Analysis of German Credit Dataset

Members

Zeynep Cevik, <zcevik@ryerson.ca>

I[redacted]S[redacted]Naupada,[redacted]

Summary: Goal of the project is to develop a strategy for the bank managers that help them in making decisions on loan approval to potential customers. The analysis involves deciding which attributes should be considered by the bank managers to access and predict the creditability.

As a part of data preparation along with the exploratory analysis the variable Account Balance and Value Savings are reassigned using SQL, to make it an ordinal data.

Supervised machine learning models were applied on the dataset to analyze the dataset in depth : Decision Trees and Naïve Bayes. In the case of entropy and gini grow types of decision trees, the same metric values are calculated. Recall is nearly 88% which implies 88% of the time, model predicts positive values as positive correctly. Error rate is approximately 23% which means 23% of the time the model predicts positive values as negative and vice versa.

Based on entropy and gini decision tree 4 variables are important in predicting the creditability: Account balance is the most important followed by, duration of credit, purpose, payment status of previous credit and value savings are the other important variables respectively.

Along with this, the Naïve Bayes method is also used to analyze the error rate and the precision, recall, accuracy of prediction of the model. The Naïve Bayes gives a very high error rate compared to decision tree. So, for the analysis of the dataset, Decision trees are the best method of ML model

As a result, considering all the ML model results the bank managers should consider account balance of a customer as the most important attribute to be considered for the approval of loan followed by payment credit history, duration of credit and value savings.

[redacted]

FINAL PROJECT

By Zeynep Burcu CEVIK and Lakshmi Sravanthi NAUPADA

ABSTRACT:
The approval of a loan application can be a life changing decision for credit seeking customers. Importantly analyzing credit worthiness of potential clients becomes more than just a decision point as a worthy customer should not be denied of getting a mortgage. The following analysis of the dataset and the methodologies help loan officers to make the right decisions in approving loan to the worthy customers. Goal of the project is to develop a strategy for the bank managers that help them in making decisions about loan approval to potential customers. Dataset used for the analysis is the German Credit Data that contains 20 attributes with 1000 rows and the class attribute showing a good or a bad credit risk.

Tools and languages used: SQL, SAS, R, Python.

DATA PREPERATION:
In order to come up with a solution analyzing each of the attributes in the data set and determining the influence of the attributes on the class attribute i.e the final decision in loan approval is very important. The following shows the data preparation of the dataset.

Section 1: Attribute Types

Determining the attribute type is very important, as not all methodologies and strategies are suitable for all types of variables. For instance, a scatter plot cannot be built for two qualitative variables, as it is appropriate for quantitative variables. In this section, we will briefly talk about the attribute types.

There is a class attribute, credibility, which is dependent variable in our analysis. This class attribute is a binary variable: 1 represents good credit and 0 represents bad credit. Hence this attribute is a qualitative, specifically ordinal variable.

There are twenty more attributes in the file which will be analyzed in detail. The irrelevant variables will be omitted, and the variables of interest are used for the advanced analysis. Although all the attributes seem to be numeric in the .csv file (see Table 1), the description of the attributes says not all of them are quantitative variables.

Account Balance is a qualitative variable. In addition, it is somewhat ordinal. By using SQL, we have changed the variable so that the value of the variable increases, the balance of the account also increases.

Before,

$$Account\ Balance(AB) = \begin{Bmatrix} 1\ if\ AB < 0 \\ 2\ if\ 0 \leq AB < 200 \\ 3\ if\ AB > 200 \\ 4\ if\ no\ account \end{Bmatrix}$$

After manipulation,

$$Account\ Balance(AB) = \begin{Bmatrix} 0\ if\ no\ account \\ 1\ if\ AB < 0 \\ 2\ if\ 0 \leq AB < 200 \\ 3\ if\ AB > 200 \end{Bmatrix}$$

Now, with the new format of Account Balance, we can say it is an ordinal variable.

Duration of Credit is a quantitative variable, in particular discrete according to the data file.

Payment Status of Previous Credit shows credit history and is a qualitative variable, nominal in specific.

Purpose, Sex and Marital Status, Guarantors are qualitative variables, nominal type.

Credit Amount and Instalment Percent are quantitative variables. They can be continuous but there exist only integers in the data file.

Value Savings/Stocks is a qualitative variable. Like Account Balance, it is somewhat ordinal. With a little manipulation, we have changed the variable.

Before,

$$Value\ Savings\ (VS) = \begin{Bmatrix} 1\ if\ VS < 100 \\ 2\ if\ 100 \leq VS < 500 \\ 3\ if\ 500 \leq VS < 1000 \\ 4\ if\ VS \geq 1000 \\ 5\ if\ no\ savings \end{Bmatrix}$$

After manipulation,

$$Value\ Savings\ (VS) = \begin{Bmatrix} 0\ if\ no\ savings \\ 1\ if\ VS < 100 \\ 2\ if\ 100 \leq VS < 500 \\ 3\ if\ 500 \leq VS < 1000 \\ 4\ if\ VS \geq 1000 \end{Bmatrix}$$

Now, it is an ordinal variable, so that when the savings increase, the value of the variable also increases.

Length of current employment and duration in current address are qualitative variables. They are ordinal.

Duration in current address is a qualitative variable and ordinal.

Most valuable available asset, concurrent credits and Type of apartment are qualitative variables and nominal.

Age in years, number of existing credits and number of dependents are quantitative variables and is discrete.
Occupation is qualitative and ordinal.

Telephone and foreign worker are qualitative nominal variables.

Table 1: Attribute Types:

Qualitative Variable		**Quantitative Variable**	
Nominal	**Ordinal**	**Discrete**	**Continuous**
Account Balance	Credibility (class a.)	Duration of Credit	Credit Amount
Payment Status of Previous Credit	Length of Current Employment	Age in years	Instalment Percent
Purpose	Value Savings/ Stocks	No. of existing credits at the bank	
Sex & Marital Status	Duration in current address	No. of dependents	
Guarantors	Occupation		
Most valuable available asset			
Concurrent credit			
Type of apartment			
Telephone			
Foreign worker			

Handling missing values: Missing values occur in the data set when there is no data value is stored for the variable in an observation. If the missing values are not handled properly, the resulting inferences might be inaccurate. So, it is very important to look for missing values. For this data there are no missing value for all the attributes. (Table 2.)

Section 2: Descriptive Statistics

In this section, we are going to examine the descriptive statistics of all attributes, try to find out the outliers and examine the balance of the attributes.

An outlier in a dataset is a data point that differs significantly from other observations. Some of the statistics tools such as mean are sensitive to outliers and it is very important to detect the outliers as they might harm further analysis . Here we use box plots for detection of outliers.

There is no logic in investigating the descriptive statistics of qualitative variables, but only the balance. On the other hand, when we observe the duration of the credit, we see that the customers demand loan as long as 6 years and as low as 4 months and average of 20 months. Also, since the standard deviation is a little bit high, there may be outliers. When we look at its boxplot, we see outliers in the plot. (Table 5-left)

The lowest credit amount is 250 DM (deutsche mark) and highest one is 18424 DM. The average of the loans is 3271.25 DM, and standard deviation is large which implies the existence of outliers. Also, from the boxplot, we can safely say that there are so many outliers. (Table 5-right)

If we look into the Age attribute, which is quantitative, the minimum age of the customers is 19 and maximum age is about 75 years. The average age shows 35 years and has a deviation of 11 years approximately. This suggests a possibility of outliers. (Figue 1-below right)

Installment percent varies from 1 to 4 and mean is 2.97. Also, the standard deviation is 1.1 which shows a low possibility for outliers. From its boxplot, we can conclude that there are no outliers. (Figure:1- below left)

After determining the attribute types and outliers, the next step is to observe the balance of attributes. From figure 3, we can say that none of the attributes has a balanced attribute except the Most Valuable Asset which is somewhat balanced.

Section 3: Correlation and Determination of Variables

In this section, we will try to observe the direction and strength of the variables so that if one independent variable is highly correlated with the other, we may skip one of them in our analysis. However, as you can see from the correlation matrix, represented by the circles, there are no highly correlated variables. The highest correlation coefficient in absolute term, nearly 0.63, is between Credit Amount and Duration of the Credit. The strength of this coefficient is not enough to omit one of them.

The method of the correlation in the table is Spearman correlation coefficient. The reason is the existence of qualitative variables. As shown in the above table, there are only six quantitative variables out of 21. Since we are interested in examining the relationship between all variables, we prefer to use Spearman correlation which is advised in the literature when there are qualitative variables. The correlation matrix table is an R output (library corrplot). The size and the darkness of the circles show the strength of the relationship, on the other hand the color represents the direction of the relationships as shown in the right scale. Blue points out positive relationship whereas red represents negative relationship. The diagonal of the correlation matrix is always dark blue and full circle, indicating the value 1 (perfect correlation), because the correlation of an attribute with itself is always 1.

Hence, as a result of the correlation matrix, we do not have enough evidence to remove a variable from our analysis.

PREDICTIVE MODELLING:

Section 1: Decision Tree

Decision trees are one of the most popular decision-making process techniques in supervised learning. By investigating the evaluation metric, we can test how well our data mining algorithm is performing. In this section, our aim is to discuss our decision tree results and decide which variables are important in credibility of that person so that the bank employee may decide whether to approve the loan or not.

1.a. Entropy-Decision Tree:

According to the SAS output, the values in the confusion matrix for the entropy case are as follows:

f++ (True Positive) = 619
f-+ (False Positive) = 157
f- -(True Negative) = 143
f+- (False Negative) = 81
Total (T) = 1000

According to the above numbers, some results of the evaluation metrics are:

Accuracy = (f++ + f- -)/T = 632/1000=0.632
Recall (TPR) = f++/(f++ + f+-)=619/700= 0.884
Precision (P+) = f++/(f-+ +f++) = 619/776=0.798
Error Rate (ER) = (f+- + f-+)/T = 238/1000=0.238

There are 21 variables including the class attributes. From the SAS output, we can say that there are 4 important variables in predicting the credibility attribute. Account Balance is the most important variable whereas duration of credit, purpose, payment status of previous credit and value savings are the other important variables respectively. In line with this finding, the most important variable in predicting the credibility of the person depends on his/her belongings including the cash with her/his history. In addition, the fit statistic gives a quite low level of gini coefficient but above average level of entropy.

1.b. Gini-Decision Tree:

According to the SAS output, the values in the confusion matrix for the Gini case are as follows:

f++ (True Positive) = 619
f-+ (False Positive) = 157
f- -(True Negative) = 143
f+- (False Negative) = 81
Total (T) = 1000

According to the above numbers, some results of the evaluation metrics are:

Accuracy = (f++ + f- -)/T = 632/1000=0.632
Recall (TPR) = f++/(f++ + f+-)=619/700= 0.884
Precision (P+) = f++/(f-+ +f++) = 619/776=0.798
Error Rate (ER) = (f+- + f-+)/T = 238/1000=0.238

Among 20 independent variables (attributes), there are 4 important variables in predicting the credibility attribute. Account Balance is also the most important variable and Duration of credit, payment status of previous credit and values savings are the important variables too respectively. When we look at the fit statistics, we also see a quite low average error and Gini coefficient but relatively high entropy.

1.c. Comparison:

In this part, we will compare two decision trees differentiated in the grow: Entropy and Gini and decide which one is better if we can.

We want higher accuracy, recall and precision but lower levels of error rates. Accuracy, precision, and recall are trying to find out the performance about true predictions of the dataset, whereas error rates measure the false predictions. When we look at the evaluation metrics, They are the same for all the evaluation metrics.

Accuracy (Entropy)=Accuracy (Gini)
Recall (Entropy) = Recall (Gini)
Precision (Entropy) = Precision (Gini)
Error Rate (Entropy) = Error Rate (Gini)

The fit statistics such as sensitivity, specificity, entropy and gini are all the same in both cases: entropy and gini. Hence, we are indifferent between Gini and Entropy.
This indifference would change if we used the original version of our dataset. In other words, as you can remember, we changed some variables: Account Balance, and Value Savings. We also tested the original version of the dataset and some of the findings were different. In that case, Entropy performs better than Gini also, the findings of entropy are the same with the current dataset. This implies that our reassigning the values of Account Balance and Value Savings will not change the decision tree performance. This is an important observation in our study.

1.d. Re-evaluation of the Selected Variables:

According to the findings, there are four important variables in predicting the credibility.

The most important variable is still the account balance which means that if the account balance of the person is high, the credibility of that person will also be high. When we observe the subtree figure, we see that if the person has a higher than 200 DM in his/her account balance or no-account balance in that bank, the credibility is good. If that person has a negative account balance or lower than 200 DM in his/her account balance, that person's credibility may be bad, but it is not a final conclusion because there are some other factors in this branch so that, that person's credibility may be good.

The duration of the credit is another important variable. The threshold duration is 21.68 months. If the duration is smaller than this number, it is a good signal towards a good credit but not final decision. The final decision depends on that person's previous credit payment status. If no credits are taken or all credits are paid back duly, the credibility of that person is bad. On the other hand, if the person has existing credits being paid back duly, delay in paying off or critical account, that person's credibility is good. The last branch does not give fully meaningful results. Because if a person has a critical account or some delays in his/her payment, it should not be a good signal in credibility.

On the other hand, if the duration is higher than the threshold level, it is a bad signal towards a bad credit. The last decision depends on that person's value savings. If person has an unknown savings or higher than 1000 DM in his/her savings account, the credibility is good. If the savings are less than 1000 DM, then the credibility is bad. It makes sense because the person already does not have enough account balance, together with not enough savings, the credibility will not be good.

When we look at the evaluation metric of this decision tree, all the statistics are the same with part a and b.

Section 2: Naïve Bayes

In this section, we will discuss 'Naïve Bayes' method which is used in supervised learning and text classification. We implemented this method in Python. The screenshots are in Appendix as all other table and figures.

According to Python output, the confusion matrix is as follows:

f++ (True Positive) = 200
f-+ (False Positive) = 71
f- -(True Negative) = 46

f+- (False Negative) = 23
Total=340

According to the above numbers, some results of the evaluation metrics are:

Accuracy = (f++ + f- -)/T = 246/340=0.723
Recall (TPR) = f++/(f++ + f+-)=200/223=0.897
Precision (P+) = f++/(f-+ +f++) = 200/271=0.738
Error Rate (ER) = (f+- + f-+)/T = 238/1000=0.344

The most striking result is the high level of error rate. When we compare these results with the decision tree results, all the evaluation metrics are pretty much the same except error rate. Error rate is way too high compared to decision tree's error rate which makes this method undesirable for this dataset.

Although Naïve Bayes is not so appropriate due to high error rate, important variables in predicting credibility are pretty much the same with the Decision Tree method. Account Balance is the most important variable, whereas Duration of Credit and Payment Status of Credit are the other important variables respectively.

CONCLUSIONS AND RECOMMENDATIONS:

In this study, we are trying to find out the indicators of predicting credibility. The dataset belongs to a German bank. It is composed of 1000 observations with 21 attributes including the class attribute. The class attribute is credibility which is one when the person has a good credibility and zero when the person has a bad credibility. We are trying to find out the important variables which significantly affect and predict the credibility. After observing the attributes in detail and changes some of them, we applied decision tree methods and naïve bayes method. According to the performance of these procedures, we choose the decision tree method regardless of the grow type, because both Gini and Entropy give the same evaluation metrics and variable importance table. We do not prefer Naïve Bayes Method due to its high error rate.

According to the decision tree method, account balance is the most important variable. When a person wants an approval from the bank for a loan, if that person has more than 200 DM in his/her account, it will be stated as good credibility and has a higher chance of approval regardless of his/her other attributes. On the other hand, if that person has a lower account balance than 200 DM, the bank looks at other attributes such as his/her payment credit history, duration of credit and value savings. If the requested credit loan duration is lower than 21.7 months, the bank will check his/her payment history, however if the requested credit loan is higher than 21.7 months than the bank checks its value savings. If the savings are bigger than 1000 DM then that person's credibility is good, otherwise bad. These results are robust, because we re-run this dataset only with the important variables and class attribute, it gives the same variables as important and their sequence is also the same. The performance is also the same.

The other important finding that we come up from this study is that the results does not change, even if we manipulated some of the variables to some extent. For instance, the previous version of the Value Savings and Account Balance attributes were not fully ordinal, we made them ordinal by changing some assigned values' position. The decision tree method when grow is Entropy gives similar results whereas Gini gives completely different results. However, when we compare the evaluation metrics, the superiority of Entropy is unambiguous. The Entropy evaluation

performance, the important variables and their sequence are not different from the manipulated version of the data.

As a conclusion, the bank should check the account balance, payment history, value savings and credit duration in order to make a final decision about a loan request.

APPENDIX – SCREENSHOTS

Table 2: SAS Output for Missing Value Detection

The MEANS Procedure

Variable	N	N Miss
Creditability	1000	0
Account_Balance	1000	0
Duration_of_Credit__month_	1000	0
Payment_Status_of_Previous_Credi	1000	0
Purpose	1000	0
Credit_Amount	1000	0
Value_Savings_Stocks	1000	0
Length_of_current_employment	1000	0
Instalment_per_cent	1000	0
Sex___Marital_Status	1000	0
Guarantors	1000	0
Duration_in_Current_address	1000	0
Most_valuable_available_asset	1000	0
Age__years_	1000	0
Concurrent_Credits	1000	0
Type_of_apartment	1000	0
No_of_Credits_at_this_Bank	1000	0
Occupation	1000	0
No_of_dependents	1000	0
Telephone	1000	0
Foreign_Worker	1000	0

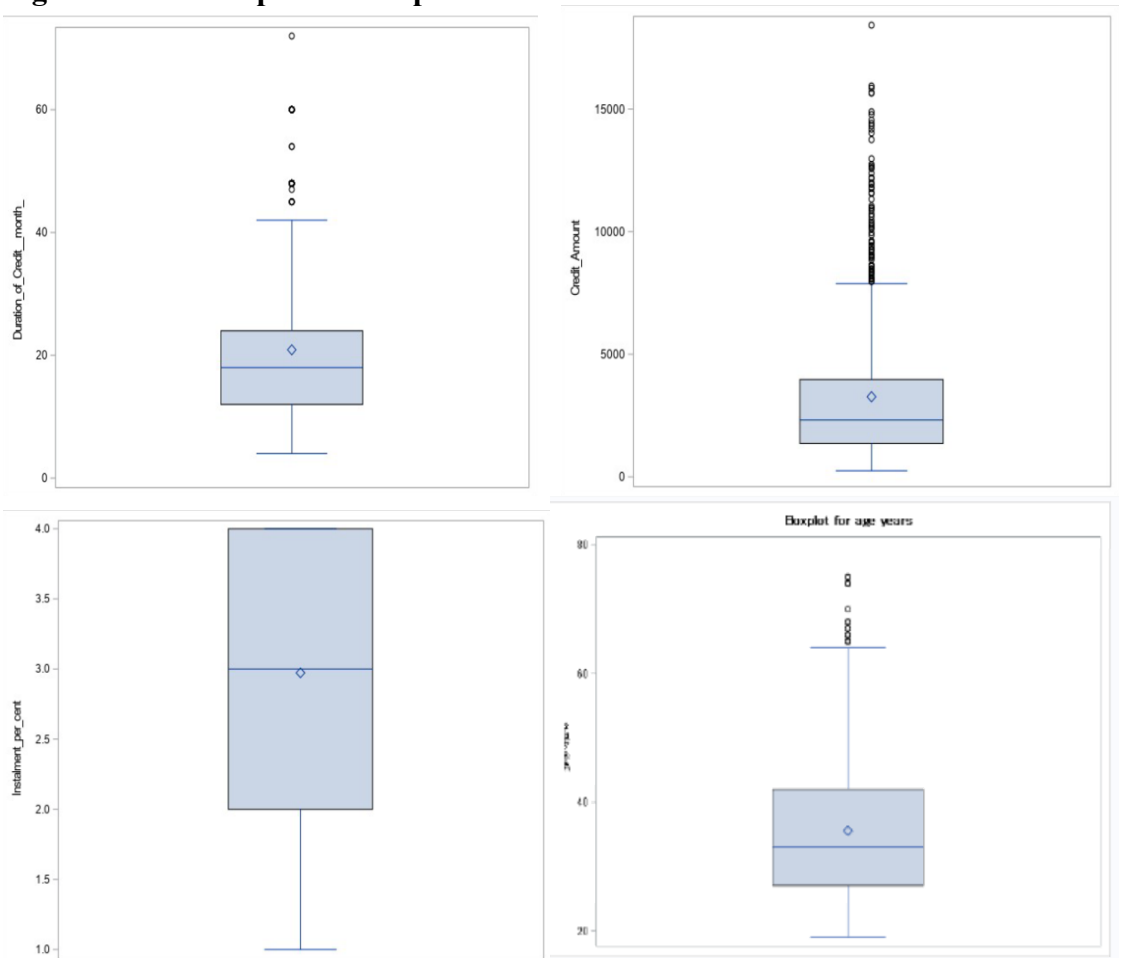
Table 3. SAS Output for Descriptive Statistics

The SAS System

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
Creditability	1000	0.7000000	0.4584869	0	1.0000000
Account_Balance	1000	1.0010000	0.9570344	0	3.0000000
Duration_of_Credit__month_	1000	20.9030000	12.0588145	4.0000000	72.0000000
Payment_Status_of_Previous_Credi	1000	2.5450000	1.0831196	0	4.0000000
Purpose	1000	2.8280000	2.7444395	0	10.0000000
Credit_Amount	1000	3271.25	2822.75	250.0000000	18424.00
Value_Savings_Stocks	1000	1.1900000	0.9668686	0	4.0000000
Length_of_current_employment	1000	3.3840000	1.2083063	1.0000000	5.0000000
Instalment_per_cent	1000	2.9730000	1.1187147	1.0000000	4.0000000
Sex___Marital_Status	1000	2.6820000	0.7080801	1.0000000	4.0000000
Guarantors	1000	1.1450000	0.4777062	1.0000000	3.0000000
Duration_in_Current_address	1000	2.8450000	1.1037179	1.0000000	4.0000000
Most_valuable_available_asset	1000	2.3580000	1.0502090	1.0000000	4.0000000
Age__years_	1000	35.5420000	11.3526701	19.0000000	75.0000000
Concurrent_Credits	1000	2.6750000	0.7056011	1.0000000	3.0000000
Type_of_apartment	1000	1.9280000	0.5301859	1.0000000	3.0000000
No_of_Credits_at_this_Bank	1000	1.4070000	0.5776545	1.0000000	4.0000000
Occupation	1000	2.9040000	0.6536140	1.0000000	4.0000000
No_of_dependents	1000	1.1550000	0.3620858	1.0000000	2.0000000
Telephone	1000	1.4040000	0.4909430	1.0000000	2.0000000
Foreign_Worker	1000	1.0370000	0.1888562	1.0000000	2.0000000

Figure 1: SAS Output for Boxplots



Figue:2. R output for Correlation Matrix* (Spearman Correlation)

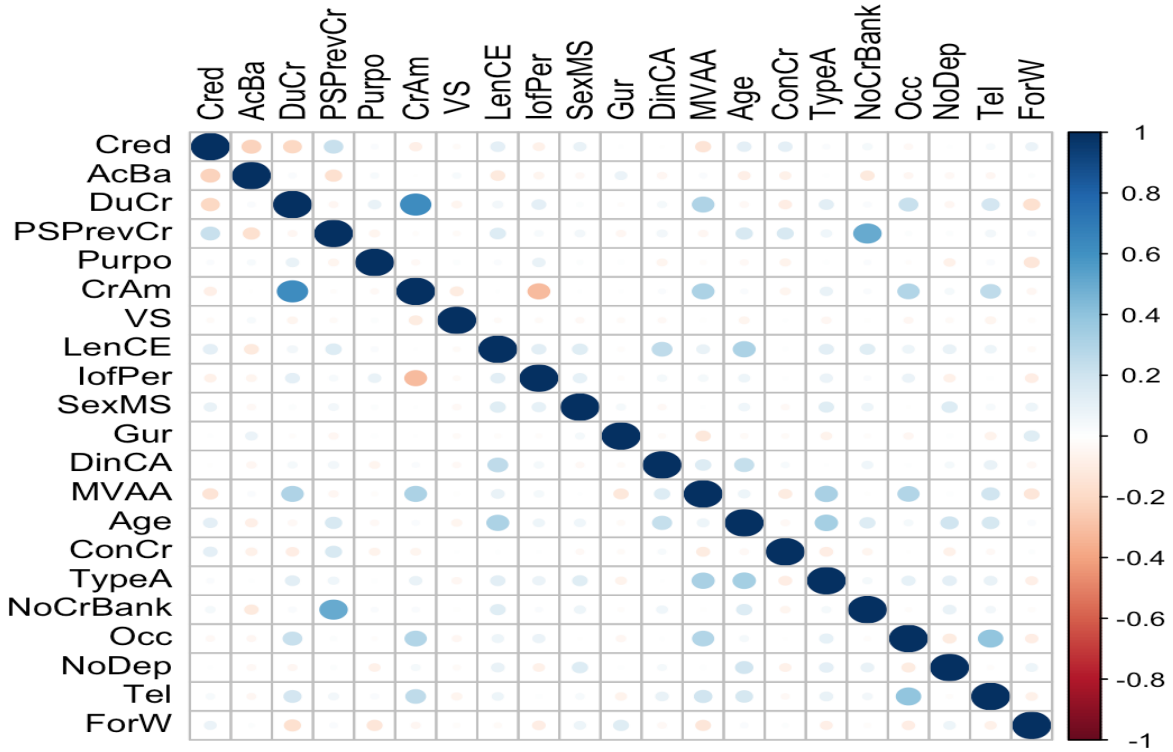


Figure 3: The Balance of the Attributes by Histogram (R plot)

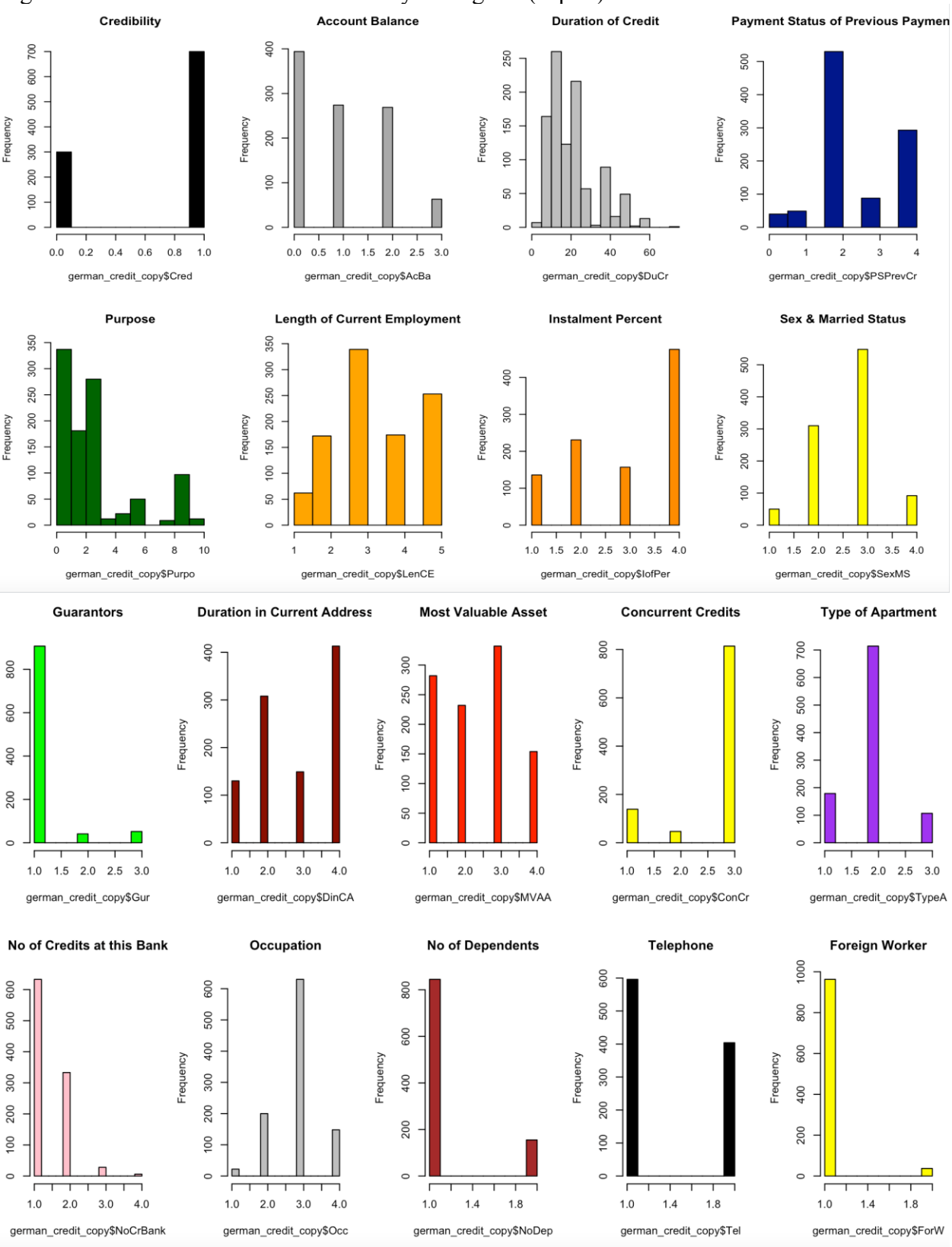


Figure 4: SAS Output-Decision Tree- Entropy

The HPSPLIT Procedure

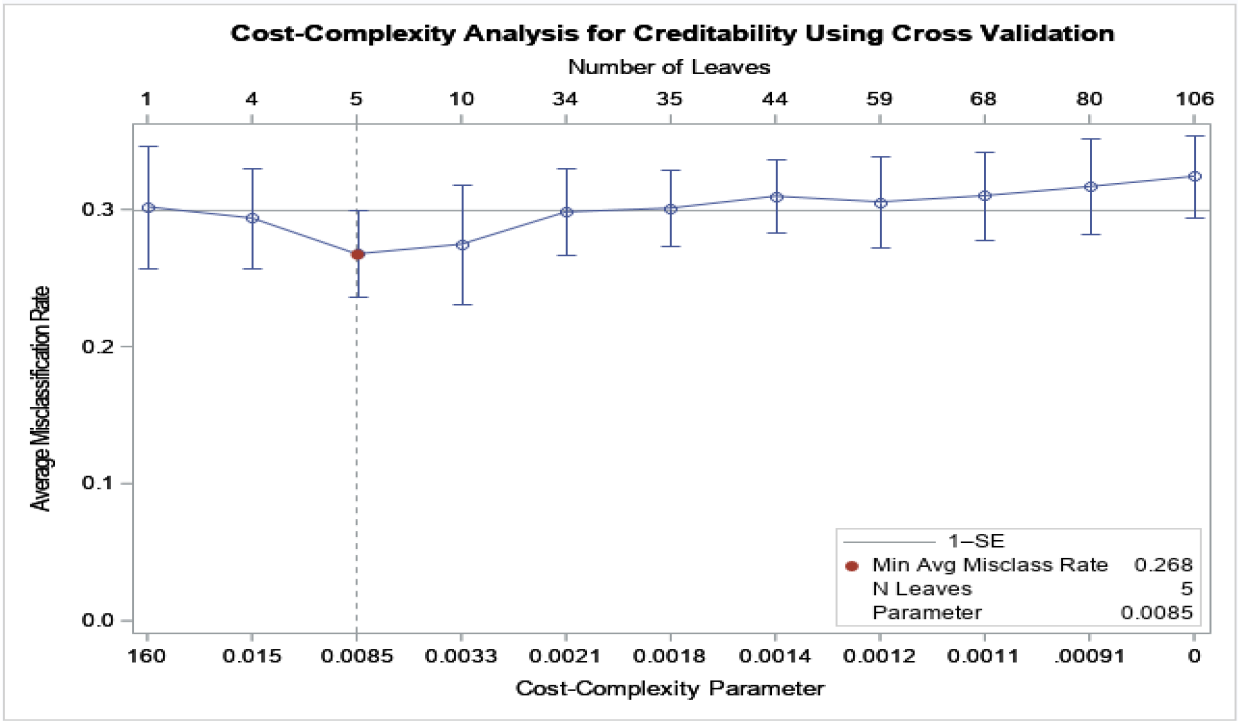
Performance Information	
Execution Mode	Single-Machine
Number of Threads	2

Data Access Information			
Data	Engine	Role	Path
WORK.GERMAN_BANK	V9	Input	On Client

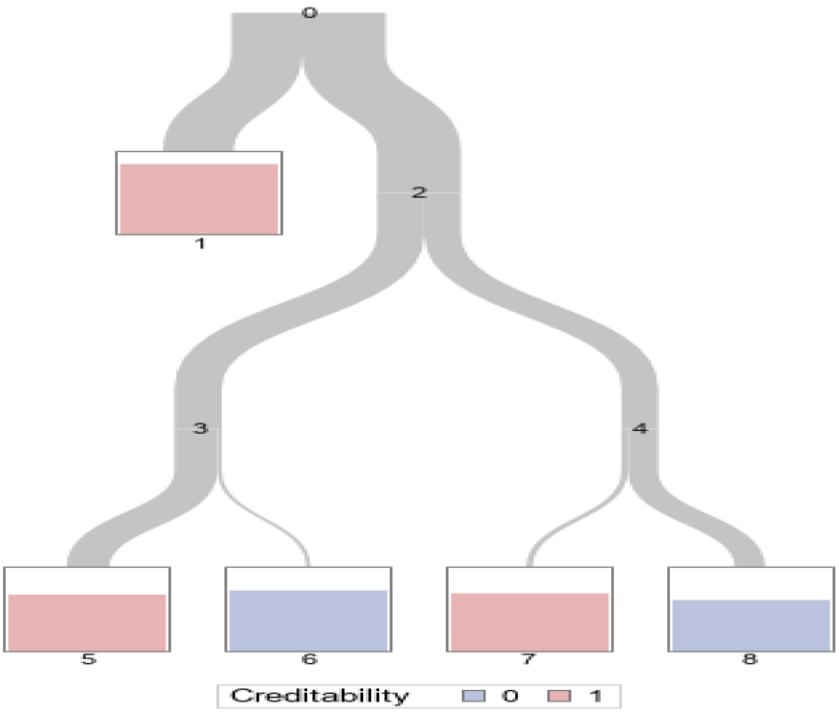
Model Information	
Split Criterion Used	Entropy
Pruning Method	Cost-Complexity
Subtree Evaluation Criterion	Cost-Complexity
Number of Branches	2
Maximum Tree Depth Requested	10
Maximum Tree Depth Achieved	10
Tree Depth	3
Number of Leaves Before Pruning	131
Number of Leaves After Pruning	5
Model Event Level	0

Number of Observations Read	1000
Number of Observations Used	1000

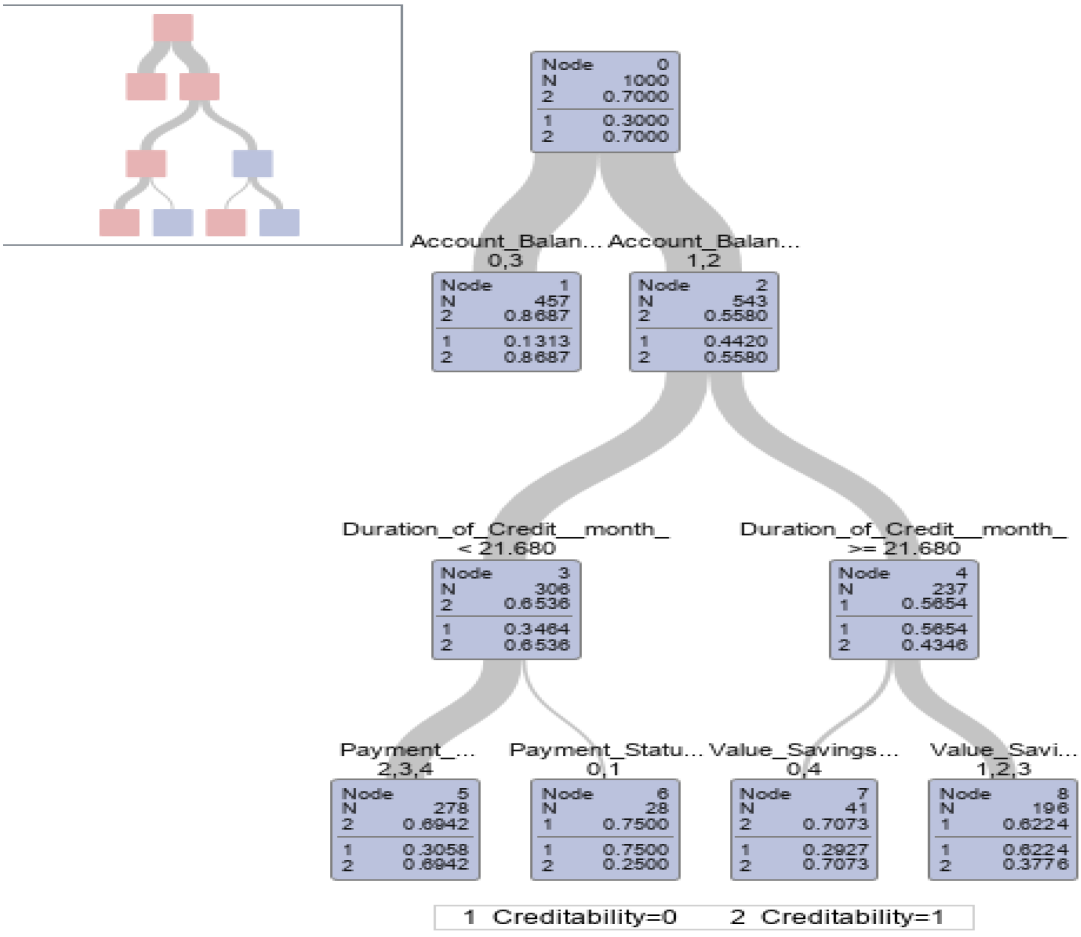
The HPSPLIT Procedure



Classification Tree for Creditability



Subtree Starting at Node=0

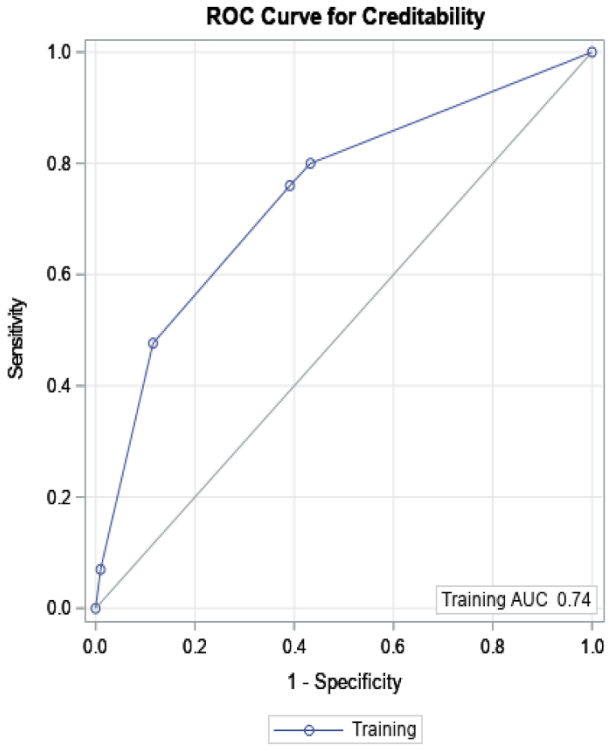


The SAS System

The HPSPLIT Procedure

Model-Based Confusion Matrix			
	Predicted		
Actual	0	1	Error Rate
0	143	157	0.5233
1	81	619	0.1157

Model-Based Fit Statistics for Selected Tree								
N Leaves	ASE	Mis-class	Sensitivity	Specificity	Entropy	Gini	RSS	AUC
5	0.1709	0.2380	0.4767	0.8843	0.7492	0.3419	341.9	0.7425



Variable Importance			
	Training		
Variable	Relative	Importance	Count
Account_Balance	1.0000	6.9217	1
Duration_of_Credit_month_	0.5171	3.5792	1
Payment_Status_of_Previous_Credi	0.4578	3.1687	1
Value_Savings_Stocks	0.3923	2.7156	1

Figure 5: SAS Output-Decision Tree- Gini

Performance Information	
Execution Mode	Single-Machine
Number of Threads	2

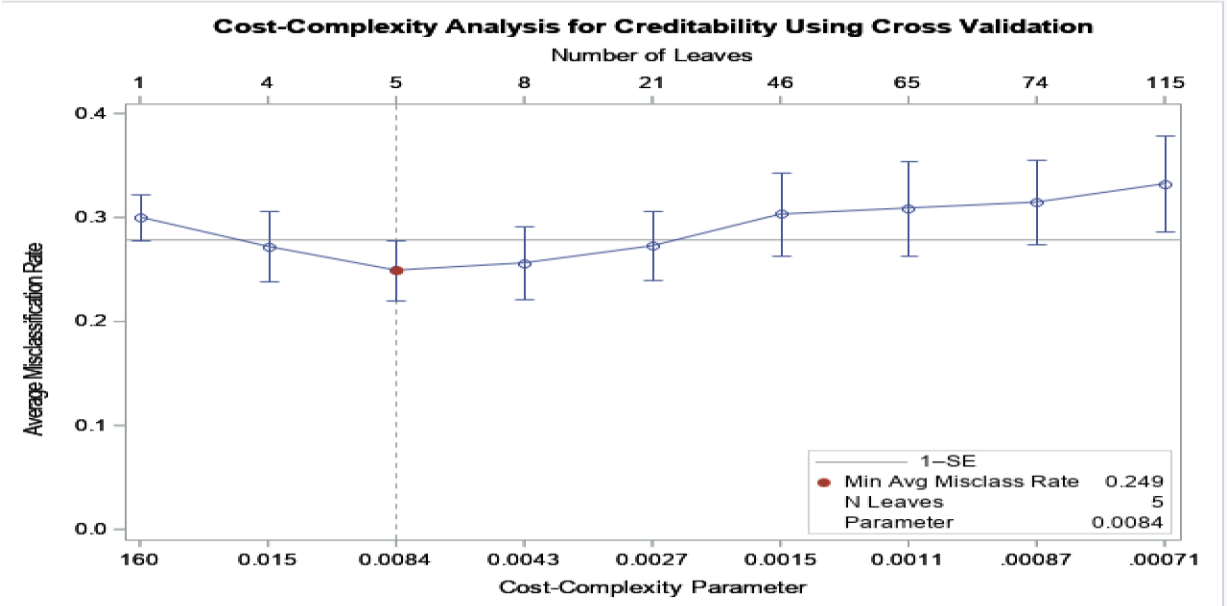
Data Access Information			
Data	Engine	Role	Path
WORK.GERMAN_BANK	V9	Input	On Client

Model Information	
Split Criterion Used	Gini
Pruning Method	Cost-Complexity
Subtree Evaluation Criterion	Cost-Complexity
Number of Branches	2
Maximum Tree Depth Requested	10
Maximum Tree Depth Achieved	10
Tree Depth	3
Number of Leaves Before Pruning	142
Number of Leaves After Pruning	5
Model Event Level	0

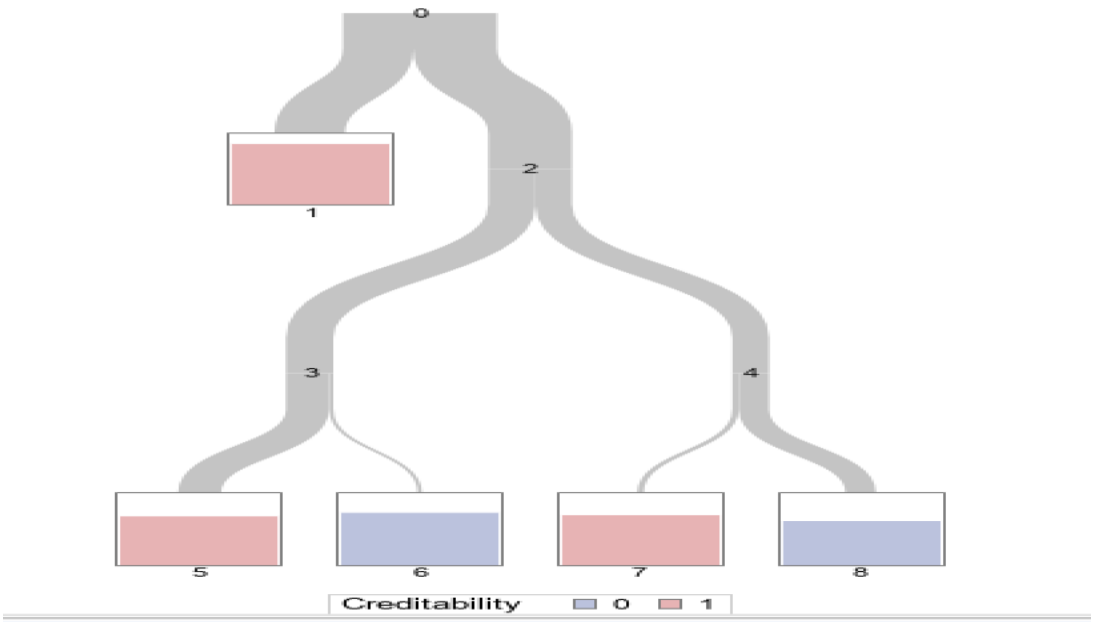
Number of Observations Read	1000
Number of Observations Used	1000

The SAS System

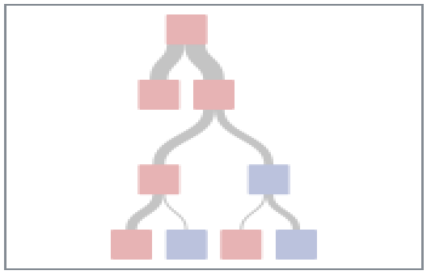
The HPSPLIT Procedure

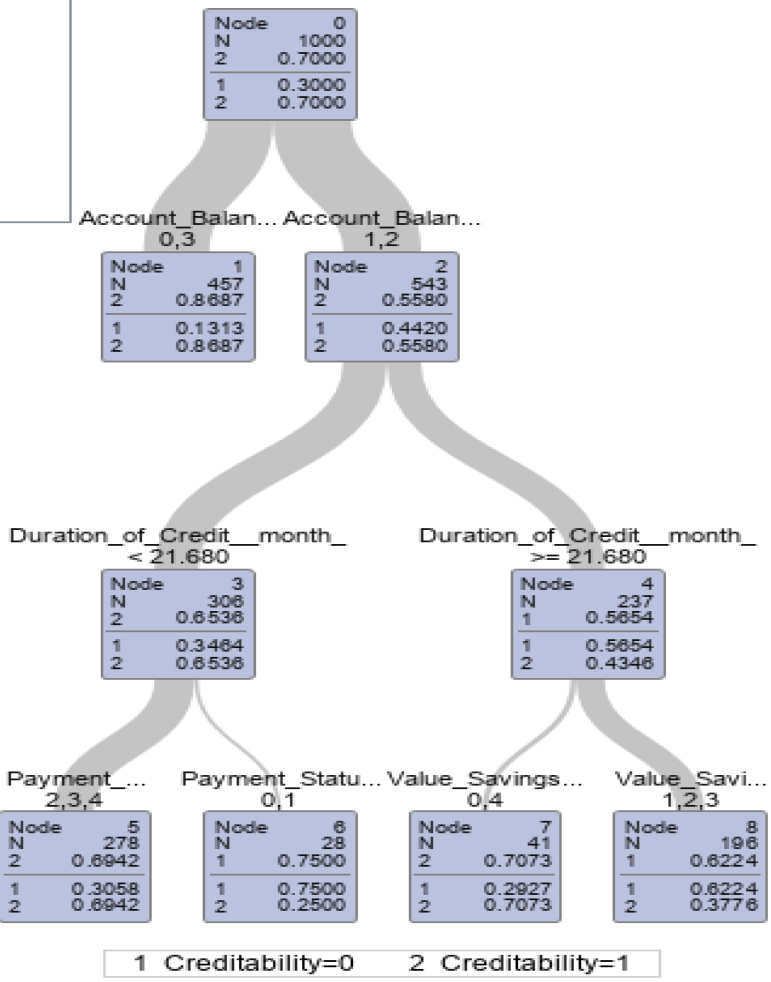


Classification Tree for Creditability



Subtree Starting at Node=0



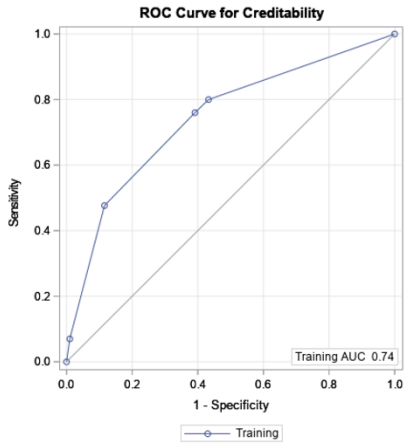


Model-Based Confusion Matrix

Actual	Predicted 0	Predicted 1	Error Rate
0	143	157	0.5233
1	81	619	0.1157

Model-Based Fit Statistics for Selected Tree

N Leaves	ASE	Mis-class	Sensitivity	Specificity	Entropy	Gini	RSS	AUC
5	0.1709	0.2380	0.4767	0.8843	0.7492	0.3419	341.9	0.7425



Variable Importance

Variable	Training Relative Importance	Training Count
Account_Balance	1.0000	6.9217	1
Duration_of_Credit__month_	0.5171	3.5792	1
Payment_Status_of_Previous_Credi	0.4578	3.1687	1
Value_Savings_Stocks	0.3923	2.7156	1

Figure 6: Python Output- Naïve Bayes

Naive Bayes Classifier

Attribute	Class 0 (0.28)	Class 1 (0.72)
Account Balance		
1	83.0	98.0
2	63.0	106.0
3	10.0	33.0
4	31.0	244.0
[total]	187.0	481.0
Duration of Credit (month)		
mean	24.2171	19.4318
std. dev.	13.1694	11.2681
weight sum	183	477
precision	1.931	1.931
Payment Status of Previous Credit		
0	16.0	14.0
1	21.0	18.0
2	107.0	238.0
3	17.0	45.0
4	27.0	167.0
[total]	188.0	482.0
Purpose		
0	61.0	100.0
1	12.0	64.0
2	35.0	81.0
3	33.0	144.0
4	3.0	4.0
5	7.0	8.0
6	14.0	22.0
8	2.0	9.0
9	22.0	48.0
10	4.0	7.0
[total]	193.0	487.0
Credit Amount		
mean	3979.4268	3142.7067
std. dev.	3551.6677	2596.0129
weight sum	183	477
precision	24.9712	24.9712
Value Savings/Stocks		
1	130.0	249.0
2	21.0	52.0
3	9.0	40.0
4	4.0	28.0
5	24.0	113.0
[total]	188.0	482.0

Duration in Current address		
1	24.0	61.0
2	54.0	143.0
3	27.0	69.0
4	82.0	208.0
[total]	187.0	481.0

Most valuable available asset		
1	37.0	147.0
2	39.0	103.0
3	68.0	166.0
4	43.0	65.0
[total]	187.0	481.0

Age (years)		
mean	33.8554	35.8884
std. dev.	11.6106	11.0512
weight sum	183	477
precision	1.0769	1.0769

Concurrent Credits		
1	35.0	54.0
2	12.0	23.0
3	139.0	403.0
[total]	186.0	480.0

No of Credits at this Bank		
mean	1.3825	1.4382
std. dev.	0.5689	0.5745
weight sum	183	477
precision	1	1

Occupation		
1	7.0	8.0
2	39.0	100.0
3	107.0	305.0
4	34.0	68.0
[total]	187.0	481.0

No of dependents		
mean	1.1803	1.1488
std. dev.	0.3845	0.3559
weight sum	183	477
precision	1	1

Telephone		
1	115.0	278.0
2	70.0	201.0
[total]	185.0	479.0

Foreign Worker		
1	183.0	457.0
2	2.0	22.0
[total]	185.0	479.0

```
# Time for evaluation on the test set
evl_nb = Evaluation(train)
evl_nb.test_model(nb, test)
print(evl_nb.summary())
```

```
Correctly Classified Instances         255               75      %
Incorrectly Classified Instances        85               25      %
Kappa statistic                          0.3959
Mean absolute error                      0.2972
Root mean squared error                  0.4181
Relative absolute error                 68.9931 %
Root relative squared error             87.1731 %
Total Number of Instances              340
```

```
Classes at different positions are  @attribute Creditability {0,1}
confusion Matrix
[[ 46.  71.]
 [ 23. 200.]]

Evaluation from the perspective of class at position 0
TP  0.39316239316239315
FP 0.1031390134529148
Precision  0.6666666666666666
Recall  0.39316239316239315

Evaluation from the perspective of class at position 1
TP  0.8968609865470852
FP 0.6068376068376068
Precision  0.7380073800738007
Recall  0.8968609865470852
```

```
# attributes: 3
attributes: [1 2 3 0]
result string:


=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 133
        Merit of best subset found:    0.076

Attribute Subset Evaluator (supervised, Class (nominal): 1 Creditability):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 2,3,4 : 3
                     Account Balance
                     Duration of Credit (month)
                     Payment Status of Previous Credit
```

```
@relation 'german_credit-weka.filters.unsupervised.attribut

@attribute Creditability {0,1}
@attribute 'Account Balance' {1,2,3,4}
@attribute 'Duration of Credit (month)' numeric
@attribute 'Payment Status of Previous Credit' {0,1,2,3,4}

@data
1,1,18,4
1,1,9,4
1,2,12,2
1,1,12,4
1,1,12,4
1,1,10,4
1,1,8,4
1,1,6,4
1,4,18,4
1,2,24,2
```