STQD6134: Business Analytics

Classification – Loan (Credit score example)

Materials from: Analysis of German Credit Data | STAT 897D (psu.edu)

Introduction

Scenario:

A bank trying to increase its customer base for some of its credit card offerings. The credit card manager wants to attract new customers who will not default on credit card loans. The bank manager might want to build a model from a similar set of past customer data that resembles the set of target customers closely. This model will be used to assign a credit score to the new customers, which in turn will be used to decide whether to issue a credit card to a potential customer. There might be several other considerations aside from the calculated credit score before a final decision is made to allocate the card.

Binary data

The bank manager might want to view and analyze several variables related to each of the potential clients in order to calculate their credit score, which is dependent on variables such as the customer's age, income group, profession, number of existing loans, and so on. The credit score here is a dependent variable, and other customer variables are independent variables. With the help of past customer data and a set of suitable statistical techniques, the manager will attempt to build a model that will establish a relationship between the dependent variable (the credit score in this case) and a lot of independent variables about the customers, such as monthly income, house and car ownership status, education, current loans already taken, information on existing credit cards, credit score and the past loan default history from the federal data bureaus, and so on.

There may be up to 500 such independent variables that are collected from a variety of sources, such as credit card application, federal data, and customers' data and credit history available with the bank. All such variables might not be useful in building the model. The number of independent variables can be reduced to a more manageable number, for instance 50 or less, by applying some empirical and scientific techniques. Once the relationship between independent and dependent variables is established using available data, the model needs validation on a different but similar set of customer data.

When a bank receives a loan application, based on the applicant's profile the bank has to make a decision regarding whether to go ahead with the loan approval or not. Two types of risks are associated with the bank's decision –

- If the applicant is a good credit risk, i.e. is likely to repay the loan, then not approving the loan to the person results in a loss of business to the bank
- If the applicant is a bad credit risk, i.e. is not likely to repay the loan, then approving the loan to the person results in a financial loss to the bank

To minimize loss from the bank's perspective, the bank needs a decision rule regarding who to give approval of the loan and who not to. An applicant's demographic and socio-economic profiles are considered by loan managers before a decision is taken regarding his/her loan application.

The German Credit Data contains data on 20 variables and the classification whether an applicant is considered a Good or a Bad credit risk for 1000 loan applicants. Here is a link to the German Credit data (*right-click and "save as"*). A predictive model developed on this data is expected to provide a bank manager guidance for making a decision whether to approve a loan to a prospective applicant based on his/her profiles.

Data cleaning and pre-processing

Predictor (Categorical)	Levels and Proportions				
Account Balance	No Account	None	Below 200 DM	200 DM or Above	
(%)	27.40%	26.90%	6.30%	39.40%	
Payment Status	Delayed	Other Credits	Paid Up	No Problem with current credits	Previous Credits Paid
(%)	4.0%	4.9%	53.0%	8.8%	29.3%
Savings/Stock Value	None	Below 100 DM	[100, 500)	[500, 1000)	Above 1000
	60.3%	10.3%	6.3%	4.8%	18.3%
Length of Current Employment	Unemployed	< 1 Year	[1, 4)	[4, 7)	Above 7
	6.2%	17.2%	33.9%	17.4%	25.3%
Installment %	Above 35%	(25%, 35%]	[20%, 25%)	Below 20%	
	13.6%	23.1%	15.7%	47.6%	
Occupation	Unemployed, unskilled	Unskilled permanent resident	Skilled	Executive	
	2.2%	20%	63%	14.8%	
Sex and Marital Status	Male, Divorced	Male Single	Male Married/ Widowed	Female	
	5.0%	31.0%	54.8%	9.2%	
Duration in Current Address	< 1 Year	[1, 4)	[4, 7)	Above 7	
	13.0%	30.8%	14.9%	41.3%	
Type of Apartment	Free	Rented	Owned		
	17.9%	71.4%	10.7%		
Most Valuable Asset	None	Car	Life Insurance	Real Estate	
	28.2%	23.2%	33.2%	15.4%	
No. of Credits at Bank	1	2 or 3	4 or 5	Above 6	
	63.3%	33.3%	2.8%	0.06%	
Guarantor	None	Co-applicant	Guarantor		
	90.7%	4.1%	5.2%		
Concurrent Credits	Other Banks	Dept Stores	None		
	13.9%	4.7%	81.4%		
No of Dependents	3 or More	Less than 3			
	84.5%	15.5%			
Telephone	Yes	No			
	40.4%	59.6%			
Foreign Worker	Yes	No			
	3.7%	96.3%			

Purpose of Credit									
New car	Used car	Furniture	Radio/TV	Appliances	Repair	Vacation	Retraining	Business	Other
10.3%	18.1%	28%	1.2%	2.2%	5.0%	0.9%	9.7%	1.2%	23.4%

highlighted red = recategorize

Depending on the cell proportions given in the one-way table above two or more cells are merged for several categorical predictors. We present below the final classification for the predictors that may potentially have any influence on Creditability

- Account Balance: No account (1), None (No balance) (2), Some Balance (3)
- Payment Status: Some Problems (1), Paid Up (2), No Problems (in this bank) (3)
- Savings/Stock Value: None, Below 100 DM, [100, 1000] DM, Above 1000 DM
- Employment Length: Below 1 year (including unemployed), [1, 4), [4, 7), Above 7
- Sex/Marital Status: Male Divorced/Single, Male Married/Widowed, Female
- No of Credits at this bank: 1, More than 1
- Guarantor: None, Yes
- Concurrent Credits: Other Banks or Dept Stores, None
- ForeignWorker variable may be dropped from the study
- Purpose of Credit: New car, Used car, Home Related, Other

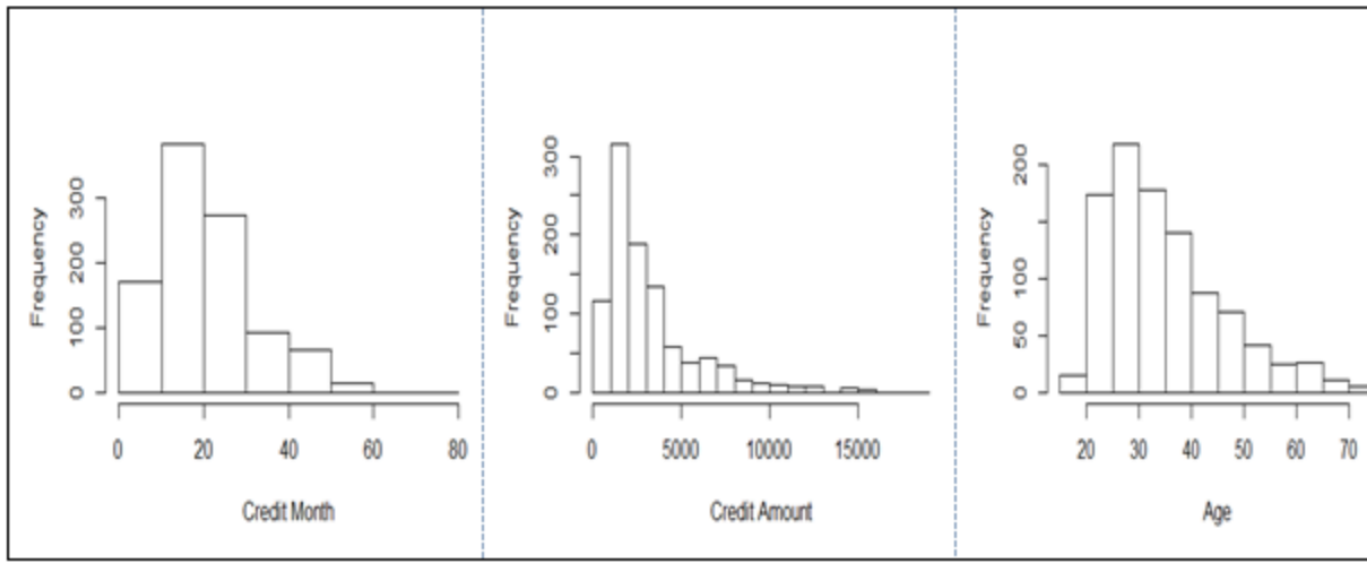
Creditability	Account.Balance 1	2	3	Row Total
0	135 0.5	105 0.4	60 0.1	300
1	139 0.5	164 0.6	397 0.9	700
Column Total	274 0.3	269 0.3	457 0.4	1000

Creditability	Payment.Status.of.Previous.Credit 1	2	3	Row Total
0	53 0.6	169 0.3	78 0.2	300
1	36 0.4	361 0.7	303 0.8	700
Column Total	89 0.1	530 0.5	381 0.4	1000

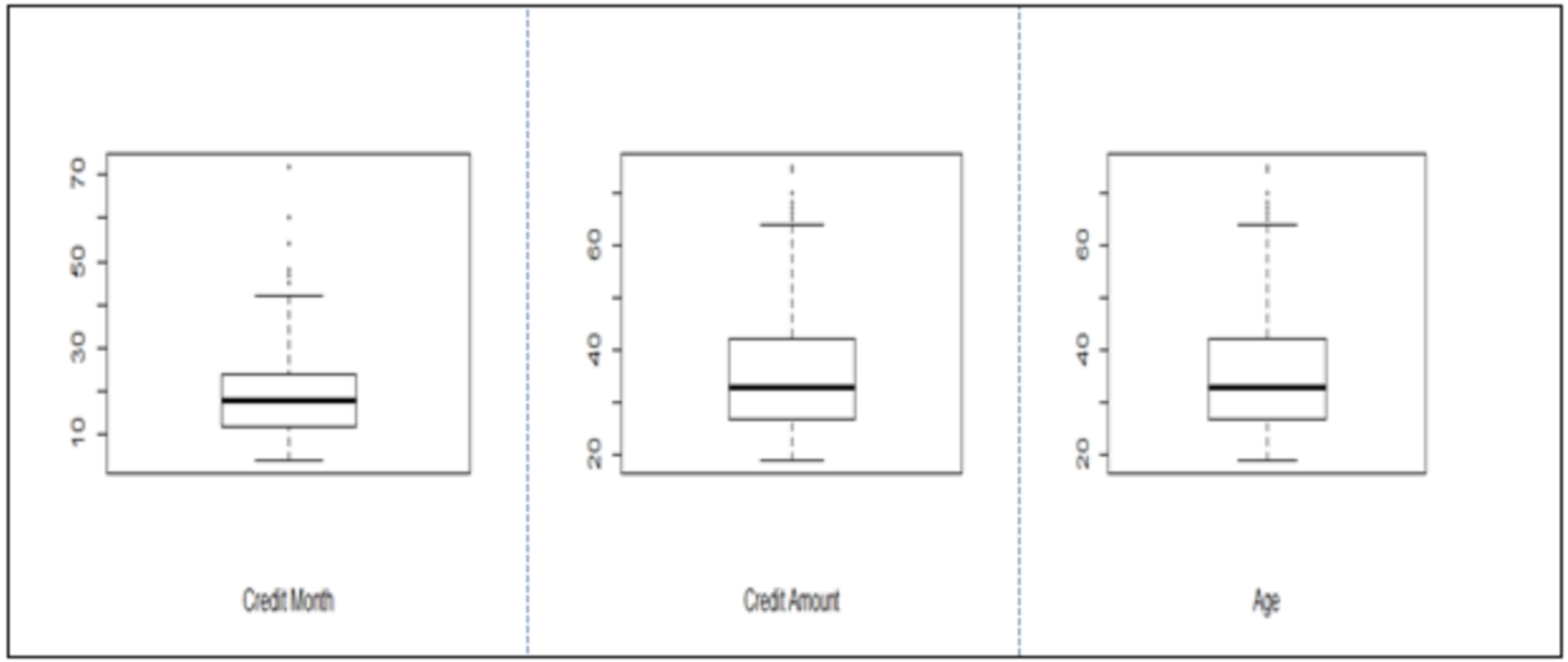
Summary for the continuous variables:

Predictor (Continuous)	Min	Q1	Median	Q3	Max	Mean	SD
Duration of Credit (Month)	4	12	18	24	72	20.9	12.06
Amount of Credit (DM)	250	1366	2320	3972	18420	3271	2822.75
Age (of Applicant)	19	27	33	42	75	35.54	11.35

Distribution of the continuous variables:



All the three variables show marked positive skewness. Boxplots bear this out even more clearly.



Logistic regression model

Predictor	Chi-square P-value
Account Balance (Nominal)	< 0.001
Payment Status (Nominal)	< 0.001
Purpose (Nominal)	< 0.001
Savings/Stock Value	< 0.001
Length of Current Employment	< 0.001
Installment %	0.14
Sex and Marital Status (Nominal)	0.01
Duration in Current Address (Nominal)	0.86
Type of Apartment	< 0.001
Most Valuable Asset (Nominal)	< 0.001
No of Credits at Bank	0.15
Guarantor	0.98
Occupation	0.42
Concurrent Credits (Nominal)	< 0.001
No of Dependents	0.92
Telephone	0.28

Predictors	Mean (Creditworthy Group)	Mean (Noncreditworthy Group)	P-value (T-test)
Duration of Credit (Month)	19.0	24.9	< 0.001
Amount of Cedit (DM)	3928.1	2985.4	< 0.001
Age	33.9	36.2	0.003

```
indexes = sample(1:nrow(German.Credit), size=0.5*nrow(German.Credit)) # Random sample of 50% of r
```

```
Train50 <- German.Credit[indexes,] # Training data contains created indices
```

```
Test50 <- German.Credit[-indexes,] # Test data contains the rest
```

```
# Using any proportion, other than 0.5 above and size Training and Test data can be constructed
```

```
LogisticModel50 <- glm(Creditability ~ Account.Balance + Payment.Status.of.Previous.Credit + Purpose + Value.Savings.Stocks +
Length.of.current.employment + Sex...Marital.Status + Most.valuable.available.asset + Type.of.apartment + Concurrent.Credits +
Duration.of.Credit..month.+ Credit.Amount + Age..years.,
family=binomial, data = Train50)

LogisticModel50final <- glm(Creditability ~ Account.Balance + Payment.Status.of.Previous.Credit + Purpose +
Length.of.current.employment + Sex...Marital.Status, family=binomial, data = Train50)

fit50 <- fitted.values(LogisticModel50S1)
Threshold50 <- rep(0,500)
for (i in 1:500)
if(fit50[i] >= 0.5) Threshold50[i] <- 1

CrossTable(Train50$Creditability, Threshold50, digits=1, prop.r=F, prop.t=F, prop.chisq=F, chisq=F, data=Train50)

perf <- performance(pred, "tpr", "fpr")
plot(perf)
```

Coefficients of Base Model				
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.81	1.12	0.73	0.47
Account.Balance2	0.28	0.30	0.95	0.34
Account.Balance3	1.30	0.31	4.25	0.00
Payment.Status.of.Previous.Credit2	0.70	0.45	1.54	0.12
Payment.Status.of.Previous.Credit3	1.61	0.48	3.38	0.00
Purpose2	-1.08	0.57	-1.91	0.06
Purpose3	-1.24	0.53	-2.34	0.02
Purpose4	-1.67	0.51	-3.25	0.00
Value.Savings.Stocks	0.25	0.11	2.34	0.02
Length.of.current.employment	0.19	0.12	1.58	0.11
Sex...Marital.Status2	0.64	0.27	2.42	0.02
Sex...Marital.Status3	0.45	0.41	1.12	0.26
Most.valuable.available.asset2	-0.45	0.35	-1.28	0.20
Most.valuable.available.asset3	-0.20	0.31	-0.67	0.51
Most.valuable.available.asset4	-1.04	0.59	-1.77	0.08
Type.of.apartment2	0.07	0.31	0.21	0.83
Type.of.apartment3	0.34	0.68	0.49	0.62
Concurrent.Credits	0.56	0.30	1.86	0.06
Instalment.per.cent	-0.32	0.12	-2.68	0.01
No.of.Credits.at.this.Bank	-0.48	0.33	-1.47	0.14
Duration.of.Credit..month.	-0.02	0.01	-1.40	0.16
Credit.Amount	0.00	0.00	-2.64	0.01

Coefficients of Final Model				
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.85	0.88	0.96	0.34
Account.Balance2	0.27	0.29	0.91	0.36
Account.Balance3	1.30	0.30	4.32	0.00
Payment.Status.of.Previous.Credit2	0.90	0.44	2.04	0.04
Payment.Status.of.Previous.Credit3	1.55	0.45	3.40	0.00
Purpose2	-1.17	0.56	-2.08	0.04
Purpose3	-1.32	0.53	-2.47	0.01
Purpose4	-1.75	0.52	-3.37	0.00
Value.Savings.Stocks	0.28	0.11	2.57	0.01
Sex...Marital.Status2	0.72	0.26	2.80	0.01
Sex...Marital.Status3	0.40	0.41	0.99	0.32
Most.valuable.available.asset2	-0.41	0.35	-1.18	0.24
Most.valuable.available.asset3	-0.27	0.30	-0.89	0.37
Most.valuable.available.asset4	-0.73	0.39	-1.86	0.06
Concurrent.Credits	0.48	0.30	1.61	0.11
Instalment.per.cent	-0.34	0.12	-2.97	0.00
Credit.Amount	0.00	0.00	-4.01	0.00

Coefficients of Final Model				
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.55	0.77	2.01	0.04
Account.Balance2	0.28	0.29	0.94	0.35
Account.Balance3	1.30	0.30	4.36	0.00
Payment.Status.of.Previous.Credit2	1.08	0.43	2.52	0.01
Payment.Status.of.Previous.Credit3	1.69	0.45	3.77	0.00
Purpose2	-1.19	0.56	-2.14	0.03
Purpose3	-1.32	0.53	-2.49	0.01
Purpose4	-1.76	0.51	-3.41	0.00
Value.Savings.Stocks	0.29	0.11	2.68	0.01
Sex...Marital.Status2	0.71	0.26	2.78	0.01
Sex...Marital.Status3	0.40	0.40	0.98	0.33
Most.valuable.available.asset2	-0.40	0.34	-1.17	0.24
Most.valuable.available.asset3	-0.27	0.30	-0.91	0.36
Most.valuable.available.asset4	-0.75	0.39	-1.91	0.06
Instalment.per.cent	-0.34	0.12	-2.93	0.00
Credit.Amount	0.00	0.00	-4.03	0.00

Null deviance: 598.536 on 499 degrees of freedom
Residual deviance: 464.01 on 477 degrees of freedom
AIC: 510.01

Null deviance: 598.53 on 499 degrees of freedom
Residual deviance: 472.12 on 483 degrees of freedom
AIC: 506.12

Null deviance: 598.53 on 499 degrees of freedom
Residual deviance: 474.67 on 484 degrees of freedom
AIC: 506.67

Test Data		50% Threshold		75% Threshold		40% Threshold	
		Creditable	Non-Creditable	Creditable	Non-Creditable	Creditable	Non-Creditable
Creditable	343	274	69	147	196	308	35
Non-Creditable	157	128	29	77	80	143	14
Total	500	Accuracy = (274+29)/500 = 60.6%		Accuracy = (147+80)/500 = 45%		Accuracy = (308+14)/500 = 64%	

performance

From ROCR v1.0-1 by Tobias Sing

99.99th Percentile

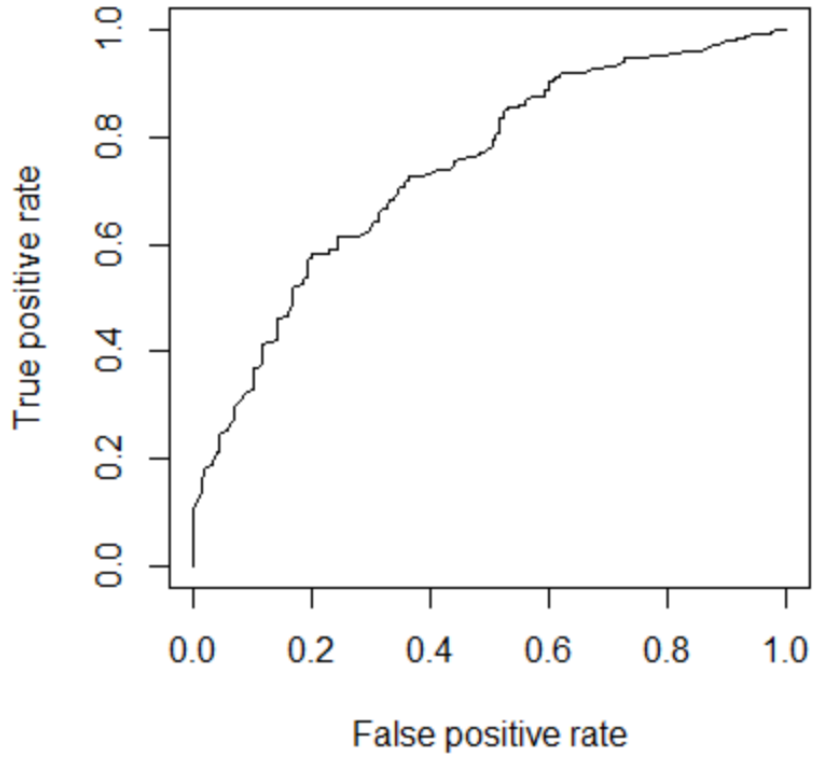
Function To Create Performance Objects

All kinds of predictor evaluations are performed using this function.

Keywords classif

Usage

```
performance(prediction.obj, measure, x.measure="cutoff", ...)
```



A Receiver Operator Characteristic (ROC) curve is a graphical plot used to show the diagnostic ability of binary classifiers. The ROC curve shows the trade-off between sensitivity (or TPR) and specificity (1 – FPR). Classifiers that give curves closer to the top-left corner indicate a better performance. As a baseline, a random classifier is expected to give points lying along the diagonal (FPR = TPR). The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

Tree-based method

```
library(tree)

Train50_tree <- tree(Creditability ~ Account.Balance+Duration.of.Credit..month.+Payment.Status.of.Previous.Credit+
Purpose+Credit.Amount+Value.Savings.Stocks+Length.of.current.employment+Instalment.per.cent+Sex...Marital.Status+
Guarantors+Duration.in.Current.address+Most.valuable.available.asset+Age..years.+Concurrent.Credits+
Type.of.apartment+No.of.Credits.at.this.Bank+Occupation+No.of.dependents+Telephone, data=Train50, method="class")

summary(Train50_tree)
plot(Train50_tree)
text(Train50_tree, pretty=0,cex=0.6)
Test50_pred <- predict(Train50_tree, Test50, type="class")
table(Test50_pred, Test50$Creditability)

Train50_prune8 <- prune.misclass(Train50_tree, best=8)
Test50_prune8_pred <- predict(Train50_prune8, Test50, type="class")
table(Test50_prune8_pred, Test50$Creditability))
```

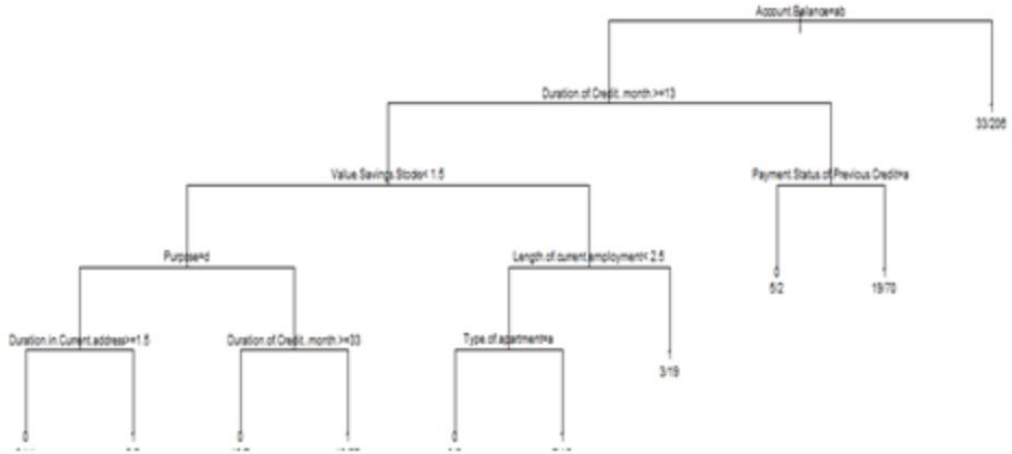
	Classified	
Test Data	Creditable	Non-Creditable
Creditable	291	93
Non-Creditable	52	64
	Accuracy = (291+64)/500 = 71%	

```
n= 500

node), split, n, loss, yval, (yprob)

* denotes terminal node

1) root 500 143 1 (0.28600000 0.71400000)
2) Account.Balance=1,2 261 110 1 (0.42145594 0.57854406)
4) Duration.of.Credit..month.>=13 165 79 0 (0.52121212 0.47878788)
8) Value.Savings.Stocks< 1.5 111 43 0 (0.61261261 0.38738739)
16) Purpose=4 45 9 0 (0.80000000 0.20000000)
32) Duration.in.Current.address>=1.5 38 4 0 (0.89473684 0.10526316) *
33) Duration.in.Current.address< 1.5 7 2 1 (0.28571429 0.71428571) *
17) Purpose=1,2,3 66 32 1 (0.48484848 0.51515152)
34) Duration.of.Credit..month.>=33 26 7 0 (0.73076923 0.26923077) *
35) Duration.of.Credit..month.< 33 40 13 1 (0.32500000 0.67500000)
70) No.of.Credits.at.this.Bank< 1.5 28 12 1 (0.42857143 0.57142857)
140) Instalment.per.cent>=2.5 17 7 0 (0.58823529 0.41176471) *
141) Instalment.per.cent< 2.5 11 2 1 (0.18181818 0.81818182) *
71) No.of.Credits.at.this.Bank>=1.5 12 1 1 (0.08333333 0.91666667) *
9) Value.Savings.Stocks>=1.5 54 18 1 (0.33333333 0.66666667)
18) Length.of.current.employment< 2.5 32 15 1 (0.46875000 0.53125000)
36) Type.of.apartment=1 10 2 0 (0.80000000 0.20000000) *
37) Type.of.apartment=2,3 22 7 1 (0.31818182 0.68181818) *
19) Length.of.current.employment>=2.5 22 3 1 (0.13636364 0.86363636) *
5) Duration.of.Credit..month.< 13 96 24 1 (0.25000000 0.75000000)
10) Payment.Status.of.Previous.Credit=1 7 2 0 (0.71428571 0.28571429) *
11) Payment.Status.of.Previous.Credit=2,3 89 19 1 (0.21348315 0.78651685) *
3) Account.Balance=3 239 33 1 (0.13807531 0.86192469)
6) Purpose=4 72 18 1 (0.25000000 0.75000000)
12) Concurrent.Credits< 1.5 11 4 0 (0.63636364 0.36363636) *
13) Concurrent.Credits>=1.5 61 11 1 (0.18032787 0.81967213) *
7) Purpose=1,2,3 167 15 1 (0.08982036 0.91017964) *
```



	Classified	
Test Data	Creditable	Non-Creditable
Creditable	315	114
Non-Creditable	28	43
	Accuracy = (315+43)/500 = 71.6%	

```
n= 500

node), split, n, loss, yval, (yprob)
  * denotes terminal node

1) root 500 143 1 (0.2860000 0.7140000)
2) Account.Balance=1,2 261 110 1 (0.4214559 0.5785441)
4) Duration.of.Credit..month.>=13 165 79 0 (0.5212121 0.4787879)
8) Value.Savings.Stocks< 1.5 111 43 0 (0.6126126 0.3873874)
16) Purpose=4 45 9 0 (0.8000000 0.2000000)
32) Duration.in.Current.address>=1.5 38 4 0 (0.8947368 0.1052632) *
33) Duration.in.Current.address< 1.5 7 2 1 (0.2857143 0.7142857) *
17) Purpose=1,2,3 66 32 1 (0.4848485 0.5151515)
34) Duration.of.Credit..month.>=33 26 7 0 (0.7307692 0.2692308) *
35) Duration.of.Credit..month.< 33 40 13 1 (0.3250000 0.6750000) *
9) Value.Savings.Stocks>=1.5 54 18 1 (0.3333333 0.6666667)
18) Length.of.current.employment< 2.5 32 15 1 (0.4687500 0.5312500)
36) Type.of.apartment=1 10 2 0 (0.8000000 0.2000000) *
37) Type.of.apartment=2,3 22 7 1 (0.3181818 0.6818182) *
19) Length.of.current.employment>=2.5 22 3 1 (0.1363636 0.8636364) *
5) Duration.of.Credit..month.< 13 96 24 1 (0.2500000 0.7500000)
10) Payment.Status.of.Previous.Credit=1 7 2 0 (0.7142857 0.2857143) *
11) Payment.Status.of.Previous.Credit=2,3 89 19 1 (0.2134831 0.7865169) *
3) Account.Balance=3 239 33 1 (0.1380753 0.8619247) *
```

Summary

- Classification techniques
  - Logistic regression
  - Tree based method