Data Analytics: Data Mining

Amelia Handley [40326169]

[1] Edinburgh Napier University
40326169@live.napier.ac.uk

1 Introduction

The aim of this coursework was to conduct an exploratory data mining of a file used to assess the observations made for past applications for credit at a German bank. Data mining is a process for extracting useful information from a dataset. It is used to find meaning relationships and patterns from the attributes [1].

So, in this case, each record of an application submitted to the bank was examined to evaluate how the bank assesses successful and unsuccessful loans. The loans provided by the bank gather information for several factors of the application to assess whether or not an applicant will pay back the loan. In this case the bank had gathered 10 attributes for the applications including: checking current status, past credit history, the purpose of the loan, the amount the applicant requires, the saving status of the applicant, the current employment status, personal status (i.e. single, married etc.), age, the type of job the applicant has and finally whether or not it is safe to provide a loan to the client. To ensure the data could provide appropriate patterns and relationships of the dataset provided from the file required data cleaning and preparation tools to be used within OpenRefine and Weka.

2 Data Preparation

2.1 Data Cleaning

The first step, before conducting any exploration data mining techniques, required the data set to be cleaning. The data was imported to OpenRefine, a tool used for working with messy data by cleaning and transforming it [2]. However, before it was uploaded the data required a row inserted to include the headings for each of the variables (which was done in Microsoft Excel). The file was then uploaded to OpenRefine.

When in OpenRefine, each of the variables was checked using the appropriate facet wrap. Some of the attributes contained parentheses (‘) which were removed accordingly. There was nothing to be cleaned for the case number, checking status, personal status or saving status. The table below highlights all the errors corrected (Table 1).

Table 1: Data Cleaning of the Cleaned Dataset.

Attribute	Case Number	Before	Transformation
Purpose	819	ather	other
	155, 574, 659	busines	business
	740, 862, 942	busness	business
	175	Eduction	education
	58, 121	Radio/Tv	radio/tv
credit_amount	432	111,328,000	111,328
	560	19,280,000	19,280
	595	13,580,000	13,580
	648	13,860,000	13,860
	660	63,610,000	63,610
	444	7,190,000	7,190
	452	5,180,000	5,180
	514	5,850,000	5,850

2

age	66, 80	222	22
	174	333	33
	26	6	60
	54	1	19
	192	-34	34
	233	-35	35
	280	-29	29
	305	0.44	44
	333	0.24	24
	448	0.35	35
Job	65, 69	Yes	skilled

A text facet wrap was used on the purpose attribute and it was found that there were 10 instances of incorrect data. These mistakes were either spelling mistakes or using capital letters for the input. It was noted that there were no applicants who applied for the loan with the purpose of a “vacation”

The credit amount variable was then checked using a numeric facet wrap, and it was discovered that there were significantly high amounts of money being requested by applicants. For instance, the highest amount was 111,328,000 euros (€). These amounts were viewed as an input error so, for the 7 instances found, the data was divided by 1000, so 111,328,000€ would become 111,328€.

There was a numeric facet wrap on the age. There were instances of outliers, two of the results being “222” and one being “333”. The inputs were assumed to be incorrect and changed to “22” and “33” respectively. In addition, 2 of the results from the age facet were low at “1” and “6” which was considered too low for an applicant for a loan so was changed to 19 and 60 respectively. There were 3 cases of negative age numbers, such as “-34” which was changed to 34. Finally, 3 of the applicants had decimal ages, for instance “0.44” which was changed to 44.

Lastly, a test facet wrap was performed on the job type. It was found that there were only 2 cases of input error, with the applicant selecting “Yes”. This was then assumed to the applicant meaning they were a skilled worker and so it was changed to “skilled”.

This produced a cleaned version of the initial dataset. This dataset was the used in the data conversion stage to convert the data in nominal and numerical values.

2.2 Data Conversion

2.2.1. Nominal Conversion

To convert the dataset to allow for association and certain classification algorithms to work required transforming the cleaned dataset numerical variables into nominal. To do this the edit cells > transform function was used on OpenRefine.

First the “credit_amount” was edited. Using the General Refine Expression Language (GREL) to transform the credit amount requested by an applicant into 4 credit “bands”. The code used to transform the column is shown below:

```
if(value < 2000, "0<X<2000",value)
if(and(value>=2000,value<6000),"2000<=X<6000",value)
if(and(value>=6000,value<10000),"6000<=X<10000",value)
if(value >= 10000 , "X>=10000", value)
```

This separates the credit amount into 4 separate categories: “Less than 2000”, “Between 2000 and 6000”, “Between 6000 and 10000” and “More than 10000”.

The age of the participant was also transformed to nominal values. The GREL code is shown below:

```
if(value <=30, "X<=30", value)
if(and(value>=31, value<=45), "31<=X<=45", value)
if(and(value>=46, value<=60, "46<=X<=60", value)
if(value >=61, "X>=61", value)
```

The age was separated into 4 categories: less than 30 years old (<=30), between 31 and 45 (>=31 and <=45), between 46 and 60 (>=46 and <=60) and greater than 61 (>=61).

2.2.2. Numerical Conversion

Using the cleaned dataset attributes there was an instance of nominal data which required to be transformed to allow for numerical conversion. This is because algorithms used for clustering required numerical attributes.

To convert the dataset the “class” attribute was converted from good and bad to 1 and 0. This was to enable the class to give a percentage of the chance an applicant was given a loan from the bank. To do this Jython was used to transform the values as shown below:

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

So now good has a value of 1 and bad has a value of 0.

3 Data Analytics

3.1 Classification

In order to use the classification algorithm (ID3) required the use of the cleaned nominal dataset. Classification is used to predict the outcome using nominal attributes [2]. This meant using the adapted dataset which transformed the numeric variables into nominal ones and removing the case number column (as it is numerical).

The ID3 algorithm was used to generate a decision tree for the dataset. To generate the tree the algorithm uses a greedy search from a top-down approach where each attribute is tested. The attribute selected is the best for classification from the data set. Entropy is used to decide which of the variables is used to construct the tree [3].

Initially, the cross-validation was used initially to produce the tree. This gave 607 (60.7%) correctly classified instances of the applicants whereas there was 265(26.5%) incorrectly classified. The training set was used to see if the results could be improved upon, which they were [6]. The tree produced correctly classified 988 (98.8%) instances making the number of incorrectly classified instances 12. No other settings were changed.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 699   1 |   a = good
  11 289 |   b = bad
```

Image 1: Confusion Matrix of Credit Dataset

The confusion matrix showed that 699 (69.9%) were classified as good applicants for the loan whilst 289 were classified as bad (with the 12 remaining being the incorrectly classified instances) (see image 1)- with 11 false positives and one false negative.

Rule 1:

```
IF checking_status = <0 AND credit_history=critical/other existing credit AND pur-
pose = furniture/equipment AND credit_amount = 0<X<2000 THEN credit = good
```

If the checking status is less than 100 with credit history of critical or there being other existing credit, the purpose being for furniture or equipment and the credit amount less than 2000 then the applicant has a good chance of receiving a loan.

Rule 2:

```
IF checking_status = <0 AND credit_history=critical/other existing credit AND pur-
pose = new car AND employment = >=7 AND age 31<=X<=45 AND job=skilled AND
credit_amount = 2000<=X<6000 THEN credit = bad
```

When the checking status is less than 100, with a credit history of critical or there being existing credit with the purpose being a new car. As well as the applicant having been employed for more than 7 years in a skilled job or an official, the age between 31 and 45 and the amount of money requested being between 2000 and 6000 then the applicant has a bad chance of receiving a loan.

Rule 3:

```
IF checking_status = no checking AND purpose = radio/tv AND credit_history = crit-
ical/other existing credit AND job = unskilled resident AND saving_status = no
known savings AND personal_status = female div/dep/mar THEN credit = bad
```

With a checking status of “no checking”, and the purpose being for a radio or tv, with existing credit, a job as an unskilled resident as well as no known savings and a personal status of a female who is divorces, separated or married. Then the chances of the applicant receiving a loan are bad.

Rule 4:

```
IF checking_status = 0<=X<200 AND saving_status = <100 AND purpose = radio/tv AND
job = skilled AND credit_history = existing paid AND employment = 1<=X<4 AND
credit_amount = 0<X<2000 THEN credit = good
```

If there between 0 and 200 for checking status, the saving status is <100, the purpose for applying for a loan is a radio/tv, the applicant has been employed in a skilled job for between 1

to 4 years, the credit history has been duly paid and the credit amount is less than 2000 then the chances of getting a loan are good.

Rule 5:

```
IF checking_status = no checking AND purpose = new car AND credit_amount =
2000<=X<6000 AND credit_history = existing paid AND saving_status = <100 AND em-
ployment = 1<=X<4 AND personal_status = female div/dep/mar THEN credit = bad
```

If the applicant currently has no account with the bank, the purpose of the loan is for a new car. They have an existing loan, have less than 100 savings and been employed between 1 and 4 years. Also, the personal status of the applicant is a female who is divorced, separated or married then the chances of getting a loan are bad.

Rule 6:

```
IF checking_status = no checking AND purpose = business AND employment = 1<=X<4
AND credit_history = existing paid AND credit_amount = 0<X<2000 THEN credit = good
```

If the applicant currently has no account with the bank, has the purpose of the loan for business, has been employed for between 1 and 4 years, has paid off any existing loans and the credit amount is less than 2000 then the chances of getting a loan are good.

From the rules described above there can be some observations made of the dataset. Using the classification algorithm showed that typically applicants that asked for less than 2,000€ were accepted for a loan application. Also, it the applicant has already paid off an existing loan then they are likely to be accepted.

On the other hand, it can also be said that applicants with the purpose of a new car are rejected for a loan. Additionally, if the personal status is a female who is divorced, separated or married then they are also likely to be rejected.

3.2 Clustering

Clustering helps divide datasets into groups, known as clusters, based on relationships and the patterns in the data [4]. In this instance, it would enable the bank to make decisions based on the clusters of data that already exist.

SimpleK-means clustering is an algorithm used to find the clusters which have not been labelled within the dataset [5]. Each of the observations made to the cluster use the Euclidean distance to calculate the distance so it can see if a cluster is closer to another's "centroid". The centroid is the point the algorithm picks out to represent each of the clusters). After assigning the data points to a cluster, it calculates the clusters score by totaling up the Euclidean distances between each of the data points [7].

To use this algorithm meant using the numerical dataset which was prepared beforehand. The case number column was removed beforehand as it was not seen as necessary. The algorithm was set to find 6 clusters- as there are 6 rules being evaluated from the dataset- and used "training set". The rest of the settings remained the same.

Table 2: Clustering Results from Credit Dataset using SimpleK-means

Cluster	Cluster 0 (202.0)	Cluster 1 (83.0)	Cluster 2 (97.0)	Cluster 3 (252.0)	Cluster 4 (184.0)	Cluster 5 (182.0)
Checking_status	No checking	<0	0<=X<200	No checking	<0	0<=x<200
Credit_history	Critical/other existing credit	Critical/other existing credit	Existing paid	Existing paid	Existing paid	Existing paid
Purpose	New car	Used Car	Radio/tv	Radio/tv	New car	Radio/tv
Credit_amount	2925.20	4626.65	4812.33	3249.48	3690.95	3006.86
Saving_status	<100	<100	<100	No known savings	<100	<100
Employment	1<=X<4	Unemployed	>=7	>=7	>=7	1<=X<4
Personal_status	Female div/dep/mar	Male single	Male single	male single	Male single	Female dvi/dep/mar
Age	33	42	36	38	37	30
Job	Skilled	High qualif/self emp/mgmt.	Unskilled resident	Skilled	Skilled	Skilled
Class	0.89	0.83	0.51	0.94	0.30	0.59

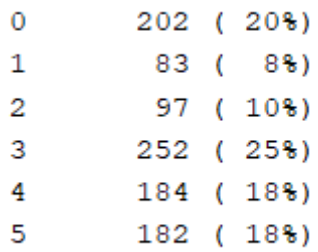

Image 2: Clustered Instances of the Credit Dataset

From table 2 we can see the results of the 6 clusters which form the 6 rules produced from the dataset. There is only one case of a cluster producing a poor result and that is in cluster 4 which includes 18% cluster instances of the overall total of applicants (Image 2).

Cluster 4 has a checking status of less than 0 and a debit history of existing debt which has been paid back duly until now. The cluster shows that the applicants have applied for a 3690.95€ loan with the purpose being a new car. The saving status is less than 100 with employment being over or equal to 7 years. The personal status is a single male who is a skilled worked of 37 years old. Resulting in the cluster receiving a 30% chance of receiving a loan.

The most common cluster was cluster 3 with it having 252 of the applicants in the cluster (or 25%). The cluster shows that the checking status of it is “no checking” meaning that there is no current account with the bank. The applicants have had existing loans but paid them back duly till now. They also applied for a radio/tv with a requested credit amount of 3249.95€. The saving status of the clusters was “no known savings” with employment greater

than or equal to 7 years. Additionally, the cluster was a 38 year old single male who works as a skilled employee or official.

The general trend of the clustering results showed that the checking status of “no checking” produces a higher chance of receiving a loan from the bank. Those with a credit history of “critical/other existing credit” have much higher chances of receiving a loan at 89% (cluster 0) and 83% (cluster 1). Applicants applying for a loan for the purpose of a radio/tv have at least a 51% chance of being accepted as well as if a skilled female who is divorced, separated or married has at least a 59% chance of receiving a loan. Additionally, if the applicant is in management, self-employed or an officed they have an 83% chance of receiving a loan.

3.3 Association

Association algorithms are used to see which data has a relationship with another. This can be thought of as and IF-THEN relationship. For instance, if a customer were to buy nappies then they will likely also buy wipes.

The Apriori algorithm generates association rules based on frequent itemset, i.e. each set of data. It is a tool which can highlight the most frequent trends in the dataset [8].

The nominal dataset was used as Apriori can only nominal variables. This meant removing the case number column before running the algorithm. The number of rules produced from the algorithm was changed to 6 - as there were 6 rules to be produced
Changed number of rules to 6. The rest of the settings remained the same as the default.

```
Best rules found:
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)
```

When the checking status is “no checking”, the purpose for the application is a radio/tv then the applicant is likely to get a loan. This is likely for 94% of the applicants.

```
2. 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)
```

If the checking status of the loan is “no checking”, the credit history is critical or there is other existing credit then the applicant has a 93% of receiving the loan.

```
3. checking_status=no checking employment=>=7 115 ==> class=good 107
<conf:(0.93)> lift:(1.33) lev:(0.03) [26] conv:(3.83)
```

When the checking status is “no checking” and employment is greater than or equal to 7 years then the applicant has a 93% chance of receiving the loan

```
4. checking_status=no checking credit_amount= 0<X<2000 169 ==> class=good 156
<conf:(0.92)> lift:(1.32) lev:(0.04) [37] conv:(3.62)
```

When the checking status is “no checking” and the credit amount being requested is less than 2000 then the chance of receiving a loan is 92%

8
```
5. checking_status=no checking credit_amount= 0<X<2000 job=skilled 114 ==>
class=good 105    <conf:(0.92)> lift:(1.32) lev:(0.03) [25] conv:(3.42)
```

If the checking status is “no checking”, the credit amount is less than 2000 and the job is skilled then the applicant has a 92% chance of receiving a loan.

```
6. 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)
```

If the checking status is “no checking” and the applicant is a single male with a skilled job. This applicant has a 92% chance of a receiving a loan.

The general trend for the Apirori algorithm found that the most common checking status is “no checking”. The best chance for the applicant receiving a loan is when they also apply for a radio or tv (Rule 1-94%).

4 Conclusion

By using three different algorithms- classification, clustering and association- provided an in-depth exploration of the dataset was conducted. The general trend appeared that applicants were likely to receive a loan. If the applicant has a checking status of “no checking”. This could be due to that fact that the applicant currently does not have an account with the bank so the bank will probably want their custom.

Their chances of receiving a loan increased if the applicant asked for less than 2,000€ as seen from the classification rules 1,4 and 6 and association rules 4 and 5. This could be due to it being a lesser amount of money the applicant has to pay back so the bank would feel more confident about getting the money and interest back. Employment also seems to be an important factor as those employed for seven or more years have greater chances of receiving a loan (see clustering - clusters 2 and 3, association - rule 3). Also, if an applicant has paid off an existing loan that also increases their chances of a loan (see classification rule - 4,6 and clustering clusters 2,3,5). A bank would feel confident giving money to someone with a stable job for more than seven years as the applicant would be most likely receiving income to pay off the loan and an applicant who has already paid one off.

Also, two factors which seem to be significant together included a single male applying for a loan with a skilled job from rule 6 in association and from clusters 1,2 and 3. When a female that was divorced, separated or married applied for a loan it appeared that those with an “unskilled” job were rejected (classification - rule 3) whilst ifs the female applicant has a skilled job they are more likely to be accepted (clustering - clusters 0 and 5).

Overall, the results of the SimpleK-means algorithm, the clustering technique, was the best tool for the analysis of the dataset. This is because the observations made by the algorithm resulted in a vector representing an applicant, meaning that all the attributes are represented. Additionally, from a visual and analytical viewpoint it was easier to evaluate the data, compared to the results created from, for instance, classification which printed out the whole tree. To conclude, clustering highlighted the importance some factors - in particular - the checking status of “no checking” and how the attributes “single male” and “skilled” job related to each other, This information was further highlighted when implementing the other algorithms to provide a better insight into the dataset.

References

1. S.Neelamegam, E.Ramaraj. Classification algorithm in Data mining: An Overview. International Journal of P2P Network Trends and Technology (2013).
2. OpenRefine Homepage, https://openrefine.org/, last accessed 2019/11/24
3. The Learning Machine. https://www.thelearningmachine.ai/tree-id3, last accessed 2019/11/26
4. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/ , last accessed 2019/11/26
5. Trevino, A., Introduction to K-Means Clustering https://blogs.oracle.com/datascience/introduction-to-k-means-clustering, last accessed 2019/11/26
6. Waikato., https://www.cs.waikato.ac.nz/ml/weka/mooc/dataminingwithweka/transcripts/Transcript2-2.txt last accessed 2019/11/28
7. BioTuring. https://blog.bioturing.com/2018/10/17/k-means-clustering-algorithm-and-example/, last accessed 2019/11/28
8. Hackerearth., https://www.hackerearth.com/blog/developers/beginners-tutorial-apriori-algorithm-data-mining-r-implementation/, last accessed 2019/11/29