Published on *STAT 897D* (https://onlinecourses.science.psu.edu/stat857)

Analysis of German Credit Data

Data mining is a critical step in knowledge discovery involving theories, methodologies and tools for revealing patterns in data. It is important to understand the rationale behind the methods so that tools and methods have appropriate fit with the data and the objective of pattern recognition. There may be several options for tools available for a data set.

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

Objective of Analysis:

Minimization of risk and maximization of profit on behalf of 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 Files for this case (*right-click and "save as"*) :

- German Credit data - german_credit.csv [1]
- Training dataset - Training50.csv [2]
- Test dataset - Test.csv [3]

The following analytical approaches are taken:

- Logistic regression: The response is binary (Good credit risk or Bad) and several predictors are available.
- Discriminant Analysis:
- Tree-based method and Random Forest

Sample R code for Reading a .csv file

GCD.1 - Exploratory Data Analysis (EDA) and Data Pre-processing

Before getting into any sophisticated analysis, the first step is to do an EDA and data cleaning. Since both categorical and continuous variables are included in the data set, appropriate tables and summary statistics are provided.

Sample R code for creating marginal proportional tables

Proportions of applicants belonging to each classification of a categorical variable are shown in the following table (below). The pink shadings indicate that these levels have too few observations and the levels are merged for final analysis.

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%

Since most of the predictors are categorical with several levels, the full cross-classification of all variables will lead to zero observations in many cells. Hence we need to reduce the table size. For details of variable names and classification see Appendix 1.

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

Cross-tabulation of the 9 predictors as defined above with Creditability is shown below. The proportions shown in the cells are column proportions and so are the marginal proportions. For example, 30% of 1000 applicants have no account and another 30% have no balance while 40% have some balance in their account. Among those who have no account 135 are found to be Creditable and 139 are found to be Non-Creditable. In the group with no balance in their account, 40% were found to be on-Creditable whereas in the group having some balance only 1% are found to be Non-Creditable.

Sample R code for creating K1 x K2 contingency table.

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

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

	Value.Savings.Stocks				
Creditability	1	2	3	4	Row Total
0	217	34	17	32	300
	0.4	0.3	0.2	0.2	
1	386	69	94	151	700
	0.6	0.7	0.8	0.8	
Column Total	603	103	111	183	1000
	0.6	0.1	0.1	0.2	

	Length.of.current.employment				
Creditability	1	2	3	4	Row Total
0	93	104	39	64	300
	0.4	0.3	0.2	0.3	
1	141	235	135	189	700
	0.6	0.7	0.8	0.7	
Column Total	234	339	174	253	1000
	0.2	0.3	0.2	0.3	

	Sex & Marital.Status			
Creditability	1	2	3	Row Total
0	129	146	25	300
	0.4	0.3	0.3	
1	231	402	67	700
	0.6	0.7	0.7	
Column Total	360	548	92	1000
	0.4	0.5	0.1	

	No.of.Credits.at.this.Bank		
Creditability	1	2	Row Total
0	200	100	300
	0.3	0.3	
1	433	267	700
	0.7	0.7	
Column Total	633	367	1000
	0.6	0.4	

	Guarantors		
Creditability	1	2	Row Total
0	272	28	300
	0.3	0.3	
1	635	65	700
	0.7	0.7	
Column Total	907	93	1000
	0.9	0.1	

	Concurrent.Credits		
Creditability	1	2	Row Total
0	76	224	300
	0.4	0.3	
1	110	590	700
	0.6	0.7	
Column Total	186	814	1000
	0.2	0.8	

	Type.of.apartment			
Creditability	1	2	3	Row Total
0	70	186	44	300
	0.4	0.3	0.4	
1	109	528	63	700
	0.6	0.7	0.6	
Column Total	179	714	107	1000
	0.2	0.7	0.1	

	No.of.dependents		
Creditability	1	2	Row Total
0	254	46	300
	0.3	0.3	
1	591	109	700
	0.7	0.7	
Column Total	845	155	1000
	0.8	0.2	

	Purpose				
Creditability	1	2	3	4	Row Total
0	17	58	96	129	300
	0.2	0.3	0.3	0.4	
1	86	123	268	223	700
	0.8	0.7	0.7	0.6	
Column Total	103	181	364	352	1000
	0.1	0.2	0.4	0.4	

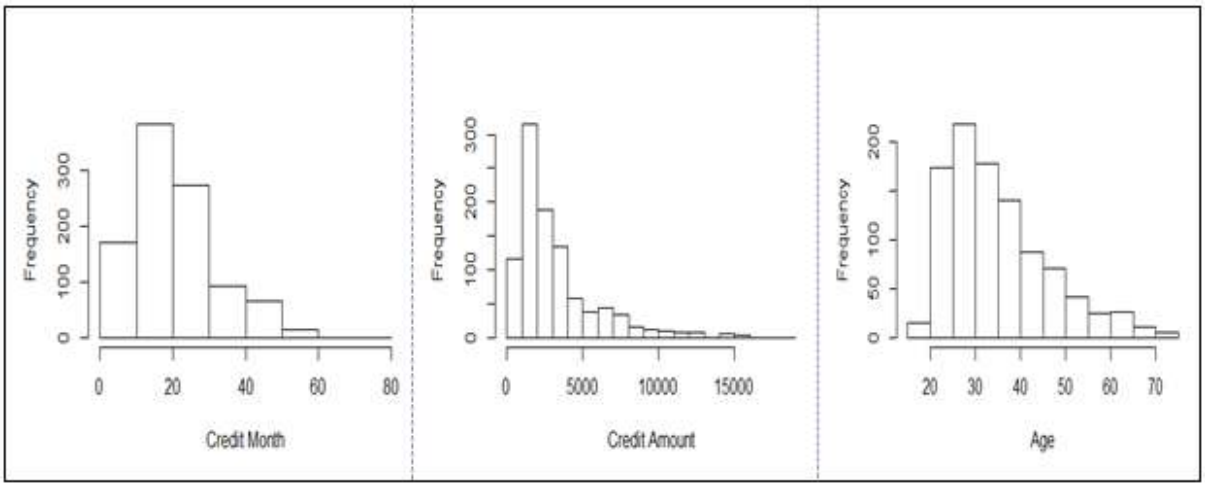
	Instalment.per.cent				
Creditability	1	2	3	4	Row Total
0	34	62	45	159	300
	0.2	0.3	0.3	0.3	
1	102	169	112	317	700
	0.8	0.7	0.7	0.7	
Column Total	136	231	157	476	1000
	0.1	0.2	0.2	0.5	

Summary for the continuous variables:

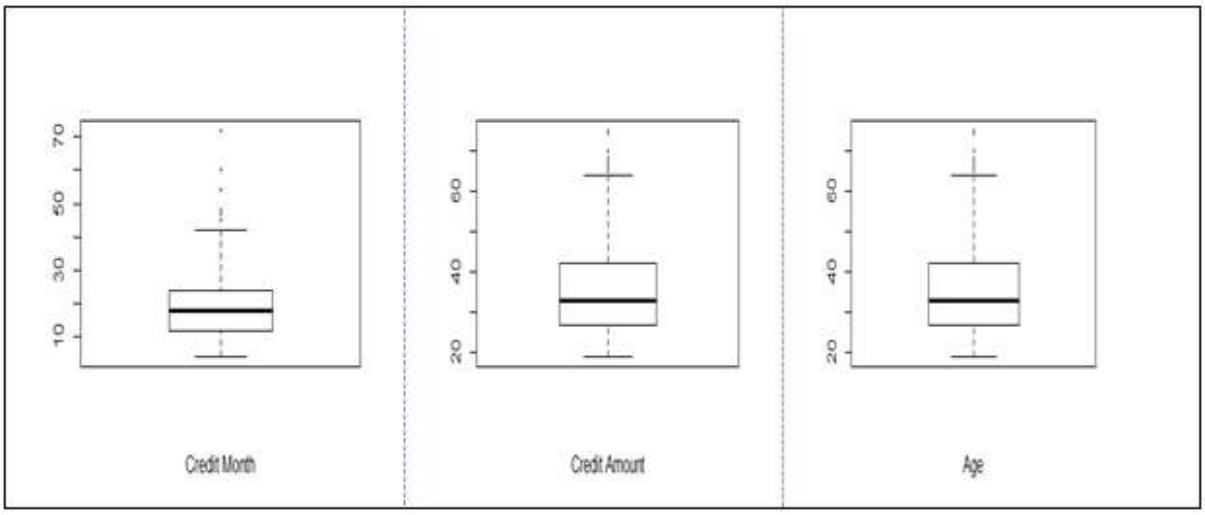
Sample R code for Descriptive Statistics.

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.



In preparation of predictors to use in building a logistic regression model, we consider bivariate association of the response (Creditability) with the categorical predictors.

GCD.2 - Towards Building a Logistic Regression Model

Since the number of predictors in this problem is not very high, it is possible to look into the dependency of the response (Creditability) on each of them individually. The following table summarizes the chi-square *p*-values for each contingency table. Note that among the sample of size 1000, 700 were Creditable and 300 Non-Creditable. This classification is based on the Bank's opinion on the actual applicants.

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

Only significant predictors are to be included in the logistic regression model. Since there are 1000 observations 50:50 cross-validation scheme is tried:

Model Building with 50:50 Cross-validation

Sample R code for 50:50 cross-validation data creation.

1000 observations are randomly partitioned into two equal sized subsets – Training and Test data. A logistic model is fit to the Training set.

Results are given below, shaded rows indicate variables not significant at 10% level.

Sample R code for for Logistic Model building with Training data and assessing for Test data.

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

R output:

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

Removing the nonsignificant variables a second logistic regression is fit to the data.

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

R output:

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

Need to remove another variable to come up with a model where all predictors are significant at 10% level.

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

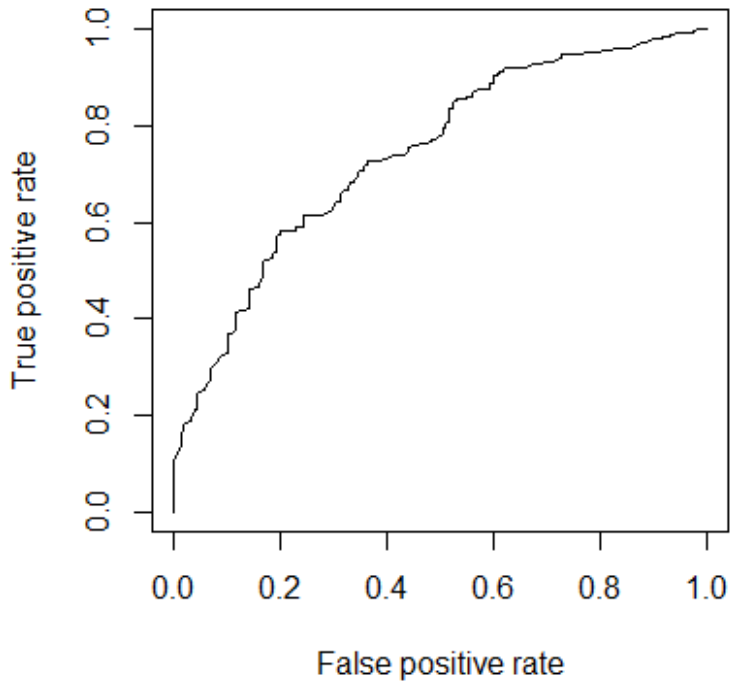
R output:

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

This model is recommended as the final model based on the Training Data. Final performance of a model is evaluated by considering the classification power. Following are a few tables defined at different thresholds of classification.

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%	

Following figure shows the performance of the classifier through ROC curve.



GCD.3 - Applying Discriminant Analysis

For discriminant analysis all the predictors are not used. Only the continuous variables and the ordinal variables are used as for the nominal variables there will be no concept of group means and linear discriminants will be difficult to interpret. The predictors are assumed to have a multivariate normal distribution.

Sample R code for Discriminant Analysis.

Prior probability was taken as observed in the Training sample:

71.4% Creditable and 28.6% Non-creditable

Group Means	Creditable	Non-Creditable
Value.Savings.Stocks	2.00	1.57
Length.of.current.employment	2.53	2.26
Concurrent.Credits	1.87	1.75
Instalment.per.cent	2.95	3.09
No.of.Credits.at.this.Bank	1.37	1.32
Credit.Amount	2897.58	4102.17
Duration.of.Credit..month.	19.93	25.59
Duration.in.Current.address	2.81	2.87
Age..years.	35.93	34.29

Linear Discriminant Analysis

Variables	Coefficients of Linear Discriminants
Value.Savings.Stocks	0.41
Length.of.current.employment	0.31
Concurrent.Credits	1.06
Instalment.per.cent	-0.32
No.of.Credits.at.this.Bank	0.22
Credit.Amount	0.00
Duration.of.Credit..month.	-0.03
Duration.in.Current.address	-0.22
Age..years.	0.01

	Classified	
Test Data	Creditable	Non-Creditable
Creditable	290	53
Non-Creditable	143	14
	Accuracy = (290+14)/500 = 60.8%	

Quadratic Discriminant Analysis

	Classified	
Test Data	Creditable	Non-Creditable
Creditable	266	77
Non-Creditable	132	25
	Accuracy = (266+25)/500 = 58.2%	

Neither logistic regression nor discriminant analysis is performing well for this data. The reason DA may not do well is that, most of the predictors are categorical and nominal predictors are not used in this analysis.

GCD.4 - Applying Tree-Based Methods

Sample R code for Tree method.

Both categorical and continuous predictors are used for binary classification. Using `rpart{library=rpart}`, the following tree is obtained without any pruning.

R output:

```
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) *
```

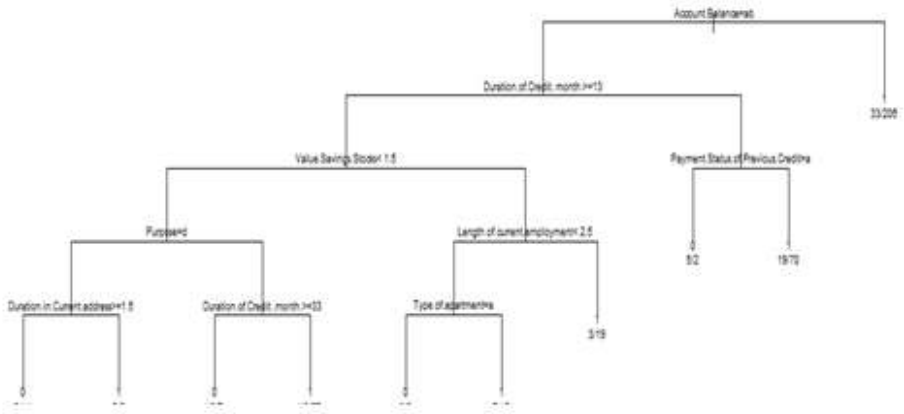
Applying the procedure on Test data, classification probability shows improvement.

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

The CP table is as follows:

	CP	nsplit	rel error	xerror	xstd
1	0.058	0.000	1.000	1.000	0.071
2	0.049	3.000	0.825	1.000	0.071
3	0.021	5.000	0.727	0.902	0.068
4	0.010	9.000	0.643	0.895	0.068
5	0.010	13.000	0.601	0.930	0.069

Following is the result for pruning the above tree for cross-validated classification error rate 90%.



```
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) *
```

There is minor improvement in accuracy % also

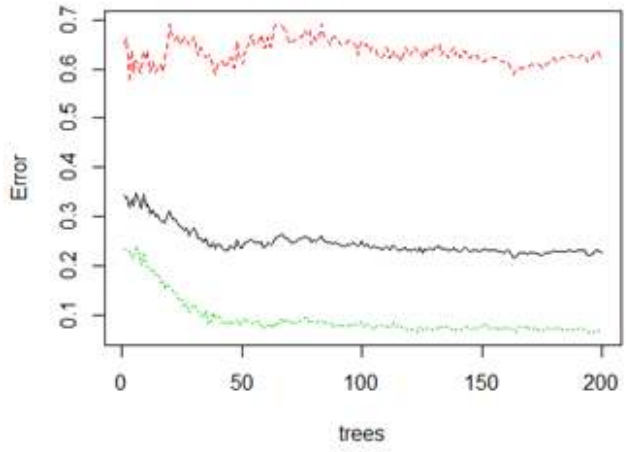
	Classified	
Test Data	Creditable	Non-Creditable
Creditable	315	114
Non-Creditable	28	43
	Accuracy = (315+43)/500 = 71.6%	

Conclusion: For this data set tree-based method seems to be working better than logistic regression or discriminant analysis.

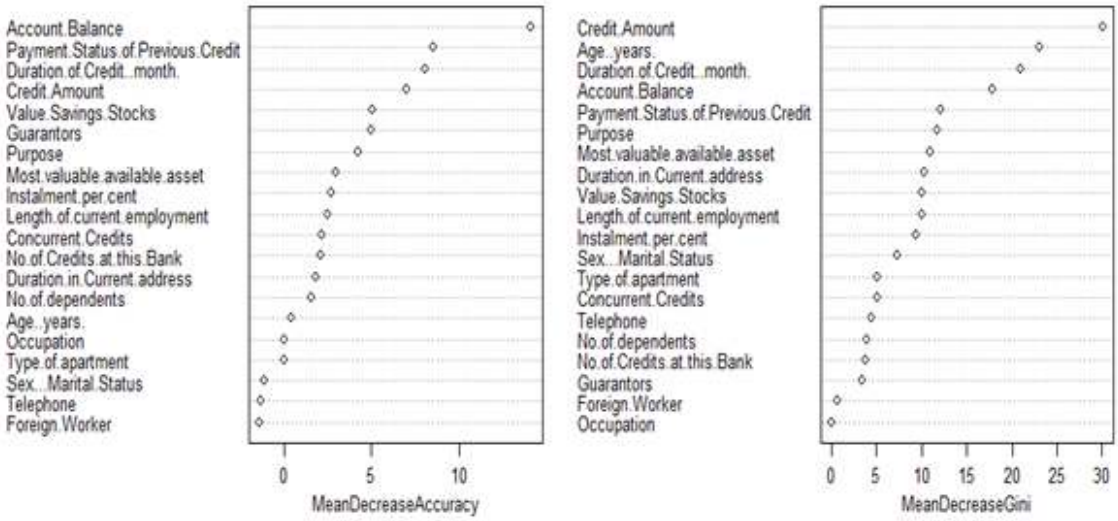
GCD.5 - Random Forest

Sample R code for Random Forest.

Completely unsupervised random forest method on Training data with *ntree* = 200 leads to the following error plot:



Importance of predictors are given in the following dotplot.



which gives rise to the following classification table:

	Classified	
Test Data	Creditable	Non-Creditable
Creditable	321	111
Non-Creditable	22	46
	Accuracy = (321+46)/500 = 73.4%	

With judicious choice of more important predictors, further improvement in accuracy is possible. But as improvement is slight, no attempt is made for supervised random forest.

GCD.6 - Cost-Profit Consideration

Ultimately these statistical decisions must be translated into profit consideration for the bank. Let us assume that a correct decision of the bank would result in 35% profit at the end of 5 years. A correct decision here means that the bank predicts an application to be good or credit-worthy and it actually turns out to be credit worthy. When the opposite is true, i.e. bank predicts the application to be good but it turns out to be bad credit, then the loss is 100%. If the bank predicts an application to be non-creditworthy, then loan facility is not extended to that applicant and bank does not incur any loss (opportunity loss is not considered here). The cost matrix, therefore, is as follows:

		Predicted	
		Creditworthy	Non-Creditworthy
Actual	Creditworthy	+0.35	0
	Non-Creditworthy	- 1.00	0

Out of 1000 applicants, 70% are creditworthy. A loan manager without any model would incur [0.7*0.35 + 0.3 (-1)] = - 0.055 or 0.055 unit loss. If the average loan amount is 3200 DM (approximately), then the total loss will be 1760000 DM and per applicant loss is 176 DM.

Logistic regression model performance:

Actual	Prediction by Logistic Regression			Prediction by DA		Prediction by Tree	Random Forest
	50% Threshold	75% Threshold	40% Threshold	Linear	Quadratic		
	Creditable	Creditable	Creditable	Creditable	Creditable	Creditable	Creditable
Creditable	0.548	0.294	0.616	0.58	0.532	0.63	0.642
Non-Creditable	0.256	0.154	0.286	0.286	0.264	0.056	0.044
Per applicant profit	-0.0642	-0.0511	-0.0704	-0.083	-0.0778	0.1645	0.1807

Tree-based classification and random forest show a per unit profit; other methods are not doing well.

GCD - Appendix - Description of Dataset

Variable	Description	Categories	Score	rel. frequency in % for

Data and additional description may be found here:

http://www.stat.uni-muenchen.de/service/datenarchiv/kredit/kredit_e.html [4]

Source URL: https://onlinecourses.science.psu.edu/stat857/node/215

Links:
[1] https://onlinecourses.science.psu.edu/stat857/sites/onlinecourses.science.psu.edu.stat857/files/german_credit.csv
[2] https://onlinecourses.science.psu.edu/stat857/sites/onlinecourses.science.psu.edu.stat857/files/Training50.csv
[3] https://onlinecourses.science.psu.edu/stat857/sites/onlinecourses.science.psu.edu.stat857/files/Test50.csv
[4] http://www.stat.uni-muenchen.de/service/datenarchiv/kredit/kredit_e.html