Data Analytics Coursework 2

Steven Pyper
40319882@live.napier.ac.uk

Abstract. This report is an analysis of the dataset provided for coursework 2 of Data Analytics. The aim of this report is to prepare and clean the dataset provided so that it may be analyzed to find any underlying patterns and relationships. When the dataset has been cleaned with Openrefine with all errors either corrected or discarded multiple datasets will then be created with different data types so that they can be used with different algorithms. Once there are multiple cleaned datasets, they will then be analyzed with different techniques and algorithms depending on their datatypes. There will be 3 different types of algorithms used which are classification, association and clustering. These will show patterns inside the data which can be used to show what kind of applicants are safer to offer loans to and which aren't. it may also show attributes of an applicant that give them an advantage or disadvantage over others. Finally, the information found in the analysis will be discussed and the algorithms will be compared to show which techniques were the most effective for finding these patterns in the data.

1 Data Preparation

1.1 Data Cleaning

Before putting the dataset into OpenRefine modifications are required as there are no column names, this would mean that OpenRefine would assume that the first row of entries are the column names. By using Excel, it is easy to add a row at the very top and then add in the column names taken from the metadata provided. Once loaded into OpenRefine cleaning of the data can begin. Many columns have apostrophes surrounding each of the entries which can be removed by using a text facet and simply editing each entry, this will simplify future edits to the entries.

There are entries in the Purpose column which have misspellings and capital letters which do not fit with the metadata, the following are incorrect entries which need to be edited: busines and business should become business, ather should become other, Education should become education. These can be corrected as the true meaning of the data is clear to see as they are simple errors.

The age column had many values which would be considered invalid, three of these were negative numbers and had to be changed to become positive, the values make sense as a positive number (-29, -34, -35) and as such can be changed rather than discarded. Similarly, other values were between 0 and 1 which if multiplied by 100 would also be reasonable entries (0.24, 0.35, 0.44). Both errors could be put down to human error, but it is easy to see what the number should have been as there are both digits there. There is are also age entries of 1 and 6 which do not provide enough information to make an educated guess about what the value should be therefor will be discarded. The final issue with the Age column is that there were two values which were over 100 and are repeating characters (222,333), if the final digit is removed from these entries, they can become valid and it was likely human error.

The case number column can also be removed as it serves no purpose and it does not give any information it is simply a way of keeping track of each user.

Finally, there are also issues with the Job column as there are two entries which have the job "yes", these do not fit in the meta data and do not give any information to that would be useable to predict what they should be for and as such will be removed.

The result of all these changes will be stored in DatasetClean.csv.

1.2 Data Conversion

DatasetNominal.csv will focus on converting the clean dataset into a more useable nominal dataset. To do this there are two main columns that need to be changed. The credit amount column currently has 919 different options which makes using algorithms like id3 impossible as it does not accept numeric options but also algorithms such as j48 which tends to prune it because it cannot get enough accuracy to make use of it. For this dataset it is required to lower the amount of choices for the credit amount, therefor the following transform will be completed:

```
if value <= 1000 : return "between 0 and 1000"
if (value > 1000 and value <=3000) : return "between 1001 and 3000"
if (value > 3000 and value <= 6000) :return "between 3001 and 6000"
if (value > 6000 and value <= 9000) : return "between 6001 and 9000"
if (value > 9000 and value <= 12000) : return "between 9001 and 12000"
if (value > 12000 and value <= 15000) : return "between 12001 and 15000"
if (value > 15000 and value <= 18000) : return "between 15001 and 18000"
if value > 18000 : return "larger than 18000"
```

This will create 8 separate choices, rather than the 919 previously, which while it does lose some of the accuracy, specifically with the larger than 18000 option, it will allow for more generalization in analysis and allow this column to be used in algorithms which prefer nominal data.

The other column which needs to be changed over into nominal data is age. To make this nominal data it will be split up into 4 age groups. 4 groups were chosen as less than 4 led to a lot of accuracy being lost making the outputs from algorithms to be meaningless, whereas more than 4 created to many small groups which would also lose out in terms of generalization. This column will also have to be transformed into numeric to make this change. To do this the following transform code will be required

```
if value <= 25:return "young adult"
if value > 25 and value <= 35: return "adult"
if value > 35 and value <= 55: return "older adult"
if value > 55:return "old age"
```

This dataset is then exported and saved as “DatasetNominal.csv” and will be used for classification and association-based algorithms

A final dataset will need to be created for use of clustering, by using the DatasetNominal.csv as a base the class column will be changed to be a 1 for good and 0 for bad, this way when using clustering algorithms it will be possible to get a mean value for how likely it is for someone in that cluster to be given a loan or not which will give a lot more accuracy compared to just being given a yes or no. to do this the following transform will be used:

```
if value = good: return 1
if value = bad: return 0
```

This column will then be transformed into numerical using the quick transform method, so that the algorithms will know that the mean can be a decimal point number to give the accuracy discussed previously. This dataset will be saved as DatasetNominalClass0-1.csv. All of these csv’s will be loaded into Weka and will be saved as .arff files with the same name as they already have.

2 Data Analytics

2.1 Classification(predictive)

By using the OneR Algorithm we can find the best single predictor for the data. While this tends not as accurate than other methods, it allows for producing an easy to understand rule which shows the most influential single class as explained by the Easy strength on the OneR R project webpage. (von Jouanne-Diedrich, 2017) By using datasetnominal.csv dataset and Cross-Validation with 10 folds the Algorithm will give the following rules

If Credit_History=Critical/other existing credit Then Class = Good

If Credit_History=existing paid Then Class = good

If Credit_History=delayed previously Then Class = good

If Credit_History=no credits/all paid Then Class = bad

If Credit_History=all paid Then Class = bad

These rules have a 71.68% Accuracy over the whole dataset while using the training set. This means that out of 996 instances, 714 were correctly classified but 282 were incorrectly classified. Showing by the confusion matrix the rules over pick instances as being good to give a loan too with 246 entries being classified as good when they should have been bad which is further expanded when you look at precision per class, as good had 72.6% but bad had 59.6%. Due to this it is difficult to draw to many conclusions however with a 71% accuracy which mostly comes from correct classifications of good candidates, it can be said that the credit history of the person does have a huge impact, on whether a person is accepted but not weather they are declined as that must come from other attributes. The reason this algorithm likely over picks good rather than bad is likely because there is more yes's than no's in the dataset, so it finds it easier to calculate class as equaling yes.

To follow this up using the J48 Algorithm the previous rules could refined or disproven, changing the confidence factor to 0.15 the tree will be pruned more which means it give less detailed understanding for certain areas but should show the more basic rules. In this case the following rules were selected from the output which seem to give the most useful information or gave more information about the output:

If Checking_status = <0 And Credit_History = critical/other existing credit
Then Class = good (67.0/18.0)

If Checking_status = <0 And Credit_History = delayed previously
Then Class = bad (12.0/3.0)

If Checking_status = <0 And Credit_History = no credits/all paid
Then class = bad (13.0/3.0)

If Checking_status = <0 And Credit_History = all paid
Then class = bad (22.0/6.0)

If Checking_status = 0<=X<200 Then class = good (269.0/105.0)

If Checking_status = no checking Then class = good (391.0/45.0)

If Checking_status = >=200 Then class = good (63.0/14.0)

It should be noted that Credit_History = existing paid was ignored from these rules as it had a much deeper tree but to small datasets to provide enough information to analyse.

The overall accuracy of this algorithm was only marginally better at roughly 74% when using the training set however when looking at the specific accuracy of some of the rules provided then a clear pattern appears. With all the rules shown apart from rule 5 there was a much higher level of accuracy from all the rules as shown by the following accuracies in order as shown above (excluding rule 5):74%,75%,77%,73%88%,78% and increased the precision per class to 76% for good and 66% for bad. When looking at these rules alone it has a higher accuracy of around 78%. These rules give a much better understanding of the credit history from the first section, it shows that the Credit_History does have an impact on the likelihood of getting a loan but only if the checking status <0 and that if there is no checking account at all then you have a really high chance of getting a loan as the bank will be looking for new customers. However, with rule 5 a clear weakness is shown as it was classified as being good even though 40% of the grouping was misclassified. As such more detailed analysis must be done for checking status 0<=X<200 as there will be other patterns underlying that set.

2.2 Association(descriptive)

With Apriori-type algorithms it is easy to find rules which contain one or two IF statements which can be used to find an outcome, for out purposes we are looking for one or two if statements which result in the class equaling good or bad. As we want to see specific information about whether a person is safe to give a loan to, we will want to use the DatasetNominalclass0-1.csv dataset so that the algorithm can output a number between 0-1 which represents the chance of getting a loan. Due to many of the rules not being applicable the maximum number of rules will be increased to 100, when this was done there were still only 11 rules found, this was due to the confidence from each of the rules being set to 90%, this will lowered from 90% to 70% as these rules would still give a good chance of being useful at showing who would be good to give loans to. Finally, enabling CAR (class association rules) will allow only rules which equal a class to be given out and once the rules are found disabling it to find the lift and conviction for each rule chosen. One consistent IF statement that came up with these rules was checking status which as discussed in the first section was found to be a rough indication for weather a person should receive a loan or not, likely due to the fact that taking a risk on new people gave the chance for the bank to gain long term if they did pay the loan back, These rules expand on this and give better confidence by adding additional if statements which improves the generalization of the rules. The following useful rules were found:

If Checking_status = “no checking” AND employment = >=7 Then Class=good

This rule had a 115 people falling into the If category and had a result of 107 people being classified as good, this gives it a confidence value of 93% which is very high with a coverage of just over 10%. This shows that if someone has been employed for a long period of a time and does not have a checking account with the bank they are very safe to give loans to, this will be due to the consistent income which is unlikely to change so they will be able to repay the loan and the bank will have a consistent new customer. With a lift of 1.33 this rule has a high chance of being accurate at predicating if someone is good to give a loan to in future datasets and is not likely to just be a coincidence.

If Job=skilled Then Class=Good

While this rule has a lower confidence of only 70% it still shows that if a person has a skilled job, they are more likely to be given a loan than other professions. With a much larger coverage with around 627 people being skilled and 442 of them being given a loan shows the strength of this rule which could be expanded with my ifs to make it more applicable and improve its confidence. With a lift and conviction of only 1.01 this rule should be used carefully in new datasets as the confidence of 70% may decline or increase as the rule may just be coincidental.

If Checking_status=”no checking” Then Class=good

This rule shows that a large portion of customers who do not have a checking account are likely to get a loan, this is likely due to the fact that the bank wants more customers to be using their accounts as that is where they earn money. With this rule there are 391 who fit into if side and 346 people who are given a loan, this is a confidence rate of 88% which for one rule is very high with a sizeable coverage showing its accuracy. With a lift of 1.26 this is another rule which should be used for future datasets as it has a much higher chance of being accurate.

If Checking_status = “no checking” and job = skilled then class = good

This rule expands on the previous rules by merging the rules which increased the confidence. This lowers the coverage to 263 people out of which 237 are given a loan, this gives this rule a confidence value 90% which is a massive improvement over the previous job rule and shows a clear relationship that job and checking status combined have a big impact on whether or not a loan is given. Due to the skilled job which provides constant good income and the bank wants new customers that are likely to pay back the interest on loans the bank this rule provides a very safe choice.

If Checking_status = “no checking” AND Personal status = “male single” AND Job = skilled Then Class=good

This is an extension of the previous rules which further improve its generalization by making the even more specific which increases the confidence to 93% although it does lower the coverage. It shows that single males with a skilled job are much more likely to be given a loan with 138 out of 149 being given one. This shows that if the person with a skilled job has no checking account with the bank and is single male they are more likely to get a loan, compared to the previous rule as the person is guaranteed to single this means they are less likely to be providing for two people which means they will have more money to use to repay the loan increasing the safety of the loan.

If age = “older adult” then class=good

This is another simple rule however still has decent coverage with around 257 out of 339 people being given a loan which also gives a confidence of 76% which for a single rule is still fairly high, this gives the impression that the older a person is the more likely it is they are to get a loan, other factors likely cause this such as them working for longer periods of time or them just being more trustworthy due to age(may have previously had debts).

2.3 Clustering(descriptive)

When using the Expectation Maximisation Algorithm it was found that using 7 clusters was most effective, lower values did not give enough insight into the clusters as they would tend to have multiple very small clusters that seemed to fit into clusters and one large cluster that was filled with all the good outcome but did not split it up and give enough uniqueness. By using 7 clusters it was found that cluster 3 was the largest and contained a 99% accuracy rate of finding people who were excellent candidates for giving a loan whereas cluster 1 and 7 found most of the people who would be riskier candidates for loans. All the clusters will be explained in the table below along with the if they would be good candidates or not and the algorithm will be tested using the training set.

Cluster (Instances)	Explanation
1 (124)	A group of individuals who had little saved inside their bank account, many of these people had taken loans before which they had paid. There was a mixture of reasons for them to take the loan but mainly focused around buying a new tv/radio or a new car. The loans they were looking to take out was mostly below €6000 with a large majority looking for between €1000 and €3000. Around of these

	people had little to no savings and were likely to have been employed for a very short amount of time. Most of these people were between 25 to 35 and were female in a skilled job. Overall this cluster gave the lowest chance of being given a loan with only a 34% chance. This is likely due to the lack of savings and uncertainty over income based on the fact they have not been working for that long but they do have a chance based on the fact that they are looking for smaller loans and do have a skilled job that should pay ok.
2 (40)	Another cluster where most people have very little in their checking account and little savings. Many of them have existing credit which has not been paid off yet. Many of these people were looking to buy a car (new or used) with a few looking to buy furniture or invest in a business. Most of these people were looking to take out smaller loans between €1000 and €3000 but some were looking for larger loans of between €3000 and €9000. This group had mostly been employed in skilled jobs for large amounts of time and were mostly single males who were over 25. This group was given a better chance of getting loans at roughly 66%. This is likely because they have been employed in skilled work for long periods of time meaning that they likely have a stable (and high) income. However, the reason this chance is not higher is because they already have loans out and/or are looking to take larger loans out. With a deviation of 47% it is expected that some of the people that may be given a loan would carry a risk. This deviation is likely due to the grouping being so small but due because the group is very distinct conclusions are still useful.
3 (526)	The largest Grouping had a vast majority of people without a checking account and had no known savings or very little with that bank. Many of whom either already had credit or had paid of a previous loan. Most of these people were looking to buy a tv or radio with some looking to buy a car. Most of the loans requested were small although there were a few larger loans. Very few of the people in this group were unemployed or and the vast majority were in skilled jobs and many had been employed for a long time. were over 55. This group had an almost guaranteed chance of getting a loan at roughly 99.9%, this group could be considered as the most risk free for a loan as they have a consistent income which likely pays well which means they should be able to pay off loans easily. The low Standard Deviation of only 1.7% shows that this cluster is a consistently good choice for the bank to give a loan too as there is very little risk being carried.
4 (97)	Like cluster 2 this group had a large portion of people with low amounts saved in their savings and checking accounts with a previous history of having credit. However, unlike cluster 2 they were more interested in buying furniture or tv’s. Favored towards taking smaller loans but again some larger loans. Much of this group had been employed for between 1 and 4 years however some had been employed for less than 1. This grouping is large majority under 25 females who are in skilled work. This grouping was given roughly a 60% chance of being given a loan, the reason this likely isn’t higher due to the fact that although they have been in skilled work they have not been working for that long and are quite young so it is more of a risk, they also lack savings which would further increase the risk. The same as cluster 2 there are some individuals who would be better than the average to give loans to but there are also a few that would be riskier.

5 (75)	Focuses on a group who have no checking account but have previously had debt which has been paid off. They have very little if anything in savings. Most of these people are looking to buy a radio/tv. The vast majority of these are looking for loans between €1000 and €6000. They have been working for between 1 and 4 years in skilled jobs. Most are single males between the age of 25 and 35. This group could be considered mostly save to give loans too(roughly 91%) as they are working in skilled jobs and have been for a period of time and have previously paid of debt showing that they had a good enough income to pay it off previously. However as shown by the deviation rate of 28% there are still some people in this grouping who are less likely to be able to pay it off than others.
6 (34)	A group of people who have some money in their checking account and have paid off all their previous debts. They are looking to buy a car and looking to take loans of between 3000 and 9000. This group has a range of money in their savings accounts with some having a larger amount but some with very little. Many were also unemployed, but some had been employed for many years in jobs that require high qualifications. They are a mix of male and female adults aged roughly between 25 and 45. This group is small and has a lot of mix in many of the attributes and as such it is difficult to tell if they should be given a loan or not as shown by the 58% loan chance and the deviation of 49%, this bracket therefor should not be used in deciding if loans should be given out or not as it is difficult to draw conclusions from such a mixed grouping.
7 (100)	Consists mostly of people who already paid off previous debt and have a little in their checking account and with little savings. These people were also looking to buy cars with loans between 1000 and 9000. They are mostly single males who are over between the age of 35 and 55. Many of which are in skilled jobs however this grouping was only given a 38% chance of getting a loan, this is likely due to the fact that there is little in the savings accounts and even though they do have an income from their skilled job they may struggle to pay of the larger loans. There is a large devotion with this grouping though suggesting that many in this grouping would be fine with receiving a loan but most of them would be risky.

3 Conclusion

3.1 Findings

It was discovered that the most influential attributes for whether a person was given a loan was if a person did not have a checking account with the bank, this is likely because if they do not have a checking account an extra risk is worthwhile to get more customers using the bank.

It was also discovered that if the person looking for the loan was skilled; they had a much higher chance of getting a loan than other types of jobs likely due to the increased income allowing them to repay the loans.

Also a relationship appeared between employment time and being given a loan, they longer a person was employed for the more likely they were to be given a loan, alternatively people who were unemployed or had been employed for short period of time were likely to be declined.

If was also found that single males improved the chance of getting a loan if the person was also in skilled work likely due to having less expenditure than if they were in a relationship and the only person getting income.

The final interesting finding was that as the persons age increased, they were also more likely to be given a loan, this is likely due to them having been in work for longer or having more history about having loans and being more mature.

3.2 Algorithms

Out of all the algorithms used, Expectation Maximization seemed to be the most effective as the clusters that were created were able to give a unique understanding of the different groups and there chances of being given loans with a very high accuracy, as the value returned was between 0 and 1 it not only showed groups that good or bad for giving loans too but also showed groups that in a middle ground and were more risky but still might be worth giving a loan too but perhaps at a higher interest rate. However when using any clustering method it takes more time to do analysis as you have to go through smaller numbers of cluster and increase it until one is found which efficiently splits the clusters into useable size's but also give enough detail and uniqueness to make the analysis accurate. Clustering was particularly good at picking clusters that were given a loan rather than not, this is likely due to the fact that the dataset contained more people that were given a loan, this is one of the main strengths of clustering as written by Earl Cox, "cluster analysis has the virtue of strengthening the exposure of patterns and behaviors as more and more data becomes available" (Cox, 2005), meaning that if a larger dataset was given more interesting patterns could be found improving the accuracy of the clusters.

The outputs from Apriori were also very useful as it very quickly gives a lot of single and double rules which give lots of insight into the underlying data and with only a few changes you get lots of rules and see the confidence values attached to them, with this you can find high confidence short rules which means that they are very general and can be combined with other attributes to become more specific and further increase the confidence meaning that most of the rules that can be picked from are very accurate(with all the rules being above 70%,with one reaching 93%). However, as Rajul wrote "post analysis is required to obtain interesting rules as many of the generated rules are useless" (Anand, Vaud, & Kumar Singh, 2009) which shows one of the main weaknesses of association-based algorithms, they take longer after the output is created to get useful rules and even after this some of the rules will have lower lift and conviction meaning that even if they have high confidence with this dataset care must be taken with them in future

The j48 algorithm was able to show some very basic rules about the data which can be used as a starting point as the outputs were not as accurate as the previous methods with a roughly 74% accuracy. However, this method is very fast and does not take lots of time to understand the underlying data showing some of the most influential attributes and show which of the rules performed better than the others, such as one with 88%.

Finally, the OneR algorithm performs poorly in terms of accuracy as it can only pick the most influential single attribute, this does however mean that you can very quickly find a starting point for an investigation, in this instance it had a habit of over picking people as being good for loans even when they should have been classified as being poor choices as such this algorithm should not be used alone.

References

Anand, R., Vaud, A., & Kumar Singh, P. (2009). *Association rule mining using multi-objective evolutionary algorithms: Strengths and challenges.* Coimbatore, India: IEEE.

Cox, E. (2005). *Fuzzy Modeling and Genetic Algorithms for Data Mining and Exploration.* San Francisco: Morgan Laufmann Publishers.

von Jouanne-Diedrich, H. (2017, 05 05). *OneR - Establishing a New Baseline for Machine Learning Classification Models*. Retrieved from cran.r-project: https://cran.r-project.org/web/packages/OneR/vignettes/OneR.html