Universidad Carlos III de Madrid

MSc in Statistics for Data Science

Modelling competition

Regression models

Authors:

David de la Fuente López

Diego Perán Vacas

Huijin Yu

Katherine Fazioli

Adrian White

19^{th} ***January 2021***

The purpose of this assignment is to build a model to predict a binary score for a customer's credit value (good or bad) based on the provided data set. The model's intended use would be for a bank or a similar financial institution. This is an area of particular importance, since the global financial crash of 2008 was attributed in large part to mass defaults on sub-prime loans, particularly mortgage payments for housing loans. Subsequent regulations passed in the United States and internationally have imposed strict criteria on loans and it is thus in the interest of financial institutions to invest in their credit assessment methodologies.

Beyond financial regulation, it is in the interest of profit seeking financial institutions to improve their systems of credit assessment in order to expand their customer base while minimizing the risk of defaults on loans. Research through regression analysis could point to previously unused information as statistically significant for credit determination and open lending opportunities to potential customers incorrectly assessed as low credit. Thus, safely enlarging the company's customer base.

Our analysis will focus on the data set provided to build the best model for predicting a good or bad credit score. We will begin with the construction of our model, including our exploratory analysis, justification for variable selection, test for goodness of fit, etc. Since there are many possible approaches, we will defend our choice of methodology. Lastly, we will test our model's predictive power, consider cases of mis-classification, and the characteristics of our predicted values.

First, we load our data set and consider the variables provided. It is composed of 23 categorical variables (some binary) and 6 that for now are considered continuous. The last variable, binary, we will take as the response variable.

Model-building

Thus, we are going to transform the categorical variables as factors and we eliminate the first column since is simply the number of the observation.

As a first step we are going to try a Generalized Additive Models (GAMs), for if in our example we are dealing with a data set in which some predictors are non-linear.Let's check this with continuous predictors since linearity with categorical predictors is meaningless.

The problem is that most of the variables that are considered continuous do not take more than 10 values, so it makes no sense to study their linearity.

In the case of the DURATION and AGE variables, they take 33 and 53 different values correspondingly, that is, a very small number. So we decided to discretize these variables, to find the most efficient model.

Exploratory analysis

Let us try to transform into categorical variables. After several tests and a long study we came to the conclusion that the best option is to transform DURATION and AMOUNT to categorical and kept AGE as numerical. So, we are going to start with a brief visual analysis to know the distribution of these variables and their frequencies when categorizing them.

Visual analysis

AGE

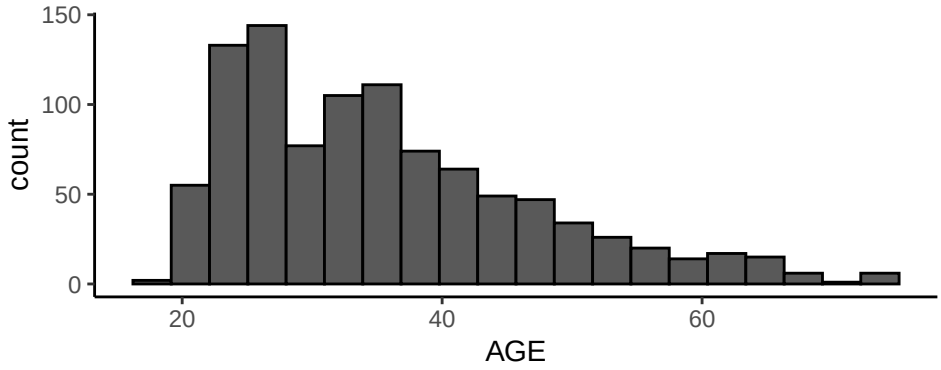


Table 1: Frequency table AGE

	Var1	Freq
1	1	240
2	2	243
3	3	260
4	4	257

DURATION

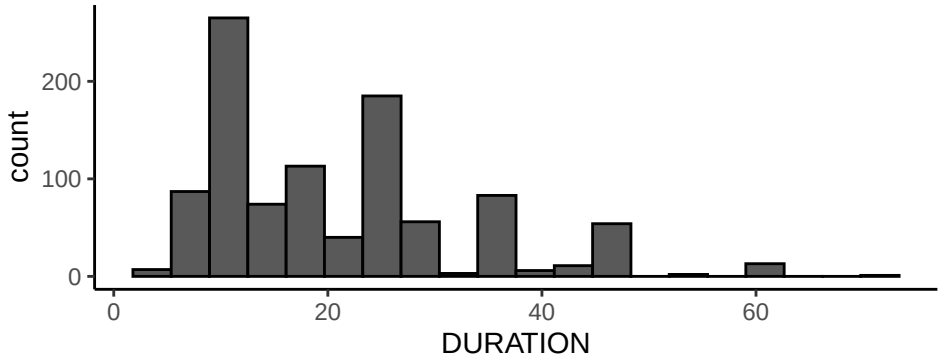


Table 2: Frequency table DURATION

	Var1	Freq
1	1	180
2	2	253
3	3	153
4	4	201
5	5	213

AMOUNT

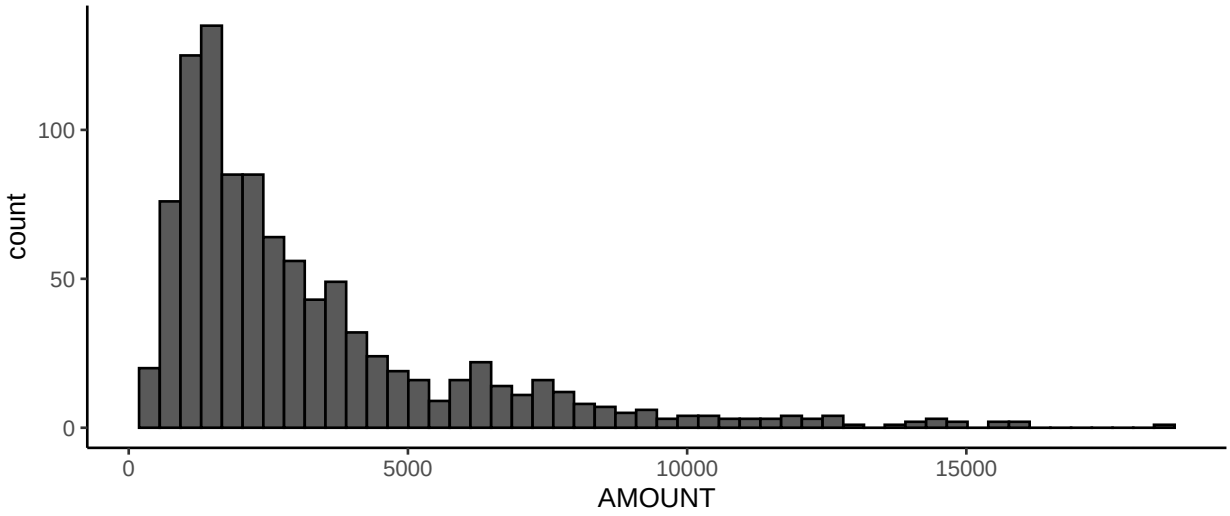


Table 3: Frequency table AMOUNT

	Var1	Freq
1	1	222
2	2	251
3	3	293
4	4	234

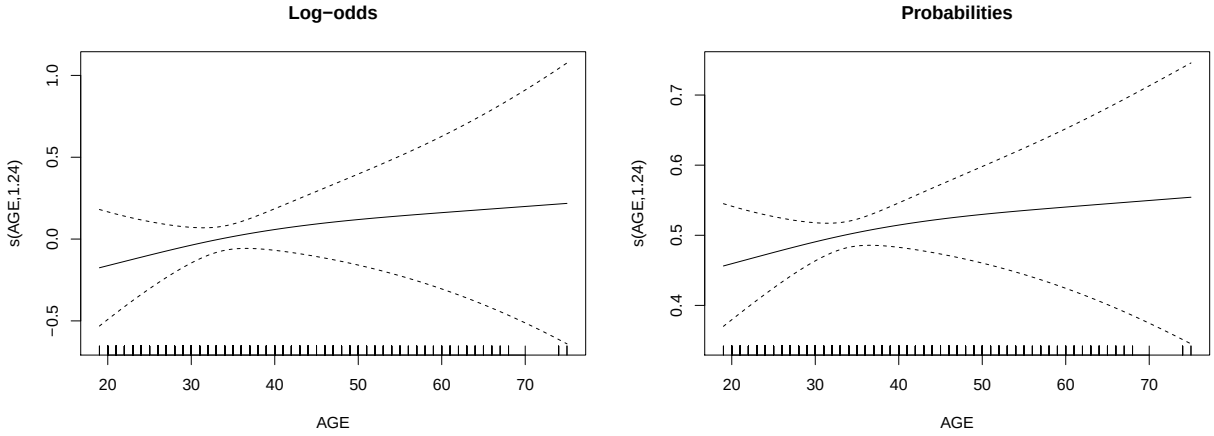
Model fitting

GAM with AGE numerical

We proof the model with all possible predictors, as we see in the code below.

```
library(mgcv)
GL_PROOF=gam(RESPONSE~s(AGE)+CHK_ACCT+HISTORY+NEW_CAR+USED_CAR+FURNITURE+
               credit$"RADIO/TV"+EDUCATION+RETRAINING+
               SAV_ACCT+EMPLOYMENT+INSTALL_RATE+MALE_DIV+MALE_SINGLE+
               credit$MALE_MAR_or_WID+credit$"CO-APPLICANT"+GUARANTOR+
               PRESENT_RESIDENT+REAL_ESTATE+PROP_UNKN_NONE+OTHER_INSTALL
             +RENT+OWN_RES+NUM_CREDITS+JOB+NUM_DEPENDENTS+TELEPHONE
             +FOREIGN+AMOUNT_cat+DURATION_cat,family = binomial,data=credit)
summary(GL_PROOF)
```

Visualizing smoothing, p value for smoothing term for age is 0.511



Second GAM with no smoothing terms (GLM)

And then another example, without any smoothing term. We omit both outputs since they have no relevance in the final conclusion of the investigation.

```
GL_PROOF_2=gam(RESPONSE~AGE+CHK_ACCT+HISTORY+NEW_CAR+USED_CAR+FURNITURE
               +credit$"RADIO/TV"+EDUCATION+RETRAINING+SAV_ACCT+EMPLOYMENT+INSTALL_RATE+
                 MALE_DIV+MALE_SINGLE+credit$MALE_MAR_or_WID+credit$"CO-APPLICANT"+GUARANTOR+
                 PRESENT_RESIDENT+REAL_ESTATE+PROP_UNKN_NONE+
                 OTHER_INSTALL+RENT+OWN_RES+NUM_CREDITS+JOB+NUM_DEPENDENTS+TELEPHONE+
                 FOREIGN+AMOUNT_cat+DURATION_cat,family = binomial,data=credit)
summary(GL_PROOF_2)
```

Comparing models, p=0.2163 We now compare both models that we have previously formulated. Since the p-value is considerably high, we reject the option to use GAM and then we will use GLM. That is, we have no evidence to reject the null hypothesis, which is the model without smoothing terms. It is encoded in R as follows: anova(GL_PROOF_2,GL_PROOF, test="Chisq").

```
## [1] 0.2162513
```

GLM with AMOUNT and DURATION as numerical variables

Now we use a new data frame to proceed (remove AMOUNT and DURATION numerical variables and AGE categorical variable).

Then we proof GLM again but calling on GLM function and new data set. We obtain the same result as before.

```
GL=glm(RESPONSE~.,data=credit_GL,family="binomial")
summary(GL)
```

GLM considering interactions

We tried to use GLM now considering all possible interactions, but it is computationally difficult. So we should somehow change our data set as it contains too many variables, from this model we can simplify the method.

```
## Warning: glm.fit: algorithm did not converge
```

Note removed the PURPOSE variable

Assess for Multi-colinearity

Multi-colinearity may not affect the bias of the model, but it significantly inflates the variance of our estimators. Since we have a number of predictors, we must test for this. We assess for multi-colinearity in our table using the condition number. Our number is very high, this indicates severe multi-colinearity. It is possible that this is due to the prevalence of dummy variables. We will nevertheless investigate this issue. The Kappa value below indicates the Condition Number.

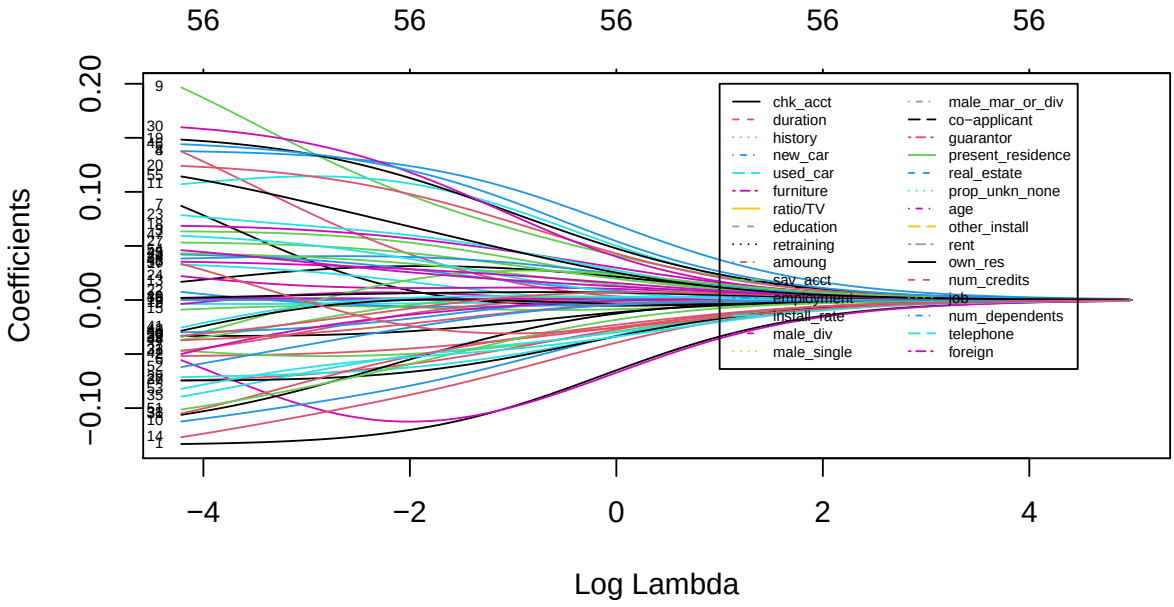
Kappa

```
## [1] 97180.13
```

VIF We use the Variance Inflation Factor, which checks for R squared values close to one, to investigate which variables may be the cause for our inflated Condition Number. We see that our VIF values are below 10, except for DURATION_cat and AGE_cat. However, we know from their construction that these are dependent on the Duration and Age variables. This may explain the high Condition Number.

```
##                       GVIF Df GVIF^(1/(2*Df))
## CHK_ACCT          1.455872  3        1.064602
## DURATION          9.318970  1        3.052699
## HISTORY           2.480530  4        1.120258
## NEW_CAR           4.137921  1        2.034188
## USED_CAR          2.335437  1        1.528214
## FURNITURE         3.842254  1        1.960167
## `RADIO/TV`        4.169996  1        2.042057
## EDUCATION         1.966915  1        1.402467
## RETRAINING        2.659513  1        1.630801
## AMOUNT            4.764540  1        2.182783
## SAV_ACCT          1.453211  4        1.047831
## EMPLOYMENT        2.798076  4        1.137254
## INSTALL_RATE      1.484448  1        1.218379
## MALE_DIV          1.228499  1        1.108377
## MALE_SINGLE       1.677438  1        1.295159
## MALE_MAR_or_WID   1.301303  1        1.140746
## `CO-APPLICANT`    1.094158  1        1.046020
## GUARANTOR         1.101730  1        1.049633
## PRESENT_RESIDENT  1.929987  3        1.115816
## REAL_ESTATE       1.274292  1        1.128846
## PROP_UNKN_NONE    3.174475  1        1.781706
## AGE               7.709814  1        2.776655
## OTHER_INSTALL     1.178211  1        1.085454
## RENT              5.247084  1        2.290651
## OWN_RES           6.610192  1        2.571029
## NUM_CREDITS       1.749203  1        1.322574
## JOB               2.577134  3        1.170908
## NUM_DEPENDENTS    1.259293  1        1.122182
## TELEPHONE         1.436172  1        1.198404
## FOREIGN           1.085546  1        1.041895
## AGE_cat          10.339940  3        1.476000
## DURATION_cat     11.806046  4        1.361486
## AMOUNT_cat        6.479290  3        1.365383
```

Ridge regression Lastly, we use the Ridge Regression to visually assess for multi-colinearity. The Ridge Regression effectively shrinks the predictors' coefficients close to each other as the ridge parameter increases. In the generated plot, te see that the majority of curves do not converge early. This confirms our findings from the VIF values that we do not have significant issues with multi-colinearity.



Variable Selection

We begin by identifying potentially important interactions between our variables to include in the model. We do so by considering all iterations of the model including two predictors and their interaction. For example, we construct a model including AGE, CHK_ACCT, and their interaction; we repeat for all possible combinations of predictors. Then, we select interactions in which appeared statistically significant in the individual models using a threshold of 0.05. We note that we did not formally compare the models by calculating the change in deviance between the models with and without the interaction terms using the likelihood ratio test. However, this crude approach allows us to efficiently identify the interactions that may be of most impact in our full model. We completed this step using the for loop provided below which was obtained and modified from a previously published code.

```
##  [1] "CHK_ACCT"          "HISTORY"           "NEW_CAR"           "USED_CAR"
##  [5] "FURNITURE"         "RADIO_TV"          "EDUCATION"         "RETRAINING"
##  [9] "SAV_ACCT"          "EMPLOYMENT"        "INSTALL_RATE"      "MALE_DIV"
## [13] "MALE_SINGLE"       "MALE_MAR_or_WID"   "CO_APPLICANT"      "GUARANTOR"
## [17] "PRESENT_RESIDENT"  "REAL_ESTATE"       "PROP_UNKN_NONE"    "AGE"
## [21] "OTHER_INSTALL"     "RENT"              "OWN_RES"           "NUM_CREDITS"
## [25] "JOB"               "NUM_DEPENDENTS"    "TELEPHONE"         "FOREIGN"
## [29] "RESPONSE"          "DURATION_cat"      "AMOUNT_cat"

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge
```

We included all of the predictors and significant interactions we detected into an initial model.

```
model_full=glm(RESPONSE ~ .+CHK_ACCT:EMPLOYMENT+CHK_ACCT:JOB+CHK_ACCT:NUM_DEPENDENTS+
                 HISTORY:PRESENT_RESIDENT+HISTORY:OWN_RES+NEW_CAR:OTHER_INSTALL+
                 RADIO_TV:EMPLOYMENT+RETRAINING:OWN_RES+SAV_ACCT:NUM_CREDITS+
                 EMPLOYMENT:PRESENT_RESIDENT+EMPLOYMENT:PROP_UNKN_NONE+
                 EMPLOYMENT:OWN_RES+INSTALL_RATE:PRESENT_RESIDENT+
                 INSTALL_RATE:NUM_CREDITS+INSTALL_RATE:NUM_DEPENDENTS+MALE_DIV:PRESENT_RESIDENT+
                 PRESENT_RESIDENT:RENT+PRESENT_RESIDENT:JOB+PROP_UNKN_NONE:OWN_RES+
                 PROP_UNKN_NONE:DURATION_cat+OTHER_INSTALL:PRESENT_RESIDENT+DURATION_cat:NEW_CAR+
                 DURATION_cat:REAL_ESTATE+DURATION_cat:JOB+DURATION_cat:NUM_DEPENDENTS+
                 DURATION_cat:AMOUNT_cat+AMOUNT_cat:NUM_CREDITS+
                 CHK_ACCT:REAL_ESTATE+CHK_ACCT:OTHER_INSTALL+CHK_ACCT:DURATION_cat+
                 HISTORY:OTHER_INSTALL+SAV_ACCT:NUM_DEPENDENTS+INSTALL_RATE:TELEPHONE+
                 PROP_UNKN_NONE:SAV_ACCT+AGE:PROP_UNKN_NONE+NUM_DEPENDENTS:TELEPHONE+
                 TELEPHONE:DURATION_cat+AMOUNT_cat:NUM_DEPENDENTS+
                 AMOUNT_cat:TELEPHONE+HISTORY:RENT+RADIO_TV:CHK_ACCT+
                 RETRAINING:AGE+SAV_ACCT:CHK_ACCT+SAV_ACCT:RETRAINING+INSTALL_RATE:OWN_RES+
                 AGE:EMPLOYMENT+RENT:RETRAINING+NUM_CREDITS:RENT+
                 TELEPHONE:FOREIGN+AMOUNT_cat:SAV_ACCT,
               family=binomial,data=credit_GL)
```

We then use stepAIC to find the model that minimizes AIC. We leave this code commented, since it is computationally very expensive. The resulting model is below.

```
model_st=glm(RESPONSE ~ CHK_ACCT + HISTORY + NEW_CAR + USED_CAR +
    RADIO_TV + RETRAINING + SAV_ACCT + EMPLOYMENT + INSTALL_RATE +
    MALE_DIV + MALE_SINGLE + MALE_MAR_or_WID + CO_APPLICANT +
    GUARANTOR + PRESENT_RESIDENT + REAL_ESTATE + PROP_UNKN_NONE +
    AGE + OTHER_INSTALL + RENT + OWN_RES + NUM_CREDITS + JOB +
    NUM_DEPENDENTS + TELEPHONE + FOREIGN + DURATION_cat + AMOUNT_cat +
    CHK_ACCT:NUM_DEPENDENTS + HISTORY:OWN_RES + NEW_CAR:OTHER_INSTALL +
    RETRAINING:OWN_RES + SAV_ACCT:NUM_CREDITS + EMPLOYMENT:PRESENT_RESIDENT +
    EMPLOYMENT:PROP_UNKN_NONE + EMPLOYMENT:OWN_RES +
    INSTALL_RATE:PRESENT_RESIDENT +
    INSTALL_RATE:NUM_CREDITS + MALE_DIV:PRESENT_RESIDENT+
    PROP_UNKN_NONE:DURATION_cat +
    PRESENT_RESIDENT:OTHER_INSTALL + NEW_CAR:DURATION_cat +
    REAL_ESTATE:DURATION_cat +
    NUM_DEPENDENTS:DURATION_cat + CHK_ACCT:REAL_ESTATE + CHK_ACCT:DURATION_cat +
    HISTORY:OTHER_INSTALL + SAV_ACCT:PROP_UNKN_NONE + NUM_DEPENDENTS:TELEPHONE +
    TELEPHONE:DURATION_cat + NUM_DEPENDENTS:AMOUNT_cat + TELEPHONE:AMOUNT_cat +
    CHK_ACCT:RADIO_TV + RETRAINING:AGE + CHK_ACCT:SAV_ACCT +
    RETRAINING:SAV_ACCT + EMPLOYMENT:AGE, family=binomial,data=credit_GL)
summary(model_st)
```

Predictive power and Miss-classification rate

First, we separate our dataset into two parts: training and test set. So that the training set represents 70 % of the total sample. We also define our final model that we have chosen for the reasons we have explained above.

In this way we use the model obtained previously, with the training set. Since the model is too long, we highlight the value of the deviance AIC, as parameters that help us to know the effectiveness of the model.

```
model_st_final=glm(RESPONSE ~ CHK_ACCT + HISTORY + NEW_CAR + USED_CAR +
    RADIO_TV + RETRAINING + SAV_ACCT + EMPLOYMENT + INSTALL_RATE +
    MALE_DIV + MALE_SINGLE + MALE_MAR_or_WID + CO_APPLICANT +
    GUARANTOR + PRESENT_RESIDENT + REAL_ESTATE + PROP_UNKN_NONE +
    AGE + OTHER_INSTALL + RENT + OWN_RES + NUM_CREDITS + JOB +
    NUM_DEPENDENTS + TELEPHONE + FOREIGN + DURATION_cat + AMOUNT_cat +
    CHK_ACCT:NUM_DEPENDENTS + HISTORY:OWN_RES + NEW_CAR:OTHER_INSTALL +
    RETRAINING:OWN_RES + SAV_ACCT:NUM_CREDITS + EMPLOYMENT:PRESENT_RESIDENT +
    EMPLOYMENT:PROP_UNKN_NONE + EMPLOYMENT:OWN_RES +
    INSTALL_RATE:PRESENT_RESIDENT +
    INSTALL_RATE:NUM_CREDITS + MALE_DIV:PRESENT_RESIDENT+
    PROP_UNKN_NONE:DURATION_cat +
    PRESENT_RESIDENT:OTHER_INSTALL + NEW_CAR:DURATION_cat +
    REAL_ESTATE:DURATION_cat +
    NUM_DEPENDENTS:DURATION_cat + CHK_ACCT:REAL_ESTATE + CHK_ACCT:DURATION_cat +
    HISTORY:OTHER_INSTALL + SAV_ACCT:PROP_UNKN_NONE + NUM_DEPENDENTS:TELEPHONE +
    TELEPHONE:DURATION_cat + NUM_DEPENDENTS:AMOUNT_cat + TELEPHONE:AMOUNT_cat +
    CHK_ACCT:RADIO_TV + RETRAINING:AGE + CHK_ACCT:SAV_ACCT +
    RETRAINING:SAV_ACCT + EMPLOYMENT:AGE, family=binomial,data=credit_GL_train)
```

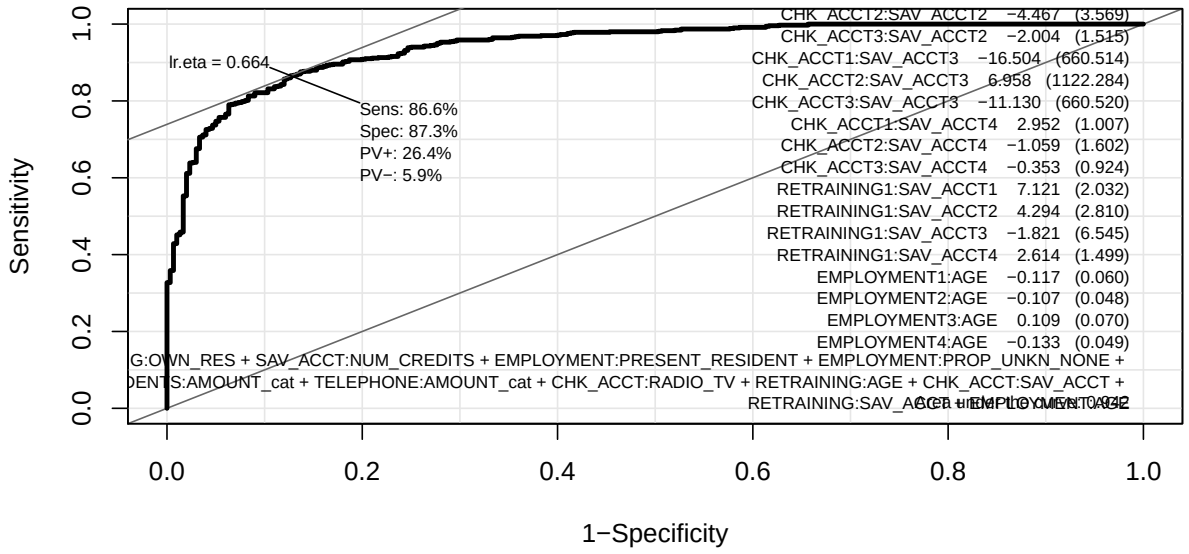
```
## [1] "AIC:616.291360026158"
```

```
## [1] "Deviance:284.291360026158"
```

So, we create predictions from our model and transform these predictions into probabilities, using the inverse of the logit function. That is $p = \frac{e^{X\beta}}{1+e^{X\beta}}$

We are going to check that cutoff is the optimal one, that is, with which a higher precision is obtained with this model. We test with all values between 0.001 and 0.999.

Before calculating the total precision of the model we are going to use the ROC curve, since it also gives us indications about the quality of the model. In addition, it also gives us a value for the cut-off that may turn out to be the optimal one.



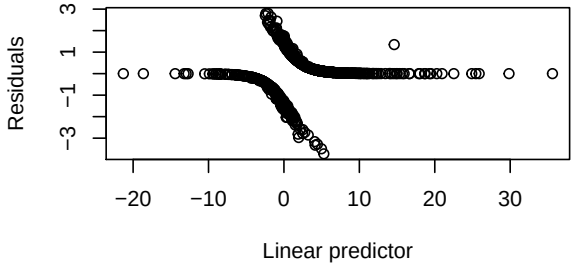
The results are very satisfactory as it achieves very high specificity and sensitivity values, as well as a 0.942 AUC. However, we use the previously estimated cut-off to check the final precision of the model.

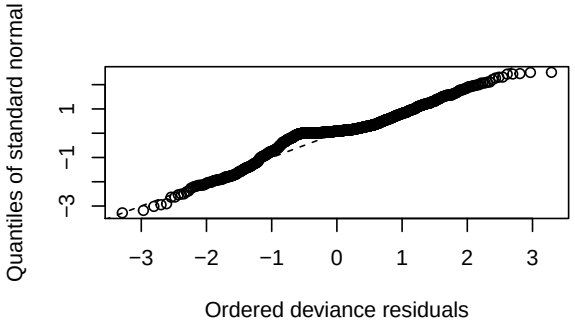
```
## [1] 0.23
```

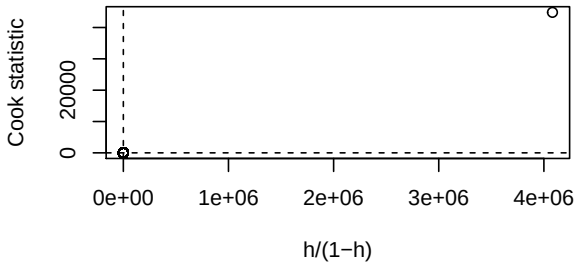
Model Diagnositcs

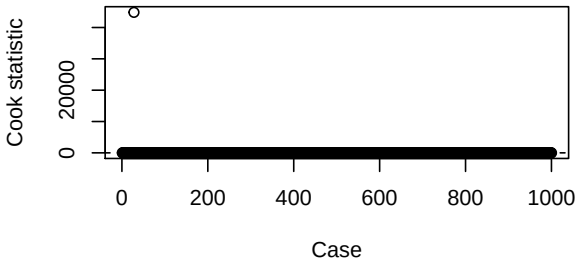
Hosmer and Lemeshow goodness of fit (GOF) test, p-value = 0.4307.

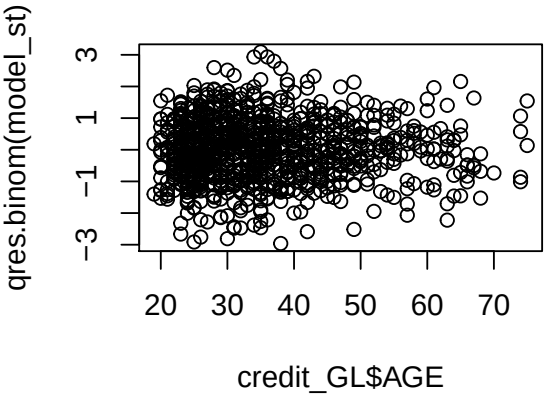
```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  credit_GL$RESPONSE, predict(model_st, type = "response")
## X-squared = 8.0288, df = 8, p-value = 0.4307
```

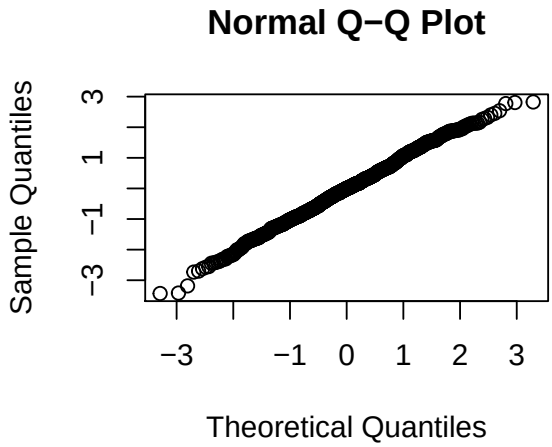












Model Characteristics

Our model is quite complex and uses a number of predictors and their interactions. Many of our predictors are categorical variables that have been converted into sets of dummy variables by the logistic regression model. We see from our coefficients that the magnitude of each of these predictors and interactions is not the same. Some of our coefficients are very close to zero, while others are above 30.

We know from logistic regression analysis that our coefficients measure the change in the log of the odds that our response variable takes a value of 1. This is best understood by the Odds Ratio, the ratio of the odds of an event occurring in one group to the odds of it recurring in another group. The Odds Ratio can be expressed as the exponential of the coefficient of the predictor in question, $OR = e^{\beta_i}$. Thus, by looking at the magnitude of the coefficients, we can assess the predictor's effect on the odds of our response.

For example, looking at the Checking Account predictor, we notice that CHK_ACCT2 has a coefficient of 10, much higher than the other Checking Account levels. This implies that high levels in checking accounts increase the odds of good credit, more so than a smaller levels. We see an even greater coefficient for SAV_ACCT3, which has a coefficient of approximately 30. Savings accounts are generally more stable and representative of the client's available financial resources, since checking accounts are typically used for transactions. A high level of savings often means a strong net worth. Our model has assigned a very strong weight to this predictor for good credit. Interestingly, the interaction term CHK_ACCT1:SAV_ACCT3 has a high negative coefficient of -20. This implies that while high savings are important, if it is accompanied by a low or absent checking account the odds of good credit will decrease.

Another strong predictor is Employment, where EMPLOYMENT2 has a coefficient of 15 and EMPLOYMENT4 of 31. This implies a significant increase in the odds of good credit given stable employment. However, Employment seems to have a significant compounded effect with Residence (PRESENT_RESIDENT), where a number of statistically significant interaction terms are included with negative coefficients. Specifically, EMPLOYMENT4, which points to employment over seven years, has very low coefficients of -23, -30, and -25 for its interactions with PRESENT_RESIDENT2, PRESENT_RESIDENT3, PRESENT_RESIDENT4, higher than the interaction with shorter term employment. Our model suggests that long term employment coupled with long term residence decrease the odds of good credit.

Another interesting predictor are applications for the purpose of Retraining (RETRAINING), which on its own has a coefficient of -3, thus lowering the odds of good credit. However, when interacting with medium level savings accounts (RETRAINING1:SAV_ACCT1 and RETRAINING1:SAV_ACCT2), we see high interaction coefficients of 11 and 21. Our model suggests that taking a loan for the purpose of retraining will on its own lower the odds of good credit; however, with decent savings this will contribute to an increase in the odds.

These are some of the most prominent predictors and interactions in our model, but there are more and we will not detail them all. It is a complex model, but compensates for this in terms of its predictive accuracy.