Introduction

Bankruptcy is a legal process through which people or other entities who cannot repay debts to creditors may seek relief from some or all of their debts. In most jurisdictions, bankruptcy is imposed by a court order, often initiated by the debtor.

The purpose of this report is to use regularised regression models in order to identify the variables which predict whether a company will go bankrupt within the medium term (4 years). For this purpose, I will be construction models using ridge and lasso logistic regression, as well as elastic nets and assess which model is best at predicting bankruptcy.

Data

The data contains observations from 700 companies across 32 variables. In total, 350 companies from the data set have been declared bankrupt. We can take a look at each variable's data type as per Fig. 1.

```
'data.frame':    700 obs. of  32 variables:
 $ Var1 : num  0.6175 0.0139 0.1018 -0.0325 0.0503 ...
 $ Var2 : num  0.819 0.986 0.611 0.344 0.222 ...
 $ Var3 : num  0.6175 0.0185 0.1204 0.0207 0.0945 ...
 $ Var4 : num  0.2389 0.0354 0.0732 0.0252 0.0351 ...
 $ Var5 : num  0.6175 0.0185 0.1204 -0.0408 0.0503 ...
 $ Var6 : num  0.6175 0.0185 0.1204 -0.0408 0.0503 ...
 $ Var7 : num  0.1814 0.0139 0.389 0.6562 0.7777 ...
 $ Var8 : num  10.9 0 20.9 28.7 14 ...
 $ Var9 : num  0.6138 0.0185 0.1338 0.0176 0.0937 ...
 $ Var10: num  0.2129 0.0265 0.035 -0.012 0.0142 ...
 $ Var11: num  0.6808 0.1991 0.6109 0.1417 0.0224 ...
 $ Var12: num  3.25 2.33 3.94 3.87 4.3 ...
 $ Var13: num  0.407 0.1528 0.253 0.0723 -0.1834 ...
 $ Var14: num  -0.021 -0.124 0.105 0.226 0.216 ...
 $ Var15: num  0.2129 0.0354 0.0413 -0.0113 0.0223 ...
 $ Var16: num  0.6238 0.0185 0.1338 0.0168 0.1188 ...
 $ Var17: num  2.901 0.523 2.965 2.708 3.562 ...
 $ Var18: num  0.819 0.986 0.701 0.465 0.366 ...
 $ Var19: num  0.215 0.0354 0.04596 0.00622 0.03354 ...
 $ Var20: num  0.21158 0.0354 0.04596 0.00652 0.02647 ...
 $ Var21: num  43.5 61.4 57.2 70.7 33.4 ...
 $ Var22: num  32.6 61.4 36.3 42 19.4 ...
 $ Var23: num  0.5381 0.0185 0.0409 -0.0913 0.0197 ...
 $ Var24: num  0.18548 0.0354 0.01406 -0.03372 0.00556 ...
 $ Var25: num  0.1814 0.0139 0.2986 0.5005 0.5265 ...
 $ Var26: num  723 33 2218 537 -3669 ...
 $ Var27: num  0.215 0.0354 0.03469 0.00622 0.03354 ...
 $ Var28: num  0.787 0.965 0.965 1.015 0.973 ...
 $ Var29: num  0.0497 0 0.1646 0 -0.0373 ...
 $ Var30: num  22.82 9.69 37.43 67.47 54.25 ...
 $ Var31: num  0.6175 0.0185 0.1204 -0.0408 0.0503 ...
 $ class: int  0 1 0 0 1 1 0 0 1 1 ...
```

Figure 1. Variable data types.

All variables are numerical, apart from the variable "class" which is our response and takes integer values. The response takes value 1 if a company was declared bankrupt and value 0 otherwise. For this reason, I will have to use a logistic model for our data. The table in Figure 2 offers information on each variable and what they represent.

Variable Name	Description	Variable Name	Description
Var1	net profit / total assets	Var2	equity / total assets
Var3	(gross profit + extraordinary items + financial expenses) / total assets	Var4	(gross profit + depreciation) / sales
Var5	(gross profit + interest) / total assets	Var6	gross profit / total assets
Var7	total liabilities / total assets	Var8	(inventory * 365) / sales
Var9	profit on operating activities / total assets	Var10	net profit / sales
Var11	(equity - share capital) / total assets	Var12	logarithm of total assets
Var13	working capital / total assets	Var14	(total liabilities - cash) / sales
Var15	(gross profit + interest) / sales	Var16	profit on sales / total assets
Var17	total sales / total assets	Var18	constant capital / total assets
Var19	profit on sales / sales	Var20	profit on operating activities / sales
Var21	rotation receivables + inventory turnover in days	Var22	(receivables * 365) / sales
Var23	EBITDA (profit on operating activities - depreciation) / total assets	Var24	EBITDA (profit on operating activities - depreciation) / sales
Var25	short-term liabilities / total assets		
Var26	working capital	Var27	(sales - cost of products sold) / sales
Var28	total costs /total sales	Var29	retained earnings / total assets
Var30	(short-term liabilities *365) / sales	Var31	EBIT / total assets
class	Class where 1 indicates a bankruptcy		

Figure 2. Description for each variable.

Exploratory analysis

In order to explore the data further, I have created summaries of the data using the summary() function. The variables take values within different ranges, therefore the variables will need to be centered and scaled. The summaries of the data can be found in the R script.

Further, I have created some plots in order to get a visual idea about the relationship between different variables. As seen in Figure 3 and Figure 4, we have nearly perfect linear relationships between some of the variables. In order to inspect this issue further I decided to look at the correlation between variables using the function cor(). Indeed, some variables have nearly perfect correlation. This could cause the issue of multicollinearity in our data.

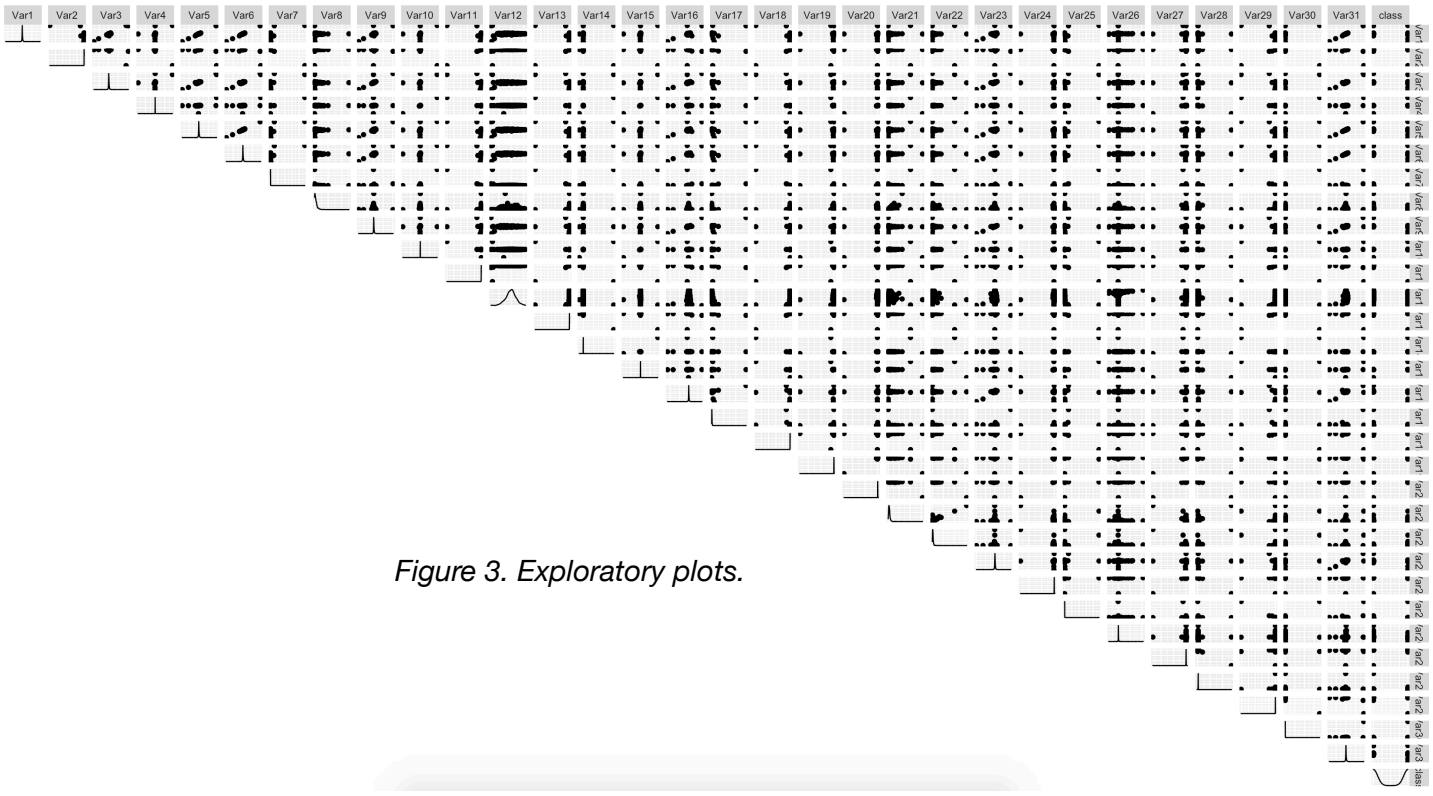

Figure 3. Exploratory plots.

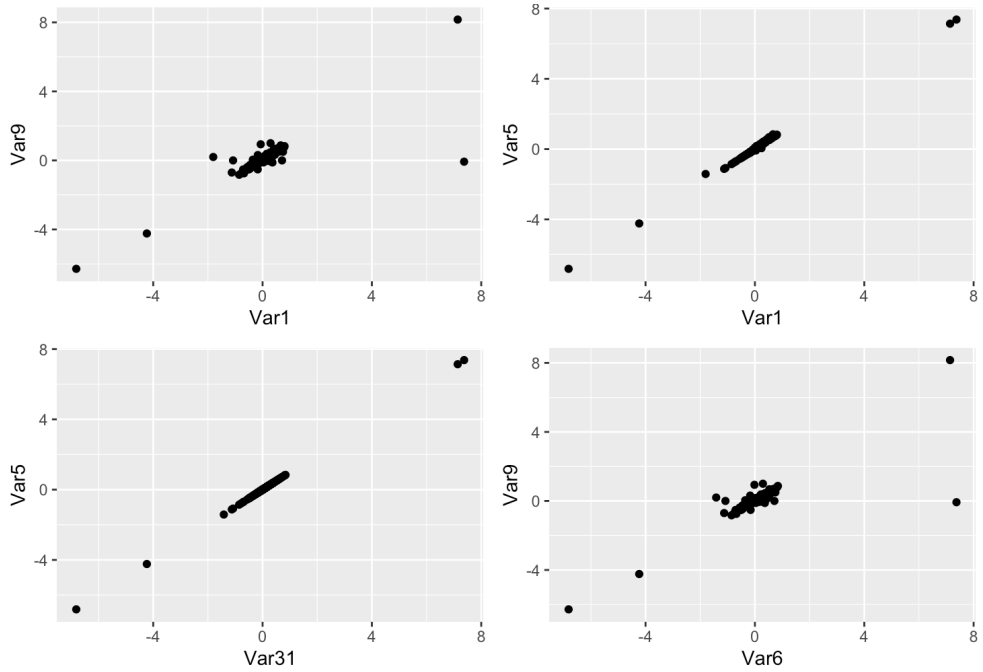

Figure 4. Exploratory plots.

I have also produced a series of box plots between the response and different variables, however, they were not very informative and I have decided to leave them out of the report.

Before I proceeded further with the analysis, I have split the data into training and testing sets. 70% of the observations went into the training data set, and the rest of 30% went into the test data set. Following this, I have standardised all the variables except for the response using the mean and variance fitted to the training data set in order to avoid data leakage.

Logistic ridge regression

Ridge regression is a model tuning method that is used to analyse data that suffers from multicollinearity, which we have seen it is the case with our data in the exploratory analysis section. This method performs L2 regularisation. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values. (Ashok, 2020)

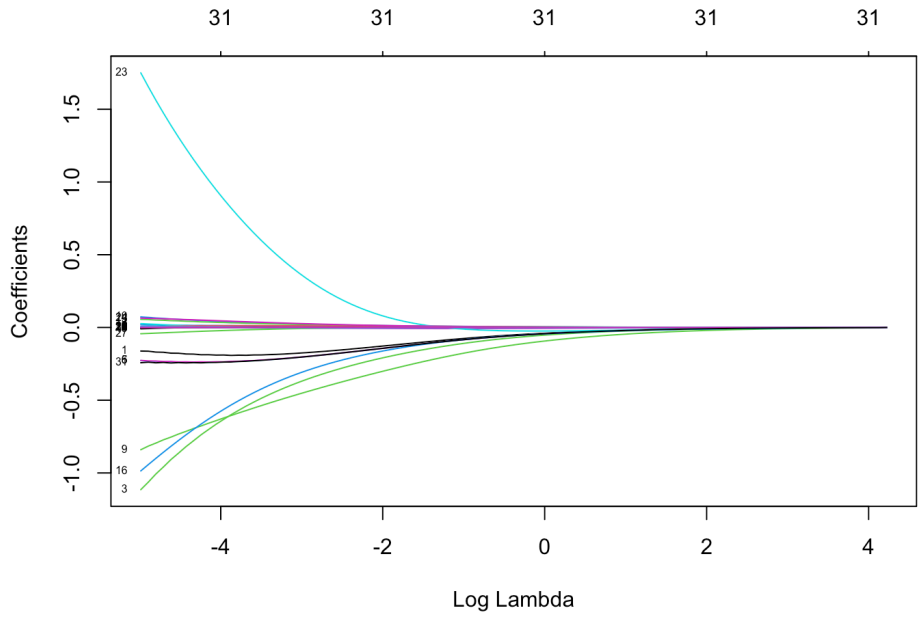

Figure 5. Coefficients against log of lambda.

I will begin the analysis by fitting a logistic ridge regression model to our data using the function glmnet() with alpha set to 0. As such, I have produced the plot presented in Figure 5, which illustrates the coefficients' values as a function of log of lambda .
On the plot we can see that when the value of log lambda is 4, all coefficients are approximately 0. The more we decrease lambda, the coefficients grow away from 0 until they reach a point

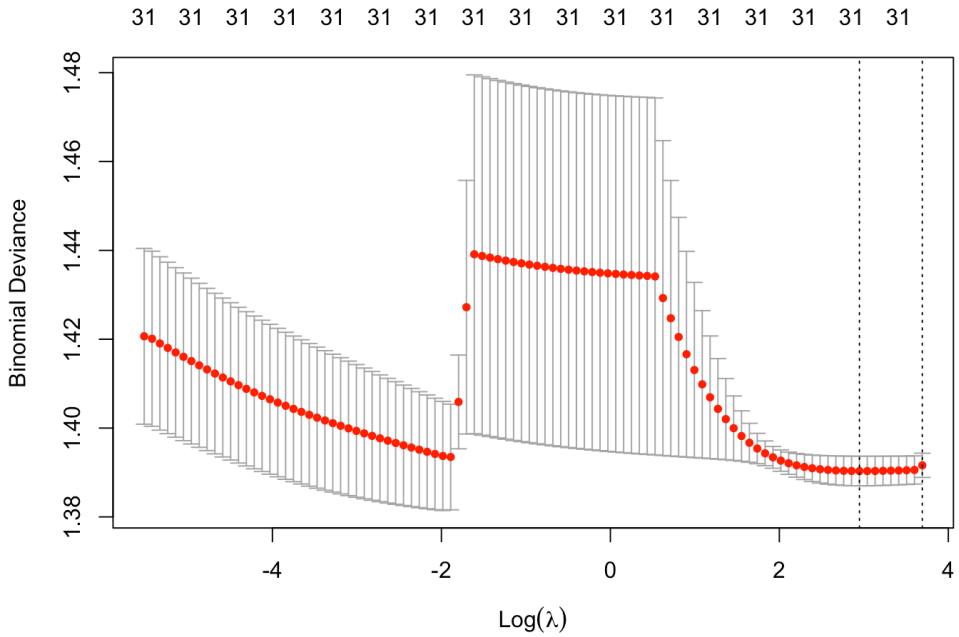

Figure 6. Binomial deviance against values of log of lambda.

where they are effectively unregularised.

Next, I have used cross validation in order to select the best value of lambda for our model. Figure 6 illustrates a graph of the deviance against the values of the logarithm of lambda.

As we can see, the deviance is smallest at larger values of log of lambda. The two vertical dotted lines represent the minimum deviance and one standard deviation of the minimum deviance and the latter represents the more restricted model since the value of lambda is larger. The value of lambda for which the deviance is minimised is 19.06635, and the value of lambda that is within one standard deviation away from the minimum deviance is 40.13284.

```
32 x 1 sparse Matrix of class "dgCMatrix"
                      s0
(Intercept) -8.453218e-02
Var1        -1.024631e-03
Var2        -1.530306e-05
Var3        -1.218464e-03
Var4         1.203376e-04
Var5        -1.086557e-03
Var6        -1.086557e-03
Var7         1.529986e-05
Var8        -1.528234e-06
Var9        -2.504947e-03
Var10        1.286281e-04
Var11       -1.520279e-05
Var12       -2.681343e-04
Var13       -1.539117e-05
Var14        2.836455e-06
Var15        1.279163e-04
Var16       -1.264104e-03
Var17        4.686607e-05
Var18       -1.547140e-05
Var19        6.649426e-05
Var20        8.369334e-05
Var21        3.055960e-07
Var22        6.192497e-07
Var23       -1.248086e-03
Var24        9.536146e-05
Var25        1.533075e-05
Var26       -2.873430e-08
Var27        5.837989e-05
Var28       -7.993152e-05
Var29       -1.424423e-05
Var30        7.032712e-09
Var31       -1.086557e-03
```

Figure 7. Coefficients of the ridge regression model.

I have then fit a logistic ridge regression model with 40.13284 as the value of lambda. Figure 7 presents the coefficients of the fitted model.
As we can see that some coefficients have been minimised (variable 1, variable 10, variable 15), whereas others have a significantly higher weight than others, such as variable 24, variable 20 and variable 30.

Logistic Lasso regression

Lasso regression is a regularisation technique which uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models. This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection. Lasso regression uses the L1 regularisation parameter. (Kumar, 2021)

The first step I took was to fit a lasso model with the function glmnet() with parameter alpha equalling 1 and to plot the coefficients' values as a function of log of lambda. The plot is illustrated in Figure 8.

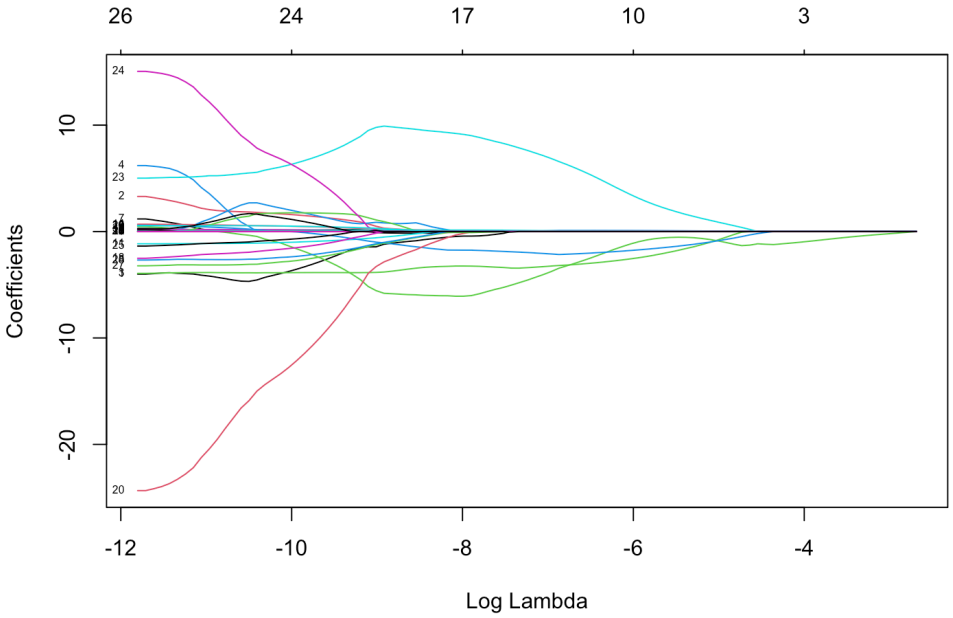

Figure 8. Coefficients for Lasso logistic regression.

We can see that, as lambda decreases, more coefficients enter the model.

The next step is to use cross validation to select the most appropriate value of lambda. Figure 9 illustrates the binomial deviance against different values of log of lambda.

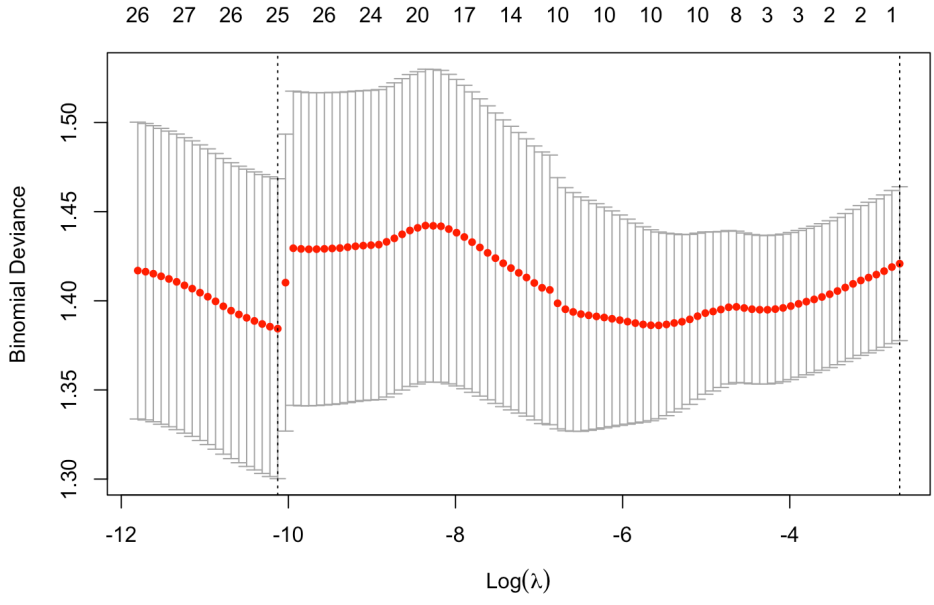

Figure 9. Deviance against log of lambda for lasso regression.

As we can see from the plot, the deviance is lowest when lambda takes a smaller value, with 25 parameters in the model, as opposed to the value of the deviance at one standard deviation

```
32 x 1 sparse Matrix of class "dgCMatrix"
                      s0
(Intercept)  2.854389e+00
Var1        -4.169845e+00
Var2         1.668814e+00
Var3        -3.801410e+00
Var4         .
Var5         2.490170e-01
Var6         2.768933e-16
Var7         .
Var8        -1.696384e-03
Var9        -1.258156e+00
Var10        2.506048e+00
Var11       -1.040681e+00
Var12        1.698234e-01
Var13        3.768064e-02
Var14        5.114826e-01
Var15        1.459367e+00
Var16        .
Var17        8.562288e-02
Var18       -1.745731e+00
Var19        .
Var20       -1.325438e+01
Var21        .
Var22        1.851006e-04
Var23        6.178616e+00
Var24        6.738521e+00
Var25       -8.830320e-01
Var26       -5.840827e-06
Var27       -2.802674e+00
Var28       -2.370156e+00
Var29        5.311501e-01
Var30        .
Var31        1.256945e+00
```

Figure 10. Coefficients for lasso regression.

where just one parameter remains in the model and the deviance is larger. The value of lambda for which the deviance is at minimum is 0.00003996064, and this is what I will use to fit our lasso regression model.

Once I have fitted the logistic lasso regression model, we have the following coefficients (illustrated in Figure 10). We can see that only 25 variables remain in the model, as for 6 of them the coefficients are now 0.

Elastic Net

Elastic net is a regularized regression method that linearly combines the L1 and L2 penalties of the lasso and ridge methods.

In order to select the best value of our tuning parameters alpha and lambda, I have used the function train() from the caret package and got the following results:

alpha <dbl>	lambda <dbl>
1	3.15226e-05

It appears that the best value of alpha is 1, so lasso regression is the preferred option. The value of lambda is even smaller that what I previously used for lasso regression. Therefore, it would be interesting to take a look at our coefficients. As per Figure 11, 28 coefficients remain in the model, 3 of them having been reduced to 0.

Predictions and results

The next step is to evaluate the performance of our three models for predictions on the test data set.

The logistic ridge regression model predicts very low probabilities for all data points, therefore it only predicts class 0 (meaning that the company doesn't go bankrupt) throughout. As such, the correct classification rate for this model is only 0.450237, which means that it performs worse than even random chance (~0.50). Based on this, I conclude that it is not a good model at all.

```
32 x 1 sparse Matrix of class "dgCMatrix"
                      s1
(Intercept)  3.032992e+00
Var1        -4.443591e+00
Var2         1.774958e+00
Var3        -3.877719e+00
Var4         7.182384e-03
Var5         1.981808e-02
Var6         1.930669e-15
Var7         .
Var8        -1.800761e-03
Var9        -5.706627e-01
Var10        2.624024e+00
Var11       -1.080976e+00
Var12        1.813031e-01
Var13        5.382231e-02
Var14        5.517091e-01
Var15        1.686961e+00
Var16        1.516568e-01
Var17        7.930371e-02
Var18       -1.836864e+00
Var19        .
Var20       -1.473025e+01
Var21        .
Var22        7.325640e-05
Var23        5.654671e+00
Var24        7.672051e+00
Var25       -8.947343e-01
Var26       -5.725557e-06
Var27       -3.036214e+00
Var28       -2.566153e+00
Var29        5.408652e-01
Var30        5.944292e-05
Var31        1.576737e+00
```

Figure 11. Coefficients for Elastic Net.

The logistic lasso regression model is more balanced, as it predicts both classes 0 and 1 (indicating bankruptcy). The correct classification rate for this model is 0.5734597, meaning that the model has the same performance as random chance. I would say that the predictive power of the model is rather weak, but it does perform better than the ridge regression model.

The Elastic Net model performs the best out of the three model with a correct classification rate of 0.6445498. However, although the performance is superior compared to the other models, it still doesn't do a great job at predicting the correct class. Below are the cross classification table and and class specific classification rates.

```
  predictions.net
    0  1
0 63 32
1 43 73
```

Figure 12. Cross classification table.

```
  predictions.net
           0         1
0 0.6631579 0.3368421
1 0.3706897 0.6293103
```

Figure 13. Classification rates.

As we can see from the tables, the Elastic Net model does a better job at predicting class 0 (non-bankruptcy) than it does at predicting class 1 (bankruptcy).

Discussion

Based on the out-of-sample performance, I would say that the Elastic Net model is the best, and this is the model I would choose. However, I do have some concerns, since the parameter lambda is so small for this case of lasso regression, and 28 out of 31 coefficients remain in the model. Ideally, we would end up with less parameters and a simpler model. However, the first case with a larger value of the tuning parameter lambda yielded such poor results on the test data, I think it would be sensible to choose the Elastic Net model.

Furthermore, I am concerned about the issue of multicollinearity, as I am unsure to what extent this has been dealt with, considering that such a high number of variables remain in the model.

Conclusion

Upon performing the exploratory data analysis, I have noticed that a large number of variables are almost perfectly correlated with each other, which would pose significant issues for logistic regression, since it introduces multicollinearity in the data. Furthermore, we also have a very high number of variables. As such, regularised regression approaches would be most suitable for modelling the data.

I proceeded and fit three models to the data - a logistic ridge regression model, a logistic Lasso regression model and an Elastic Net. The logistic ridge regression model with tuning parameter lambda equalling 40.13284 minimised some of the coefficients, but it did not perform well on the test data.

The logistic Lasso regression model performed slightly better on the test data, and removed five variables from the model, however the overall classification rate was still lacking.

The Elastic Net results yielded Lasso regression as the best option, but with a smaller value of lambda than what I had previously used. The overall classification rate was the best out of the lot, however it was still not very good. Furthermore, since we had a significant amount of multicollinearity in the data and the fact that a large number of variables (28 out of 31) still remain in the model, I have doubts about the overall validity of the model.

References

Kumar, D., 2021. *A Complete understanding of LASSO Regression*. [online] Great Learning. Available at: <https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/> [Accessed 21 June 2022].

Ashok, P., 2020. *What is Ridge Regression?*. [online] Great Learning. Available at: <https://www.mygreatlearning.com/blog/what-is-ridge-regression/> [Accessed 21 June 2022].