Modelling Competition - Credit Scoring

Cesar Conejo Villalobos - Xavier Bryant (Group glmnet)

20/11/2021

1. Introduction

A credit score is a value assigned to a person that helps financial institutions to identify customers who a have a good history of, or good potential to, pay their debt on time and in its full amount. Our goal here is to find the best classification model to train our data based on customers who have a good credit and a bad credit, and then use this model classify whether a customer is likely to have good or bad credit. This is of course fundamental to some financial lenders, especially if credit scores are unavailable to them or a customer has a limited credit history, as they need to be able to predict, with accuracy, whether a person will be a reliably pay their debts. Also, financial lenders could use predictions such as these in combination with credit scores and help them determine whether a customer will likely pay back debts.

For this project, the key metrics for the our implemented classification models are the following:

- Global accuracy (General predictive power of the model)
- Sensitivity: In order to lower the risk, the focus of the models must be to correctly detect the largest possible number of customers with bad credit.

Also, the nominal values of good and bad customers misclassified can be useful metrics in the performances of the models.

2. Description Data set

The data set consists of 1000 observations of past applicants. The original data set had a set of 30 predictor variables, but we have combined some variables of the data to now consist of 20 predictor variables. There is 1 response variable, for those with good and bad credit scores, making a total of 21 variables in the data. We classified the predictor variables into four groups:

1. Social Background

In this group, we identify the social profile of the applicants, as the age, number of dependents, sex, and marital status. Also, this group of variables includes the immigration status of the applicants, especially the period of residence `PRESENT_RESIDENT` and the nationality of the worker `FOREIGN`. Also, there is the information of the nature of the job, with a reference level of 0 associated with an unemployed and unskilled worker to (3) highly qualified employee. A summary of the variables in this group is available in Table 1.

In the case of `MALE_STATUS`, this variable was created from three separate binary variables: MALE_DIV, MALE_SINGLE, and MALE_MAR_WID to indicate, instead in a factor, if someone is male and their martial status. We can see in `MALE_STATUS` that a large portion of the sample is represented by OTHER because likely a large amount of females considered were categorized there. This is a bit concerning because, with the series of original binaries, it is not obvious if someone does not receive a "1", for a certain variable, because they are female or because they don't fall into the particular category, for example, divorced, single or married/widowed. Therefore, this information is lost from the analysis.

Table 1: Social background covariates

AGE	NUM_DEPENDENTS	MALE_STATUS	JOB	PRESENT_RESIDENT	FOREIGN	TELEPHONE
Min. :19.00	Min. :1.000	OTHER :310	0: 22	0:130	No :963	No :596
1st Qu.:27.00	1st Qu.:1.000	SINGLE :548	1:200	1:308	Yes: 37	Yes:404
Median :33.00	Median :1.000	MAR_WID : 92	2:630	2:149		
Mean :35.55	Mean :1.155	DIVORCED: 50	3:148	3:413		
3rd Qu.:42.00	3rd Qu.:1.000					
Max. :75.00	Max. :2.000					

Table 2: Economic background covariates

EMPLOYMENT	PROP_RSTATE	RESIDENCE	CHK_ACCT	SAV_ACCT
0: 62	OTHER :564	OTHER :108	0:274	0:603
1:172	NO_OWNS_PROP:154	RENT :179	1:269	1:103
2:339	OWNS_RS :282	OWN_RESID:713	2: 63	2: 63
3:174			3:394	3: 48
4:253				4:183

2. Economic Background

The variables corresponding to his group can be found in Table 2 and represent the economic profile of the applicants.

The `PROP_RSTATE` factor variable was created from two binary variables in the original data set, REAL_ESTATE and PROP_UNKN_NONE. These variables represented, respectively, whether someone owned real estate or property. The REAL_ESTATE and PROP_UNKN_NONE variables coding, we have found, is slightly flawed as there is sizable portion of the sample that falls into OTHER. One can either own property/real estate or not therefore there should only be two categories, and these original binaries should be completely inversely correlated. However, the coding does not make it clear if property is defined as real estate or not. We recognize those defined as OTHER could be potentially missing values. There could be potential strategies to fill in these possible missing values, however, since these definitions are not clear we will maintain the original structure. These strategies were also not in the scope of this course.

`RESIDENCE` was created from two binaries, for RENT, representing if the applicants rents where they live, and OWN_RES, indicating if someone owns their place of residency. Similarly, to PROP_RSTATE, there are some values coded as OTHER which are not clearly defined. If someone does not rent or own their residence, it is not clear what their living situation is. Living for free with friends or relatives is a possibility but this is not clear. These also present an opportunity for missing values.

A further note - these two variables, PROP_RSTATE and RESIDENCE, for example, could be co-linear, as those who own real estate will likely live in it, but as is the case for the rest of our data set, since we are predicting, co-linearity is not a significant issue. We have determined that it is best to leave the variables to see their power of prediction and remove them later, testing their statistical significance and contribution to fit through LRT.

Finally, variables related to the checking account status (`CHK_ACCT`) and Average balance in savings accounts (`SAV_ACCT`) can give valuable insights related to the behavior of the customer. In the case of `CHK_ACCT` the breakdown of the variables is the following : 0 : < 0 DM; 1: $0 < \ldots < 200$ DM; 2 : $=> 200$ DM; and 3: no checking account.

For the variable `SAV_ACCT`, the breakdown is the following it's: 0 : < 100 DM; 1 : $100 <= \ldots < 500$ DM; 2 : $500 <= \ldots < 1000$ DM; 3 : $=>1000$ DM; and 4 : unknown/ no savings account

Table 3: Credict products covariates

DURATION	AMOUNT	INSTALL_RATE	PURPOSE_CREDIT	GUARANTEES
Min. : 4.0	Min. : 250	Min. :1.000	OTHER : 55	NONE :907
1st Qu.:12.0	1st Qu.: 1366	1st Qu.:2.000	RADIO_TV :280	CO_APPLICANT: 41
Median :18.0	Median : 2320	Median :3.000	NEW_CAR :234	GUARANTOR : 52
Mean :20.9	Mean : 3271	Mean :2.973	USED_CAR :103	
3rd Qu.:24.0	3rd Qu.: 3972	3rd Qu.:4.000	FURNITURE :181	
Max. :72.0	Max. :18424	Max. :4.000	EDUCATION : 50	
			RETRAINING: 97	

Table 4: Credit History covariates

HISTORY	NUM_CREDITS	OTHER_INSTALL
0: 40	Min. :1.000	No :814
1: 49	1st Qu.:1.000	Yes:186
2:530	Median :1.000	
3: 88	Mean :1.407	
4:293	3rd Qu.:2.000	
	Max. :4.000	

3. Credit Products

The variables of this group are given in Table 3 below. They explain the nature of credits products available in the bank such as the duration and the amount of the credits.

The variable `GUARANTEES` express the guaranties of the credit products. We notice how the major part of the credits lack of guarantees.

The `PURPOSE_CREDIT` can be interpreted as the special lines of products offered by the financial institutions. This new category can be give insight if some specific credit products are more riskier than others.

4. Credit History

The co-variates available in this group can be found in Table 4. At first, this variable can be considered as the first estimator of the behavior of the customers.

The variable `HISTORY` represents the past credit behaviour and is broken down as follows: 0: no credits taken; 1: all credits at this bank paid back duly; 2: existing credits paid back duly till now; 3: delay in paying off in the past; and 4: critical account. We will discuss and describe this variable further below, because, credit history is often vital in determining credit scores in reality.

5. Response

Lastly, our outcome variable is divided into those who have good credit, representing 70% of the data, and those who have bad credit, being 30% (Table 5). That would be 700 and 300 each, for a total of 1000. This mix of data is fine for our prediction as it doesn't differ too far from a half-half split, which would be optimal. There is also a sizable amount of observations, which will allow for a more efficient prediction.

Table 5: Distribution of response (Good/Bad) applicants

	x
Bad	300
Good	700

3. Exploratory Analysis

We have chosen several graphs that we expect are quite relevant to our prediction and help us to understand more about the breakdown of those who have good credit and those who don't based on some predictor variables. We have selected only a few plots of the predictor variable as it would be difficult to show many of them, but we can still gather some insight from the below.

For example, the left panel in Figure 1 looking at the breakdown of `HISTORY`, we actually see less of a representation of those with bad credit in the bad credit history columns 3 and 4. Column 0 is no credits taken, which is logical for people with bad credit, but columns 1 and 2 actually represent good credit history. Since, HISTORY, is a fundamental variable for the analysis, from a logical perspective at least, this casts some concern on the quality that our analysis could present.
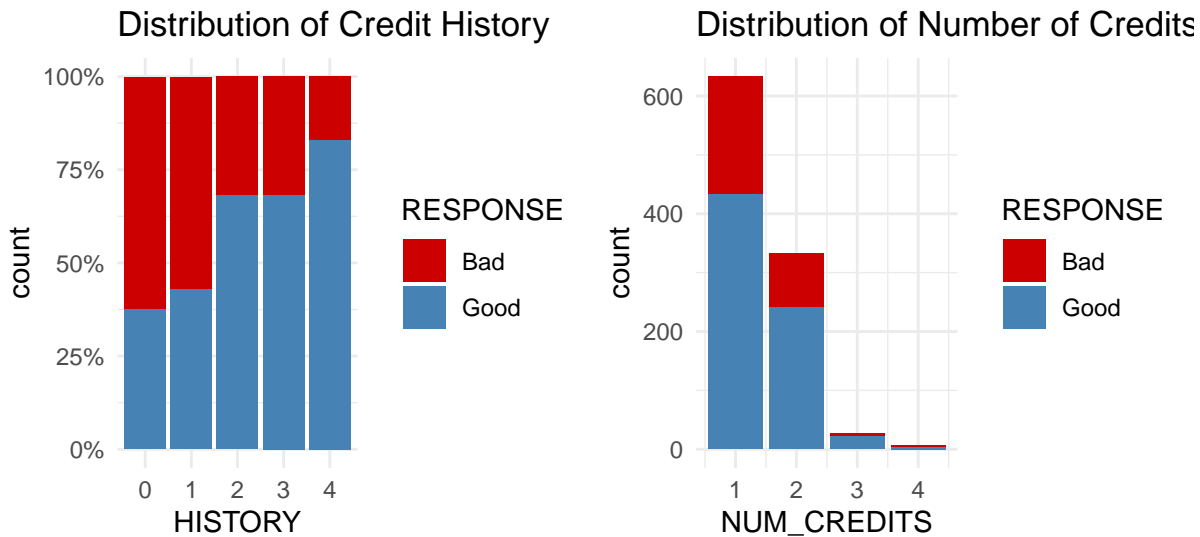

Figure 1: Important Variable Applicants - Credit History

In the right panel of Figure 1, the number of credits the customer has at the bank, does show that clients with poor credit do make up a larger proportion of those with fewer credits. This is logical as those with more credits have likely be approved more in the past for loans.

In relation with the Economic Background variables, Figure 2 reflects that people with poor credit have less money in their checking (right panel) and savings account (left panel). We can see distinctly that the proportion of those on the lower end of the spectrum who have less savings and a smaller balance on their checking account, are made up far more by people with worse credit.
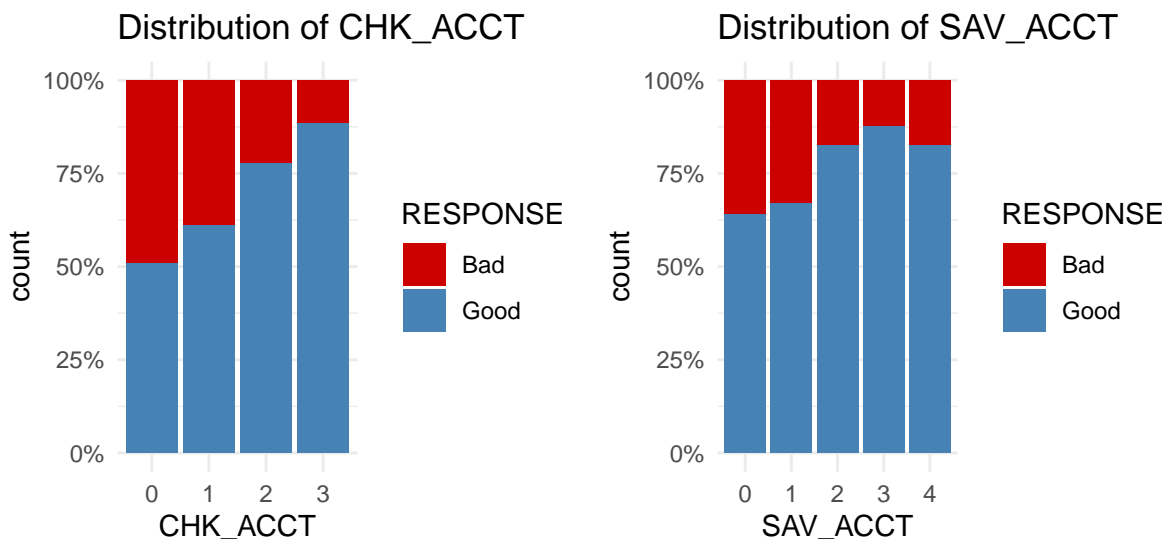
Figure 2: Important Variable Applicants Economic Background

4. Modeling Fitting

4.1 Spliting dataset training and testing

We first split the data set into testing and training using the 0.8 threshold, a fairly standard value in these types of analyses. The rather large sample size permits us to have a threshold, such as 0.8, because there remains still a large number of observations in the testing set, 200, for verification. The testing and training data is selected randomly to avoid any chance of the ordering of the original data set to create unbalanced groupings of the predictors in the training and testing set.

4.2 Prepocessing training set (Candidates models calibration)

The models below are our training models, which means this data used to create the model, is used to test it. We will then using a testing data set to check if our model is not just fitted to the data that created it.

The important outputs are in the lower left hand side and upper right hand side of the confusion matrix, which compares the predicted results, good or bad, versus the actual data. Theses values are predicted incorrectly, as a false positives, in the lower left hand side, and false negatives, in the upper right hand side. The value that is likely to be of most important for a financial institution is false positives, a person who is predicted as having good credit but in reality has bad, as they then could be lending money to someone who may not pay it back. However, there is a trade off between false positives and false negatives - but in the case of a bank, a false negative would likely be of less importance, despite losing a potentially good customer. Therefore minimizing false positives is our key goal, along with the general accuracy of the model.

It is important to note that since this is our training data, and the predicted data and the verification true values come from the same data set, it is likely to be more accurate than testing on other data, although not necessarily, as over fitting data is not the case with all models, particularly well designed ones. To avoid over fitting, a method of cross validation is used to create the model which separates, through iteration, the training data set into separate groups and tests the data against these other groups. We use cross validation in the `caret` package, which at default, is 5-fold cross validation meaning that the data is separated in 5 segments and tested against these other segments, 5 times randomly. Moreover, the optimization process in Caret focus in to optimize the sensibility of the possible fitted values, in order to detect the highest number of bad customers.

We now can go into our models:

4.2.1 Base Line Model: Logistic regression.

First, we started with a baseline logistic classification model. This model is a simple logistic regression that includes all the variables without interaction. This model will be compared to the other methods below that take more complex approaches.

The feature importance of the variables in this model is given in figure 3. We will discuss this briefly before moving on to the model's confusion matrices. We can see, for example as we discussed above, that HISTORY does not have a strong affect as the grouping, rather contradicting our intuition. It's surprisingly that credit history does not play a vital role, since this is largely what credit scores are based on in reality. The variables that did appear very significant are CHK_ACCT and, less so, SAV_ACCT where we did see distinct groups forming earlier. The most influential variables, CHK_ACCT3 and SAV_ACCT4, represented whether an observation had a relatively high balance on their savings and checkings account. This has some logic to it, that people with money in their savings or credit account, will pay their bills as they have the money to do so. It appears that INSTALL_RATE plays a role, the value of the recurrent charge for a loan as a percentage of income, and the AMOUNT, of the loan, has an important affect - indicating that the financial burden of the loan plays an important role as to whether the client will have bad or good credit. PRESENT_RESIDENT1, or if someone has been a resident for less than 1 year, has notable influence in predicting credit score, possibly indicating that scores are very sensitive when someone has less domestic credit history. In a later model, LRT Subset, we try to select variables based on significance and affect.

In Figure 4 showing the confusion matrix for the baseline logistic model, we can see that a fair number of results have been misclassified, that being 21.6% or 100% minus the accuracy shown at 78.4%. The misclassified values are in orange. This is quite high for a financial institution as a 21.6% margin for error is difficult to work around. Two other important criterion are the specificity and the sensitivity which we eluded to above. The specificity is the rate of correctly identifying true positives, or correctly identifying those who have good credit. The sensitivity is the rate of correctly identifying true negatives, or correctly identifying those who have bad credit. We can see that our specificity is fairly high, or the rate of true positives at 88.4%, however our sensitivity is quite low at 55%. Referring to our previous discussion, for a creditor the rate of sensitivity is important, as those would be customers with low credit scores being identified as having good scores is a large risk for a lender. Details of the confusion matrix are given in Figure 4. Another metric we included was the detection rate or the true positives divided by the entire sample size, or the training set size, of 800. The detection rate for this model is 16.5%, and we know that in the original data, the number of people with bad credit was 30%. Therefore, the detection rate is another way of interpreting the sensitivity rate. About 55% of the negatives were correctly identified approximately equivalent to $16.5/30 = 55\%$. The Bad Misclasification size is also included which is the number of false positives in the model.

4.2.2. Penalized Methods: Ridge regression

Second we look at Kernel Ridge Regression, or Ridge Regression for classification. Details of the confusion matrix are given in Figure 5. Kernel Ridge Regression utilizes techniques to create a ridge regression from a binary outcome variable. Ridge regression is a technique to avoid multi-collinearity, as linear estimates can be unbiased but as their variances are large, the fitted values may be far from their originals. Therefore ridge regression introduces bias to lower variance. The net effect, raising bias but lowering variance, may produce more reliable estimates. We can see in our data that this is perhaps not the effective for our data. The ridge regression, in fact, marginally decreases the accuracy from 78.4% to 78.2%, and the misclassification rate increased from 21.6% to 21.8%. Similarly, it decreases the sensitivity to 53.7% and marginally increases the specificity to 88.7%. The results are not better for our data than for logistic model as sensitivity and accuracy decrease. This is likely because variance is not a significant issue in our model, and possibly bias is, therefore trading off bias for variance does not work.
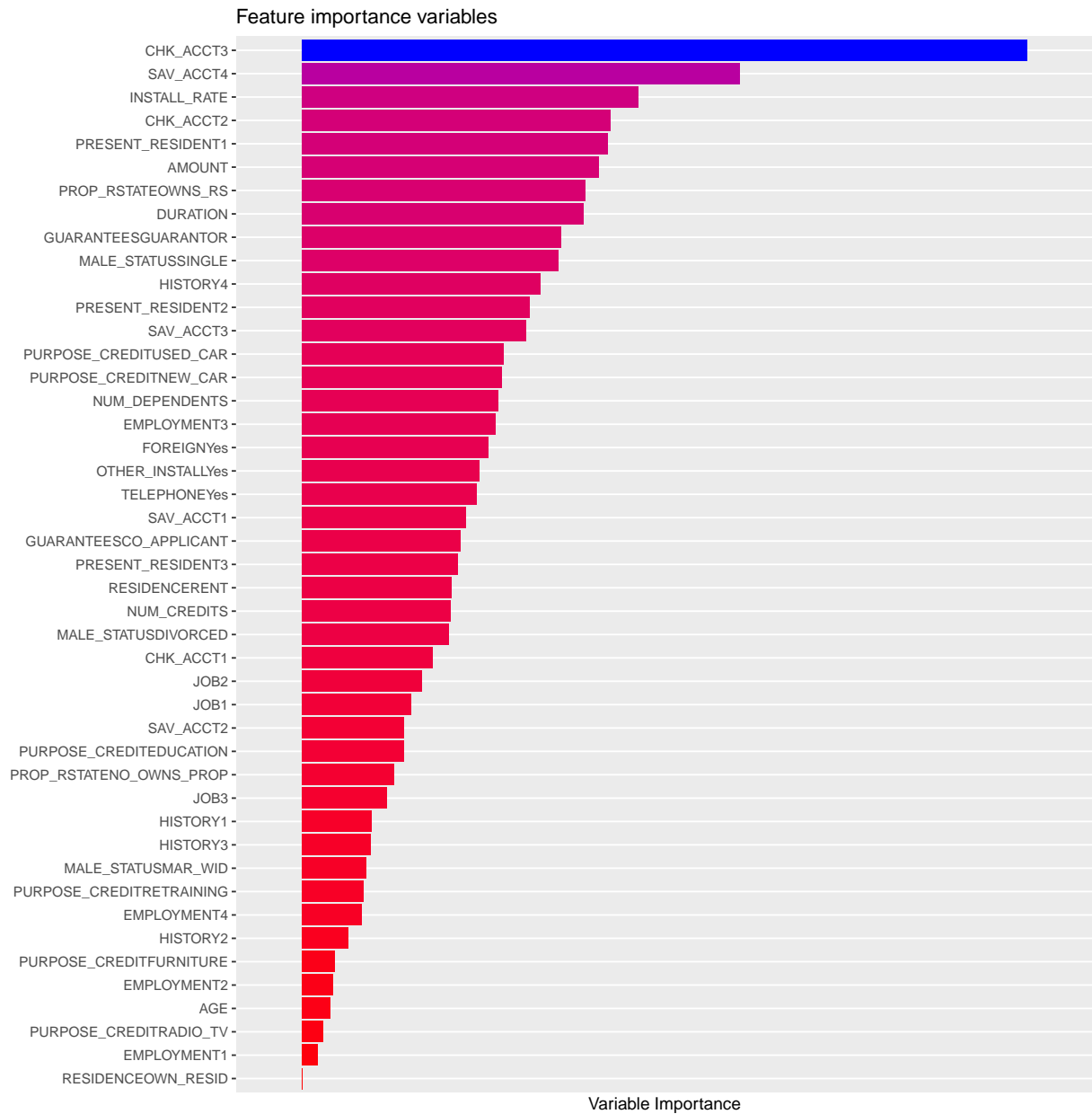
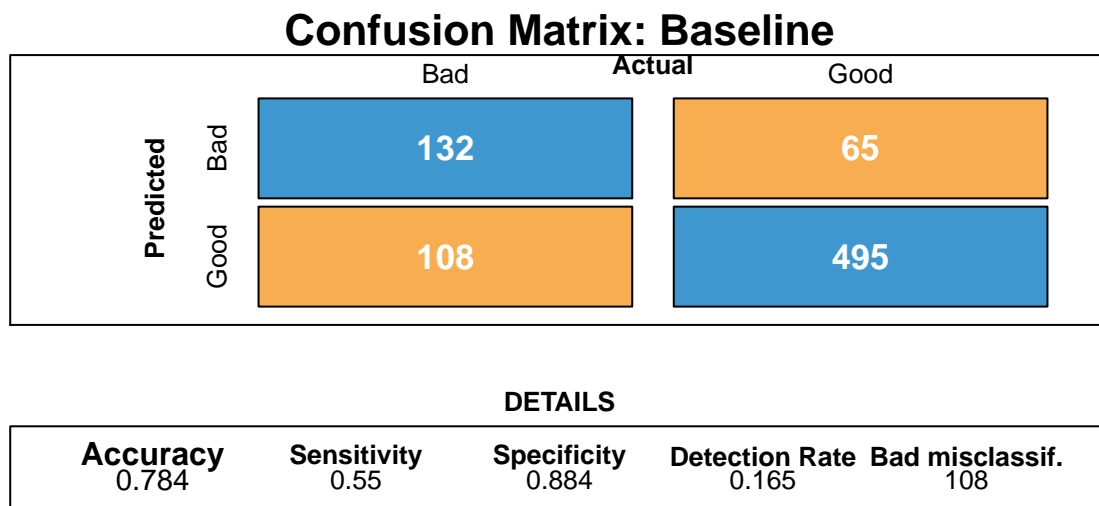Figure 3: Feature importance Base line model

Confusion Matrix: Baseline

Actual

		Bad	Good
Predicted	Bad	132	65
	Good	108	495

DETAILS

Accuracy	Sensitivity	Specificity	Detection Rate	Bad misclassif.
0.784	0.55	0.884	0.165	108

Figure 4: Confusion Matrix Baseline Model in train data

Confusion Matrix: Ridge

Actual

		Bad	Good
Predicted	Bad	129	63
	Good	111	497

DETAILS

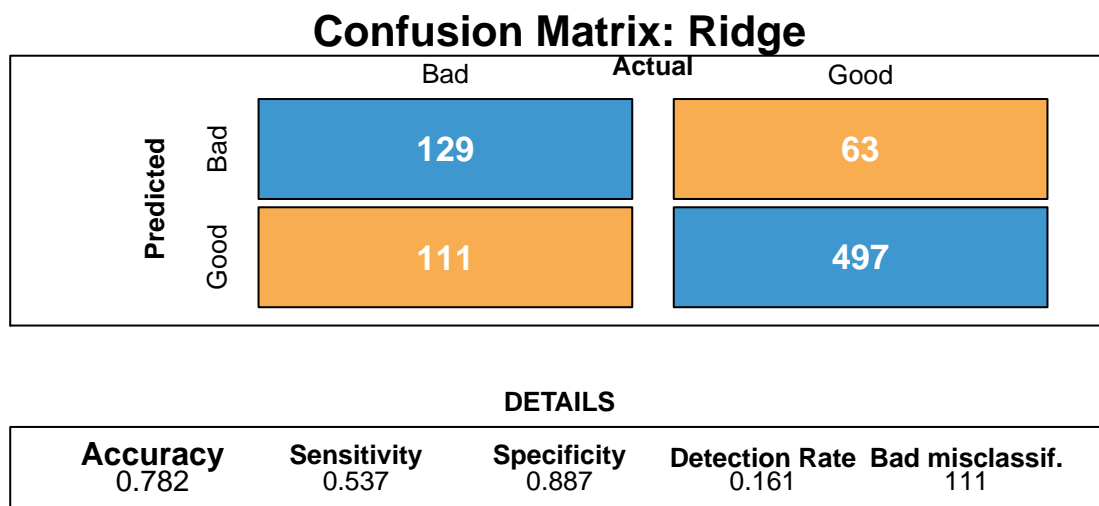| Accuracy	Sensitivity	Specificity	Detection Rate	Bad misclassif.
0.782	0.537	0.887	0.161	111

Figure 5: Confusion Matrix Ridge regression in train data

4.2.3. Best subset of predictor with LRT

Third we look at the LRT subset of predictors logistic regression model in Figure 6.

In order to the create the LRT model, before cross validation, we used training data to select our best subset of predictors. This model uses those predictors. We used an LRT method to select our predictors, which periodically selects the independent variables that have the least significant effect and then comparing the iterative models total fit and statistical significance against the model with the entire set of data. We unfortunately could not verify our results with a stepped AIC as the model that was created using the AIC criterion was not converging properly, possibly due to an unbalanced outcome variable in that model, although this is bizarre considering that the balance of 70-30 is not too extraordinary and our sample size is fairly large.

The variables that we found were most significant in the LRT model are:

RESPONSE ~ DURATION + AMOUNT + INSTALL_RATE + NUM_DEPENDENTS + CHK_ACCT + SAV_ACCT + GUARANTEES + PRESENT_RESIDENT + PROP_RSTATE + OTHER_INSTALL + RESIDENCE + TELEPHONE + AMOUNT:TELEPHONE + NUM_DEPENDENTS:TELEPHONE + IN-STALL_RATE:PRESENT_RESIDENT + INSTALL_RATE:PROP_RSTATE + NUM_DEPENDENTS:GUARANTEES

This is 9 variables of the original 20, with about 5 interactions, for a total of 16 variables. Referring to our groups of predictors, the LRT model selected variables of which 3 out 7 are from Social Background; 1 of 3 are related to Credit History; 4/5 are from Economic Background; and 5/6 are Credit Products. The only criteria in Credit Products to not be selected is the PURPOSE_CREDIT, referring to purpose of the loan, which is notable, indicating that people with bad credit do not necessarily prioritize certain payments. EMPLOYMENT, was the only factor that wasn't significant in the LRT model as part of the Economic Background, which is also surprising, as EMPLOYMENT is often a key consideration for loans.

However, selecting specific variables through LRT, has not changed the accuracy significantly from the baseline log regression, only by 1%, but specificity has now raised to 90%, a 1.6% improvement from 88.4%. Sensitivity, our key target, has unfortunately dropped by 0.4%, although this is rather marginal. In terms of our models, at least in terms of our training set, the LRT model appears to have the best results. This could indicate that reducing our variables and adding interactions helped to lower variance more than the corresponding gain in bias from removing predictors. However, the preference of LRT over our other models is minimal and verifying our models on the test data will be important.

Confusion Matrix: Subset LRT

	Actual Bad	Actual Good
Predicted Bad	131	56
Predicted Good	109	504

DETAILS

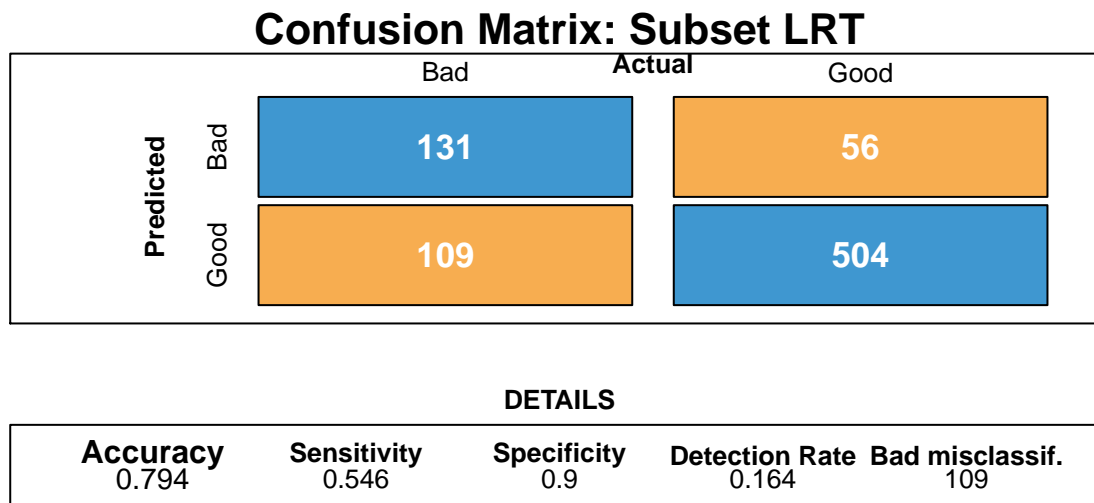| Accuracy	Sensitivity	Specificity	Detection Rate	Bad misclassif.
0.794	0.546	0.9	0.164	109

Figure 6: Confusion Matrix subser predictors in train data

4.3 Testing the models

We are now using the test data set to verify our results with a new set of data from the training set used to create the models. We'll see that the values are in total smaller, as the test data is only 200 operations, but what is significant is the proportions we discussed earlier, particularly the accuracy and the sensitivity.

4.3.1. Predictions with test data: Baseline Log

Compared to the baseline for the test data, the accuracy dropped by about 4%, specificity as well as sensitivity dropped similarly. This is to be expected as the the model was created with the train data and since now we are using test data, a different set of data we can expect less accuracy. Also, since the train data set is also smaller we can expect that to slightly effect accuracy as well. Details of the confusion matrix are given in figure 7.
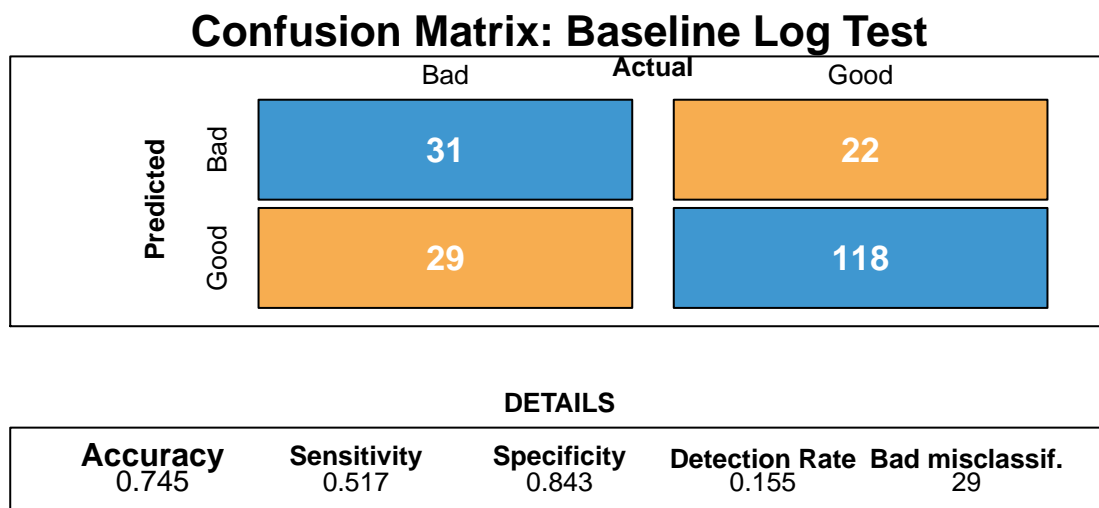
Confusion Matrix: Baseline Log Test

		Bad	**Actual**	Good
Predicted	Bad	31		22
	Good	29		118

DETAILS

Accuracy	Sensitivity	Specificity	Detection Rate	Bad misclassif.
0.745	0.517	0.843	0.155	29

Figure 7: Confusion Matrix Baseline Model in test data

4.3.2. Predictions with test data: Ridge

We can see that, similar to the baseline, sensitivity, specificity and accuracy drop between 4-5% the test data set and the train. Comparing the baseline and ridge on the test data, the ridge model appears worse, rather than equal as in the train, the misclassification rate is higher by 1% and the sensitivity lower by about 3%. Sensitivity, our most important metric, decreases for the test ridge model in comparison to the log model. Therefore the baseline log model appears superior so far on the test data. Details of the confusion matrix are given in figure 8.

4.3.3. Predictions with test data: Subset LRT

Details of the confusion matrix are given in figure 9. In comparison to the log and ridge results, the LRT subset results decrease much more substantially from its test model. The LRT sensitivity drops by about 10%, which is unfortunate for our model. This indicates the LRT subset may be overfit, as it selected specific variables, relating possibly to why its sensitivity dropped much more than other models The LRT model's specificity actually maintains the same value compared to our test results, performing better in this regard than the log and ridge results. The accuracy drops by a similar 5% overall from the test simulation like the

Confusion Matrix: Ridge Test

		Actual		
		Bad	Good	
Predicted	Bad	29	22	
	Good	31	118	

DETAILS

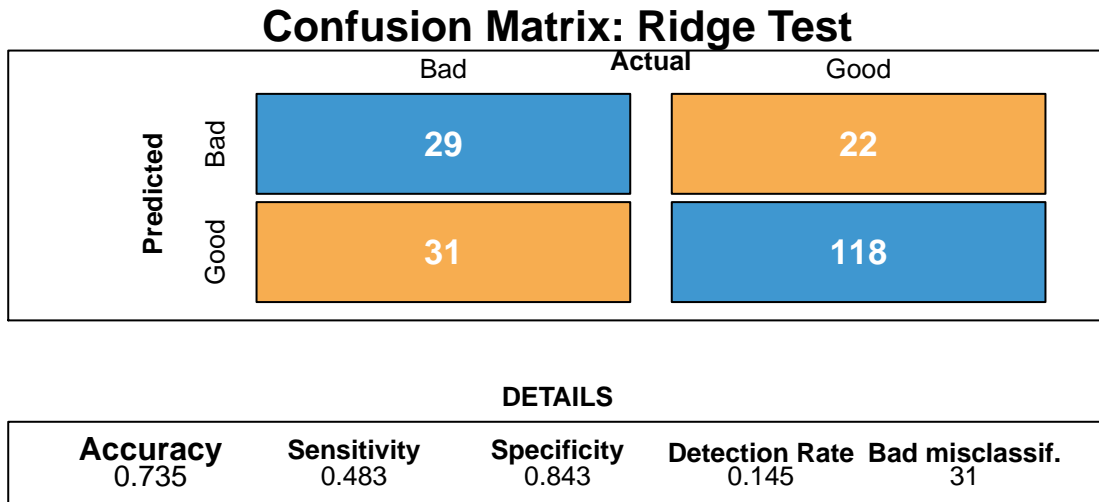| Accuracy	Sensitivity	Specificity	Detection Rate	Bad misclassif.
0.735	0.483	0.843	0.145	31

Figure 8: Confusion Matrix Ridge regression in test data

baseline and ridge models. The accuracy is slightly higher in comparison to the other models, being 0.5% greater than the logistic test model.
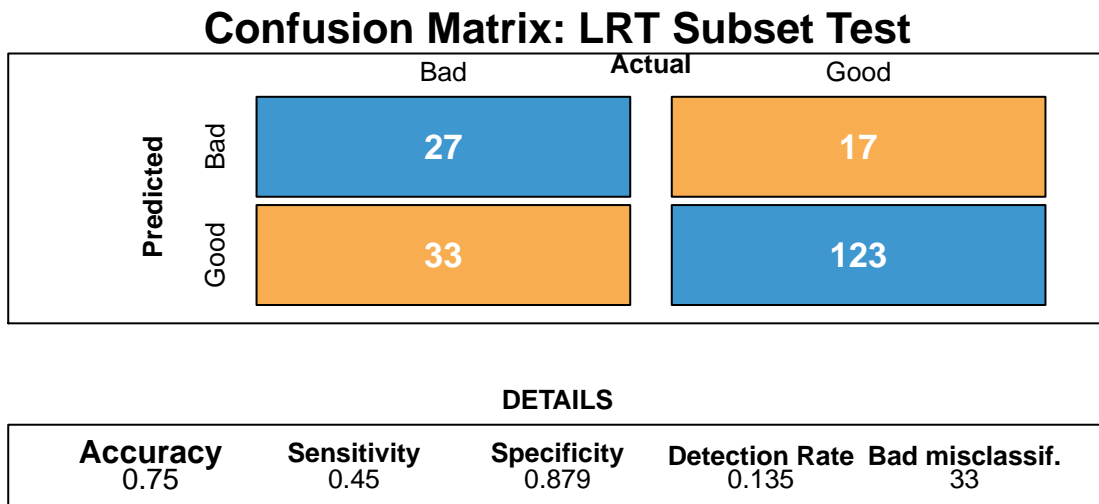
Confusion Matrix: LRT Subset Test

		Actual		
		Bad	Good	
Predicted	Bad	27	17	
	Good	33	123	

DETAILS

Accuracy	Sensitivity	Specificity	Detection Rate	Bad misclassif.
0.75	0.45	0.879	0.135	33

Figure 9: Confusion Matrix subser predictors in test data

5. Conclusions

Looking overall at our values, it appears that our best model, in terms of efficiency in predicting for the test data, is a baseline log model with no interactions or variables removed. However, it's accuracy is only 74.5% and misclassifcation rate is 25.5%. The LRT Subset model that selects predictor variables, is marginally more accurate at 75% and has a specificity 4% higher. However, the LRT test's sensitivity is almost 6% lower. Therefore, we select the baseline logistic model as sensitivity is particularly important for our model because false positives or predicting customers with bad credit as customers with good credit puts creditors at a high level of risk. This is likely worse for creditors than predicting good customers as bad, specificity in other words, because there is less risk they will lose money. They could be, however, potentially losing some profits from lost customers.

In terms of the results of our chosen model, our specificity and accuracy are not high enough and should be improved. The models, which are all Generalized Linear Models, are likely struggling to separate the groups of those with good and poor credit. A model allowing for more variance, and greater bias in effect, may be more accurate. Models from machine learning such as KNN or SVM could be good options. They would also allow for the tuning of hyper parameters allowing us to adjust the variance-bias trade off to optimize our results. Co-linearity is likely not a strong issue as the results from ridge regression do not perform exceedingly well, nor are prediction results in general influenced largely by co-linearity. Lack of data size is also likely not an issue as we already have 1000 observations. It's possible that increasing the number of poor creditors in our model could help as the step AIC analysis failed to converge, possibly indicating that lack of "1s" in the data is an issue. Adding more co-variates could work, although, this may not be realistic as there is already a number of co-variates, a total of 30 in the original set. We did notice some issues with the classes of our co-variates, particularly PROP_RSTATE and MALE_STATUS, that could help to improve our results but, as these are only two variable without large influence. This, nonetheless, will likely not have a dramatic influence.

For reference, and with the idea of understanding better the R raw code, it is available a repository in github.com/cconejov/ with all the details of the developed steps of the project.

6. Appendix

Summary of the final chosen model called *Base Line logistic*. In this case, we set seed as 42 and use the function `createDataPartition` from Caret with a proportion of `p = 0.8` for training:

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7545  -0.6304   0.3516   0.6456   2.3133
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.30416    0.11256  11.586  < 2e-16 ***
## DURATION             -0.33541    0.13127  -2.555  0.01062 *
## AMOUNT               -0.39568    0.14713  -2.689  0.00716 **
## INSTALL_RATE         -0.34772    0.11393  -3.052  0.00227 **
## AGE                   0.03080    0.11938   0.258  0.79638
## NUM_CREDITS          -0.16605    0.12311  -1.349  0.17739
## NUM_DEPENDENTS       -0.18290    0.10257  -1.783  0.07455 .
## CHK_ACCT1             0.13058    0.11017   1.185  0.23594
## CHK_ACCT2             0.29541    0.10561   2.797  0.00516 **
## CHK_ACCT3             0.84225    0.12811   6.574 4.88e-11 ***
```

```
## HISTORY1                    -0.08757    0.13851   -0.632  0.52722
## HISTORY2                     0.10452    0.24752    0.422  0.67283
## HISTORY3                     0.09606    0.15482    0.620  0.53493
## HISTORY4                     0.50446    0.23361    2.159  0.03082 *
## PURPOSE_CREDITRADIO_TV        0.03894    0.19958    0.195  0.84530
## PURPOSE_CREDITNEW_CAR        -0.33260    0.18340   -1.814  0.06975 .
## PURPOSE_CREDITUSED_CAR        0.30844    0.16850    1.831  0.06717 .
## PURPOSE_CREDITFURNITURE      -0.05246    0.17849   -0.294  0.76885
## PURPOSE_CREDITEDUCATION      -0.11416    0.12346   -0.925  0.35514
## PURPOSE_CREDITRETRAINING      0.08517    0.15246    0.559  0.57643
## SAV_ACCT1                     0.15331    0.10327    1.485  0.13764
## SAV_ACCT2                     0.09715    0.10500    0.925  0.35484
## SAV_ACCT3                     0.25195    0.12421    2.029  0.04251 *
## SAV_ACCT4                     0.47177    0.11877    3.972 7.13e-05 ***
## EMPLOYMENT1                  -0.02645    0.18589   -0.142  0.88683
## EMPLOYMENT2                   0.06382    0.22887    0.279  0.78036
## EMPLOYMENT3                   0.34611    0.19711    1.756  0.07910 .
## EMPLOYMENT4                   0.11450    0.21289    0.538  0.59067
## MALE_STATUSSINGLE             0.27691    0.11927    2.322  0.02024 *
## MALE_STATUSMAR_WID            0.06145    0.10550    0.582  0.56025
## MALE_STATUSDIVORCED          -0.12972    0.09760   -1.329  0.18379
## GUARANTEESCO_APPLICANT       -0.13249    0.09236   -1.434  0.15146
## GUARANTEESGUARANTOR           0.26544    0.11292    2.351  0.01874 *
## PRESENT_RESIDENT1            -0.44509    0.16074   -2.769  0.00562 **
## PRESENT_RESIDENT2            -0.27798    0.13448   -2.067  0.03872 *
## PRESENT_RESIDENT3            -0.24518    0.17402   -1.409  0.15886
## PROP_RSTATENO_OWNS_PROP      -0.13657    0.16362   -0.835  0.40388
## PROP_RSTATEOWNS_RS            0.29487    0.11463    2.572  0.01010 *
## OTHER_INSTALLYes             -0.15625    0.09694   -1.612  0.10699
## RESIDENCERENT                -0.28367    0.20872   -1.359  0.17411
## RESIDENCEOWN_RESID            0.00054    0.23551    0.002  0.99817
## JOB1                         -0.30977    0.31289   -0.990  0.32217
## JOB2                         -0.39618    0.36460   -1.087  0.27722
## JOB3                         -0.20851    0.27043   -0.771  0.44069
## TELEPHONEYes                  0.18112    0.11420    1.586  0.11272
## FOREIGNYes                    0.22208    0.13159    1.688  0.09147 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 977.38  on 799  degrees of freedom
## Residual deviance: 691.08  on 754  degrees of freedom
## AIC: 783.08
##
## Number of Fisher Scoring iterations: 5
```