GERMAN CREDIT DATA

Summary

Banks and lending companies need to make the right decision in determining who should or should not get loan approvals. The purpose of this analysis is to minimize financial loss from the banks prospective.
The Study analysed Credit data, applied Predictive Modelling using Classification Algorithms and Post Predictive Analysis using Association Rules to recommend to the company about the best Characteristics of Creditable and their Potential Target Customers.
An exploratory analysis was done and the attributes that give biggest information gain were taken further for analysis.
The data preparation was followed by a classification analysis using different algorithms.
Clustering and Association mining were then used on the good class to find out the characteristics of the customers who are creditable.

The tools used in the study are R and Weka.

The study was conducted by Tejinder Sahni, Francisco Andrade, Glen Alexander and Joseph Costa.

Introduction

The business problem is determining a better way to reduce the number of credit cards issued to people who have bad credit or may default on a credit card if they are issued one. Bad debt is very costly for this industry and any improvement would help save millions.

The stakeholders determined here are the bank, investors, financial advisors, any employee within the bank and possible people looking for a credit card.

Data Preparation

Attribute type: nominal, ordinal or quantitative.

Summary of the Type:

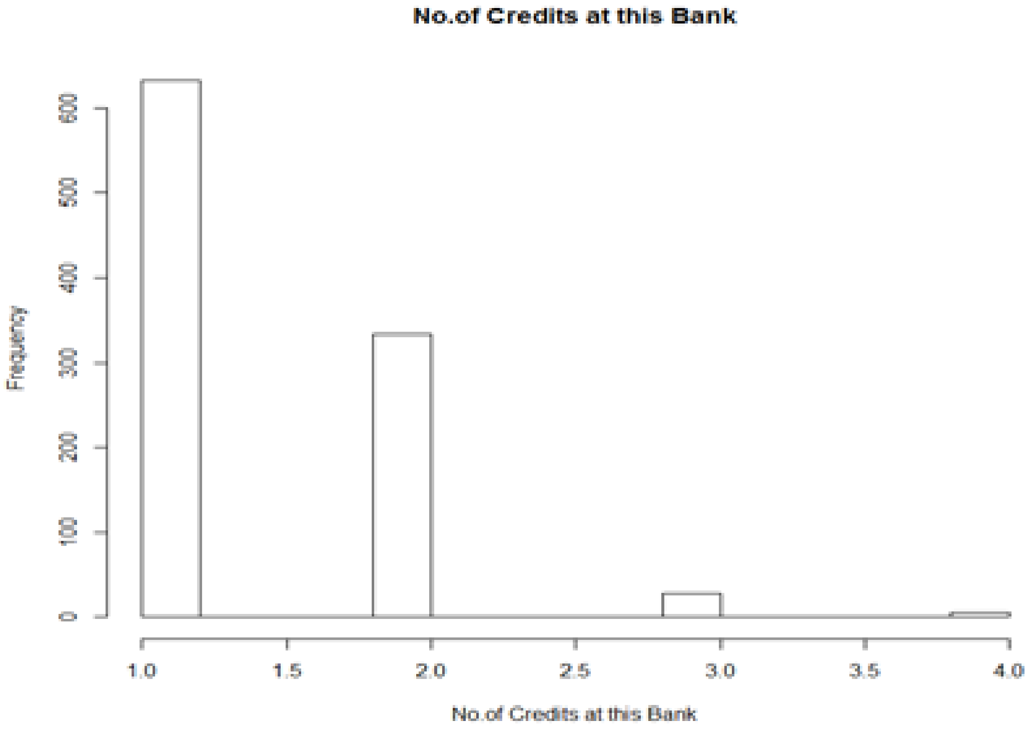
Numeric	**Nominal**
Duration of Credit	Credibility
Installment percent	Account Balance
Age	Payment status of previous credit
Credit Amount	Purpose of the loan
No. of credits at this bank	Value savings / stock
No. of Dependants	Length of current employment
	Sex and Marital status
	Guarantors
	Duration in current address
	Most Valuable available asset
	Concurrent credits
	Type of apartment
	Occupation
	Telephone
	Foreign Worker

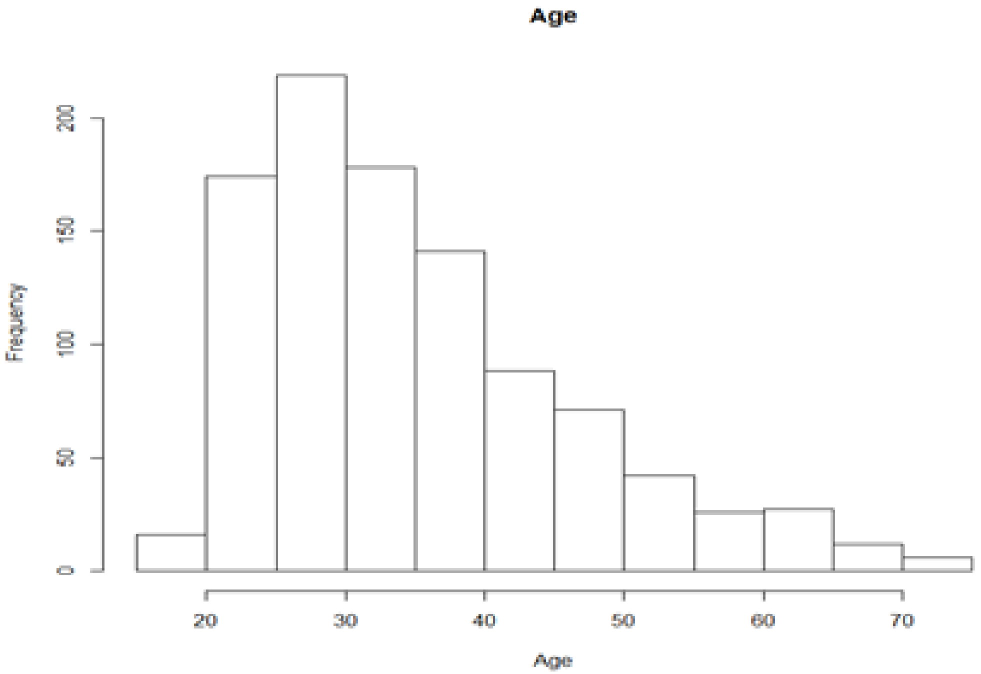
Max, Min, Mean and Standard Deviation of the numerical attributes.

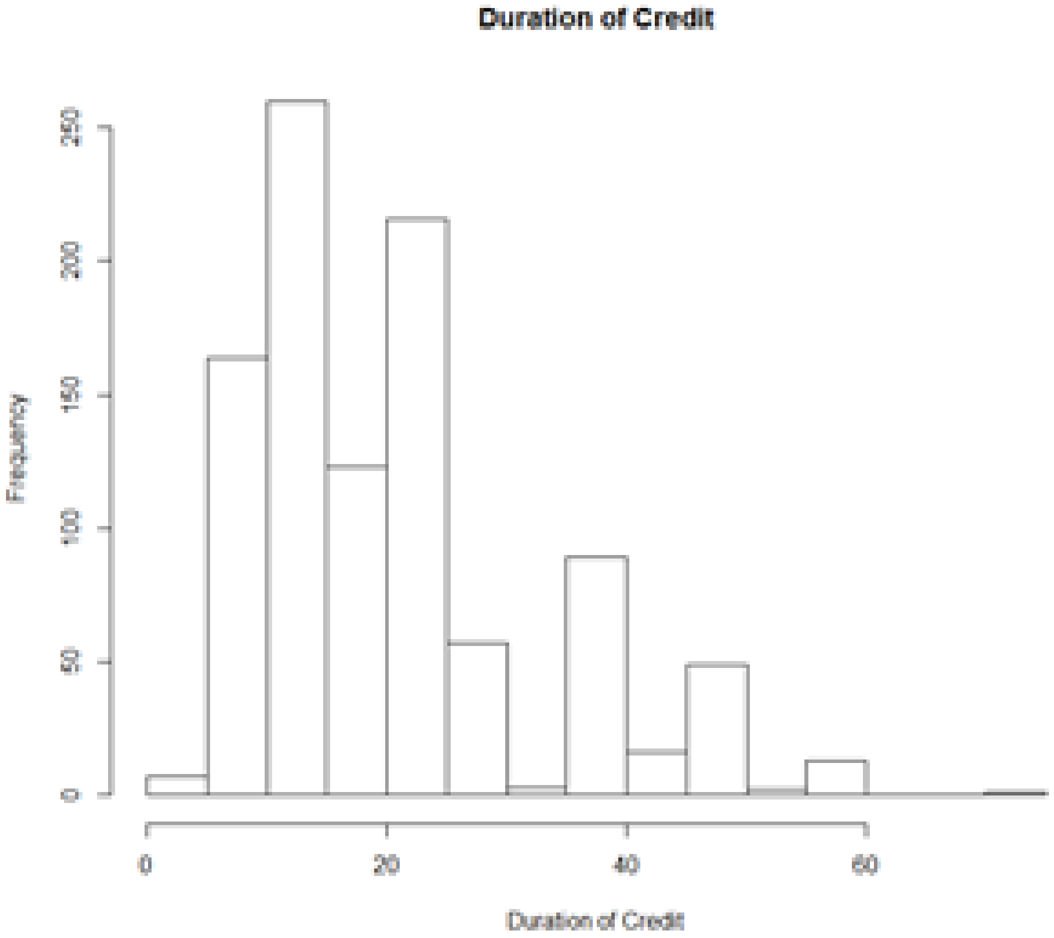
Attribute	Min	Max	Mean	Standard Deviation
Duration of Credit	4	72	20.903	12.059

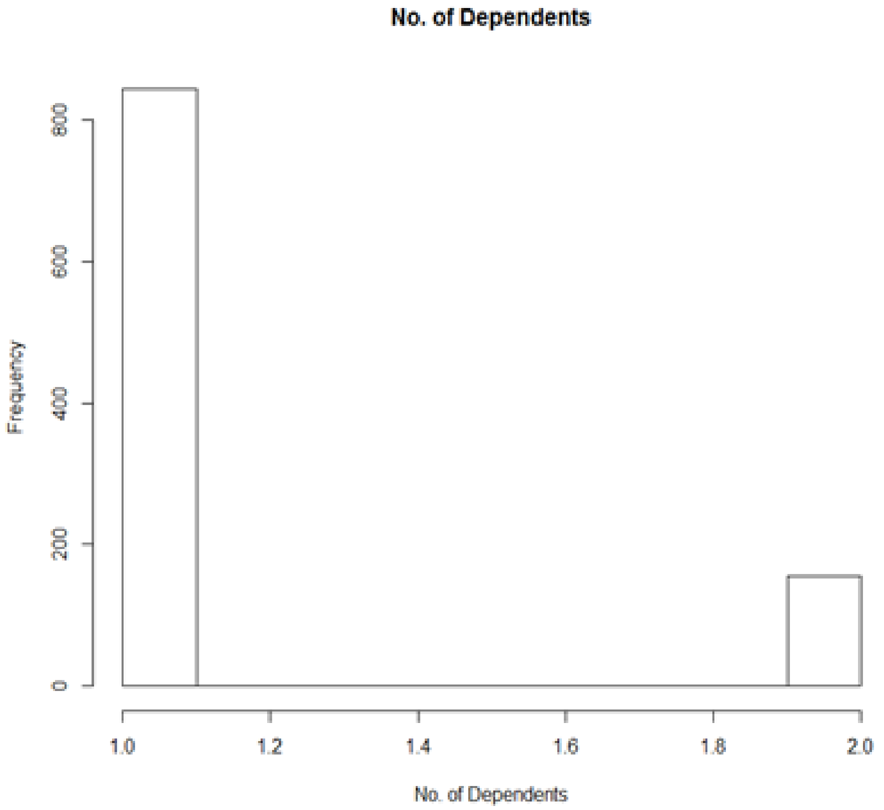
Credit Amount	250	18424	3271.25	2822.75
Installment Percent	1	4	2.973	1.119
No. of credits at this Bank	1	4	1.407	.578
Age	19	75	35.542	11.353
No. Of Dependants	1	2	1.155	0.362

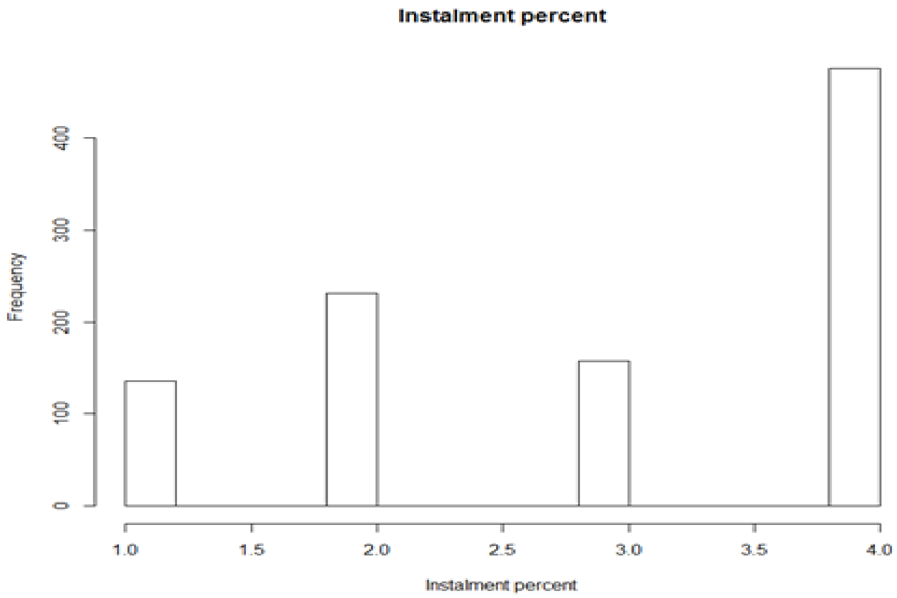
Histograms were made for the numerical attributes to see the normality of data .

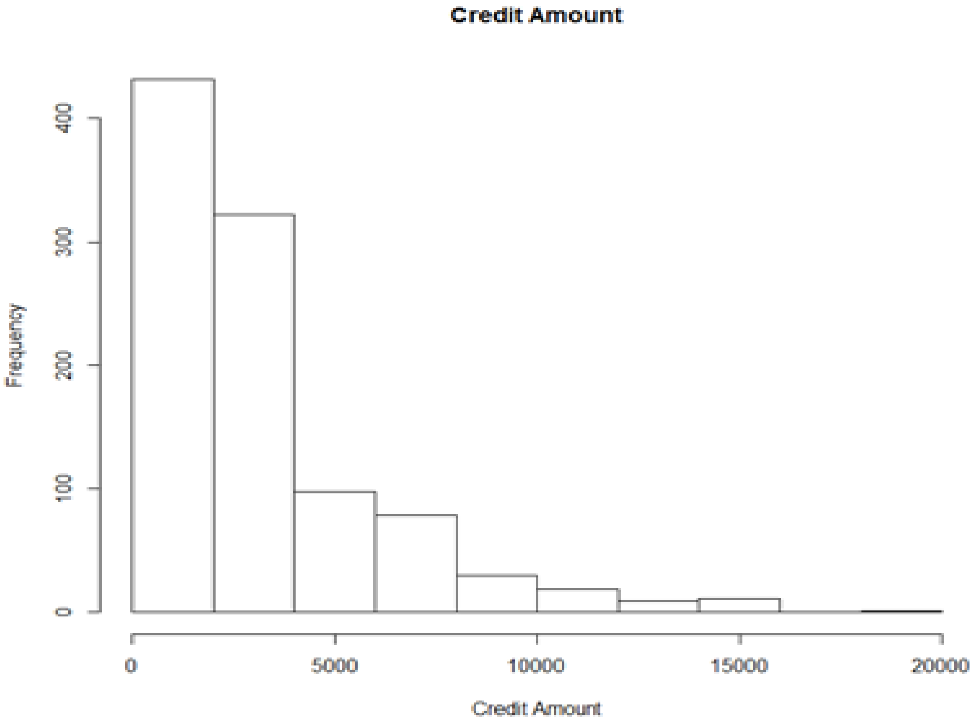



Age
Frequency
0
50
100
150
200
20
30
40
50
60
70
Age

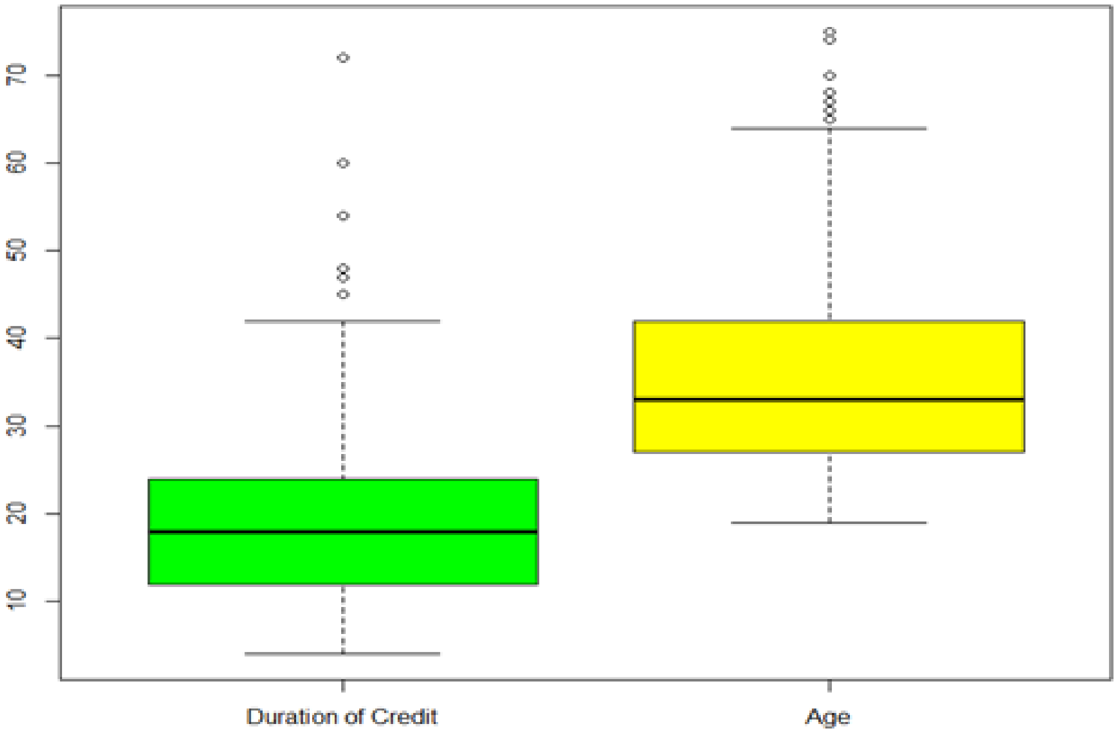

Duration of Credit
Frequency
0
50
100
150
200
250
0
20
40
60
Duration of Credit

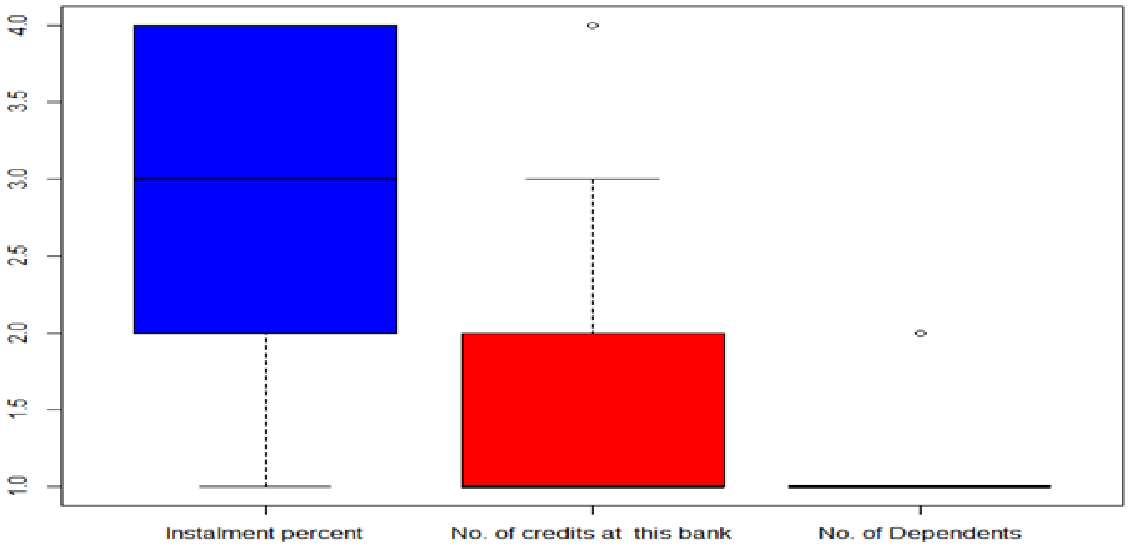

No. of Dependents
Frequency
0
200
400
600
800
1.0
1.2
1.4
1.6
1.8
2.0
No. of Dependents

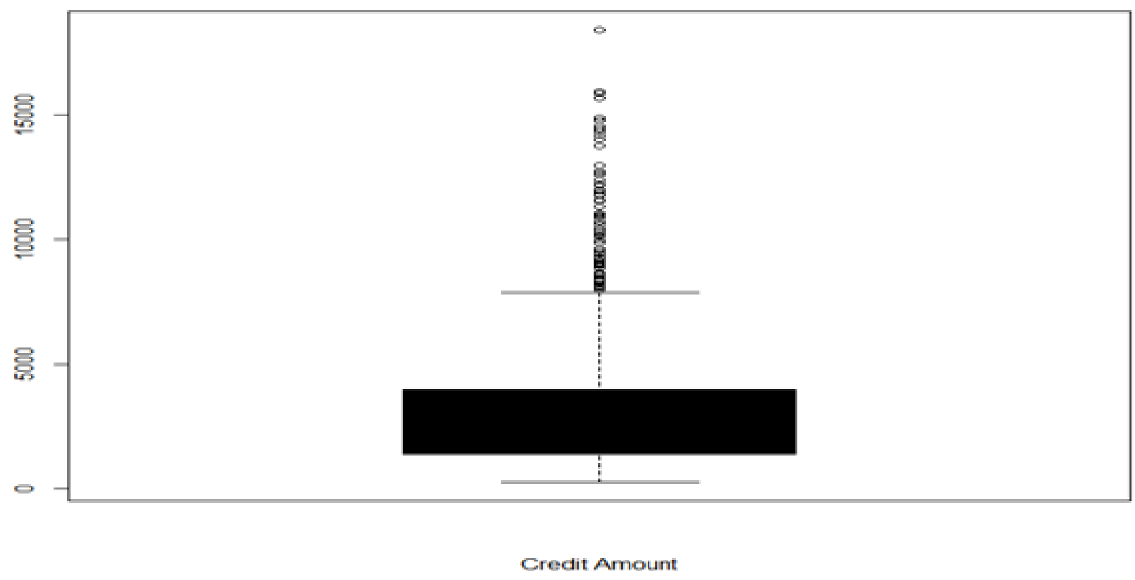

Instalment percent
Frequency
0
100
200
300
400
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Instalment percent


Credit Amount
Frequency
0
100
200
300
400
0
5000
10000
15000
20000
Credit Amount

Box plots were done to find out the outliers in the different attributes.







Here is a list of outliers in the data set:

Attribute	Outliers
Duration of Credit	>42
Credit Amount	>7500
Instalment percent	
No. of credits at this bank	>=3
Age	>50
No. of Dependents	2

The above chart shows the values of the attributes which are outliers and can be deleted from the data .After deleting the outliers the remaining data is left with 741 instances instead of 1000. Classification was run on the smaller data set with 741 instances and compared to the results of the original data set of 1000 instances but it was found that the classification results were better with the original dataset .So the original data set was selected for further analysis .

Using the Info Gain Attribute Eval in Weka , the attributes that gave the maximum gain were selected . The following attributes were taken forward to be used for the classification and post predictive analysis .

Here is the list of attributes:

1. **Account Balance** Checking account status (1: < 0 DM, 2: 0<=...<200 DM, 2 > 200 DM, 4: No checking account), where DM= Deutsche Mark (qualitative attribute).
2. **Credit history (qualitative)** 0: no credits taken, 1: all credits at this bank paid back duly, 2: existing credits paid back duly till now, 3: delay in paying off in the past, 4: critical account
3. **Duration of Credit (month)** Duration of credit in months (numerical)
4. **Value Savings/Stocks** Qualitative attribute showing average balance in savings and stocks (1 : < 100 DM, 2: 100<= ... < 500 DM, 3 : 500<= ... < 1000 DM, 4 : =>1000 DM, 5: unknown/ no savings account)
5. **Purpose** Qualitative attribute showing the purpose of the loan (0: New car, 1: Used car , 2: Furniture/Equipment, 3: Radio/Television, 4: Domestic Appliances , 5: Repairs ,6: Education ,7: Vacation, 8: Retraining ,9: Business, 10: Others)
6. **Credit Amount Numerical** value showing the credit amount
7. **Most valuable available asse**t Qualitative attribute showing valuable assets (1 : real estate 2 : savings agreement/ life insurance, 3 : car or other, 4 : unknown / no property) 14. Age (years): Numerical value showing age in years
8. **Length of current employment** Qualitative attribute showing length of employment (1 : unemployed, 2: < 1 year, 3: 1<=...<4 years, 4: 4<=...<7 years, 5:>=7years).
9. **Type of apartment** Type of housing (1 : rent, 2 : own, 3 : for free)
10. **Age (years)** Numerical value showing age in years
11. **Concurrent Credits** Installment plans (1 : bank, 2 : stores, 3 : none)
12. **Sex & Marital Status** Qualitative attribute showing gender and marital status (1: male : divorced/separated, 2: female : divorced/separated/married, 3 : male: single, 4: male : married/widowed, 5 : female : single)
13. **Foreign Worker** Qualitative attribute showing whether the person is the foreign worker or not (1: yes, 2: no)
14. **Guarantors** (Qualitative) Guarantors and co-applicants: (1 : none, 2 : co-applicant, 3 : guarantor)
15. **Occupation** Job (Qualitative) (1 : unemployed/ unskilled - non-resident, 2 : unskilled - resident, 3 : skilled employee / official, 4 : management/ self-employed/highly qualified employee/ officer)

The attributes that were dropped are below.

Following attributes can be dropped from the study:

16. **Instalment percent**: Installment rate in percentage of disposable income (numerical)
17. **Duration in Current address**: Qualitative value showing the duration in current address (1: <= 1 year, 1<...<=2 years, 2<...<=3 years, 3:>4years)
18. **No of Credits at this Bank:** Numerical value showing number of existing credits at the bank
19. **No of dependents:** Numerical value showing number of dependents
20. **Telephone:** Qualitative attribute for telephone number (1: yes, 2: No)

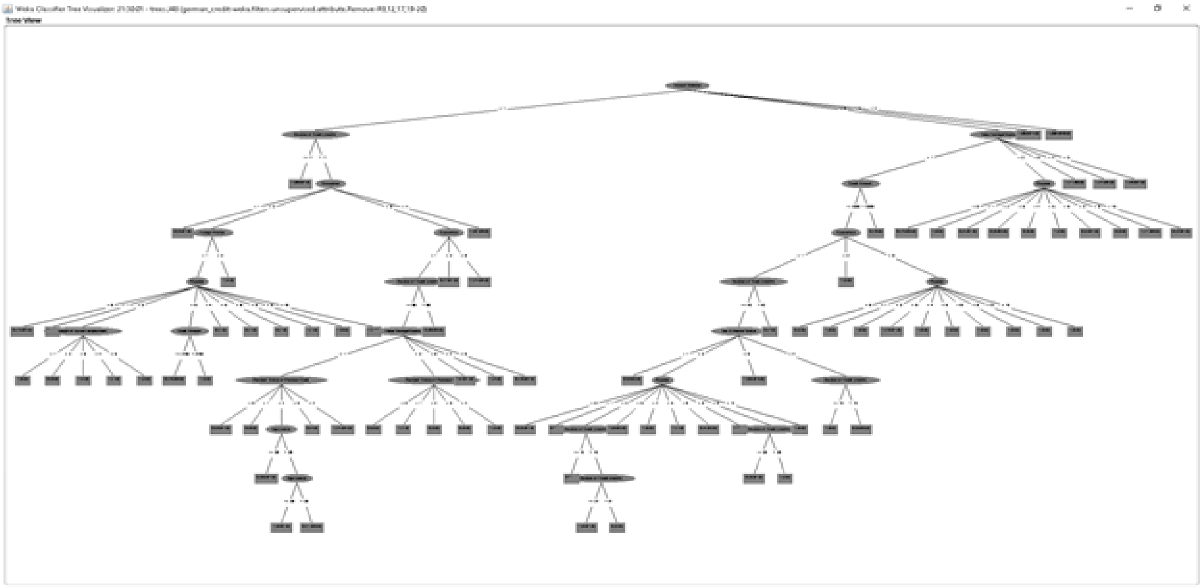
Predictive Modeling (Classification)

Naïve Bayes:

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naïve-Bayes classification algorithm is a supervised learning and a great method of classification. It is one of the most successful algorithms for learning to classify.

J48

J48 classification is a used to generate a decision tree. It can handle both continuous and discrete attributes. It is the continuation of C4.5. It has become a very popular classification to create a decision tree.


Weka Classifier Tree Visualizer: 21:30:01 - trees.J48 (german_credit-weka.filters.unsupervised.attribute.Remove-R8,12,17,19-20)
Tree View

Random Forest:

Random Forest is an ensemble learning classification and regression it builds off of multiple decision trees outputting average. It creates a classification and regression of individual trees. It is better at fitting the data them Random Tree.

The chart below shows the outcome of running the algorithms in Naive Bayes, J48 and Random Forest. We had made changes to the folds from 10 to 1000 to see if there was any effect on the Correctly classified Instances.

Tests	1000 fold	10 fold	Confidence Factor	Correctly Classified Instances (%)	True Positive	False Positive	Precision	Recall
Naive Bayes		x	Not applicable	75	0.867	0.532	0.795	0.867
	x		Not applicable	75.7	0.87	0.507	0.8	0.87
J48		x	0.25	71.9	0.841	0.507	0.776	0.841
	x		0.25	70.8	0.817	0.547	0.777	0.817
		x	0.2	72.2	0.844	0.563	0.778	0.844
Random Forest		x	Not applicable	75	0.901	0.603	0.777	0.901
	x		Not applicable	75.3	0.897	0.583	0.782	0.897

We found that the Random Forest gave us the best True Positive rate of .901 on a 10-fold testing with a 75% correctly classified instances. But the False positive rate of .603 is of concern as predict the bad customers as good customers. The choice that is considered the best is Naive Bayes with a 1000-fold cross validation. Its True positive rate is .87 which is less than the random Forest 0.901 but its false Positive rate is .507 which is much less compared to the FP of Random Forest at .603. Also, the correctly classified instances are 75.7% which is higher than Random forest rate of 75%. The best performing algorithm is Naïve Bayes with a 1000-fold cross validation.

Post-predictive Analysis

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).

Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness.

We extracted the good class instances which are 700. We applied the cluster analysis and the association rule to the good class to find out any strong patterns in the characteristics of the customers who can be extended credit.

Cluster Analysis Results

The cluster analysis was done using different number of cluster and then comparing the readings of the number of instances and the value of attributes. The number of clusters that were used 2,3,4, and 5 . The readings of the cluster 2 are given as under

```
Final cluster centroids:
                                 Cluster#
Attribute                                        Full Data      0        1
                                                 (700.0)   (484.0)   (216.0)
==========================================================================
Account Balance                                  4         4         4
Duration of Credit (month)                       12         24         12
Payment Status of Previous Credit                2         2         4
Purpose                                          3         3         2
Credit Amount                                    1258      701       1169
Value Savings/Stocks                             1         1         1
Length of current employment                     3         3         5
Sex & Marital Status                             3         3         3
Guarantors                                       1         1         1
Most valuable available asset                    3         3         2
Age (years)                                      27        26        36
Concurrent Credits                               3         3         3
Type of apartment                                2         2         2
Foreign Worker                                   1         1         1




Time taken to build model (full training data): 0.02 seconds

=== Model and evaluation on training set ===
Clustered Instances

0      484 ( 69%)
1      216 ( 31%)
```

The clusters show a strong grouping of the characteristics for the following attributes

Concurrent Credit 3

Type of Apartment 2

Length of Employment 3,5

After the clustering, good class was subjected to the association rules analysis. The attributes were changed from numeric to nominal . The association was also done to find any close patterns in the characteristics of the customers. The rules were run on different support and confidence values and the best possible association rules were deliberated.

Association Rules

Following are the best set of rules that were found. They are with a minimum support of .5 and a confidence of .7

```
 1. Type of apartment=2 528 ==> Foreign Worker=1 504    <conf:(0.95)> lift:(1) lev:(0) [0] conv:(1)
  2. Value Savings/Stocks=1 386 ==> Foreign Worker=1 368    <conf:(0.95)> lift:(1) lev:(0) [0]
conv:(0.96)
  3. Concurrent Credits=3 590 ==> Foreign Worker=1 561    <conf:(0.95)> lift:(1) lev:(-0) [-1]
conv:(0.93)
  4. Concurrent Credits=3 Type of apartment=2 444 ==> Foreign Worker=1 422    <conf:(0.95)>
lift:(1) lev:(-0) [-1] conv:(0.91)
  5. Sex & Marital Status=3 402 ==> Foreign Worker=1 377    <conf:(0.94)> lift:(0.98) lev:(-0.01) [-6]
conv:(0.73)
  6. Foreign Worker=1 667 ==> Concurrent Credits=3 561    <conf:(0.84)> lift:(1) lev:(-0) [-1]
conv:(0.98)
  7. Type of apartment=2 528 ==> Concurrent Credits=3 444    <conf:(0.84)> lift:(1) lev:(-0) [-1]
conv:(0.98)
  8. Type of apartment=2 Foreign Worker=1 504 ==> Concurrent Credits=3 422    <conf:(0.84)>
lift:(0.99) lev:(-0) [-2] conv:(0.95)
  9. Type of apartment=2 528 ==> Concurrent Credits=3 Foreign Worker=1 422    <conf:(0.8)> lift:(1)
lev:(-0) [-1] conv:(0.98)
 10. Foreign Worker=1 667 ==> Type of apartment=2 504    <conf:(0.76)> lift:(1) lev:(0) [0] conv:(1)
 11. Concurrent Credits=3 590 ==> Type of apartment=2 444    <conf:(0.75)> lift:(1) lev:(-0) [-1]
conv:(0.99)
 12. Concurrent Credits=3 Foreign Worker=1 561 ==> Type of apartment=2 422    <conf:(0.75)>
lift:(1) lev:(-0) [-1] conv:(0.98)
 13. Concurrent Credits=3 590 ==> Type of apartment=2 Foreign Worker=1 422    <conf:(0.72)>
lift:(0.99) lev:(-0) [-2] conv:(0.98)
```

It was realised that a strong association was found in the attributes concurrent credit 3, type of apartment 2.

The results of the cluster analysis and the association rules are quite close, and these characteristics were found to be strong indicators of a creditable customer. These characteristics are having no concurrent credits, owning a house and a length of employment more than 3 years. Foreign workers are found to be creditable customers.

Conclusion and Recommendations

The exploratory analysis of the data set was done, and attributes were selected based on the maximum information gain. Fifteen out of the initial twenty attributes were selected for predictive modeling.

Predictive analyses were done using three algorithms: Naive Bayes, Decision Tree and Random Forest. The Results were compared for the true and false positive rates to determine the best model. It was found that the Random forest performed the best with a True Positive of .901 at 10-fold testing. However, the false positive was high for Random forest algorithm. Naive Bayes with 1000-fold which has a True Positive of .87 and a False Positive .507. Predictability significantly improved after removing the uncorrelated attributes.

The Naive Bayes Model was selected as the Classification model.

Post Predictive analysis using the Association Rules and Clustering was performed on the creditable class instances only. It was realised that a strong association was found in the attributes concurrent credit 3, type of apartment 2 and Foreign Worker 1.

The results of the cluster analysis and the association rules were quite close, and these characteristics were found to be strong indicators of a creditable customer.