**SET09120 Data Analytics 2020/21**

**Coursework II**

**David Ciocoiu 40323308**

1. **Introduction**

A dataset about credit applications recorded by a German bank has been provided.  
We understand that banks profits due to the interests applied on their loans and that the business’s main objective is to increase revenue.   
A big issue lies between loans repayment and banks must decide whether to take the risk of awarding a loan.

Therefore, the purpose of the analysis is to provide interesting patterns through data mining techniques in order to allow the bank to make informed decision based on the applicant’s profiles.

1. **Data Preparation**

Data quality is paramount for cost reduction, increased efficiency and informed decisions (among others); before undertaking any analysis, our dataset must be cleaned using the provided metadata and prepared for future models.  
[OpenRefine](https://openrefine.org/) has been used to normalize the data by correcting misspellings, multiple types and eliminating outliers

**2.1 Data Cleaning**

The first step undertaken is to add missing headers for each column as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| case\_no | checking\_status | credit\_history | purpose | credit\_amount |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| saving\_status | employment | persona\_status | age | job | class |

Then we detect any errors and outliers in the data:

\*notice that the quotes in the various attributes have been removed for consistency in the data

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute Name** | **Error Value** | **Corrected Value** | **Info** |
| **checking\_status** |  |  |  |
|  | ‘<0’ | <0 | Extra quotes |
|  | ‘>=200’ | >=200 | Extra quotes |
|  | ‘0<=X<200’ | 0<=X<200 | Extra quotes |
|  | ‘no checking’ | no checking | Extra quotes |
| **credit\_history** |  |  |  |
|  | ‘all paid’ | all paid | Extra quotes |
|  | ‘critical/other existing credit’ | critical/other existing credit | Extra quotes |
|  | ‘delayed previously’ | delayed previously | Extra quotes |
|  | ‘existing paid’ | existing paid | Extra quotes |
|  | ‘no credits/all paid’ | no credits/all paid | Extra quotes |
| **purpose** |  |  |  |
|  | ‘domestic appliance’ | domestic appliance | Extra quotes |
|  | ‘new car’ | new car | Extra quotes |
|  | ‘used car’ | used car | Extra quotes |
|  | ather | other | Misspelling |
|  | busness | business | Misspelling |
|  | busines | business | Misspelling |
|  | Eduction | education | Misspelling, Capital letters |
|  | Radio/Tv | radio/tv | Capital letters |
| **credit\_amount** |  |  |  |
| case\_no: 432 | 111328000 | 13280 | Outlier: compared against age > 27 and purpose ‘other’, new value seems to fit in. |
| case\_no: 444 | 7190000 | 7190 | Outlier: compared against age > 40, employment >= 7 and purpose ‘education’, new value seems to fit in. |
| case\_no: 452 | 5180000 | 5189 | Outlier: compared against:  Age between 20 and 29; employment 1<=X<4 and purpose ‘radio/tv’, new value seems to fit in. |
| case\_no: 514 | 5850000 | 5850 | Outlier: compared against:  Age between 20 and 20; employment 1<=X<4 and purpose ‘radio/tv’, new value seems to fit in. |
| case\_no: 560 | 19280000 | 1928 | Outlier: compared against:  credit\_history ‘critical/other existing credit’; checking\_status 0<=X<200 and purpose ‘furniture/equipment, new value seems to fit in. |
| case\_no: 595 | 13580000 | 13580 | Outlier: compared against age > 27 and purpose ‘other’, new values seems to fit in. |
| case\_no: 648 | 13860000 | 1386 | Outlier: compared against:  saving\_status 500<=X1000, new value seems to fit in. |
| case\_no: 660 | 63610000 | 6361 | Outlier: compared against:  credit\_history ‘critical/other existing credit’; checking\_status 0<=X<200 and purpose ‘furniture/equipment, new value seems to fit in. |
| **saving\_status** |  |  |  |
|  | ‘<100’ | <100 | Extra quotes |
|  | ‘>=1000’ | >=1000 | Extra quotes |
|  | ‘100<=X<500’ | 100<=X<500 | Extra quotes |
|  | ‘500<=X<1000’ | 500<=X<1000 | Extra quotes |
|  | ‘no known savings’ | no known savings | Extra quotes |
| **employment** |  |  |  |
|  | ‘<1’ | <1 | Extra quotes |
|  | ‘>=7’ | >=7 | Extra quotes |
|  | ‘1<=X<4’ | 1<=X<4 | Extra quotes |
|  | ‘1<=X<7’ | 1<=X<7 | Extra quotes |
| **personal\_status** |  |  |  |
|  | ‘female/div/dep/mar’ | female/div/sep/mar | Extra quotes, misspelling. |
|  | ‘male div/sep’ | male div/sep | Extra quotes |
|  | ‘male mar/wid’ | male mar/wid | Extra quotes |
|  | ‘male single’ | male single | Extra quotes |
| **age** |  |  |  |
|  | 222 | 22 | Outlier: Removed extra digit on every “222” instance. |
|  | 333 | 33 | Outlier: Removed extra digit on every “333” instance. |
|  | 6 | 33 | Outlier: Underage: transformed to median 33. |
|  | 1 | 33 | Outlier: Underage: transformed to median 33. |
|  | -34 | 34 | Invalid entry: negative age not possible. |
|  | -35 | 35 | Invalid entry: negative age not possible. |
|  | -29 | 29 | Invalid entry: negative age not possible. |
|  | 0.44 | 44 | Invalid entry: fractional age not possible. |
|  | 0.24 | 24 | Invalid entry: fractional age not possible. |
|  | 0.35 | 35 | Invalid entry: fractional age not possible. |
| **job** |  |  |  |
|  | ‘high qualif/self emp/mgmt’ | high qualif/self emp/mgmt. | Extra quotes |
|  | ‘unemp/unskilled non res’ | unemp/unskilled non res | Extra quotes |
|  | ‘unskilled resident’ | unskilled resident | Extra quotes |
|  | ‘skilled’ | skilled | Extra quotes |
|  | yes | skilled | Assumed mistyping. |

Lastly no duplicates have been found when comparing against the case\_no, therefore we directly proceed removing this column as it will not be relevant for the purpose of the analysis.

**2.2 Data Conversion**

From the cleaned data set, two additional data sets have been produced.

In order to obtain the nominal data set below, many attempts have been done splitting the data in different size bins. The proposed approach has a good overall accuracy

|  |  |  |
| --- | --- | --- |
| **Nominal Data Set** | **Original Attribute** | **Transformed Attribute** |
| credit\_amount |  |  |
|  | Numerical series of integers between 392 and 18424 | [[392-1765]](javascript:%7b%7d)  [(1765-3279]](javascript:%7b%7d)  [(3279-4793]](javascript:%7b%7d)  [(4793-6308]](javascript:%7b%7d)  [(6308-7822]](javascript:%7b%7d)  [(7822-9337]](javascript:%7b%7d)  [(9337-10851]](javascript:%7b%7d)  (10851-12366]  [(12366-13880]](javascript:%7b%7d)  [(13880-15395]](javascript:%7b%7d)  [(15395-16909]](javascript:%7b%7d)  [(16909-18424]](javascript:%7b%7d) |
| age |  |  |
|  | Numeric series of integers between 19 and 75 | [[19-32]](javascript:%7b%7d)  [(32-41]](javascript:%7b%7d)  [(41-53]](javascript:%7b%7d)  [(53-64]](javascript:%7b%7d)  [(64-75]](javascript:%7b%7d) |

1. **Data Analytics**

Data mining is the extraction of implicit, previously unknown, and potentially useful information from data. (Data Mining, Witten & Frank, Preface).

This information can be used to make decisions and allows for the so-called “Knowledge Discovery in Databases”. Various methods are available, here we will adopt: Classification, Clustering an Association.

**3.1 Classification**

The purpose of classification learning is to classify examples based on a target attribute, which in our case is deciding whether to give a loan or not. There are many techniques available and each one of them has different benefits, here we will look at OneR and J48.

**3.1.1 OneR Algorithm**

The OneR algorithm generates a one level decision tree, finding the attribute which makes the fewest prediction errors.

OneR has been ran in Weka (keeping the default settings) on the nominal bank dataset obtaining quite promising statistical results, with an accuracy of 71.7% using a training set validation.

The algorithm chose “credit\_history” as the classifying attribute and over 1000 instances, 717 are correctly classified

|  |  |
| --- | --- |
| **R.1** | IF credit\_history = critical/other existing payment OR credit\_history = existing paid OR credit\_history = delayed previously THEN good |
|  |  |
| **R.2** | IF credit\_history = no credits/all paid OR credit\_history = all paid THEN bad |

The above rules are self descriptive, although it is clear that with such a simplistic approach many instances will be misclassified; for instance, why should we award a loan with someone with critical or other existing credit?

**3.1.2 J48 Algorithm**  
C4.5, available under the name of J48 in Weka, is the answer to the above questions.

This algorithm generates a pruned decision tree, not only giving us a deeper insight into the dataset (as opposed to OneR), but also simplifies the tree avoiding overfitting with important benefits on interpretability and generalization (hence why we use J48 and not ID3).

By using the training set testing option on the nominal bank data set we get an accuracy of 80.6%, a good value, with the decision tree picking the “checking\_status” as the first most useful determiner.

Rules have been picked as a trade off between highest accuracy and highest coverage, and by looking at them we can deduce that “checking\_status”, “credit\_history” and purpose play an important role in the decision making.

Additionally, the confusion matrix gives us a deeper insight: 662 values have been correctly classified as good (true positives) and 38 have been misclassified as bad (false negatives), whereas 144 values have been correctly classified as bad (true negatives) and 156 have been misclassified as good (false positives).

|  |  |  |
| --- | --- | --- |
| **a** | **b** | **<-- classified as** |
| 662 | 38 | a = good |
| 156 | 144 | b = bad |

|  |  |
| --- | --- |
| **R.1** | IF checking\_status = <0 AND credit\_history = critical/other existing credit THEN good (67.0/18.0) |
|  | **Coverage:** 67 **Accuracy:** 73.13% Out of 67 instances, 18 were misclassified.  If a client has a checking status inferior to 0 and a critical or other existing credit history then the loan can be awarded. |
| **R.2** | IF checking\_status = <0 AND credit\_history = existing paid AND saving\_status = no known savings AND job = skilled THEN bad (15.0/5.0) |
|  | **Coverage:** 15 **Accuracy: 66**.66% Out of 15 instances, 5 were misclassified.  If a client’s checking status is less than zero, with existing credits paid, no savings and a skilled job, then the loan should not be awarded. |
| **R.3** | IF checking\_status = <0 AND credit\_history = existing paid AND saving\_status = <100 AND purpose = new card THEN bad (31.0/9.0) |
|  | **Coverage:** 31 **Accuracy:** 70.96% Out of 31 instances, 9 were misclassified.  A loan should not be awarded if a client’s checking status is less than zero, there is an existing paid credit history, the savings are less than 100 and the purpose for the loan is buying a new car. |
| **R.4** | IF checking\_status = no checking THEN good (394.0/46.0) |
|  | **Coverage:** 394 **Accuracy:** 88.32% Out of 394 instances, 46 were misclassified.  A loan can be awarded if there is no checking status. |
| **R.5** | IF checking\_status = >=200 THEN good (63.0/14.0) |
|  | **Coverage:** 63 **Accuracy:** 77.77% Out of 63 instances, 14 were misclassified.  A loan can be awarded if the checking status is bigger or equal to 200. |
| **R6** | IF checking\_status = 0<=X<200 AND credit\_amount = [392-1765] AND purpose = radio/tv THEN good (36.0/8.0) |
|  | **Coverage:** 36 **Accuracy:** 77.77% Out of 36 instances, 8 were misclassified.  The bank shall award a loan if the customer’s checking status is between 0 and 200 included and the credit amount is between 392 and 1765 included, with the purpose of buying a radio or tv. |

**3.2 Association**

The aim of association is to find any correlations or associations between the attributes, hence describing frequent patterns in the data set.  
One of the simplest algorithms for association is Apriori, which given a minimum support uses an iterative approach applying breadth first search on the dataset in order to find frequent items sets and generate rules according to their confidence (cases in which the rule application is correct).  
The results seem quite promising, especially when looking at the confidence which ranges between 91% and 94%, which means that the rules are very accurate.

|  |  |
| --- | --- |
| **R.1** | checking\_status=no checking purpose=radio/tv 127 ==> class=good 120 <conf:(0.94)> lift:(1.35) lev:(0.03) [31] conv:(4.76) |
|  | Very high confidence: 94%  If a customer has no checking status and the purpose for the loan is to purchase a radio or a tv then it is safe to award a loan. |
| **R.2** | checking\_status=no checking credit\_amount=[392-1765] 145 ==> class=good 136 <conf:(0.94)> lift:(1.34) lev:(0.03) [34] conv:(4.35) |
|  | Very high confidence: 94%  If a customer has no checking status and the requested credit is between 392 and 1765 (both inclusive) then it is safe to award a loan. |
| **R.3** | checking\_status=no checking credit\_history=critical/other existing credit 153 ==> class=good 143 <conf:(0.93)> lift:(1.34) lev:(0.04) [35] conv:(4.17) |
|  | Very high confidence: 93%  If a customer has no checking status with a credit history being either critical or having other existing credit then it is safe to award a loan. |
| **R.4** | checking\_status=no checking employment=>=7 115 ==> class=good 107 <conf:(0.93)> lift:(1.33) lev:(0.03) [26] conv:(3.83) |
|  | High confidence: 93%  If a customer has no checking status and their employment status is higher or equal to 7 years, then it is safe to award a loan. |
| **R.5** | checking\_status=no checking personal\_status=male single job=skilled 151 ==> class=good 139 <conf:(0.92)> lift:(1.32) lev:(0.03) [33] conv:(3.48) |
|  | High confidence: 92%  If a customer has no checking status and they are a single male with a skilled job, then it is safe to award a loan. |
| **R6** | checking\_status=no checking age=(32-41] 117 ==> class=good 107 <conf:(0.91)> lift:(1.31) lev:(0.03) [25] conv:(3.19) |
|  | High confidence: 91%  If a customer has no checking status and their age is between 32 (exclusive) and 41 (inclusive) then it is safe to award a loan. |

**3.3 Clustering**