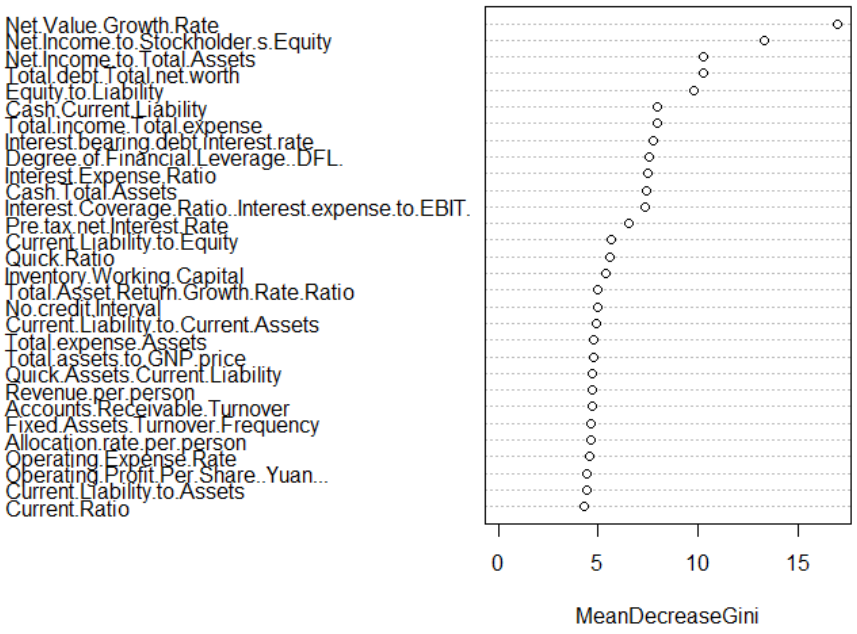

Figure 19: Feature Importance Plot - Random Forest

From the figure we can notice that the feature here selected as the "most important" one does not even appear in the more parsimonious logistic models considered before.

This is probably not due to the fact that these models were wrongly specified, but because this Random Forest specification is too prone to overfitting. Maybe by tuning also the other hyperparameters the Random Forest classifier could overcome this overfitting problem and achieve much better results on the test set. And in this case also the feature importance plot can possibly be much different with respect to the one presented here.

We left this issue open for further research.

In any case, the performances here obtained are not much higher than the ones obtained by logistic models, further sustaining the hypothesis that these "not so good" results are due to the poor predictive power of our dataset itself (and no gain is obtained by including the variables excluded by the preprocessing).

6.4 SIR approach in Supervised Dimension Reduction

Here we present a plot of the SIR approach for our dataset.

We can see from the figure that the SIR algorithm does not succeed in dividing well the two classes, and that the directions other than the first are essentially useless.

This is probably because this algorithm is not suited for classification problems and its rescal-

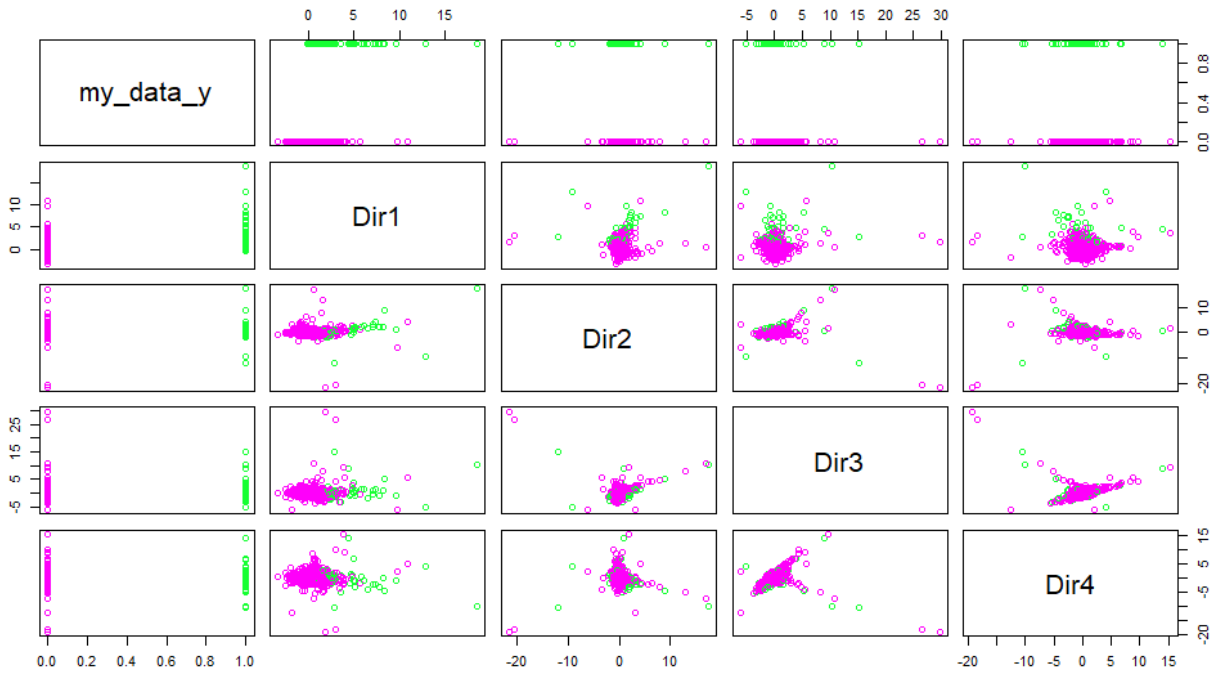

Figure 20: SIR plot

ing matrix adds only noise to the estimation by adding these additional eigenvalues that are essentially zero and contribute nothing to the explanation. Thus, we can safely discard its results.

6.5 LASSO using AUPRC as tuning parameter

Here we present for completeness the results of the λ tuning for the LASSO when AUPRC is used as tuning parameter for the (5-fold) cross-validation.

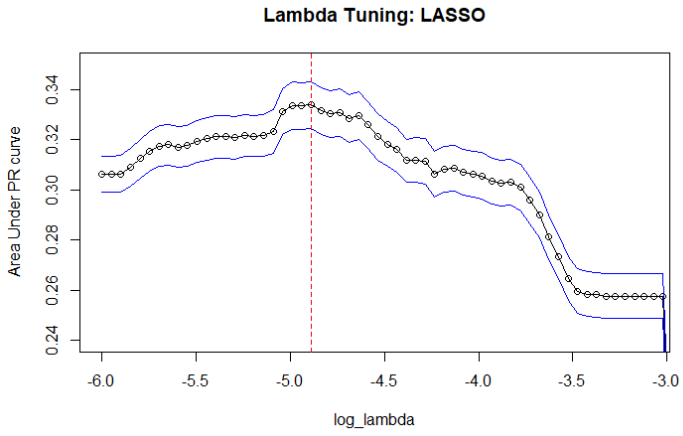

Figure 21: Lambda Tuning: LASSO (AUPRC tuning)

The min λ chosen by this model is overall not so different with respect to the one chosen by using the binomial deviance. Thus, we can safely keep the results seen before.

6.6 Complete Features Analysis

Features | Models	A	B	C	D	E	F	G
[1] "Operating.Gross.Margin"							
[2] "Pre.tax.net.Interest.Rate"							
[3] "Operating.Expense.Rate"							
[4] "Research.and.development.expense.rate"							
[5] "Interest.bearing.debt.interest.rate"							
[6] "Tax.rate..A."							
[7] "Revenue.Per.Share..Yuan..."							-*
[8] "Operating.Profit.Per.Share..Yuan..."							+
[9] "Realized.Sales.Gross.Profit.Growth.Rate"							
[10] "Regular.Net.Profit.Growth.Rate"							
[11] "Continuous.Net.Profit.Growth.Rate"							
[12] "Total.Asset.Growth.Rate"						+*	
[13] "Net.Value.Growth.Rate"							
[14] "Total.Asset.Return.Growth.Rate.Ratio"							
[15] "Cash.Reinvestment.."					+*	+*	
[16] "Current.Ratio"							
[17] "Quick.Ratio"							
[18] "Interest.Expense.Ratio"							
[19] "Total.debt.Total.net.worth"					+*		
[20] "Long.term.fund.suitability.ratio..A."							
[21] "Contingent.liabilities.Net.worth"						+*	
[22] "Accounts.Receivable.Turnover"						-	
[23] "Average.Collection.Days"							
[24] "Inventory.Turnover.Rate..times."							
[25] "Fixed.Assets.Turnover.Frequency"			+*	+*	+*	+*	
[26] "Net.Worth.Turnover.Rate..times."					-*	-*	
[27] "Revenue.per.person"							
[28] "Operating.profit.per.person"							

Table 3 continued from previous page

Features | Models	A	B	C	D	E	F	G
[29] "Allocation.rate.per.person"							
[30] "Current.Assets.Total.Assets"							
[31] "Cash.Total.Assets"			-*	-*	-*	-*	-*
[32] "Quick.Assets.Current.Liability"							
[33] "Cash.Current.Liability"			+	+			
[34] "Current.Liability.to.Assets"	+*	+*	+*	+*	+*	+*	+*
[35] "Operating.Funds.to.Liability"					+*	+*	
[36] "Inventory.Working.Capital"							
[37] "Inventory.Current.Liability"							
[38] "Long.term.Liability.to.Current.Assets"							
[39] "Total.income.Total.expense"					-*	-*	-
[40] "Total.expense.Assets"							
[41] "Current.Asset.Turnover.Rate"							
[42] "Quick.Asset.Turnover.Rate"					+*		
[43] "Cash.Turnover.Rate"					-*	-*	
[44] "Cash.Flow.to.Sales"							
[45] "Fixed.Assets.to.Assets"	+	+	+	+	+		
[46] "Current.Liability.to.Liability"						-*	-*
[47] "Current.Liability.to.Equity"						-*	
[48] "Equity.to.Long.term.Liability"	+*	+*	+*	+*	+*	+*	
[49] "Cash.Flow.to.Liability"					-*	-*	
[50] "CFO.to.Assets"					-*	-*	
[51] "Cash.Flow.to.Equity"					+*		
[52] "Current.Liability.to.Current.Assets"	+*	+*				-*	
[53] "Liability.Assets.Flag"	-	-					
[54] "Net.Income.to.Total.Assets"	-*	-*	-*	-*	-*	-*	-*
[55] "Total.assets.to.GNP.price"						-*	
[56] "No.credit.Interval"							
[57] "Net.Income.to.Stockholder.s.Equity"	+*	+*				-*	
[58] "Degree.of.Financial.Leverage..DFL."							
[59] "Interest.Coverage.Ratio..Interest.expense.to.EBIT."							

Table 3 continued from previous page

Features | Models	A	B	C	D	E	F	G
[60] "Equity.to.Liability					+*		

*significance at 10%

Table 4: Complete Features Analysis

References

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4):589–609.

Balcaen, S. and Ooghe, H. (2006). 35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems. *The British Accounting Review*, 38(1):63–93.

Balleisen, E. J. (2001). *Navigating failure: bankruptcy and commercial society in antebellum America*. Univ of North Carolina Press.

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of accounting research*, pages 71–111.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(5):849–911.

James, G., Witten, D., Hastie, T., Tibshirani, R., and Taylor, J. (2023). *An introduction to statistical learning: With applications in python*. Springer Nature.

Kumar, P. R. and Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques–a review. *European journal of operational research*, 180(1):1–28.

Lin, W.-Y., Hu, Y.-H., and Tsai, C.-F. (2011). Machine learning in financial crisis prediction: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):421–436.

Ma, Y. and Zhu, L. (2013). A review on dimension reduction. *International Statistical Review*, 81(1):134–150.

Meier, L., Van De Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(1):53–71.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of accounting research*, pages 109–131.

Prusa, J., Khoshgoftaar, T. M., Dittman, D. J., and Napolitano, A. (2015). Using random undersampling to alleviate class imbalance on tweet sentiment data. In *2015 IEEE international conference on information reuse and integration*, pages 197–202. IEEE.

Sachin, D. et al. (2015). Dimensionality reduction and classification through pca and lda. *International journal of computer Applications*, 122(17).

Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432.

UCI (2020). Taiwanese Bankruptcy Prediction. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5004D.

Zhou, Y., Jin, R., and Hoi, S. C.-H. (2010). Exclusive lasso for multi-task feature selection. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 988–995. JMLR Workshop and Conference Proceedings.