|  |  |  |
| --- | --- | --- |
| Contents  Executive Summary1  Problem Description1  Analysis1  Data Cleaning2  Data Mining2  Take Away3  Conclusion 3 |  | Modelling Small Loan Approval Process *An approach to automate process of approval for Small Loans using Data Mining* Executive Summary The BHM bank, one of the largest European banks, head-quartered in Frankfurt and has operations in all large metropolitan areas in Western and Southern Europe. The recent economic downturn and problems with some of its loans, forced BHM to reexamine its lending process, especially the small loans that are often issued without collateral. Taking advantage of the data mining techniques, can this process be automated and thereby lead to substantial savings in terms of personnel costs.  Based on the extensive data on loan applications of the 10 years or so, a data mining model has to be constructed and based on its accuracy, appropriate actions in terms of level of automation of the process can be taken.  Using Data conversion and applying some classifiers like Naive Bayesian and decision tree (ID3, J48) using k fold stratification a set of models are prepared and their performance is calculated in terms of RIS and KAPPA.  The best model is selected for recommendation along with the limitations when additional manual intervention is necessary. |

# Problem Description

BHM wants to automate small loan approval process. A sample set of 1000 loan applications has been provided. BHM wants to use data analytics to solve either one of the problems below:

* Fully automate the loan approval process for small loans
* If full automation is not possible, provide a recommendation system that can be used for cross verification.

# Analysis

Sample File Provided: credit.mdb (Worksheet: credit-original)



**ATTRIBUTES:** CHK\_ACCT, DURATION, HISTORY, NEW\_CAR, USED\_CAR, FURNITURE, RADIO/TV, EDUCATION, RETRAINING, AMOUNT, SAV\_ACCT, EMPLOYMENT, MALE\_DIV, MALE\_SINGLE, MALE\_MAR\_or\_WID, CO-APPLICANT, GUARANTOR,YEARS\_IN\_RESIDENCE,AGE, RENT, OWN\_RES, NUM\_CREDIT\_ACC, JOB, NUM\_DEPENDENT,TELEPHONE, FOREIGN

**GOAL VARIABLE:** Approve

# Data Cleaning

Using Attribute Consolidation technique, we create new attributes in CELL A which consolidate several attributes mentioned in cell B from the sample file.

|  |  |
| --- | --- |
| **CELL A** | **CELL B** |
| PURPOSE | NEW\_CAR, USED\_CAR, FURNITURE, RADIO/TV, EDUCATION, RETRAINING |
| GENDER\_MARITAL-STATUS | MALE\_DIV, MALE\_SINGLE, MALE\_MAR\_or\_WID |
| RESIDENCE | RENT, OWN\_RES |

The file is then converted to **ARRF** format for it to be used by WEKA.



# Data Mining

The following classification models were run on the cleansed sample set.

|  |  |  |
| --- | --- | --- |
| **TYPE** | **MODEL** | **FOLDS** |
| DISCRETIZED | NAIVE BAYESIAN SIMPLE | 10 |
| DISCRETIZED | NAIVE BAYESIAN SIMPLE | 15 |
| DISCRETIZED | ID3 | 10 |
| DISCRETIZED | ID3 | 15 |
| CONTINUOUS | NAIVE BAYESIAN SIMPLE | 10 |
| CONTINUOUS | NAIVE BAYESIAN SIMPLE | 15 |
| CONTINUOUS | J48 | 10 |
| CONTINUOUS | J48 | 15 |
| CONTINUOUS | J48 | 5 |
| CONTINUOUS | J48 | 8 |
| CONTINUOUS | J48 | 9 |

The models along with their performance are attached below:



# Observation

J48 with 8 folds and 9 folds jointly have the highest RIS & KAPPA.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **DISCRETIZATION** | **MODEL** | **FOLD** | **CONFUSION MATRIX** | |  | **RIS** | **KAPPA** |
|  |  |  | **a** | **b** |  |  |  |
| NO | J48 | 8 | 765 | 5 | **a** | 0.954877 | 0.9601 |
|  |  |  | 9 | 220 | **b** |  |  |
|  |  |  |  |  |  |  |  |
| NO | J48 | 9 | 764 | 7 | **a** | 0.95488 | 0.9604 |
|  |  |  | 7 | 222 | **b** |  |  |

To select one of these we had to go for stratified accuracy:

**J48—8 FOLD**

Stratified a: 99.35%

Stratified b: 96.07%

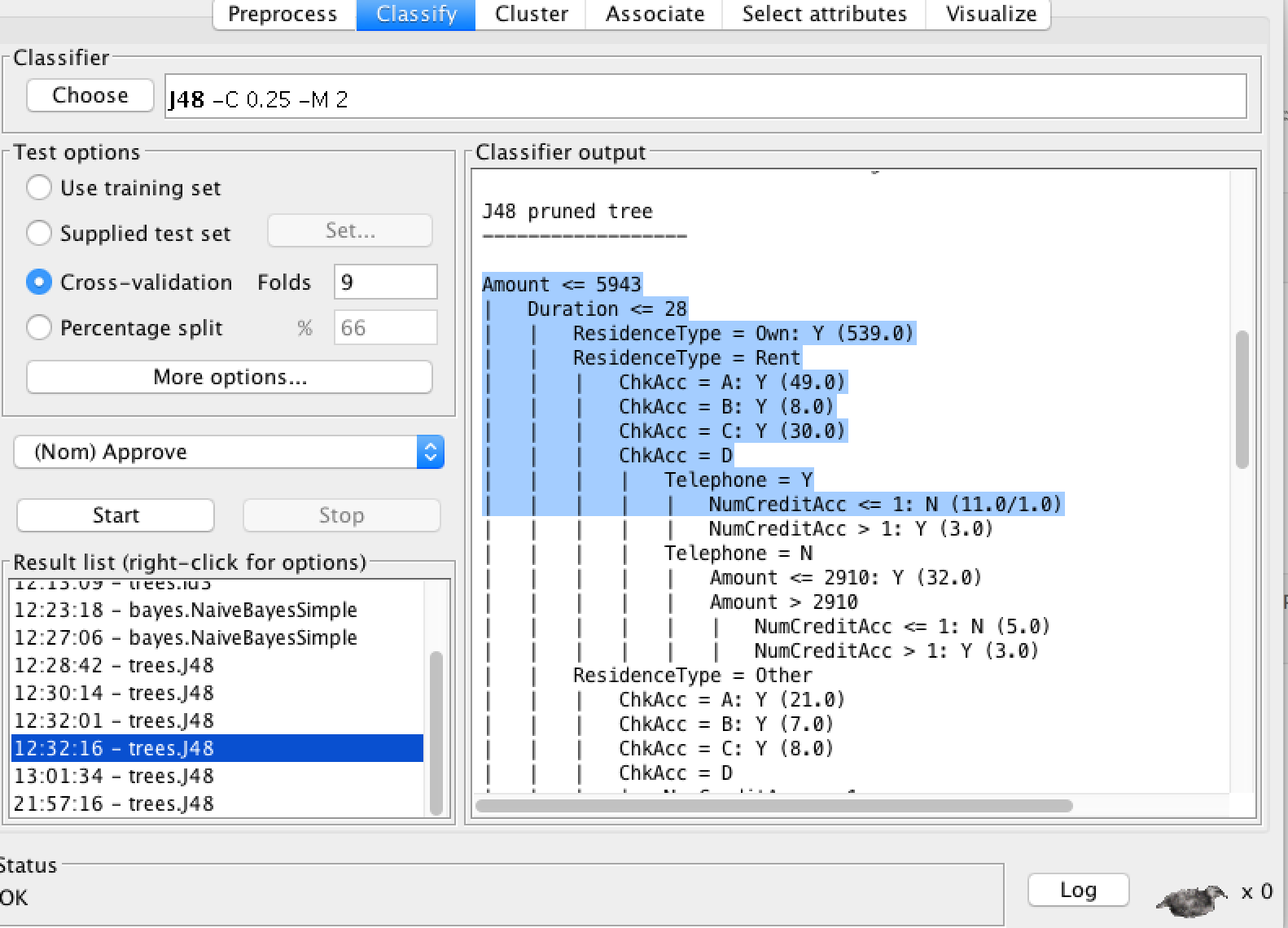
**J48—9 FOLD**

Stratified a: 99.09%

Stratified b: 96.94%

The difference between accuracy percentages of a and b in 9 fold is less than that in 8 fold. Hence the model we select is J48, 9 fold.

As indicated by the RIS & KAPPA the 9 fold model has a high accuracy (RIS: 95.4% and Kappa 96.04%). Manual intervention is needed in one specific case as demonstrated by Weka (highlighted below):



# Conclusion

Using the above sample set provided by BHM bank, J48 classifier with 9 folds provides the best model. The above system at best can be used for providing a recommendation. The reasons for that are as follows:

1. 1000 records are woefully inadequate to build a fully automated loan approval system even if it is for small loans. We at least need a sample size of 2/3rd of the existing bank data to build a model which can be used for automation.
2. Current balance in checking/savings account is not a good factor for loan approvals. There can be instances where just before applying a loan someone borrowed money from near dear ones.
3. FICO score is one attribute which is missing in the model that should be taken into account while giving loans. This takes care of how well existing credits have been re-paid by the customer.