Data Analytics - Coursework 2

Jacob Connell
40343979@napier.ac.uk
Edinburgh Napier University - SET09120 - Data Analytics

February 9, 2023

Keywords – Data Mining, Weka, Association, Clustering, Classification, Open Refine, Data Analytics

Word Count – 4468

Abstract -
This report analyses a real historical data set from a German bank using methods of machine learning in an attempt to understand the patterns within the data having an overall goal to gain knowledge. Using three different experimental methods, the report aims to answer what makes a good creditor to the bank, with the premise of making more profit. Utilising Classification, Association and Clustering in the Java based machine learning program Weka, the methods used are outlined in detail and the results interpreted to uncover interesting patterns and develop knowledge from this data set.

Contents

1 Introduction **3**

2 Data Preparation **3**
- 2.1 Data Cleaning . . . 3
  - 2.1.1 Data Completeness . . . 3
  - 2.1.2 Tuple Duplication . . . 3
  - 2.1.3 Data Consistency . . . 3
  - 2.1.4 Data Accuracy . . . 3
- 2.2 Data Selection . . . 4
- 2.3 Data Transformation . . . 4

3 Classification **4**
- 3.1 Introduction & Setup . . . 4
- 3.2 Results . . . 4
  - 3.2.1 Rule 1 . . . 4
  - 3.2.2 Rule 2 . . . 4
  - 3.2.3 Rule 3 . . . 4
  - 3.2.4 Rule 4 . . . 5
  - 3.2.5 Rule 5 . . . 5
  - 3.2.6 Rule 6 . . . 5
- 3.3 Analysis . . . 5

4 Association **5**
- 4.1 Introduction & Setup . . . 5
- 4.2 Results . . . 5
  - 4.2.1 Rule 1 . . . 5
  - 4.2.2 Rule 2 . . . 6
  - 4.2.3 Rule 3 . . . 6
  - 4.2.4 Rule 4 . . . 6
  - 4.2.5 Rule 5 . . . 6
  - 4.2.6 Rule 6 . . . 6
- 4.3 Analysis . . . 6

5 Clustering **6**
- 5.1 Introduction & Setup . . . 6
- 5.2 Clustered Instances . . . 7

6 Conclusion **7**
- 6.1 Classification . . . 7
- 6.2 Association . . . 7
- 6.3 Clustering . . . 7
- 6.4 Summary . . . 8

7 Figures **9**

8 References **9**

References **9**

9 Appendices **9**

1 Introduction

The aim of this coursework is to provide an analysis from true data collected by a German bank on credit applications. With the goal of knowledge discovery, the data must first be prepared and imported into the chosen tool for machine learning. A number or methods and algorithms will be applied to identify patterns. This report will conclude with a summary of those findings and a critical analysis of the methods used.

2 Data Preparation

Data preparation is a vital step to ensure this data is ready to mine for accurate results. It's important to ensure the data quality, data cleanliness and select the most relevant data for analysis. With higher quality data it can be expected to find higher quality patterns as a result of the analysis. The main aspects of data quality in this data set that can ensured at this stage are: accuracy, completeness, consistency, and uniqueness. Timeliness isn't as relevant given this is historic data. To ensure these quality aspects, the data must first be cleaned.

2.1 Data Cleaning

Data cleaning can be split into main sub processes and some of these will be discussed in relation to the key aspects of preparation mentioned. To complete this data preparation process, the tool Open Refine (*Open Refine*, n.d.) has been used. The comma separated file was imported into a new Open Refine project for analysis. Initially, all the headings were renamed to resemble the heading values from the provide metadata (Appendix A).

2.1.1 Data Completeness

Firstly, the first stage of this process was to ensure the data is complete for the machine leaning algorithms to accept it. Using null or empty text facets on each attribute in turn (Appendix B), it's possible to confirm there were no missing values in the data that may cause a problem for the algorithms later on.

2.1.2 Tuple Duplication

This process aims to ensure entities don't have more than one occurrence in the data set. To confirm this, the data set was imported into excel and concatenated into one field. From this, OpenRefine could then be used to filter for duplicates using facets (Appendix C).

In this case, no duplicate records were found. From the metadata, each applicant is given a numerical case reference and this has been included in the duplication check. Should the numerical case reference be removed, it may uncover different results, but this test was to ensure each applicant only had one record and that is correct.

2.1.3 Data Consistency

Data consistency is a large aspect of this cleaning process. Given the many different data types, and sources that could be encountered, it's expected to find variations in data that need correcting. For example, it could be found that within a supposedly Boolean field, there may be different variations (1/TRUE/yes/Y/true/T) all indicating the same resolve impacting consistency throughout the entire set. An example of this can be found in Appendix D.

To filter out this in the data, each column was put through a text facet and the appropriate judgment used with the aid of the metadata to ensure the values were consistent throughout the set. A detailed list of example changes made in this section can be found in Appendix E. In addition to this, the use of clustering supported the identification of values that may represent the same thing. The method used was key collision (*OpenRefine: Clustering*, n.d.) with a Daitch-Mokotoff keying function (Appendix F) while this didn't spot all of the errors, it did identify the large majority.

The large majority of corrections were due to spelling errors or the use of the incorrect data type for the attribute. Each time they have been corrected to match the provided metadata, even in some cases where this seems incorrect. For example, the personal status "female div/dep/mar" is similar to "male div/sep" and is described in the metadata as "...divorced/separated...". This implies the abbreviation "...dep..." is incorrect, but as it's consistent throughout the data set and matched the metadata supplied, this has been left unchanged.

2.1.4 Data Accuracy

Data accuracy is likely one of the more challenging aspects of this process. It involves identifying the errors and using the correct judgement to ensure the data quality. This can either involve ignorance or imputation to correct the data.

In order to assess the quality of data, each attribute was evaluated in turn to ensure it was accurate. The nominal values only had a set number of possibilities and for that reason may be significantly more accurate than those of a numerical value. An example of a nominal correction made would be in the 'Job' column where the value of 'good' was uncovered using a text facet. This does not match the metadata provided and with only one outlier, an imputation was made that good would be considered 'skilled' in comparison to the other options.

In the nominal columns, the use of numerical facets was employed to filter out the extreme values for comparison. A full list of corrections made can be found in Appendix H. Using Age as an example, the numerical values were filtered to the extreme values of over 80 using a facet (Appendix I).

It was found a few ages were extreme due to repeated digits. They were corrected by removing one digit bringing them into the realistic age range of 18-90. From there, the facet was reversed to find ages below 19. A large number of ages were found to be inaccurate. For example, the screenshot in Appendix J demonstrate this.

To correct these, where obvious, decimal places were adjusted, or negative signs removed to bring these ages into the normal range. As an example, -39 was able to be adjusted to 39 and 0.26 was adjusted to 26. The full list of changes can be found in Appendix H. Where the error was ambiguous, a more thorough approach was used.

Record 44 was recorded using the Age of 1. This is an ambiguous error because it could easily have been 21, 31, 41 etc. To resolve, text facets were used to compare this record against other similar records. It can be seen that a similar

record with almost identical attributes had the age of 31, hence, the error was adjusted to 31 (Appendix K).

In some cases, within the credit amount the values seemed extreme, however, when compared with similar records these were acceptable. The main correction made to this column was the removal of trailing zeros from the end of numerical values (Appendix L).

2.2 Data Selection

Evaluating the metadata uncovered that all the attributes are relevant to the machine learning algorithms, except one, the case number. For this, each number would be different for all records and will cause problems with the algorithms used. As such, the case number attribute will be excluded from the record.

2.3 Data Transformation

For some of the algorithms available for machine learning, they only support a specific data type dependant on the methods used. This report will explore the use of three algorithms for Association, Clustering and Classification. For the Association algorithm all the data attributes need to be nominal rather than numerical. To ensure this, a simple python script was applied to each attribute in the data set that was numerical to discretise the values into binned ranges. This has been done for age and credit amount. Age was split into bins of 10 years and credit amount was split into bins of 1000, increasing in the higher values. Examples of the Scripts used and changed can be found in Appendix M.

The machine learning tool of choice is Weka (*Weka: Machine Learning*, n.d.) which will accept inputs of ARFF files. To convert the data into the correct format, taking the .csv file from OpenRefine it could be opened in Weka's ARFF viewer.

This gave the opportunity to ensure the data headings were correct and the data types were also correct. Once complete the data set could be saved directly as ARFF and Weka would manage the conversion into the correct format for our algorithms to process. This process can be seen as screenshots in Appendix N.

Overall, the cleaning of this data was relatively seamless without the requirement for complex clustering algorithms or python scripts. This is mainly due to how constrained the data is by the metadata. There are only two columns, age and credit amount, that are free values and are not restricted to a limited number of options. For this reason, the cleaning of the data was well done by using text facets as the main tool with a few other methods as discussed.

3 Classification

3.1 Introduction & Setup

For this method of supervised machine learning the algorithm of J48 has been selected. This was selected because it will accept both nominal and ordinal data, which provides a detailed decision tree with easily interpreted results. To ensure a good coverage and accuracy, a confidence factor of 0.3 has been set along with the use of pruning to ensure the tree is relevant to the needs. This reduces the complexity while increasing accuracy. With a base of C4.5 this is a significant improvement to the earlier ID3 algorithm.

With Class as the attribute, this test has been run using both a cross-validation and a Percentage Split of varying amounts. It was determined that a 80%/20% split in the data would return the best accuracy for the results, however, with such a small test number of instances (200) this may not be an accurate reflection. Using 10-fold cross-validation with pruning provided an accuracy of 73%, while a 60/40 training/test split would provide a 74.5% accuracy. As a result, this data was tested using a 60/40 split with a 0.3 confidence factor and pruning using J48.

3.2 Results

A sample of 6 rules have been taken from the decision tree that has been created which aim to reflect the most from the data available. The final pruned decision tree can be seen in Appendix O.

3.2.1 Rule 1

Rule 1:
"IF checking_status == 'no checking' THEN class == good (394.0/46.0) "

Figure 1: Classification - Rule 1

90% of clients with no checking status were approved for a loan with 394.0 predicted correctly and 46.0 not predicted correctly. This suggests, initially, that clients with no checking are safe to be approved for a loan.

3.2.2 Rule 2

Rule 2:
"IF checking_status =¡ 0 AND credit_history == 'critical/other existing credit' THEN class == good (67.0/18.0) "

Figure 2: Classification - Rule 2

Clients with a checking status equal to or below 0 were approved for a loan if their credit history was recorded as 'critical/other existing credit'. Out of the 85 test cases, 67 of these were correctly classified (78%). This suggests clients already heavily in debt with a below 0 checking status and other credit are likely to be approved for a loan.

3.2.3 Rule 3

Rule 3:
"IF checking_status =¡ 0 AND credit_history == 'existing_paid' AND purpose IN ('new car', 'education' 'domestic appliance', 'repairs', 'retraining') THEN class == bad (56.0/20.0) "

Figure 3: Classification - Rule 3

Those with a checking status equal to or below 0, with a credit history of 'existing paid' and recorded the purpose for

the loan as: 'new car', 'education', 'domestic appliance', 'repairs' or 'retraining' were not approved. There were 76 cases of this with a 73% accuracy rate in the testing.

This rule is a concatenation of 5 rules from the analysis results. These can be seen in Appendix P, all to be sharing the same initial characteristics before Purpose. For this reason, they have been combined into a more generalised rule to help simplify the results and provide a more meaningful interpretation.

3.2.4 Rule 4

Rule 4:
"IF checking_status =¡ 0 AND credit_history IN ('all paid', 'no credit', 'delayed') THEN class == bad (47.0/12.0) "

Figure 4: Classification - Rule 4

If applicants had a checking status equal to or below 0 and the credit history was one of: 'all paid', 'no credit' or 'delayed') they were not approved for a new loan. Out of 59 test cases, this rule had an 80% accuracy. This suggests that clients with a below 0 checking status and otherwise no credit history were likely not to be approved for a loan.

Again, this has been concatenated from a number of similar rules that can be found in Appendix Q.

3.2.5 Rule 5

Rule 5:
"IF checking_status 0¡= X ¡200 AND credit_amount ¡= 9283 THEN class == good (247.0/87.0) "

Figure 5: Classification - Rule 5

Clients with a checking status of between 0 and 200 and requesting a credit amount of less than 9283 were approved for a loan. Out of 374 test cases for this rule, it yielded a 66% accuracy which is strong but could be improved.

3.2.6 Rule 6

Rule 6:
"IF checking_status =¡ 0 AND credit_history == 'existing paid' AND purpose == 'radio/tv' and employment =¡ 1 AND age ¡= 30 THEN class == good (5.0/1.0) "

Figure 6: Classification - Rule 6

Applicants that had a checking status equal to or below 0, their credit history was 'existing paid', the purpose was listed as for a 'radio/tv', they were 30 or under and had been employed for a yest or less had been approved for a loan while those older than 30 were likely not to be approved for this loan. From the dataset, this returned an 83% accuracy but from only 6 test cases. This suggests that young clients with a good credit history and only recently employed are likely to be approved for a loan for a radio or tv.

3.3 Analysis

From the above sample of 6 rules from this data set there are a few characteristics that are common between rules. In 4/6 rules, the checking status ¡= 0 was a frequent requirement in the rules. Looking at the visualisation in Appendix O, it's clear to see how the credit history = 'existing' is a large characteristic in the pruned tree which can be considered a common feature of rules.

In summary, the checking status and credit history seems to be the two largest factors for affecting a loan. Those likely to get a loan are those already critically in debt or with a good credit history and a low balance. It's very likely anyone with a checking status between 0-200 will be approved for a small loan.

The confusion matrix shown in Appendix R highlights the classification success of this algorithm after it attempted to classify the 400 test instance supplied. For the class 'good' there was a good success with only an error rate of 23%. For the 'bad' class, the error rate was significantly higher at 45%. This may be due to the fact there are a lot less 'bad' records in the data set that was used for training that there were 'good', with a larger data set to train on, this may improve.

4 Association

4.1 Introduction & Setup

Using Apriori to conduct frequent pattern analysis on this data, it first requires the importation of the nominal data set that was created during the data preparation stage where all the numerical values had been binned. The algorithm was run with a confidence of 0.9 and a minimum support of 0.11. The stated number of rules required were 6. If a rule meets the minimum support and confidence level, it will be considered strong. The rules found are multi-dimensional associations with no repeated predicates.

4.2 Results

A sample of 6 rules have been collected from the output of this algorithm and will be discussed below.

4.2.1 Rule 1

Rule 1:
"IF checking_status == 'no checking' AND purpose == 'radio/tv' (127) THEN class == good (120) ¡conf:(0.94)¿ lift:(1.35) lev:(0.03) [31] conv:(4.76) "

Figure 7: Association - Rule 1

If the checking status is no checking and the purpose is radio/tv then they will be approved for a loan. This rule has a high confidence and was valid for 94% of cases. Being a strong and effective rule, it suggests those with no checking account requesting a loan for a radio/tv are safe for a

loan.

4.2.2 Rule 2

Rule 2:
" IF checking_status == 'no checking' AND credit_history == 'critical/other existing credit' (153) THEN class == good (143) ¡conf:(0.93)¿ lift:(1.34) lev:(0.04) [35] conv:(4.17) "

Figure 8: Association - Rule 2

If the checking status is no checking and the credit history is critical/other existing credit then they will get a loan. This was true for 93% of cases and makes the rule a robust one. It can be interpreted to suggests that those with existing credit problems and no checking account are a good choice for a loan.

4.2.3 Rule 3

Rule 3:
"IF checking_status == 'no checking' AND 'personal_status' == 'male single' AND job= 'skilled' (151) THEN class == good (139) ¡conf:(0.92)¿ lift:(1.32) lev:(0.03) [33] conv:(3.48) "

Figure 9: Association - Rule 3

If the checking status is no checking and the personal status is a single male with a skilled job then they will be approved for a loan. This was correct for 92% of cases, still very high, and suggests that single males in skilled employment are a good fit for a loan if they have no current account with the bank.

4.2.4 Rule 4

Rule 4:
"IF checking_status == 'no checking' AND age == 30¡=X¡40 (147) THEN class == good (134) ¡conf:(0.91)¿ lift:(1.3) lev:(0.03) [31] conv:(3.15) "

Figure 10: Association - Rule 4

If the checking status is no checking and they are ages between 30-40 then they will be approved for a loan. This has a confidence factor of 0.91 (91%). Young adults with no current account at the bank in question are a good fit for a loan.

4.2.5 Rule 5

Rule 5:
"IF checking_status == 'no checking' AND credit_amount == 1000¡=X¡2000 (130) THEN class == good (118) ¡conf:(0.91)¿ lift:(1.3) lev:(0.03) [26] conv:(3) "

Figure 11: Association - Rule 5

If checking status is no checking and the credit amount is between 1000-2000 then they will be approved for a loan. This was correct 91% of the time in the data set. Those with no checking account in this bank requesting a relatively small loan are likely safe for approval.

4.2.6 Rule 6

Rule 6:
"IF checking_status == 'no checking' AND credit_history == 'existing paid'AND job == skilled (130) THEN class == good (117) ¡conf:(0.9)¿ lift:(1.29) lev:(0.03) [26] conv:(2.79) "

Figure 12: Association - Rule 6

If checking status is no checking and the credit amount is between 1000-2000 then they will be approved for a loan. This was correct 91% of the time in the data set. Those with no checking account in this bank requesting a relatively small loan are likely safe for approval.

4.3 Analysis

From these rules it's observed that the confidence level is consistently 90% or above making these rules very efficient at predicting the outcome of the data. From this sample of six rules, it can bee seen one common characteristic checking_status == 'no checking' that strongly supports those applying for credit. With that in mind the main reoccurring characteristics on deciding for a loan seem to be checking status, credit history and credit amount.

Meeting the minimum support and confidence level isn't the only requirement because attributes could be negatively associated in a way that suggests if a rule was met, it makes the outcome less probable than it is naturally in the data set. Using the measure of lift, anything that appears more often the expected will have a value greater than one and is hence a good rule.

5 Clustering

5.1 Introduction & Setup

Using the setup of Simple K Means and the Euclidean Distance function with a cluster set of 6, this data set was analysed using the originally prepared file containing a mix of nominal and ordinal data. A percentage split of 66/44 was used for this training analysis.

The 6 clusters found by the algorithm are presented in order in Figure: 13. Similar reoccurring attributes have been highlighted.

Attribute	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Checking_status	No Checking	<0	0<=X<200	0<=X<200	No checking	No checking
Credit_history	Critical/Other existing credit	Existing Paid	Existing Paid	Existing Paid	Critical/Other existing credit	Existing Paid
Purpose	Furniture/Equipment	Furniture/Equipment	Radio/TV	New Car	New Car	Radio/TV
Credit_Amount	4511.7879	3106.5188	3533.5225	5122.0762	2706.9459	2116.0522
Saving_Status	<100	<100	No known savings	<100	<100	<100
Personal_Status	Female div/dep/mar	Male single	Male single	Male Single	Male Single	Female div/dep/mar
Age	31.0909	31.3759	39.4234	36.3143	41.2162	31.1791
Employment	4<=X<7	1<=X<4	>=7	>=7	1<=X<4	<1
Job	Skilled	Skilled	Skilled	Skilled	Skilled	Skilled
Class	Good	Good	Good	Bad	Good	Good

Figure 13: **Cluster Output**

5.2 Clustered Instances

It can be seen from the graph in Figure: 14 that the number of instances clustered here are similar but it's not an even distribution. Clusters 4 and 5 have a much larger size in comparison, with cluster 4 being the largest. This was done using a 66/44 percentage split for training/testing meaning the sample size being visualised was 440.

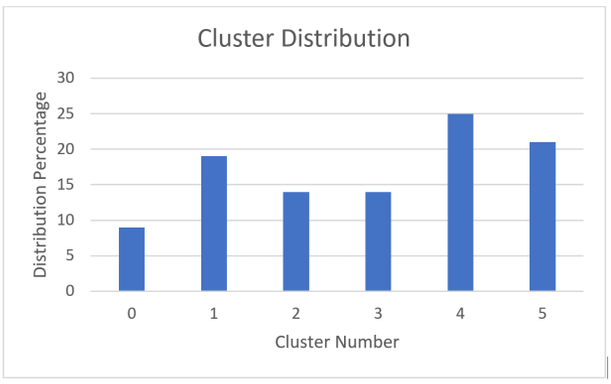


Figure 14: **Cluster Distribution**

Using the table above (Figure: 13) and the screenshot in Appendix S, cluster 3 is highly likely not to get a loan in comparison to all the others presented. The remaining 5 cluster have a significantly higher chance of being approved for a loan which can also be seen in the visualisation.

From this table it helps visualise the aspects of the data set that may be less significant in the decision process of loan applications. Just evaluating the clusters, the majority have a skilled job, a saving status of ¡100 and a single male. It's likely that these factors are less likely to affect the decision of a loan than the remaining attributes, such as purpose or employment that are widely distributed.

6 Conclusion

In this section the discussed methods with be critically reviewed with support from literature and the findings summarised for the analysis of this data set. Three methods of analysis have been used: predictive (classification), descriptive (clustering) and prescriptive (association). Throughout these three methods some of the rules and their interest have been interpreted.

6.1 Classification

Classification is a supervised learning approach allowing the algorithm to develop on training data with labels inputs and outputs. We're then able to test this learning using test data. With a test data set of 400 tuples the algorithm was successfully able to predict the outcome over 70% of the time, however, this still leaves room for 30% error. Using this information from the tree the classification rules do bring to light some interesting patterns such as if there is 'no checking' an applicant is almost always likely to get a loan.

The use of classification is generalised, specifically with a pruned tree. Using the Attribute Selection Measures (ASM) to find a root, the algorithm focuses on gain ratio which can add uncertainty to the results. C4.5 uses pessimistic pruning which will remove nodes dependant on reducing the overall error rate (KHALILOV, GÜMÜŞ, & ÖZSOY, 2015). All the rules should be exclusive and exhaustive to ensure every tuple can be classified. This method did help to visualise and understand the data but it found few clear relationships between attributes that could be used to develop knowledge. As one of the most visual methods, classification can be significantly affected by changes in data. While a large training set was used a small change in this data could drastically affect the entire decision tree, however, it's a much easier algorithm to implement and interpret than others.

6.2 Association

With the use of the Apriori to associate in Weka, it takes an unsupervised approach to find associations between data attributes. This prior knowledge means the algorithm doesn't need a training set to produce the rules. Its simple iterative approach will scan a database a number of times and is able to find all rules within large datasets, however, it can be hindered by how expensive it is to run on large corporate databases. This algorithm requires the use of nominal data to function, meaning the data had to be binned for the association analysis. This type of analysis does provide rules with measurable characteristics of the data set used such as confidence and lift, however, it doesn't provide any sort of predictability score for new data (Arora, Bhalla, & Rao, 2013).

The rules created by the Apriori algorithm in this report all offer a confidence level of over 90% with a lift greater than 1 making the rules themselves extremely strong in comparison to those discovered by classification. While not the most efficient, this prescriptive analysis method provided the clearest results compare to the other methods used. Having the value of lift gave a much clearer accuracy details for the six rules than confidence or support alone

6.3 Clustering

Clustering aims to find natural grouping in unlabelled data in order to find patterns. Its simple process offers less complex findings than other methods but they often require a more interpretation. The algorithm in use, Simple K means offers a good implementation for clustering but requires the number of clusters in advance. If K isn't known then further experiments might be need to find an appropriate K.

Using a random state the results discovered by the algorithm can differ slightly on each run which can cause issues with understanding the data (Abbas, 2008).

The K means algorithm involves clustering around a randomly initialised point, and linking each data set to it's nearest point then adjusting to the mean of all the data combined. This method struggles with busy data and is quite severely effected by any outliers in the data set. A lot of implementations of this algorithm don't support nominal data, however, the instance of SimpleKMeans in Weka supports the use of both nominal and numerical clustering. Weka can handle this by using a distance function to compare values that are nominal (*K-Means Clustering in WEKA*, n.d.). While this does work in Weka, using categorical data within clustering will never provide perfect results. For values that are clear in the meta-data to be ordinal nominal values, these could be converted into numerical values to provide clearer results.

6.4 Summary

In summation to these experiments, the question of study remains: who is safe to offer a loan to in order to increase the banks profit? From the three experiments conducted it's been concluded that having no checking status almost always guarantees the applicant a loan with credit history and credit amount affecting the decisions from there if they do have a checking account. Those with a modest credit amount are likely to be safe for a loan if it's not too large, while those with a low checking account and a bad history of credit are are more likely to get a loan than those with a good credit history. This may suggest that the banks are keen to lead to riskier clients because their able to charge a higher interest rate on the loan, resulting in more profit, or that with other existing credit they're less likely to pay it back quickly, meaning the bank makes more profit. This, however, has to be outweighed by the risk. To further analyse this, more data on clients credit history and the interest rates charged on approved loans would support further experiments into identifying the type of clients this bank feels are safe to issue loans to. Overall from the method of approach and the accuracy of the rules, the association seems to provide the best insight into the data set.

An interesting pattern represents knowledge (*Concepts of Data Mining*, n.d.). To conclude, are all these patterns found in these experiments interesting? For a pattern to be interesting it must be: easily understood buy humans; valid on test/new data; potentially useful the bank and validates a hypothesis from before the experiments. Summarising, the results in this report are clear and easy to interpret and where applicable, have been tested on a portion of the data. There was no hypothesis made at the beginning of these experiments, however, the rules found can be useful to the bank during initial screening of credit applications, specifically on the three main attributes mentioned. This could provide more useful with additional data but creates a basis for the bank to build upon their application approval processes.

7 Figures

List of Figures

1 Classification - Rule 1 . . . 4
2 Classification - Rule 2 . . . 4
3 Classification - Rule 3 . . . 4
4 Classification - Rule 4 . . . 5
5 Classification - Rule 5 . . . 5
6 Classification - Rule 6 . . . 5
7 Association - Rule 1 . . . 5
8 Association - Rule 2 . . . 6
9 Association - Rule 3 . . . 6
10 Association - Rule 4 . . . 6
11 Association - Rule 5 . . . 6
12 Association - Rule 6 . . . 6
13 Cluster Output . . . 7
14 Cluster Distribution . . . 7
15 Classification Rule 3 Combination - Source Rules . . . 16
16 Classification Rule 3 Combination - Combined Rule . . . 16
17 Classification Rule 4 Combination - Source Rules . . . 16
18 Classification Rule 4 Combination - Combined Rule . . . 16

8 References

References

Abbas, O. A. (2008). Comparisons between data clustering algorithms. *International Arab Journal of Information Technology (IAJIT)*, *5*(3).

Ali, F. M. N., & Hamed, A. A. M. (2018). Usage apriori and clustering algorithms in weka tools to mining dataset of traffic accidents. *Journal of Information and Telecommunication*, *2*(3), 231-245. Retrieved from https://doi.org/10.1080/24751839.2018.1448205 doi: 10.1080/24751839.2018.1448205

Arora, J., Bhalla, N., & Rao, S. (2013, Jul). *A review on association rule mining algorithms.* International Journal of Innovative Research in Computer and Communication Engineering. Retrieved from https://www.rroij.com/open-access/a-review-on-association-rulemining-algorithms.pdf

Concepts of data mining. (n.d.). Retrieved from http://www.industrial-electronics.com/data-mining_1b.html

Data preparation for data mining. (n.d.). Retrieved from https://www.tandfonline.com/doi/abs/10.1080/713827180?tab=permissions

Hodgson, E. (2021, Nov). *Classification vs clustering: When to use each in your business.* DotActiv. Retrieved from https://www.dotactiv.com/blog/classification-vs-clustering

KHALILOV, S., GÜMÜŞ, G., & ÖZSOY, S. (2015, Sep). *C4.5 versus other decision trees: A review.* Computer Engineering and Applications Vol. 4. Retrieved from https://core.ac.uk/reader/144285147

K-means clustering in weka. (n.d.). Retrieved from http://facweb.cs.depaul.edu/mobasher/classes/ect584/weka/k-means.html

Open refine. (n.d.). Retrieved from https://openrefine.org/

Openrefine: Clustering. (n.d.). Retrieved from https://guides.library.unlv.edu/open-refine/clustering

Weka: Machine learning. (n.d.). Retrieved from https://www.cs.waikato.ac.nz/ml/weka/

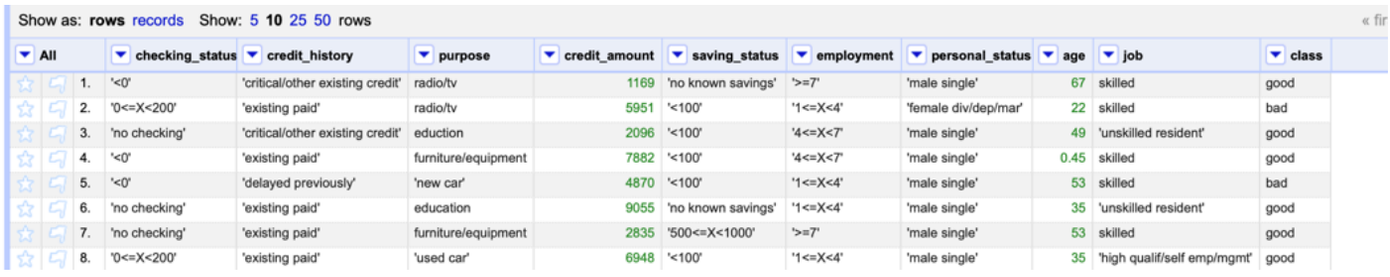
9 Appendices

List of Appendices

Appendix A - Column Heading Changes . . . 10
Appendix B - Empty/Null Text Facets . . . 10
Appendix C - Duplicate Facets and Excel Concatenation . . . 11
Appendix D - Data Consistency Example . . . 11
Appendix E - Data Consistency Changes . . . 11
Appendix F - Text Facet and Clustering Example . . . 12
Appendix F - Text Facet and Clustering Example . . . 12

Appendix H - Accuracy Changes 13
Appendix I - Age Filter Screenshot 1 13
Appendix J - Age Filter Screenshot 2 14
Appendix K - Age Investigation Screenshot 14
Appendix L - Trailing Digits - Credit Amount Comparison 14
Appendix M - Data Transformation - Python Scripts 14
Appendix N - Weka ARFF Conversion 15
Appendix O - J48 Pruned Tree 15
Appendix P - Classification - Rule 3 Combined 16
Appendix Q - Classification - Rule 4 Combined 16
Appendix R - Classification Confusion Matrix 16
Appendix S - Clustering Screenshots 17

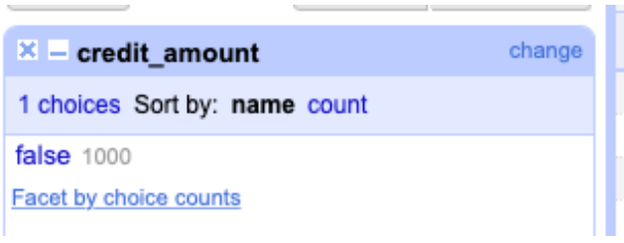
Appendix A - Column Heading Changes



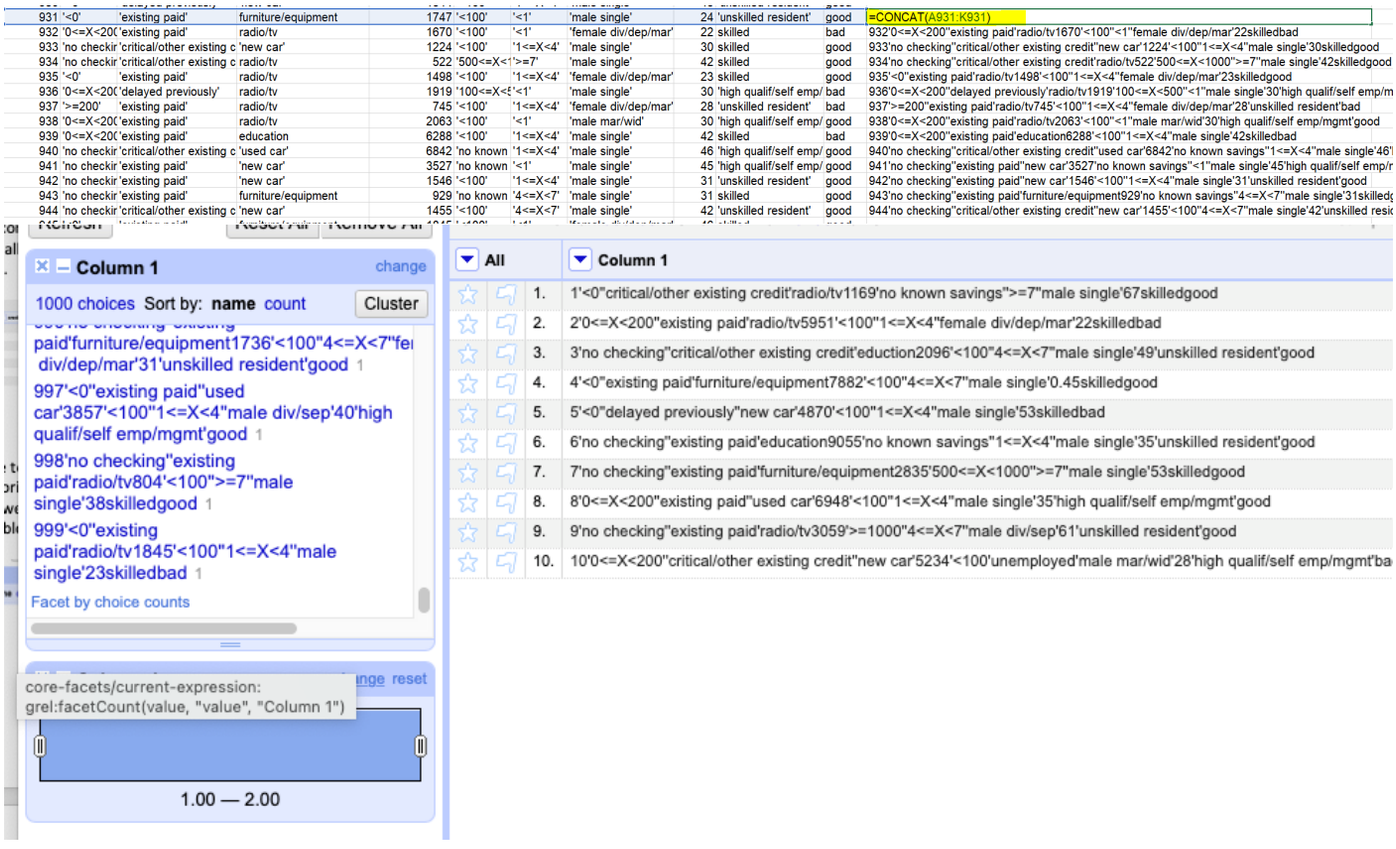
Show as: **rows** records Show: 5 **10** 25 50 rows « fir

All		checking_status	credit_history	purpose	credit_amount	saving_status	employment	personal_status	age	job	class
	1.	'<0'	'critical/other existing credit'	radio/tv	1169	'no known savings'	'>=7'	'male single'	67	skilled	good
	2.	'0<=X<200'	'existing paid'	radio/tv	5951	'<100'	'1<=X<4'	'female div/dep/mar'	22	skilled	bad
	3.	'no checking'	'critical/other existing credit'	eduction	2096	'<100'	'4<=X<7'	'male single'	49	'unskilled resident'	good
	4.	'<0'	'existing paid'	furniture/equipment	7882	'<100'	'4<=X<7'	'male single'	0.45	skilled	good
	5.	'<0'	'delayed previously'	'new car'	4870	'<100'	'1<=X<4'	'male single'	53	skilled	bad
	6.	'no checking'	'existing paid'	education	9055	'no known savings'	'1<=X<4'	'male single'	35	'unskilled resident'	good
	7.	'no checking'	'existing paid'	furniture/equipment	2835	'500<=X<1000'	'>=7'	'male single'	53	skilled	good
	8.	'0<=X<200'	'existing paid'	'used car'	6948	'<100'	'1<=X<4'	'male single'	35	'high qualif/self emp/mgmt'	good

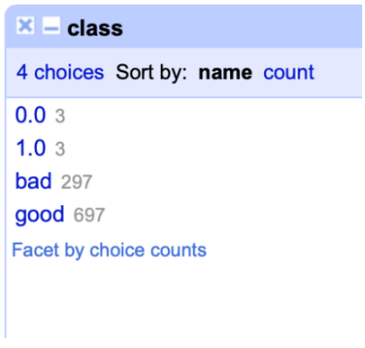
Appendix B - Empty/Null Text Facets



Appendix C - Duplicate Facets and Excel Concatenation



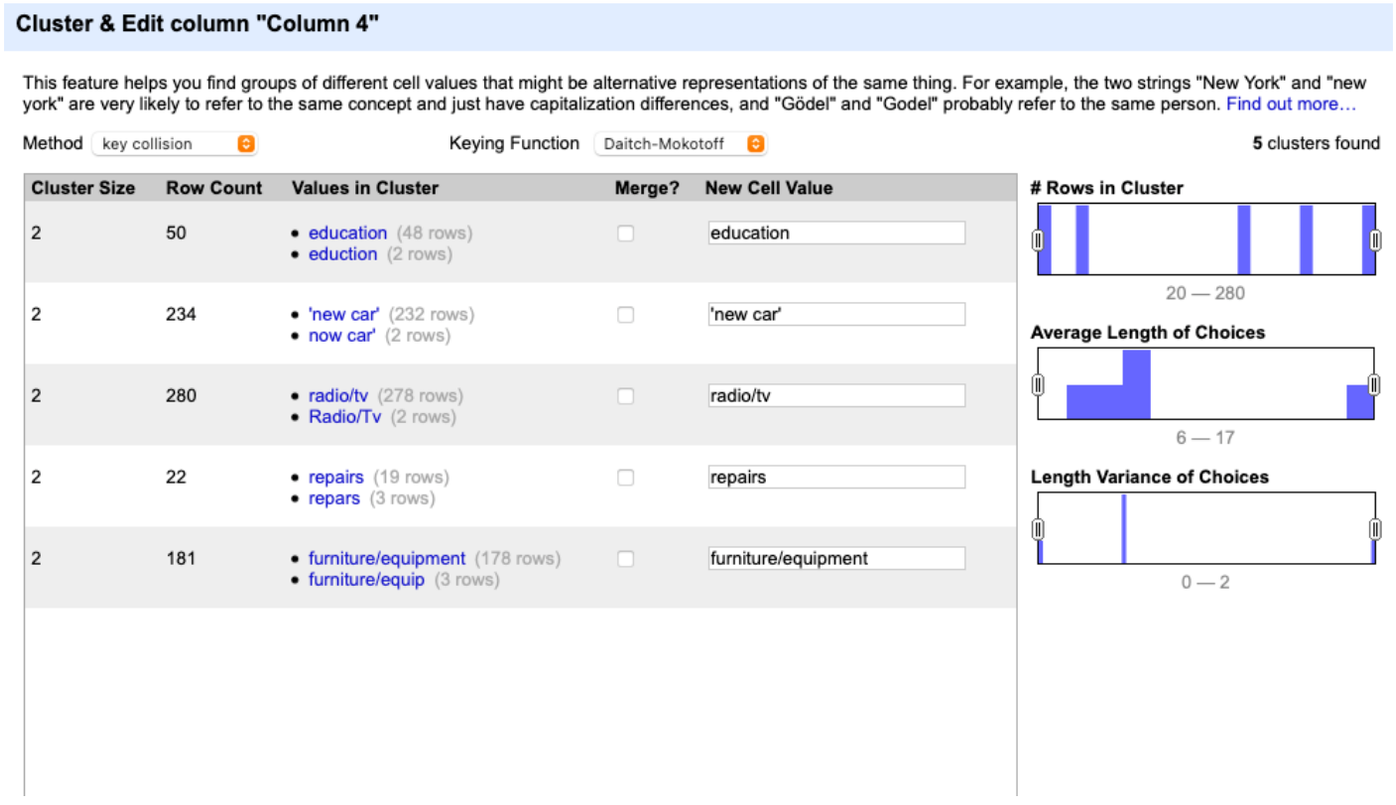
Appendix D - Data Consistency Example



Appendix E - Data Consistency Changes

Attribute	Initial Value	Final Value	Reason For Change
Purpose (Nominal)	now car	new car	Spelling Error
	furniture/equip	furniture/equipment	Abbreviation
	Radio/TV	radio/tv	Capital Letter(s)
	eduction	education	Spelling Error
Age (Numerical)	thirty	30	Data Type Error
Job (Nominal)	Unskilled resident	unskilled resident	Capital Letter(s)
Class (Boolean)	1.0	good	Data Type Error
	0	bad	Data Type Error

Appendix F - Text Facet and Clustering Example



Cluster & Edit column "Column 4"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more…

Method key collision Keying Function Daitch-Mokotoff **5** clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	50	• education (48 rows) • eduction (2 rows)	☐	education
2	234	• 'new car' (232 rows) • now car' (2 rows)	☐	'new car'
2	280	• radio/tv (278 rows) • Radio/Tv (2 rows)	☐	radio/tv
2	22	• repairs (19 rows) • repars (3 rows)	☐	repairs
2	181	• furniture/equipment (178 rows) • furniture/equip (3 rows)	☐	furniture/equipment

 # Rows in Cluster

20 — 280

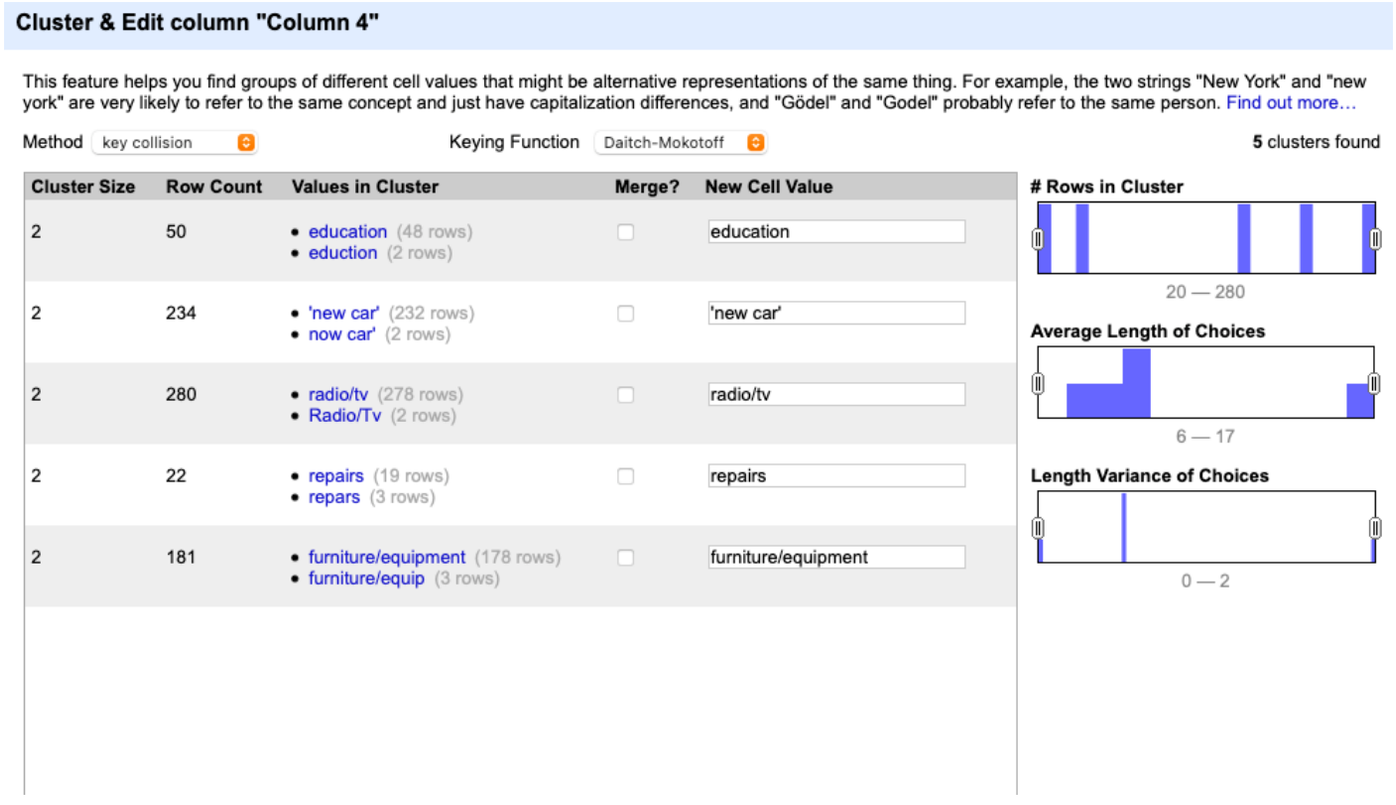
Average Length of Choices

6 — 17

Length Variance of Choices

0 — 2

Appendix F - Text Facet and Clustering Example



Cluster & Edit column "Column 4"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more…

Method key collision Keying Function Daitch-Mokotoff **5** clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	50	• education (48 rows) • eduction (2 rows)	☐	education
2	234	• 'new car' (232 rows) • now car' (2 rows)	☐	'new car'
2	280	• radio/tv (278 rows) • Radio/Tv (2 rows)	☐	radio/tv
2	22	• repairs (19 rows) • repars (3 rows)	☐	repairs
2	181	• furniture/equipment (178 rows) • furniture/equip (3 rows)	☐	furniture/equipment

 # Rows in Cluster

20 — 280

Average Length of Choices

6 — 17

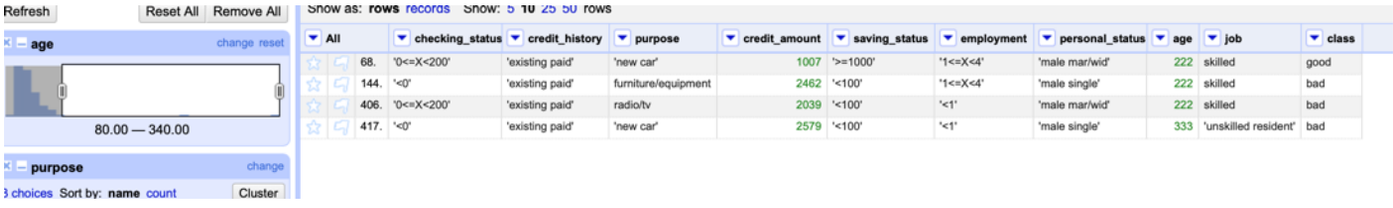
Length Variance of Choices

0 — 2

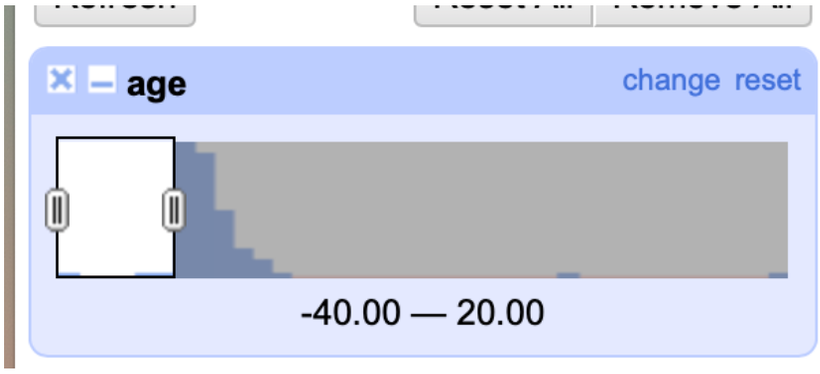
Appendix H - Accuracy Changes

Attribute	Initial Value	Final Value	Basis of Final Value
Job (Nominal)	good	skilled	Assumption that good implies skilled, rather than unskilled.
Class (Boolean)	1.0	good	Assumption that the value reflects to the Boolean 'good'
	0	bad	Assumption that the value reflects to the Boolean 'bad'
Age (Numeric)	0.45	45	Decimal place adjustment into realistic age range.
	0.26	26	
	1	31	Comparison of all other fields to similar records.
	2.3	23	Decimal place adjustment into realistic age range.
	0.26	26	
	2.3	23	
	2.3	23	
	-39	39	Inversed from negative value
	3.6	36	Decimal place adjustment into realistic age range.
	222	22	Repeated digit removed to bring into realistic age range
	222	22	
	222	22	
	333	33	
Credit_Amount (Numeric)	10530000	1053	Trailing digits removed to bring into a range similar to others for that purpose.
	46790000	4679	
	3092000	3092	
	44800000	4480	
	1237000	1237	
	12389000	12389	
	1388000	1388	
	1393000	1393	

Appendix I - Age Filter Screenshot 1



Appendix J - Age Filter Screenshot 2



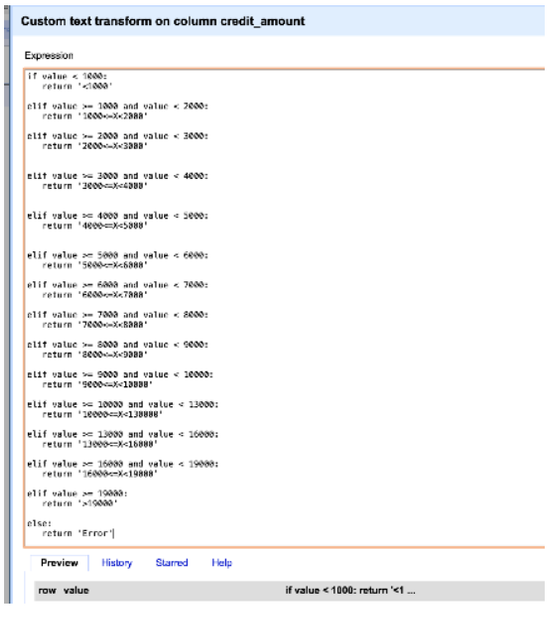
Appendix K - Age Investigation Screenshot

purpose	credit_amount	saving_status	employment	personal_status	age	job	class
'used car'	6187	'100<=X<500'	'4<=X<7'	'male mar/wid'	1	skilled	good
'used car'	6148	'100<=X<500'	'>=7'	'male mar/wid'	31	skilled	good

Appendix L - Trailing Digits - Credit Amount Comparison

All	checking_status	credit_history	purpose	credit_amount	saving_status	employment	personal_status	age	job	class
375.	'0<=X<200'	'all paid'	other	14782	'100<=X<500'	'>=7'	'female div/dep/mar'	60	'high qualif/self emp/mgmt'	bad
916.	'0<=X<200'	'no credits/all paid'	other	18424	'<100'	'1<=X<4'	'female div/dep/mar'	32	'high qualif/self emp/mgmt'	bad

Appendix M - Data Transformation - Python Scripts

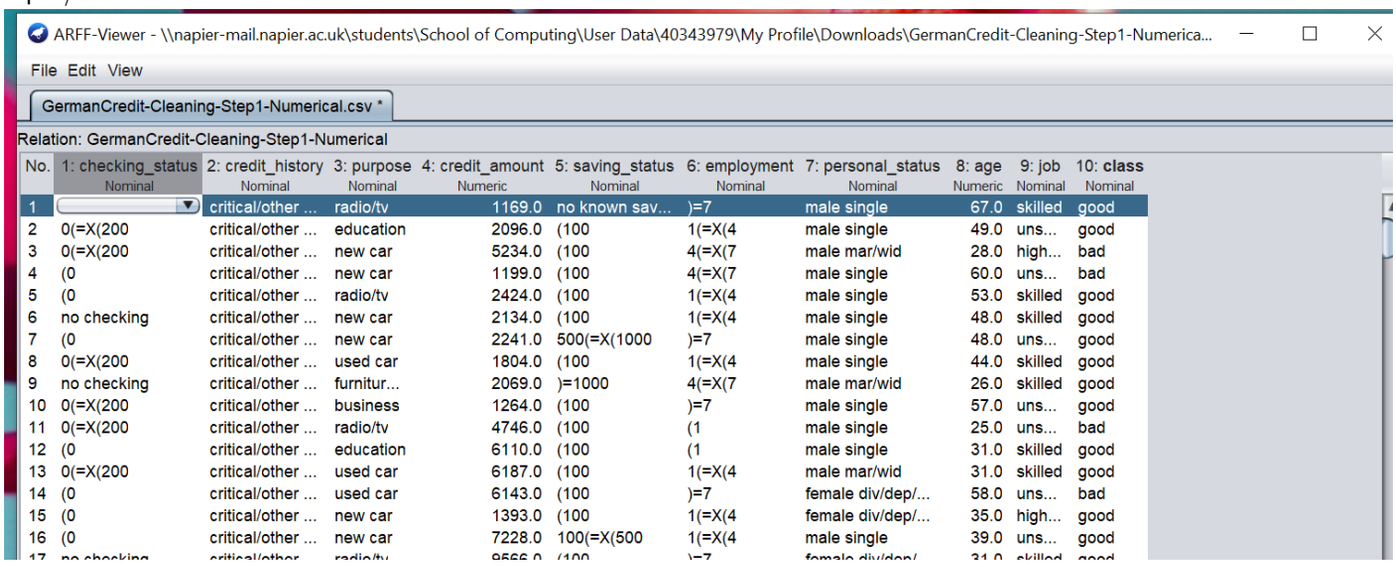


Expression

```
if value < 20:
    return '<20'
elif value >= 20 and value < 30:
    return '20<=X<30'
elif value >= 30 and value < 40:
    return '30<=X<40'
elif value >= 40 and value < 50:
    return '40<=X<50'
elif value >= 50 and value < 60:
    return '50<=X<60'
elif value >= 60 and value < 70:
    return '60<=X<70'
elif value >= 70 and value < 80:
    return '70<=X<80'
elif value >= 80 and value < 90:
    return '80<=X<90'
elif value >= 90:
    return '>90'
else:
    return 'Error'
```

Appendix N - Weka ARFF Conversion

Input/Before



No.	1: checking_status Nominal	2: credit_history Nominal	3: purpose Nominal	4: credit_amount Numeric	5: saving_status Nominal	6: employment Nominal	7: personal_status Nominal	8: age Numeric	9: job Nominal	10: class Nominal
1		critical/other ...	radio/tv	1169.0	no known sav...	)=7	male single	67.0	skilled	good
2	0(=X(200	critical/other ...	education	2096.0	(100	1(=X(4	male single	49.0	uns...	good
3	0(=X(200	critical/other ...	new car	5234.0	(100	4(=X(7	male mar/wid	28.0	high...	bad
4	(0	critical/other ...	new car	1199.0	(100	4(=X(7	male single	60.0	uns...	bad
5	(0	critical/other ...	radio/tv	2424.0	(100	1(=X(4	male single	53.0	skilled	good
6	no checking	critical/other ...	new car	2134.0	(100	1(=X(4	male single	48.0	skilled	good
7	(0	critical/other ...	new car	2241.0	500(=X(1000	)=7	male single	48.0	uns...	good
8	0(=X(200	critical/other ...	used car	1804.0	(100	1(=X(4	male single	44.0	skilled	good
9	no checking	critical/other ...	furnitur...	2069.0	)=1000	4(=X(7	male mar/wid	26.0	skilled	good
10	0(=X(200	critical/other ...	business	1264.0	(100	)=7	male single	57.0	uns...	good
11	0(=X(200	critical/other ...	radio/tv	4746.0	(100	(1	male single	25.0	uns...	bad
12	(0	critical/other ...	education	6110.0	(100	(1	male single	31.0	skilled	good
13	0(=X(200	critical/other ...	used car	6187.0	(100	1(=X(4	male mar/wid	31.0	skilled	good
14	(0	critical/other ...	used car	6143.0	(100	)=7	female div/dep/...	58.0	uns...	bad
15	(0	critical/other ...	new car	1393.0	(100	1(=X(4	female div/dep/...	35.0	high...	good
16	(0	critical/other ...	new car	7228.0	100(=X(500	1(=X(4	male single	39.0	uns...	good

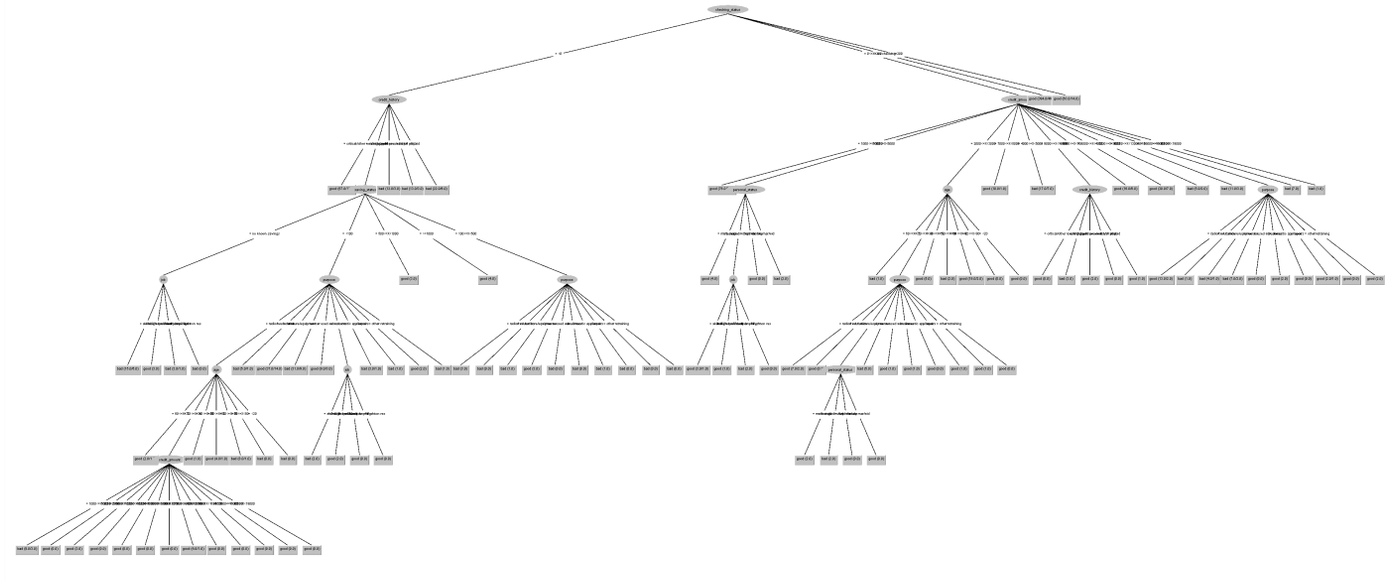
After

```
@relation GermanCredit-Cleaning-Step1-Numerical

@attribute checking_status {<0,0<=X<200,'no checking',>=200}
@attribute credit_history {'critical/other existing credit','existing paid','delayed previously','no credits/all paid','all paid'}
@attribute purpose {radio/tv,education,furniture/equipment,'new car','used car',business,'domestic appliance',repairs,other,retraining}
@attribute credit_amount numeric
@attribute saving_status {'no known savings',<100,500<=X<1000,>=1000,100<=X<500}
@attribute employment {>=7,1<=X<4,4<=X<7,unemployed,<1}
@attribute personal_status {'male single','female div/dep/mar','male div/sep','male mar/wid'}
@attribute age numeric
@attribute job {skilled,'unskilled resident','high qualif/self emp/mgmt','Unemp/unskilled non res'}
@attribute class {good,bad}

@data
<0,'critical/other existing credit',radio/tv,1169,'no known savings',>=7,'male single',67,skilled,good
0<=X<200,'existing paid',radio/tv,5951,<100,1<=X<4,'female div/dep/mar',22,skilled,bad
'no checking','critical/other existing credit',education,2096,<100,4<=X<7,'male single',49,'unskilled resident',good
<0,'existing paid',furniture/equipment,7882,<100,4<=X<7,'male single',45,skilled,good
<0,'delayed previously','new car',4870,<100,1<=X<4,'male single',53,skilled,bad
'no checking','existing paid',education,9055,'no known savings',1<=X<4,'male single',35,'unskilled resident',good
```

Appendix O - J48 Pruned Tree



Appendix P - Classification - Rule 3 Combined

Source Rules:

IF checking_status =¡ 0 AND

credit$_h$$istory == `existing_paid' AND purpose == `newcar' THEN class == bad (42.0/15.0)$

$IF checking\_status =< 0 AND credit\_history == `existing_paid' AND purpose == `domesticappliance' THEN class == bad (5.0/1.0)$

$IF checking\_status =< 0 AND credit\_history == `existing\_paid' AND purpose == `repairs' THEN class == bad (1.0)$

$IF checking\_status =< 0 AND credit\_history == `existing\_paid' AND purpose == `retraining' THEN class == bad (1.0)$

$IF checking\_status =< 0 AND credit\_history == `existing\_paid' AND purpose == `education' THEN class == bad (7.0/2.0)$

Figure 15: Classification Rule 3 Combination - Source Rules

Combined Rule:

"IF checking$_s$$tatus =< 0 AND credit_history ==$

$`existing_paid' AND purpose IN (`newcar', `education' `domesticappliance', `repairs', `retraining') THEN class == bad (56.0/20.0)$"

Figure 16: Classification Rule 3 Combination - Combined Rule

Appendix Q - Classification - Rule 4 Combined

Source Rules:

IF checking_status =¡ 0 AND credit_history == 'all paid' THEN class == bad (22.0/6.0)
IF checking_status =¡ 0 AND credit_history == 'no credit' THEN class == bad (13.0/3.0)
IF checking_status =¡ 0 AND credit_history == 'delayed' THEN class == bad (12.0/3.0)

Figure 17: Classification Rule 4 Combination - Source Rules

Combined Rule:

"IF checking_status =¡ 0 AND credit_history IN ('all paid', 'no credit', 'delayed') THEN class == bad (47.0/12.0)
"

Figure 18: Classification Rule 4 Combination - Combined Rule

Appendix R - Classification Confusion Matrix

Classification	Good (Predicated)	Bad (Predicated)
Good (Actual)	263	30
Bad (Actual)	81	26
Error Rate	23%	45%

Appendix S - Clustering Screenshots