GERMAN CREDIT

Introduction

In the present study, the aim is to develop a model of the type of applicants who can be classified according to their credit rating into good or bad credit. Our task is to identify patterns in data that lead to classification.

To this end we have used 2 classification algorithms, neural network and logistic regression.

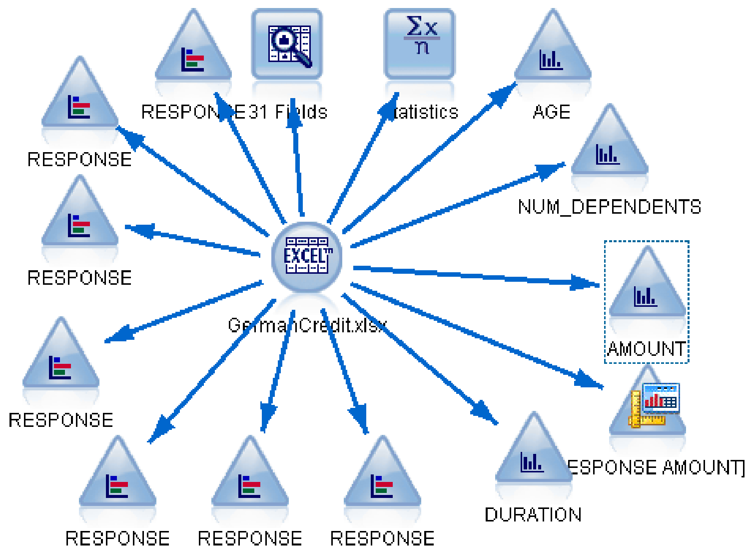
Exploratory Data Analysis

The data available for determining the German credit is a comprehensive collection of personal information of past applicants.

Of the 30 predictor variables, Age of the past credit applicants and duration of credit in months, credit amount, installment rate as a percentage of disposable income, number of existing credits at the bank and number of dependents are continuous variables.

Other predictor variables like checking account balance, savings account balance, credit history, employment, residence, nature of jobs held are categorical with 4 or 5 levels of classification.

Remaining predictor like marital status, ownership of amenities like car, TV, radio, phone, owning real estate or residence and worker legal status are binary.



Data audit shows no missing values.

Continuous variables are right skewed.

The distribution of the target variable response in the table below shows that 70% of past applicants have good credit and only 30% have bad credit.

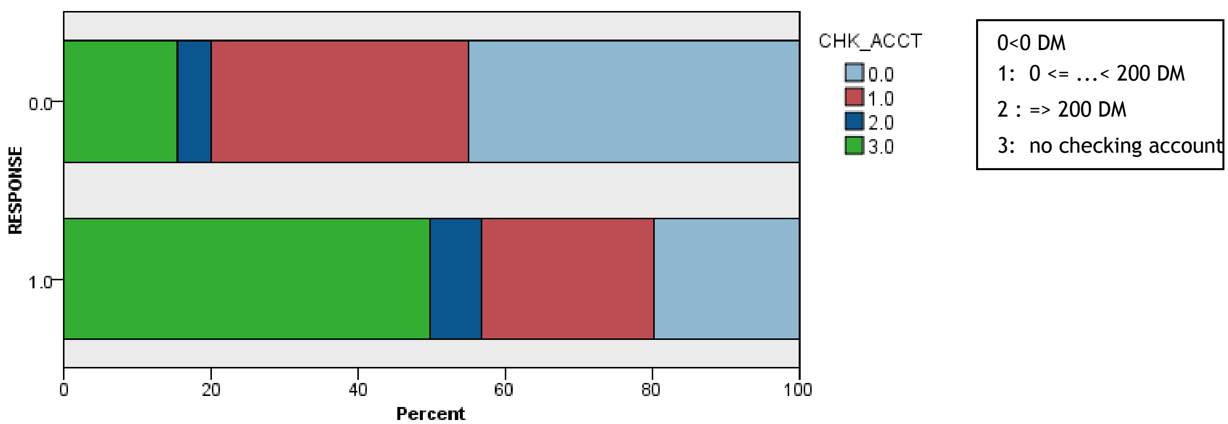
Response = 1 is classified as good credit and Response = 0 is bad credit.

DISTRIBUTION OF CREDIT

Value	Proportion	%	Count
0.000		30.0	300
1.000		70.0	700

The distributions for a few predictors are examined using the response variable as an overlay to check which predictor variable is expected to have influence in the classification model.

DISTRIBUTION OF CREDIT OVERLAYED WITH CHK_ACCT



The distribution graph shows that applicants with and without checking accounts can have good credit as well as bad credit.

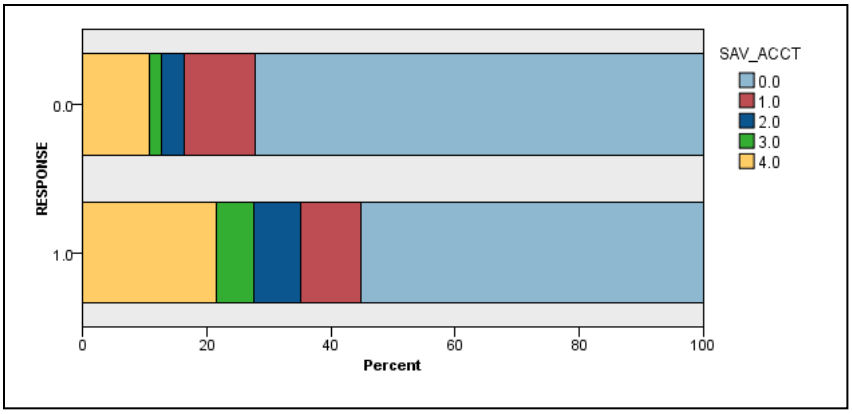
The distribution shows the surprising fact that about 50% of past applicants with good credit do not have a checking account! On the other hand only 18% of applicant with bad credit have no checking account.

About 20% of applicants with good credit have no balance in their checking accounts, however 40% of past applicants with bad credit that have no balance in checking account accounts. Thus people with no balance in checking accounts are about 20% more likely to have bad credit.

Applicants with less than 200 DM in checking account are about 20% more likely to have bad credit.

Finally, applicants with more than 200DM are only slightly more likely to have good credit.

DISTRIBUTION OF CREDIT OVERLAYED WITH SAV_ACCT



0 : < 100 DM
1 : 100<= … < 500 DM
2 : 500<= … < 1000 DM
3 : =>1000 DM
4 : unknown/ no savings acc

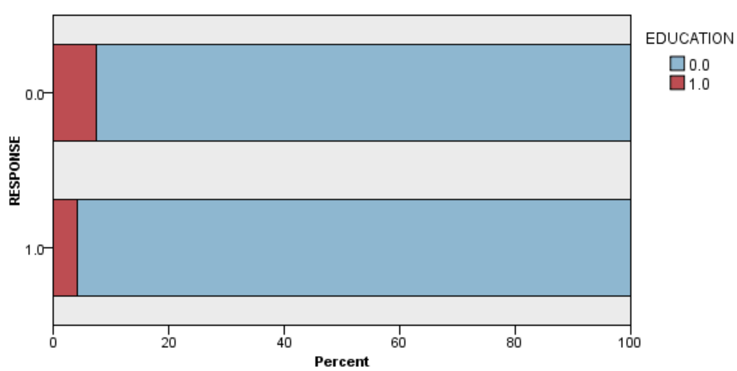
The distribution graph shows that applicants with and without saving accounts can have good credit as well as bad credit.

About 20% of applicants with no saving account balance have good credit and about 10% of applicants with no saving account balance have bad credit.

Surprisingly 55% of applicants with good credit have very low savings account balances between 0 and 100 DM compared to about 70% applicants with bad credit. Similarly only about 5% more people with balances from 100 to 500 DM have bad credit as compared to those with good credit.

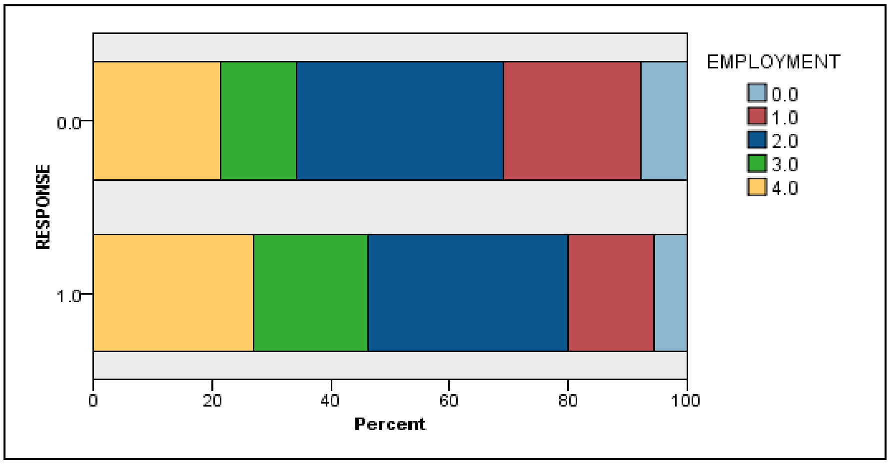
The proportion of people with good credit increases as the savings account balance increases when the balances are greater than 500 DM, as shown by bars for level 2 and 3.

DISTRIBUTION OF CREDIT OVERLAYED WITH EDUCATION



Being educated does not seem to be an important factor and does not have much influence on credit rating of applicants. Only 5% of people with good credit are educated. Similarly about 10% of people with bad credit are educated.

DISTRIBUTION OF CREDIT OVERLAYED WITH EMPLOYMENT

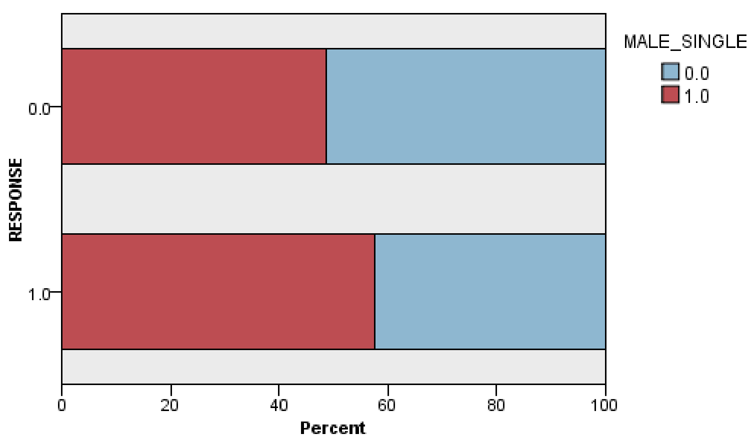


0 : unemployed
1: < 1 year
2 : 1 <= … < 4 years
3 : 4 <=… < 7 years
4 : >= 7 years

Unemployed people were only about 5 % of applicant with good credit compared to about 10% with bad credit. People unemployed between for less than a year were more likely to have bad credit. People unemployed between 1 to 4 years were equally likely to have bad or good credit. People unemployed between 4 to 7 years were 10% more likely to have good credit. People unemployed for more than 7 years were slightly more likely to have good credit.

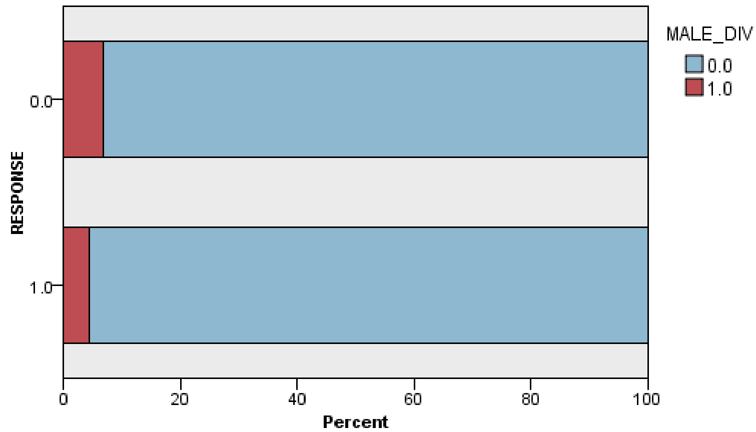
Applicants with good and bad credit have similar distributions of employment levels. So being unemployed or employed is not an important factor in the credit rating of applicants.

DISTRIBUTION OF CREDIT OVERLAYED FOR SINGLE MALES



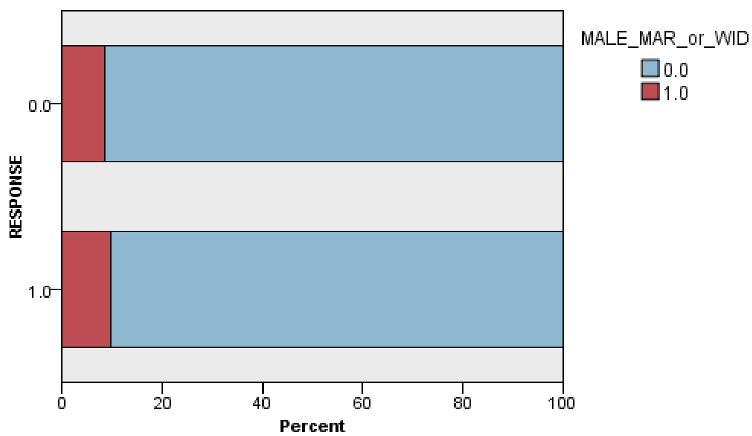
Being married is important but does not have much influence in determining credit rating. About 60% of male applicants with good credit are single compared to 50 % who have bad credit.

DISTRIBUTION OF CREDIT OVERLAYED FOR DIVORCED MALES



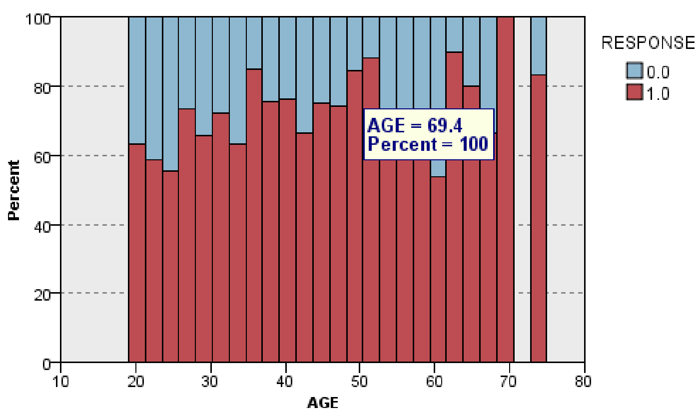
Being a divorced male has a very small influence on credit rating. Less than 10% of males who have good or bad credit are divorced.

DISTRIBUTION OF CREDIT OVERLAYED FOR MALES MARRIED OR WIDOWED



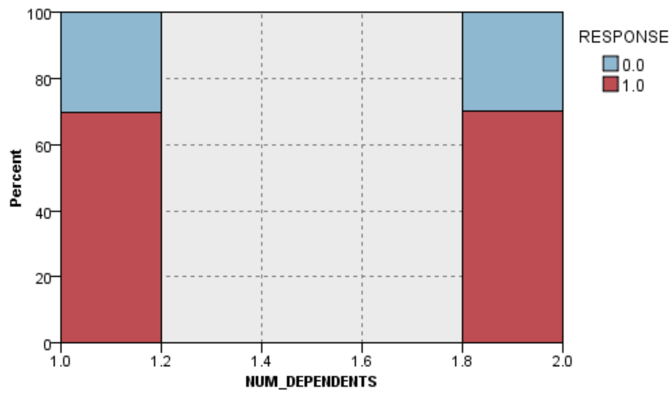
Being married or widowed does not have much influence on credit rating. Only 10% of males who have good credit are either married or widowed, with a similar result for males with bad credit.

HISTOGRAM OF AGE OVERLAYED WITH CREDIT



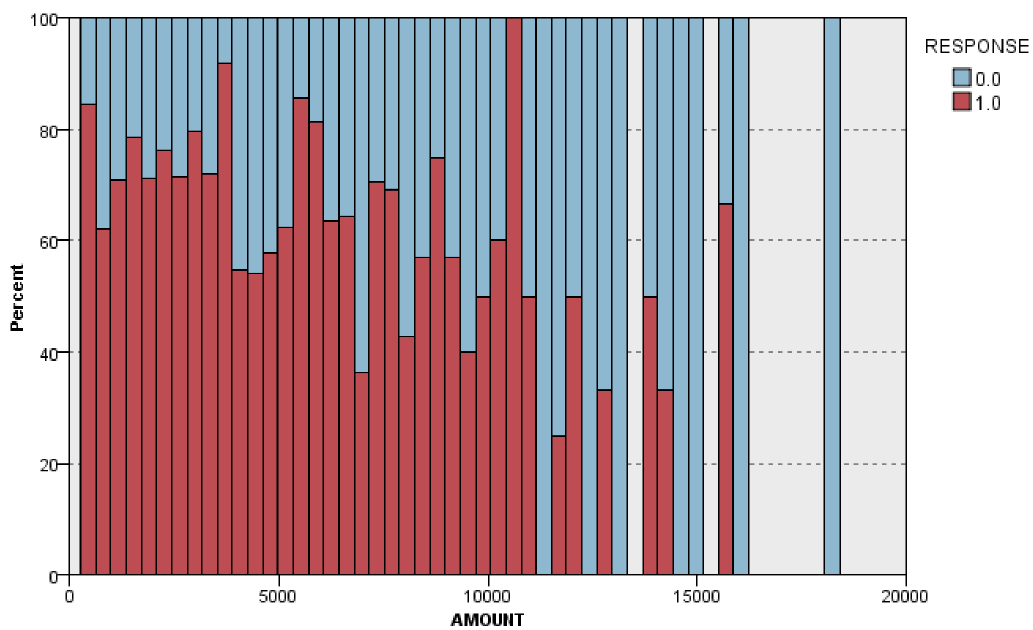
At most ages on average applicants have 60 % chance of having good credit.

HISTOGRAM OF NUMBER OF DEPENDENTS OVERLAYED WITH CREDIT



Data is available for applicants with either 1 or 2 dependents only. Applicants have 70% chance of having good credit.

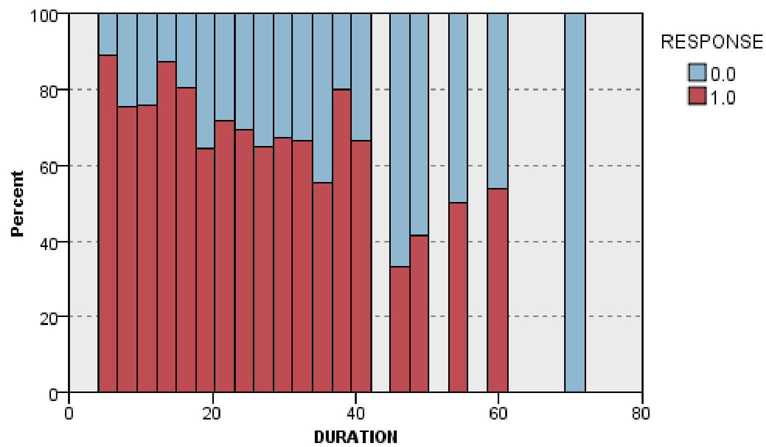
HISTOGRAM OF AMOUNT OVERLAYED WITH CREDIT



The credit rating fluctuates with amounts below $10,000.

On an average these amounts have 60% chance of having good credit. Few exceptions are

HISTOGRAM OF DURATION OVERLAYED WITH CREDIT



The histogram of duration also shows a skewed behavior similar to the other predictors.

Statistics of continuous variables

DURATION

Statistics

Count	1000
Mean	20.903
Min	4.000
Max	72.000
Range	68.000
Standard Deviation	12.059

Pearson Correlations

AMOUNT	0.625	Medium
INSTALL_RATE	0.075	Weak
AGE	-0.036	Weak
NUM_CREDITS	-0.011	Weak
NUM_DEPENDENTS	-0.024	Weak

AMOUNT

Statistics

Count	1000
Mean	3271.258
Min	250.000
Max	18424.000
Range	18174.000
Standard Deviation	2822.737

Pearson Correlations

DURATION	0.625	Medium
INSTALL_RATE	-0.271	Weak
AGE	0.033	Weak
NUM_CREDITS	0.021	Weak
NUM_DEPENDENTS	0.017	Weak

INSTALL_RATE

Statistics

Count	1000
Mean	2.973
Min	1.000
Max	4.000
Range	3.000
Standard Deviation	1.119

Pearson Correlations

DURATION	0.075	Weak
AMOUNT	-0.271	Weak
AGE	0.058	Weak
NUM_CREDITS	0.022	Weak
NUM_DEPENDENTS	-0.071	Weak

AGE

Statistics

Count	1000
Mean	35.546
Min	19.000
Max	75.000
Range	56.000
Standard Deviation	11.375

Pearson Correlations

DURATION	-0.036	Weak
AMOUNT	0.033	Weak
INSTALL_RATE	0.058	Weak
NUM_CREDITS	0.149	Weak
NUM_DEPENDENTS	0.118	Weak

NUM_CREDITS

Statistics

Count	1000
Mean	1.407
Min	1.000
Max	4.000
Range	3.000
Standard Deviation	0.578

Pearson Correlations

DURATION	-0.011	Weak
AMOUNT	0.021	Weak
INSTALL_RATE	0.022	Weak
AGE	0.149	Weak
NUM_DEPENDENTS	0.110	Weak

NUM_DEPENDENTS

Statistics

Count	1000
Mean	1.155
Min	1.000
Max	2.000
Range	1.000
Standard Deviation	0.362

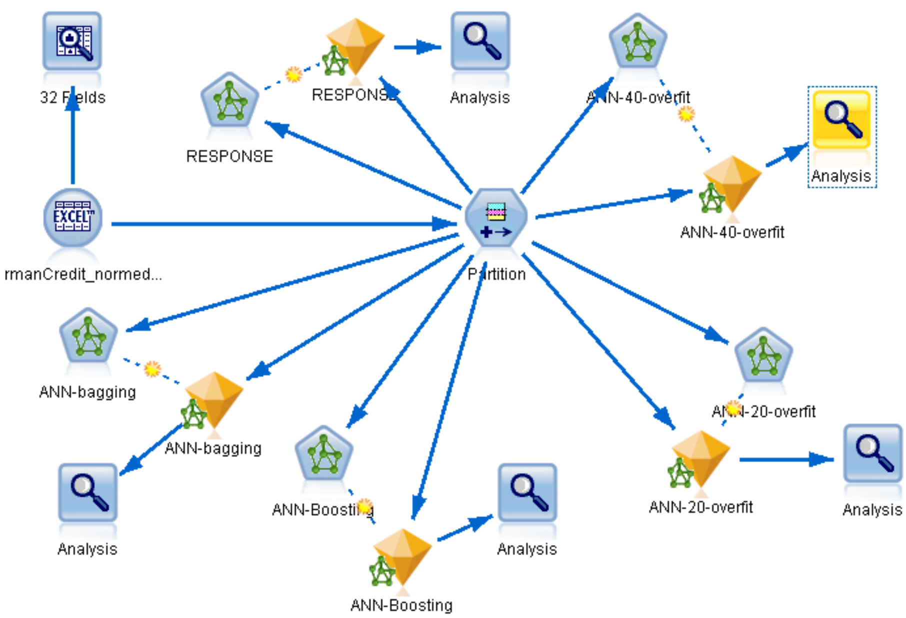
Pearson Correlations

DURATION	-0.024	Weak
AMOUNT	0.017	Weak
INSTALL_RATE	-0.071	Weak
AGE	0.118	Weak
NUM_CREDITS	0.110	Weak

Since none predictor variables are strongly correlated, we conclude all 30 predictors are important in building the classification models.

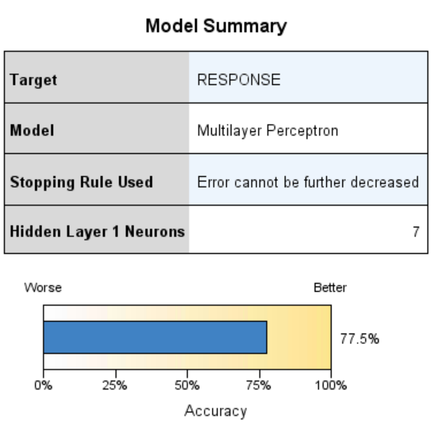
ANN Model with Min-Max normalized data

Many models with different options were generated as shown in the stream below.



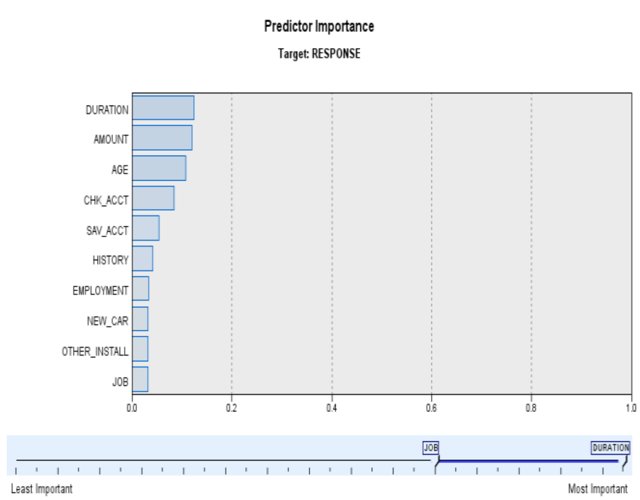
Model:

The model chosen with 20% overfit accuracy is

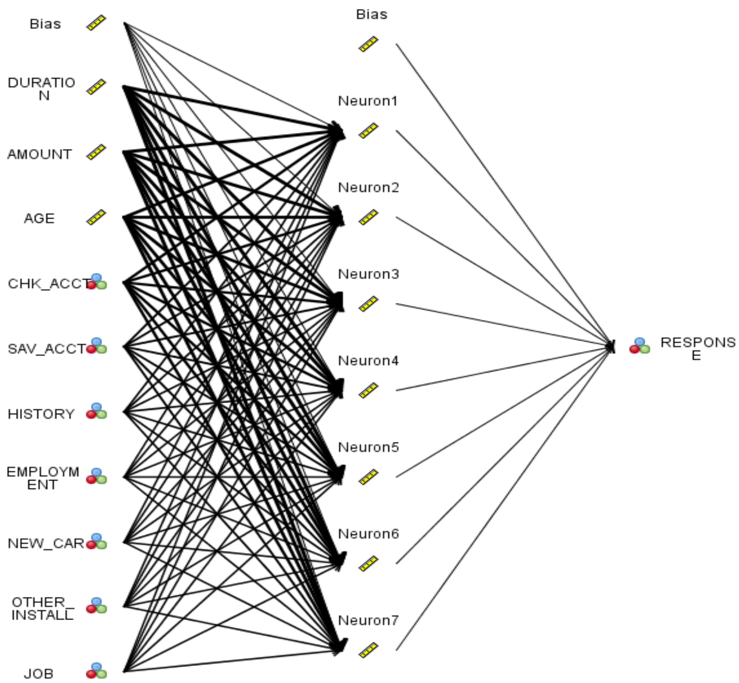


Model Summary

Target	RESPONSE
Model	Multilayer Perceptron
Stopping Rule Used	Error cannot be further decreased
Hidden Layer 1 Neurons	7



Network:



Coincidence Matrix:

The coincidence matrix is

Comparing $N-RESPONSE with RESPONSE

'Partition'	1_Training		2_Testing	
Correct	458	77.5%	304	74.33%
Wrong	133	22.5%	105	25.67%
Total	591		409	

Coincidence Matrix for $N-RESPONSE (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	86	80
1.000000	53	372

'Partition' = 2_Testing	0.000000	1.000000
0.000000	59	75
1.000000	30	245

Performance Evaluation

'Partition' = 1_Training	
0.000000	0.79
1.000000	0.135

'Partition' = 2_Testing	
0.000000	0.705
1.000000	0.13

	Predict 1	**Predict 0**
Actual 1	TP = 245	FN = 53
Actual 0	FP = 75	TN = 59

Accuracy = TP +TN / TP +TN + FP + FN = 74.3%

Other metrics are

Recall = TP/TP +FN = 82.2%

Precision = TP/TP +FP = 76.6%

Specificity = TN/TN +FP = 44.1%   1- Specificity = 56 % of false alarms.

Cost Analysis:

True negative: This represents customers correctly identified as having bad credit. The processing cost to reject the application is $20 * 59 = $1,180

True positive: This represents the customers correctly identified as having good credit. Revenue from these 245 customers = $49,000.

False negative: This represents that actually have good credit but were incorrectly identified as having bad credit. The loan rejection, cost of processing the applications is $20. In addition these 53 potential customers represent a revenue loss of: $10,60.

False positive: This represents customers with bad credit that were incorrectly identified as having good credit. These 75 customers represent a bad debt loss of $37,500.

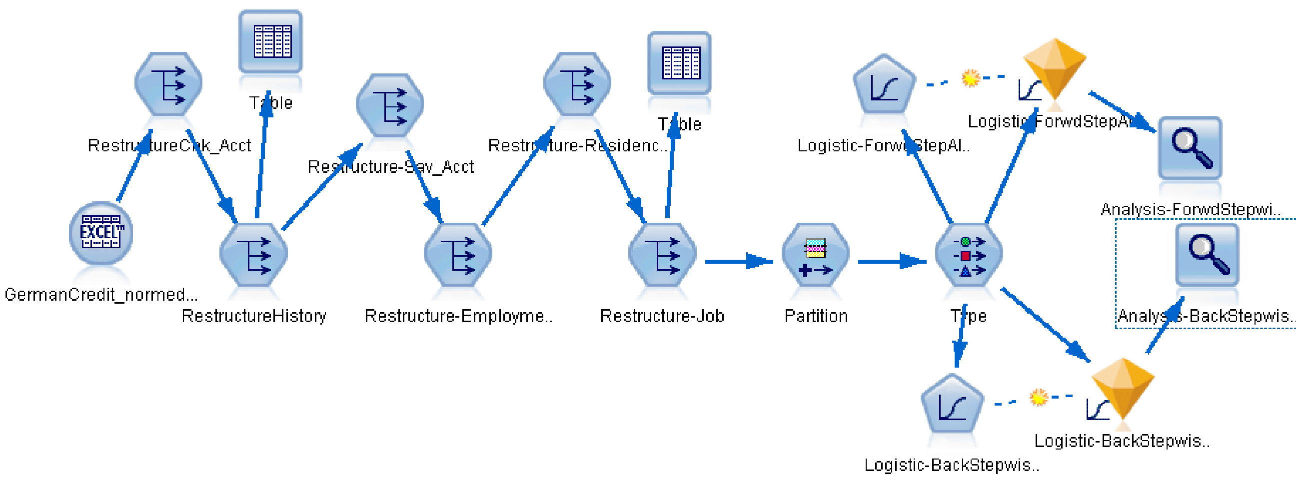
Cost summary table:

	Predict 1	**Predict 0**
Actual 1	TP = $49,000	FN = $1,060
Actual 0	FP = $37,500	TN = $1,180

The Net Profit = 9260 Euros from ANN classification model.

Logistic Regression Model with Min-Max normalization and Forwards Stepwise method

The stream used is as shown.



Model Summary:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 12[l]	DURATION	-2.704	.582	21.573	1	.000	.067
	NEW_CAR(1)	.716	.249	8.244	1	.004	2.045
	GUARANTOR(1)	-1.402	.503	7.755	1	.005	.246
	AGE	1.312	.582	5.070	1	.024	3.712
	OTHER_INSTALL(1)	.899	.272	10.885	1	.001	2.456
	RENT(1)	.524	.267	3.842	1	.050	1.689
	CHK_ACCT_1.0(1)	.554	.255	4.719	1	.030	1.740
	CHK_ACCT_2.0(1)	-1.407	.506	7.723	1	.005	.245
	CHK_ACCT_3.0(1)	2.104	.293	51.712	1	.000	8.200
	HISTORY_4.0(1)	.682	.267	6.537	1	.011	1.978
	SAV_ACCT_3.0(1)	1.799	.780	5.318	1	.021	6.046
	SAV_ACCT_4.0(1)	.972	.344	8.006	1	.005	2.643
	Constant	1.118	.789	2.004	1	.157	3.058

To check if all the above predictors deemed as important by the forward stepwise method are significant we calculate the 95% confidence intervals for each coefficient shown in the table below.

	B	S.E.	UL	LL
DURATION	-2.704	0.582	-1.56328	-3.84472
NEW_CAR(1)	0.716	0.249	1.20404	0.22796
GUARANTOR(1)	-1.402	0.503	-0.41612	-2.38788
AGE	1.312	0.582	2.45272	0.17128
OTHER_INSTALL(1)	0.899	0.272	1.43212	0.36588
RENT(1)	0.524	0.267	1.04732	0.00068
CHK_ACCT_1.0(1)	0.554	0.255	1.0538	0.0542
CHK_ACCT_2.0(1)	-1.407	0.506	-0.41524	-2.39876
CHK_ACCT_3.0(1)	2.104	0.293	2.67828	1.52972
HISTORY_4.0(1)	0.682	0.267	1.20532	0.15868
SAV_ACCT_3.0(1)	1.799	0.78	3.3278	0.2702
SAV_ACCT_4.0(1)	0.972	0.344	1.64624	0.29776
Constant	1.118	0.789	2.66444	-0.42844

Since zero is not present in the confidence intervals, all the predictors except the intercept coefficient b are significant.

Coincidence Matrix:

Comparing $L-RESPONSE with RESPONSE

'Partition'	1_Training		2_Testing	
Correct	460	77.83%	301	73.59%
Wrong	131	22.17%	108	26.41%
Total	591		409	

Coincidence Matrix for $L-RESPONSE (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	84	82
1.000000	49	376

'Partition' = 2_Testing	0.000000	1.000000
0.000000	58	76
1.000000	32	243

Performance Evaluation

'Partition' = 1_Training	
0.000000	0.81
1.000000	0.132

'Partition' = 2_Testing	
0.000000	0.677
1.000000	0.125

	Predict 1	**Predict 0**
Actual1	TP = 243	FN = 32
Actual 0	FP = 76	TN = 58

Accuracy = TP +TN / TP +TN + FP + FN = 73.6%

Other metrics are

Recall = TP/TP +FN = 88.4%

Precision = TP/TP +FP = 76.2%

Specificity = TN/TN +FP = 43.3%   1- Specificity = 56.7% of false alarms.

Cost Analysis:

Loan default (False Positive) cost 500 Euros so 500* 76 = 38,500 Euros

Revenue (True Positive) = 200 * 243 = 48,600 Euros

Loan Rejection cost (False Negative) = 32 * 50 = 1600 Euros.

False Alarms cost = 58 * 20 = 1,160 Euros

Net Profit = 7340 Euros.

True negative: This represents customers correctly identified as having bad credit. The processing cost to reject the application is $20 * 58 = $1,160

True positive: This represents the customers correctly identified as having good credit. Revenue from these 243 customers = $48,600.

False negative: This represents that actually have good credit but were incorrectly identified as having bad credit. The loan rejection, cost of processing the applications is $20. In addition these 32 potential customers represent a revenue loss of: $640.

False positive: This represents customers with bad credit that were incorrectly identified as having good credit. These 76 customers represent a bad debt loss of 500 *76 = $38,500 due to loan default.

Cost summary table:

	Predict 1	**Predict 0**
Actual 1	TP = $48,600	FN = $640
Actual 0	FP = $38,000	TN = $1,160

The Net Profit = 8800 Euros from Logistic Regression classification model.

Conclusion:

For the limited data set, the two models generated using ANN algorithm and logistic regression have almost identical accuracy. In addition cost analysis in both models are very close. It is therefore difficult to evaluate the models for the given data. Both classifiers work equally well classifying the applicants for Geman credit based on the predictors in the data set.

Inspite of the fact that ANN model building is more of a black box approach, and logistic regression modeling on the other hand, helps in interpreting the role of relevant predictors through their model coefficients, we prefer ANN because it more robust when complex dependencies exist. We would expect this as the data set becomes larger.