Modelling Competition: Credit Scoring

Cesar Conejo Villalobos - Xavier Bryant (Group glmnet)

20/11/2021

1. Introduction

Talk about credit scoring in general, why it is important?

Regression models for credit scoring. Advantanges (Easy to calibrated)

Trade-off Prediction power vs statistical Inference and interpretability.

Key metrics:

- Global Accuracy (General Predicted Power of the model)
- Bad customers sensitivity: In order to lower the risk, the focus of the models must be detect the biggest possible number of bad customers
- Number of good customer mis-classification (Nominal number of good customer predicted as bad customers)

2. Description Data set

Data consists of 1000 observations of past applicant, with 70% of the data set classified as good customer and 30% as bad customers. The original dataset has 30 predictor variables, but under some tranformations the data set consists of 20 predictor variables (+ 1 columnf for the response)

Predictor variables can be classified into four groups

1. Social Background

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

2. Economic Background

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

3. Credit Products

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	NA
3rd Qu.:24.0	3rd Qu.: 3972	3rd Qu.:4.000	FURNITURE :181	NA
Max. :72.0	Max. :18424	Max. :4.000	EDUCATION : 50	NA
NA	NA	NA	RETRAINING: 97	NA

4. Credit History

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	NA
3: 88	Mean :1.407	NA
4:293	3rd Qu.:2.000	NA
NA	Max. :4.000	NA

5. Response

	x
Bad	300
Good	700

3. Exploratory Analysis

Plots that the variables are relevant

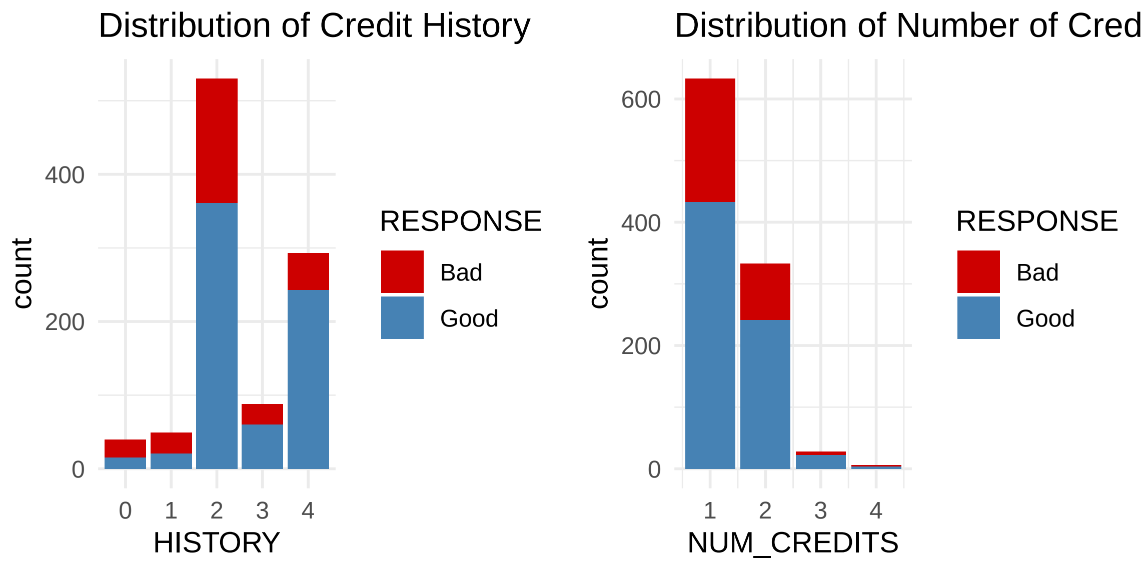


Figure 1: Important Variable Applicants credit history

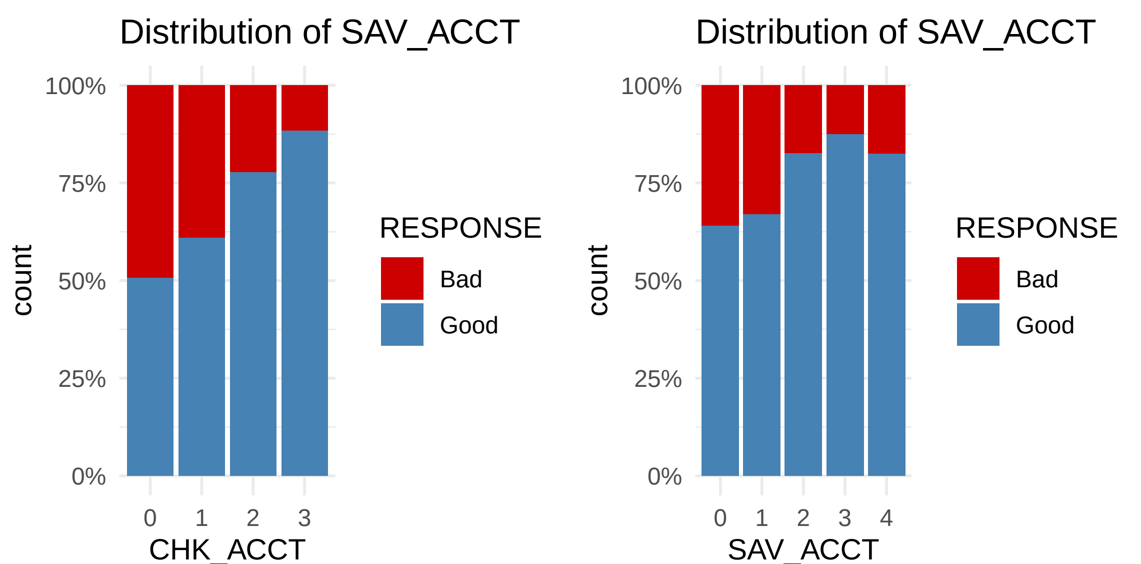


Figure 2: Important Variable Applicants Economic Backgorund

4. Modeling Fitting

4.1 Spliting dataset training and testing

```
set.seed(42)
trainRowNos <- createDataPartition(df$RESPONSE, p = 0.8, list = FALSE)
trainData <- df[trainRowNos,]
testData <- df[-trainRowNos,]
rm(df,trainRowNos)
```

4.2 Prepocessing training set (Candidates models calibration)

- Fist, using trainData for looking the best subset of predictors. LRT works, AIC did noc converge

Then, all models were calibrated using Cross-validation (Caret package)

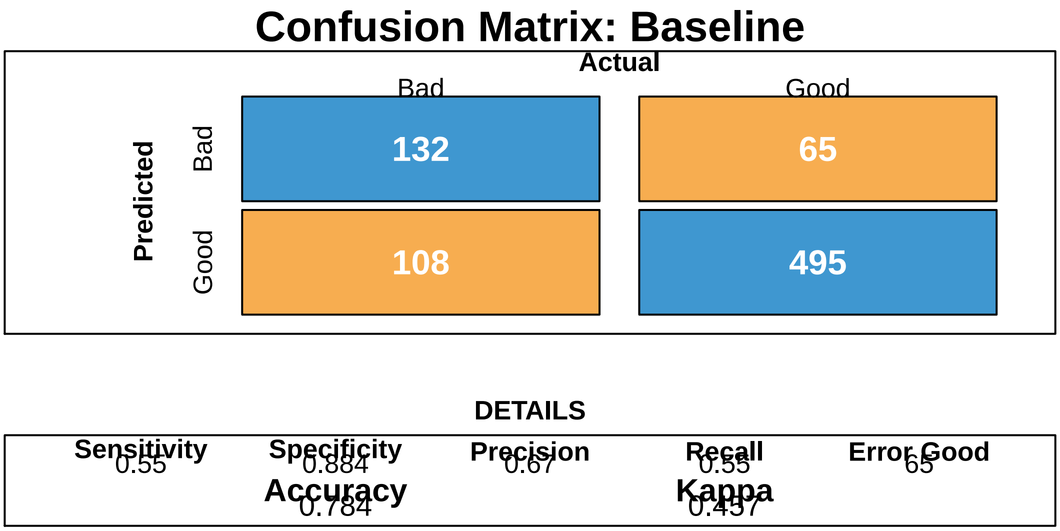
4.2.1 Base Line Model Logistic regression (simple: all variables without interactions)

4.2.2. Ridge regression Variable contributes mores to the model

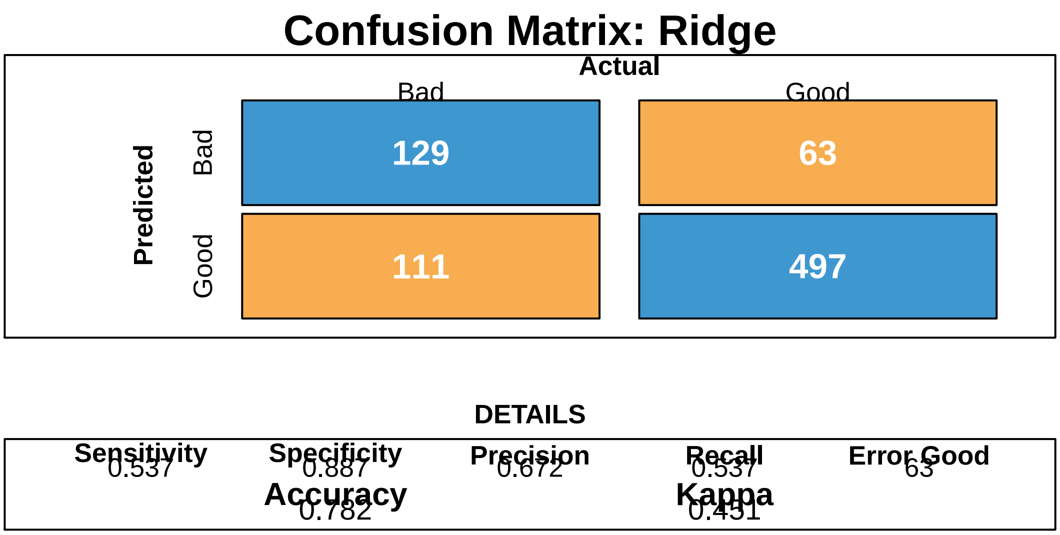
4.2.3. Best subset of predictor with lrt

```
model1_baseline <- calibrated_models[[1]]
model2_ridge <- calibrated_models[[2]]
model4_LRT <- calibrated_models[[3]]

##*******************************
# Model 1. bASElINE
##*******************************
model1_baseline.Pred.train <- predict(model1_baseline, trainData)
cm.model1_baseline.train <- confusionMatrix(model1_baseline.Pred.train, trainData$RESPONSE)

##*******************************
# Model 2. ridge
##*******************************
model2_ridge.Pred.train <- predict(model2_ridge, trainData)
cm.model2_ridge.train <- confusionMatrix(model2_ridge.Pred.train, trainData$RESPONSE)

##*******************************
## Model 4. LRT
##*******************************
model4_LRT.Pred.train <- predict(model4_LRT, trainData)
cm.model4_LRT.train <- confusionMatrix(model4_LRT.Pred.train, trainData$RESPONSE)
```

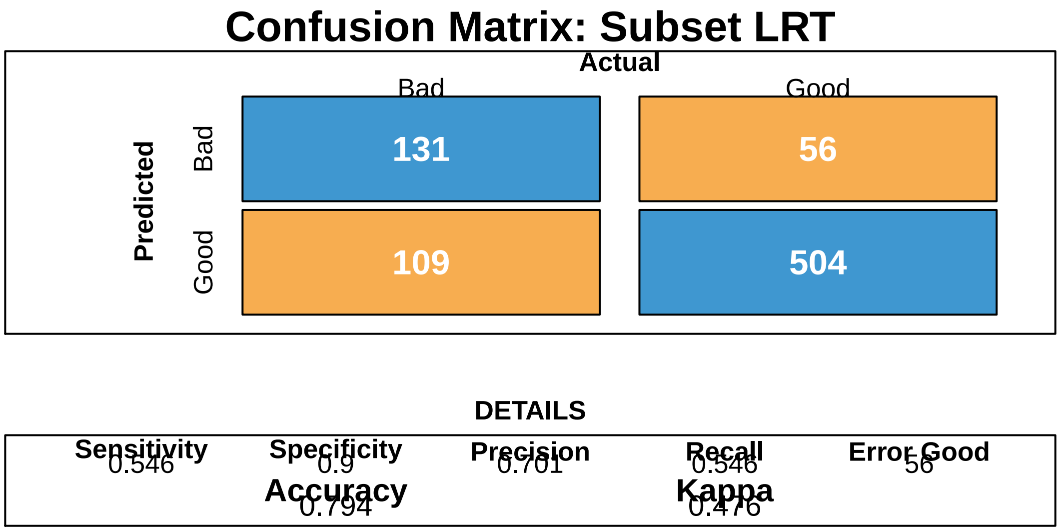
```
draw_confusion_matrix(cm = cm.model1_baseline.train, Class1 = "Bad", Class2 = "Good",title_def = 'Confu
```




```
draw_confusion_matrix(cm = cm.model2_ridge.train, Class1 = "Bad", Class2 = "Good",title_def = 'Confusion
```



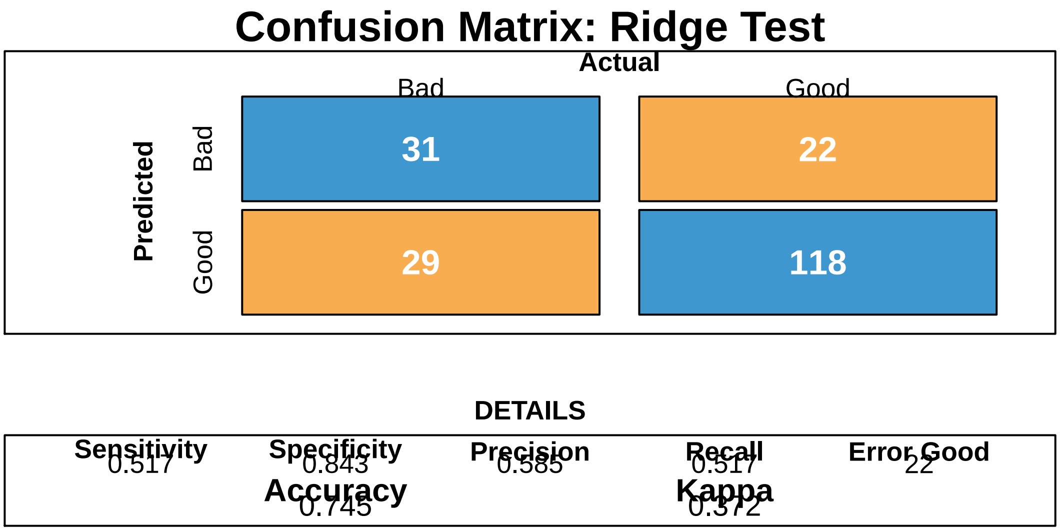

```
draw_confusion_matrix(cm = cm.model4_LRT.train, Class1 = "Bad", Class2 = "Good",title_def = 'Confusion M
```



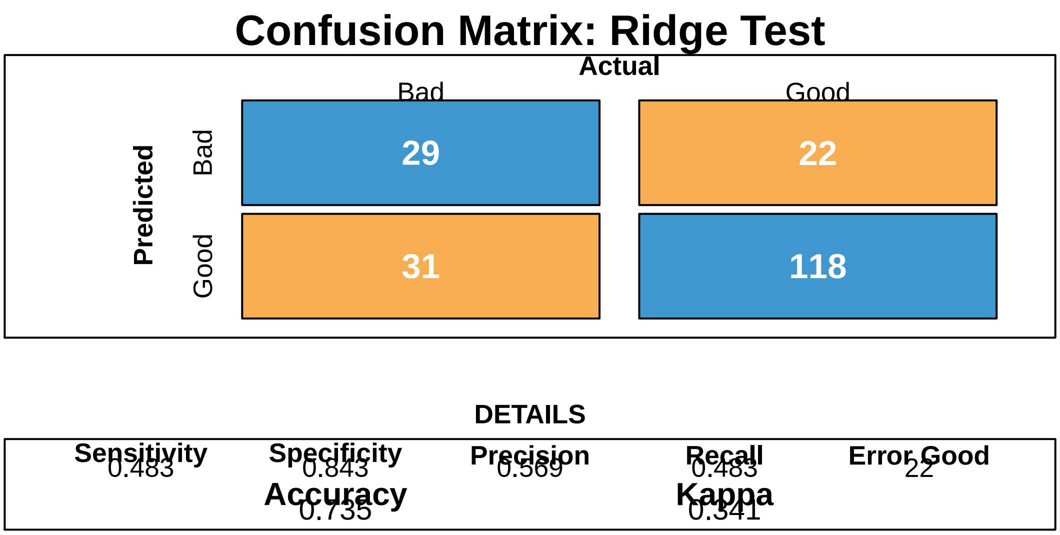
4.3 Testing the models (with testData)

4.3.1. Predictions / CM (Predicting power - Misclassificattion rate)

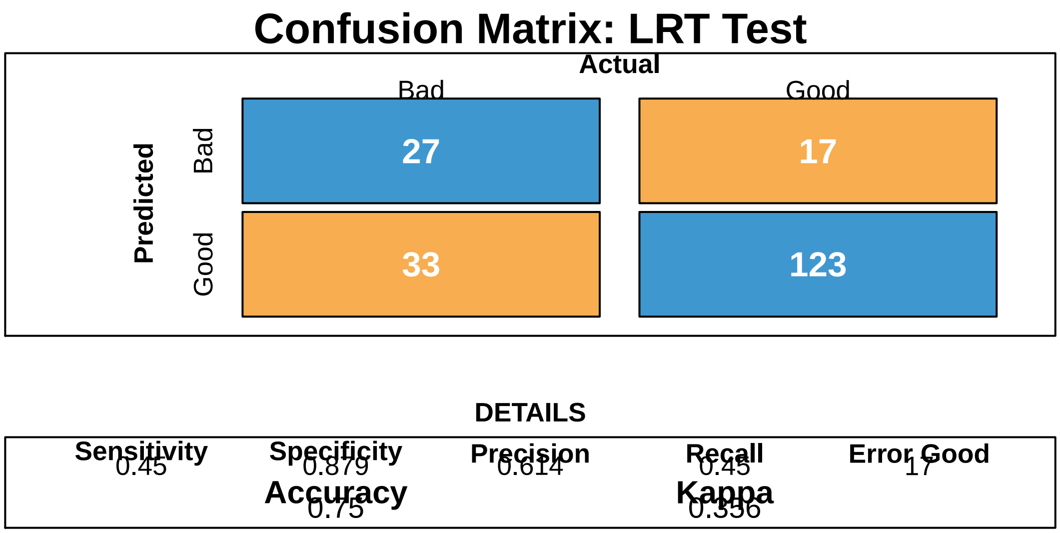
```
##********************************
## Model 1: Baseline
##********************************


model1_baseline.Pred.test <- predict(model1_baseline, testData)
cm.model1_baseline.test <- confusionMatrix(model1_baseline.Pred.test, testData$RESPONSE)


##********************************
## Model 2: Ridge
##********************************

model2_ridge.Pred.test <- predict(model2_ridge, testData)
cm.model2_ridge.test <- confusionMatrix(model2_ridge.Pred.test, testData$RESPONSE)

##********************************
# Model 4: Fitted with LRT
##********************************

model4_LRT.Pred.test <- predict(model4_LRT, testData)
cm.model4_LRT.test <- confusionMatrix(model4_LRT.Pred.test, testData$RESPONSE)

draw_confusion_matrix(cm = cm.model1_baseline.test, Class1 = "Bad", Class2 = "Good", title_def = 'Confu
```




```
draw_confusion_matrix(cm = cm.model2_ridge.test, Class1 = "Bad", Class2 = "Good", title_def = 'Confusio
```




```
draw_confusion_matrix(cm = cm.model4_LRT.test, Class1 = "Bad", Class2 = "Good", title_def = 'Confusion M
```



5. Conclusions