# 2.German credit data

## 2.1 Executive summary

Aim: To design a model to predict the defaulters for a credit company

Approach:

In order to classify customers as defaulters or non-defaulters based on factors such as income levels, job, residence, property, we used logistic regression approach. Before predicting however, we performed EDA on the data and tried various models by using the step(). After getting a good model, we then changed the cost metrics for the False Positive and False Negative classifications according to the ratio of 5:1. This is the asymmetric cost method. This is done so that our classification of defaulters as non- defaulters goes down significantly. After selecting a model, we then use the test data for prediction

Results:

One of the major observations we get is that the FNR reduces by using an asymmetric cost value. The following table compares the in sample and out of sample evaluation metrics:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **AUC** | **Symmetric MR** | **Asymmetric MR** | **FPR(p=1/6)** | **FNR(p=1/6)** |
| **In Sample** | 0.83 | 0.253 | 0.344 | 0.431 | 0.132 |
| **Out of sample** | 0.777 | 0.292 | 0.408 | 0.185 | 0.132 |
| Table 2.1: Comparison of in sample and out of sample evaluation metrics | | | | | |

The misclassification rate using the step technique (0.25) is lower than the full model MR(0.3). This is due to the fact that some variables with low predictive power were dropped from the full model to get an AIC of 738.81

## 2.2 Analysis

The German credit dataset consists of information about 1000 customers who have been classified based on their propensity to default on their loans. We are trying to model this prediction using logistic regression in this problem. We load the data at look at the summary table for this

## chk\_acct duration credit\_his purpose amount   
## A11:274 Min. : 4.0 A30: 40 A43 :280 Min. : 250   
## A12:269 1st Qu.:12.0 A31: 49 A40 :234 1st Qu.: 1366   
## A13: 63 Median :18.0 A32:530 A42 :181 Median : 2320   
## A14:394 Mean :20.9 A33: 88 A41 :103 Mean : 3271   
## 3rd Qu.:24.0 A34:293 A49 : 97 3rd Qu.: 3972   
## Max. :72.0 A46 : 50 Max. :18424   
## (Other): 55   
## saving\_acct present\_emp installment\_rate sex other\_debtor  
## A61:603 A71: 62 Min. :1.000 A91: 50 A101:907   
## A62:103 A72:172 1st Qu.:2.000 A92:310 A102: 41   
## A63: 63 A73:339 Median :3.000 A93:548 A103: 52   
## A64: 48 A74:174 Mean :2.973 A94: 92   
## A65:183 A75:253 3rd Qu.:4.000   
## Max. :4.000   
##   
## present\_resid property age other\_install housing   
## Min. :1.000 A121:282 Min. :19.00 A141:139 A151:179   
## 1st Qu.:2.000 A122:232 1st Qu.:27.00 A142: 47 A152:713   
## Median :3.000 A123:332 Median :33.00 A143:814 A153:108   
## Mean :2.845 A124:154 Mean :35.55   
## 3rd Qu.:4.000 3rd Qu.:42.00   
## Max. :4.000 Max. :75.00   
##   
## n\_credits job n\_people telephone foreign response  
## Min. :1.000 A171: 22 Min. :1.000 A191:596 A201:963 0:700   
## 1st Qu.:1.000 A172:200 1st Qu.:1.000 A192:404 A202: 37 1:300   
## Median :1.000 A173:630 Median :1.000   
## Mean :1.407 A174:148 Mean :1.155   
## 3rd Qu.:2.000 3rd Qu.:1.000   
## Max. :4.000 Max. :2.000   
##

Fig 2.1: Summary of the credit data

The data contains 21 attributes including the response variable and there are no null values in the dataset. Let us look at the individual distributions:

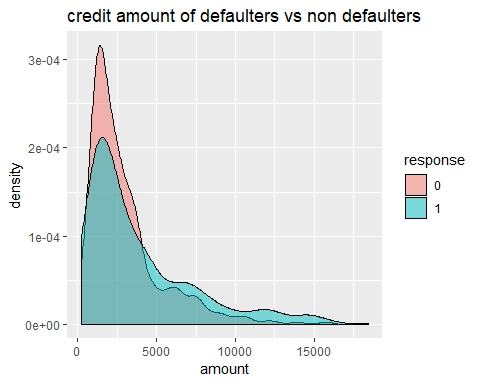
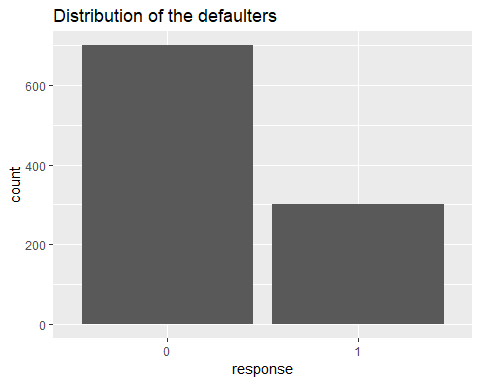


Fig 2.3: distribution of the credit amount variable

Fig 2.2: Histogram of the response variable

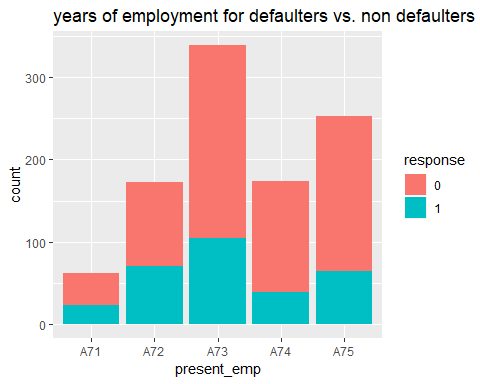
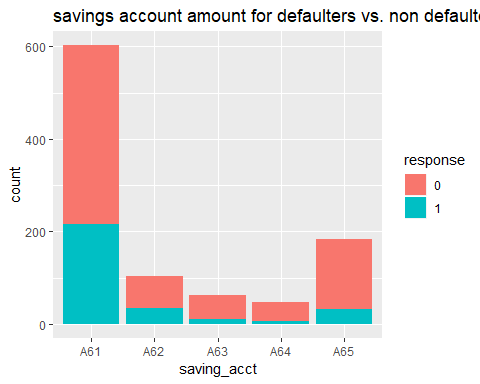


Fig 2.5: Histogram of the years of employment variable

Fig 2.4: Histogram of the savings amount variable

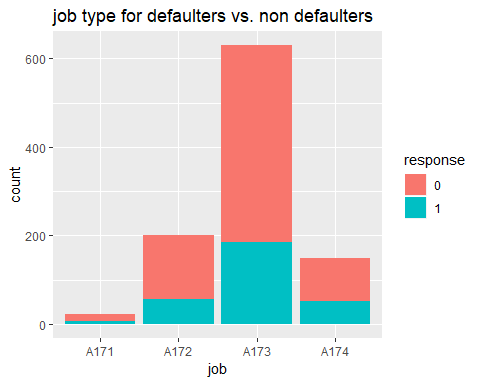
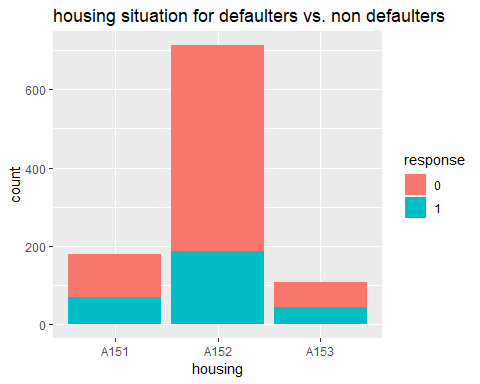


Fig 2.6: Histograms of the housing situation and job type variables variable

The data has 700 non-defaulters and 300 defaulters. Some observations from the data: 1. The defaulters seem to borrow a slightly higher amount than non-defaulters 2. Defaulters seem to have a savings account with a low balance 3. Most defaulters have been employeed for 1 to 7 years 4. Most defaulters, own their houses 5. Deaulters generally are unskilled residents or skilled employees

Let us split the data into training and testing sets and fit a logistic model on it:

##   
## Call:  
## glm(formula = response ~ ., family = binomial, data = credit.train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.5173 -0.6783 -0.3371 0.6603 2.4331   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 9.522e-01 1.222e+00 0.779 0.435725   
## chk\_acctA12 -4.628e-01 2.557e-01 -1.810 0.070367 .   
## chk\_acctA13 -1.455e+00 4.806e-01 -3.026 0.002474 \*\*   
## chk\_acctA14 -1.770e+00 2.748e-01 -6.442 1.18e-10 \*\*\*  
## duration 2.959e-02 1.119e-02 2.644 0.008191 \*\*   
## credit\_hisA31 -1.121e-01 6.228e-01 -0.180 0.857128   
## credit\_hisA32 -5.900e-01 5.019e-01 -1.176 0.239765   
## credit\_hisA33 -1.142e+00 5.465e-01 -2.089 0.036699 \*   
## credit\_hisA34 -1.832e+00 5.210e-01 -3.517 0.000437 \*\*\*  
## purposeA41 -1.807e+00 4.441e-01 -4.068 4.74e-05 \*\*\*  
## purposeA410 -2.299e+00 1.061e+00 -2.167 0.030266 \*   
## purposeA42 -6.490e-01 3.126e-01 -2.076 0.037894 \*   
## purposeA43 -9.449e-01 2.905e-01 -3.253 0.001143 \*\*   
## purposeA44 -1.368e+00 1.022e+00 -1.339 0.180453   
## purposeA45 -6.235e-01 7.034e-01 -0.886 0.375369   
## purposeA46 -1.134e-01 4.565e-01 -0.248 0.803770   
## purposeA48 -2.034e+00 1.242e+00 -1.637 0.101596   
## purposeA49 -6.358e-01 3.875e-01 -1.641 0.100826   
## amount 1.230e-04 5.015e-05 2.452 0.014202 \*   
## saving\_acctA62 -4.488e-01 3.453e-01 -1.300 0.193706   
## saving\_acctA63 -2.410e-03 4.423e-01 -0.005 0.995652   
## saving\_acctA64 -1.130e+00 6.045e-01 -1.869 0.061600 .   
## saving\_acctA65 -1.012e+00 3.048e-01 -3.320 0.000901 \*\*\*  
## present\_empA72 1.481e-01 5.183e-01 0.286 0.775148   
## present\_empA73 -1.657e-01 4.975e-01 -0.333 0.739110   
## present\_empA74 -8.216e-01 5.426e-01 -1.514 0.129966   
## present\_empA75 -2.717e-01 5.000e-01 -0.543 0.586798   
## installment\_rate 3.133e-01 1.034e-01 3.031 0.002438 \*\*   
## sexA92 -2.862e-01 4.261e-01 -0.672 0.501766   
## sexA93 -9.392e-01 4.130e-01 -2.274 0.022949 \*   
## sexA94 -3.682e-01 5.081e-01 -0.725 0.468616   
## other\_debtorA102 1.700e-01 4.981e-01 0.341 0.732946   
## other\_debtorA103 -9.009e-01 5.356e-01 -1.682 0.092540 .   
## present\_resid -3.727e-02 1.019e-01 -0.366 0.714630   
## propertyA122 2.506e-01 3.153e-01 0.795 0.426752   
## propertyA123 3.353e-01 2.769e-01 1.211 0.225888   
## propertyA124 6.897e-01 4.885e-01 1.412 0.157999   
## age -1.384e-02 1.090e-02 -1.269 0.204476   
## other\_installA142 -1.981e-01 4.624e-01 -0.429 0.668287   
## other\_installA143 -8.100e-01 2.801e-01 -2.891 0.003838 \*\*   
## housingA152 -5.698e-01 2.775e-01 -2.053 0.040058 \*   
## housingA153 -4.067e-01 5.431e-01 -0.749 0.453968   
## n\_credits 4.806e-01 2.264e-01 2.123 0.033782 \*   
## jobA172 1.263e-01 7.605e-01 0.166 0.868056   
## jobA173 3.736e-01 7.232e-01 0.517 0.605386   
## jobA174 1.552e-01 7.220e-01 0.215 0.829805   
## n\_people 1.770e-01 3.005e-01 0.589 0.555930   
## telephoneA192 -3.156e-01 2.396e-01 -1.317 0.187766   
## foreignA202 -1.077e+00 6.500e-01 -1.658 0.097365 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 912.88 on 749 degrees of freedom  
## Residual deviance: 653.23 on 701 degrees of freedom  
## AIC: 751.23  
##   
## Number of Fisher Scoring iterations: 5

Fig 2.7: Summary of the fill model (Logistic Regression)

The summary shows that the model Residual deviance value is 659.42 and the AIC value is 757.42. We can compare different models against this full model and select the variables that can be included in the analysis.

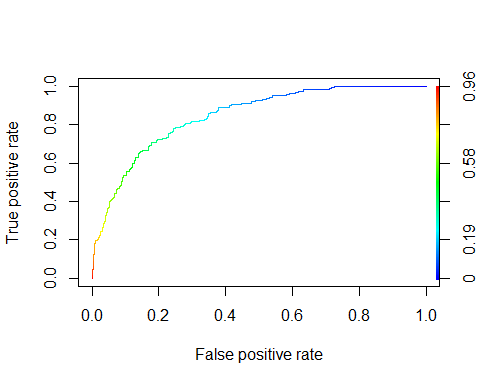


Fig 2.8: ROC curve of the full model

The ROC curve of the full model shows a rapid increase, which indicates a high predictive power of the model. The AUC value of the full model is 0.8443342. Since this is an imbalanced dataset, let us look at the Precision - Recall curve, as it gives us a better idea about the classification

## [1] 0.8443342

##   
## Precision-recall curve  
##   
## Area under curve (Integral):  
## 0.6852695   
##   
## Area under curve (Davis & Goadrich):  
## 0.6849192   
##   
## Curve for scores from 0.003078127 to 0.9579343   
## ( can be plotted with plot(x) )

Fig 2.9: Precision – Recall auc values

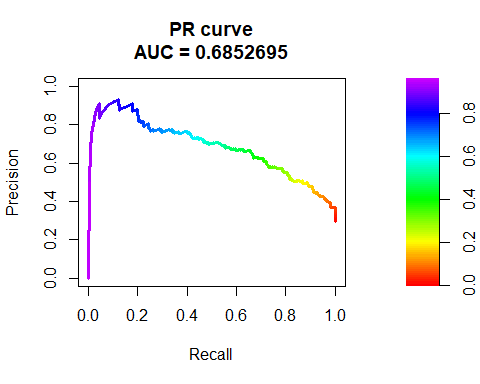


Fig 2.10: P-R curve for the full model

The above graph shows a predictive ability of the full model is only slightly better at identifying the defaulters than null models. The AUC has reduced to 0.68.

Let us take a look at the confusion matrix and the symmetric MR of the full model using the Naive cutoff of 0.29

## Predicted  
## True 0 1  
## 0 396 131  
## 1 49 174

The model has a high false positive value and a misclassification rate of 0.248, we can reduce it by either finding a better model using techniques such as step and then applying different weights to the misclassifications

##   
## Call:  
## glm(formula = response ~ chk\_acct + duration + credit\_his + purpose +   
## amount + saving\_acct + present\_emp + installment\_rate + sex +   
## age + other\_install + housing + n\_credits + foreign, family = binomial,   
## data = credit.train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.4216 -0.7008 -0.3505 0.6906 2.5011   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.455e+00 1.037e+00 1.402 0.160787   
## chk\_acctA12 -5.151e-01 2.500e-01 -2.060 0.039372 \*   
## chk\_acctA13 -1.472e+00 4.718e-01 -3.120 0.001806 \*\*   
## chk\_acctA14 -1.756e+00 2.692e-01 -6.522 6.93e-11 \*\*\*  
## duration 3.258e-02 1.082e-02 3.012 0.002592 \*\*   
## credit\_hisA31 -1.641e-01 6.088e-01 -0.270 0.787543   
## credit\_hisA32 -6.141e-01 4.988e-01 -1.231 0.218311   
## credit\_hisA33 -1.108e+00 5.456e-01 -2.030 0.042337 \*   
## credit\_hisA34 -1.885e+00 5.196e-01 -3.628 0.000285 \*\*\*  
## purposeA41 -1.866e+00 4.338e-01 -4.301 1.70e-05 \*\*\*  
## purposeA410 -2.374e+00 1.016e+00 -2.336 0.019475 \*   
## purposeA42 -6.580e-01 3.053e-01 -2.155 0.031131 \*   
## purposeA43 -1.008e+00 2.853e-01 -3.532 0.000412 \*\*\*  
## purposeA44 -1.216e+00 9.951e-01 -1.222 0.221771   
## purposeA45 -5.742e-01 6.706e-01 -0.856 0.391926   
## purposeA46 -6.663e-02 4.515e-01 -0.148 0.882699   
## purposeA48 -1.915e+00 1.197e+00 -1.600 0.109527   
## purposeA49 -7.148e-01 3.775e-01 -1.894 0.058265 .   
## amount 1.146e-04 4.624e-05 2.479 0.013168 \*   
## saving\_acctA62 -2.689e-01 3.316e-01 -0.811 0.417346   
## saving\_acctA63 5.816e-02 4.346e-01 0.134 0.893534   
## saving\_acctA64 -1.023e+00 5.790e-01 -1.766 0.077344 .   
## saving\_acctA65 -9.764e-01 2.972e-01 -3.286 0.001017 \*\*   
## present\_empA72 2.747e-01 4.442e-01 0.618 0.536384   
## present\_empA73 -8.870e-02 4.058e-01 -0.219 0.827000   
## present\_empA74 -7.665e-01 4.601e-01 -1.666 0.095722 .   
## present\_empA75 -1.778e-01 4.256e-01 -0.418 0.676124   
## installment\_rate 3.214e-01 1.002e-01 3.208 0.001336 \*\*   
## sexA92 -3.400e-01 4.163e-01 -0.817 0.414183   
## sexA93 -9.622e-01 4.011e-01 -2.399 0.016439 \*   
## sexA94 -5.437e-01 4.967e-01 -1.095 0.273664   
## age -1.755e-02 1.043e-02 -1.682 0.092529 .   
## other\_installA142 -1.794e-01 4.579e-01 -0.392 0.695202   
## other\_installA143 -7.685e-01 2.726e-01 -2.819 0.004812 \*\*   
## housingA152 -5.460e-01 2.617e-01 -2.086 0.036965 \*   
## housingA153 -5.323e-04 4.018e-01 -0.001 0.998943   
## n\_credits 4.711e-01 2.197e-01 2.144 0.032004 \*   
## foreignA202 -1.173e+00 6.451e-01 -1.818 0.069052 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 912.88 on 749 degrees of freedom  
## Residual deviance: 662.81 on 712 degrees of freedom  
## AIC: 738.81  
##   
## Number of Fisher Scoring iterations: 5

Fig 2.11: summary of the step model

The AIC value of the step model (738.81) is lower than that of full model, therefore, we select the step model as our final model

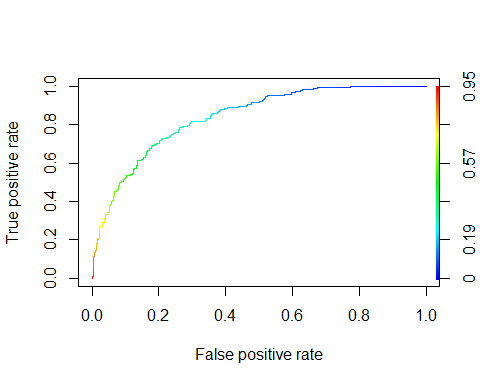
Testing this model for in sample misclassification rate: 

Fig 2.12: ROC curve for the final model(above)

The step model does seem to perform any better than the full model, it in fact gives a lower MR of 0.25. We will therefore stick to the step model for our logistic regression problem.

We can approach this problem by focusing on Precision/ Recall values instead of accuracy. We do this by assigning different weights to false positive and false negative classifications and thus calculating an asymmetric cost. For our case, the weights we assign are: weight for misclassifying a 1: 5 weight for misclassifying a 0: 1 This is equivalent to a cut off of 1/6

We now compare the train and test misclassification rates and model accuracies:

## Predicted  
## True 0 1  
## 0 292 235  
## 1 23 200

Fig 2.13: Confusion matrix for the train data with asymmetric weights

As we can see from the above values, the FNR has gone down to 0.132 as a result of using the weighted cutoff values. This means that our model now has reduced misclassification of defaulters as non-defaulters

We can now use the model for the test set data:

## Predicted  
## True 0 1  
## 0 93 80  
## 1 15 62

## FPR for test data

## 0.5147929

## MR\_test

## [1] 0.408

## FNR Test data

## [1] 0.1851852

### AUC for test data

## [1] 0.771864

Fig 2.14: Confusion matrix and other evaluation metrics for the test data with asymmetric weights

The model performs well on the test data with a reduced FNR value