

Carlos Jimenez Rodriguez

[40452913@live.napier.ac.uk](mailto:40452913@live.napier.ac.uk)

Coursework II Report -- Data Analytics – SET09120

1.- Introduction

The purpose of this coursework is to understand how to work with OpenRefine and Weka, to manage an investigation on data mining, using the dataset given, create a report about the data recognition I found. Using OpenRefine to clean the data given, then on I used Weka to visualise the data using the tools available in that software, as well it’s been used for supplying the algorithm for the data analysis, Weka is formed by machine learning algorithms, use for data mining tasks, containing tools such as classification, clustering, regression, association rules, visualization, and pre-processing.

2.- Data Preparation

Data preparations consist of processing the collecting data, cleaning, and altering it before doing the process of data analysis, usually what is done on this part is improve source data, equalize the data format, or removing outliers.

2.1 Data Cleaning.

The given dataset had a lot of errors that must be clean before the data analysis, using OpenRefine which come up with few tools which facilitate on the time of finding errors on the dataset and then on edit it, here are the errors, corrections and steps done for cleaning the dataset provided:

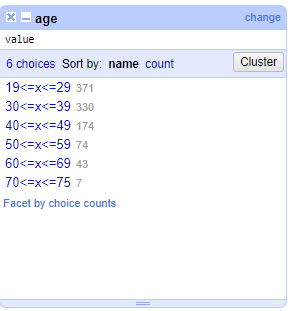
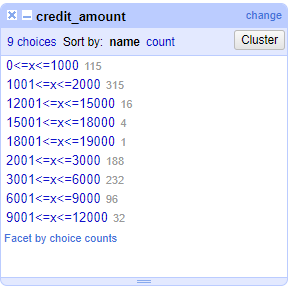
1st) The first thing I notice it was that the attribute ‘case\_no’ started with the number 2, so I had to change it to start with 1, on Edit cells > transform, on the text block I changed “value” to “value -1”, that way it changed the order of it and starts the case number with 1.

2nd) I notice as well that the names of the columns were not with their proper attribute name, so I changed each of them for their attribute name.

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute | Before | Issue | After |
| Purpose |  |  |  |
|  | ‘domestic appliance’ | Removed quotes ‘’ | domestic appliance |
|  | ‘new car’ | Removed quotes ‘’ | new car |
|  | ‘used car’ | Removed quotes ‘’ | used car |
|  | eduction | Mistyped word. | education |
|  | furniture/equip | Mistyped word. | furniture/equipment |
|  | now car’ | Mistyped word and Removed quote ‘ | new car |
|  | Radio/Tv | Changed to lowercase as the rest data. | radio/tv |
|  | repars | Mistyped word | repairs |
| credit\_amount |  |  |  |
|  | On case\_no ‘407’ credit\_amount=10530000 | Removed the zeros since they were outliers | Credit\_amount = 1053 |
|  | On case\_no ‘469’ credit\_amount=46790000 | Removed 4 zeros | Credit\_amount = 4679 |
|  | On case\_no ‘470’ credit\_amount=3092000 | Removed 3 zeros | Credit\_amount = 3092 |
|  | On case\_no ‘471’ credit\_amount=4480000 | Removed 3 zeros | Credit\_amount = 4480 |
|  | On case\_no ‘492’ credit\_amount=1237000 | Removed 3 zeros | Credit\_amount = 1237 |
|  | On case\_no ‘563’ credit\_amount=12389000 | Removed 3 zeros | Credit\_amount = 12389 |
|  | On case\_no ‘582’ credit\_amount=1388000 | Removed 3 zeros | Credit\_amount = 1388 |
|  | On case\_no ‘635’ credit\_amount=1393000 | Removed 4 zeros | Credit\_amount = 1393 |
| personal\_status |  |  |  |
|  | ‘female div/dep/mar’ | Mistyped “dep”, changed to “sep”. | ‘female div/sep/mar’ |
| age |  |  |  |
|  | case\_no ‘400’; age = -39 | Changed to a positive number | age = 39 |
|  | case\_no ‘24’; age = 0.26 | Removed the decimal | age = 26 |
|  | case\_no ‘55’; age = 0.26 | Removed the decimal | age = 26 |
|  | case\_no ‘3’; age = 0.45 | Removed the decimal | age = 45 |
|  | case\_no ‘67’; age = 222 | Mistyped number with an extra 2 | age = 22 |
|  | case\_no ‘143’; age = 222 | Mistyped number with an extra 2 | age = 22 |
|  | case\_no ‘405’; age = 222 | Mistyped number with an extra 2 | age = 22 |
|  | case\_no ‘416’; age = 333 | Mistyped number with an extra 3 | age = 333 |
|  | case\_no ‘580’  age= thirty | Changed Data type to number. | age = 30 |
|  | case\_no ‘47’; age = 2.3 | Removed Decimal point. | age = 23 |
|  | case\_no ‘58’; age = 2.3 | Removed Decimal point. | age = 23 |
|  | case\_no ‘144’; age = 2.3 | Removed Decimal point. | age = 23 |
|  | case\_no ‘566’; age = 3.6 | Removed Decimal point. | age = 36 |
|  | case\_no ‘568’; age = 3.6 | Removed Decimal point. | age = 36 |
|  | case\_no ‘43’; age = 1 | Can’t be 1 years old, be marriage and use car so I changed. | age = 21 |
|  |  |  |  |
| job |  |  |  |
|  | skilled | The other choices where with quotes, so added to be all the same. | ‘skilled’ |
|  | Case\_no (“425”,”460”,”522”)  job = good | Only 3 cases with good, so I changed to skilled. | ‘skilled’ |
| class |  |  |  |
|  | Case\_no (“143”,”242”,”952”)  class = 0 | I supposed 0 is equals to bad, changed data type to text | bad |
|  | Case\_no (“143”,”242”,”952”)  class = 1 | I supposed 1 is equals to good, changed data type to text | good |

2.2 Data Conversion.

For doing the data conversion I had to transfer from numeric to nominal value as I show on [FIG1](#FIG1), before doing the transformation I had to transform the data type to number.

By doing that it will be easier for the coalition when we analyse the data using Weka, for example in age I classified in 6 groups from one age to other depending on the value in x, in this scenario the value in age, it will set it to a group or other. Data converted to Nominal:

Data Converted to Nominal in 6 groups Data Converted to Nominal in 9 groups

***3. Data Analysis.***

In this section I will be using the program Weka, where I will be able to analyse the dataset given using the tools that are implemented already in the software, I will have to use three different ways of doing the data mining which are the followings:

***3.1 Classification.***

In this section I will be recognizing in which set of each categories using the dataset given and cleaned from before, the information in the dataset, is for known “How probable is for a person X to get a loan”, X is an exact person, so comparing it with another person who has similar credits or the same status we will be able to clarify the rules.

For finding the rules Weka generated a tree where all the data is showing, the technique used was pruning, so by using that technique reduces the size of the tree only showing the segment of the tree which has more value or quantity of information, so its delete the ones which are almost useless since there is not much information on them, the I used on Weka was J48 to generate a decision tree, which allows to anticipate the aim variable from a new data frame.

The results I got in Weka of the accuracy was 71.3% which means the classifier model is decent , using the confusion matrix table where we can get the true values, from the false negatives and false positives, the false negatives occurs when the algorithms anticipates the value as false but the actual answer is yes, so the false positive is the other way around, when the algorithms says that is a yes, while the actual answer would be no, here is the confusion matrix generated y the algorithm in Weka:

=== Confusion Matrix ===

a b <-- classified as

141 159 | a = bad

44 655 | b = good

So, in this table we can see that the algorithm from 159 False Positive and 44 False Negative, they have been wrongly categorized, so the correctly classified instances was 796, and incorrectly 203.

**Rule 1:**

*If* checking\_status **= 0<=x<200** *AND* saving\_status **= <100** *AND* credit\_amount **= 1001<=x<=2000** *AND* employment **= 1<=x<4** *THEN* **good** (11.0/3.0)

So, if the customer current credit is between 0 and 200 and the savings status is equal or less than 100 and credit amount is between 1.001 and 2.000, and they have been working for more than 1 year and no more than 4 years, then they will get the loan, since out of 11 cases that this rule is applied to , only 3 has been incorrectly classified by the algorithm.

**Rule 2:**

*If* checking\_status **= 0<=x<200** *AND* saving\_status **=100<=x<500** *AND* age **= 30<=x <=39** *THEN* **good** (19.0/5.0)

For the customers who has a current credit between 0 and 200, and a saving status between 100 and 500 and the age is between 20 and 39, these customers will get the loan, out of 19 cases only 5 of them has been incorrectly classified.

**Rule 3:**

*If* checking\_status **=<0** *AND* credit\_history **= existing paid** *AND* saving\_status **=<100** *AND* purpose **= new car** *THEN* **bad** (31.0/9.0)

In this rule for the customers who has a credit status less than 0 but has an existing payment in the history, and the saving status is equal or less than 100, and their purpose is to get a new car, they are not going to be able to get the loan, out of all the 31 clients, 9 were incorrectly classified by the algorithm.

**Rule 4:**

*If* checking status **=<0** *AND* credit\_history **= existing paid** *AND* saving\_status **=<100** *AND* purpose **= furniture/equipment**, *THEN* **good** (37.0/14.0)

Customers who have a credit status less than 0 but has an existing payment in the history, and the saving status is equal or less than 100, and their purpose is to buy furniture or equipment they are possibilities for them to get the loan, out of all the 37 clients, 14 were incorrectly classified by the algorithm.

**Rule 5:**

*If* checking\_status **= no checking**, *THEN* **good** (394.0/46.0)

In the dataset we can notice that the 88% of the people whose checking status is not checked has received the loan, out of 394 clients only 46 cases have been incorrectly categorized

**Rule 6:**

*If* checking\_status **= <0** *AND* credit\_history **= critical/another existing credit**, *THEN* **good** (66.0/18.0)

On this rule the customers who has a credit status of less than 0 but has on the history a critical credit or another existing credit, they will receive the loan, out of 66 clients only 18 where incorrectly categorized.

***3.3 Association.***

On this section I used Weka as well, suing the Apriori algorithm that’s given as a Weka tool, which function is to find the same items and some associations in the dataset.

**Rule 1:**

checking\_status = **no checking** credit\_amount **=1001<=x<=2000** 130 ==> class **= good** 118 <conf:(0.91)> lift:(1.3) lev:(0.03) [27] conv:(3)

Clients whose credit status have not been check but the credit amount is between 1001 and 2000 has possibilities to get the loan, the confidence is 0.91 which means is probable to be accurate, so 91% of all the cases with these characteristics is true all of them.

**Rule 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:(4.18)

Clients who have not been checked their status and has a credit history of a critical or other existing credit are likely to get the loan, with a 0.93 of confidence, all these clients with the same situation has a chance to be true of getting the loan, being a 93% true.

**Rule 3:**

checking\_status **=no checking** purpose **=radio/tv** 127 ==> class **=good** 120

<conf:(0.94)> lift:(1.35) lev:(0.03) [31] conv:(4.77)

While the checking status it has not been checked and the purpose is to buy a radio or a tv, for these customers there is a chance of getting the loan, with a confidence of 0.94 which is a really accurate, so is true for the 94% of the cases.

**Rule 4:**

checking\_status **=no checking** credit\_amount **=1001<=x<=2000** 130 ==> class **=good** 118 <conf:(0.91)> lift:(1.3) lev:(0.03) [27] conv:(3)

On this rule when the checking status has not been checked and the credit amount of the clients are between 1001 and 2000, they have a chance of being getting the loan, with a 0.91 of confidence, 91% of the clients has the same characteristic.

**Rule 5:**

checking\_status **=no checking** age **=30<=x<=39** 147 ==> class **=good** 134 <conf:(0.91)> lift:(1.3) lev:(0.03) [31] conv:(3.15)

For the clients whose age are between 30 and 39 and the credit status has not been checked yet, they have a chance of be getting the loan, a 0.91 of confidence, so a 91%.

**Rule 6:**

checking\_status **=no checking** employment **=>=7** 115 ==> class **=good** 107 <conf:(0.93)> lift:(1.33) lev:(0.03) [26] conv:(3.84)

Finally, on the customers who has been working for more than 7 years, and there is no data about their checking status, they have a big chance of getting the loan, the rule has a confidence of 0.93 which is very efficient.

***3.4 Clustering***

The main point of clustering is to make group of information to regulate patterns from the dataset, using Weka again, on the section of Cluster, I have used the algorithm SimpleKMeans which calculate the distance from instances and clusters so then it chose in which cluster each case should go.

Text, table

Description automatically generated[Cluster List:](#clustt)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Cluster 0 | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
| Checking\_status | No checking | <0 | 0<=x<200 | No checking | 0<=x<200 | <0 |
| Credit\_history | Existing credit | Existing paid | existing paid critical/other | Existing credit | Existing paid | Existing paid |
| Purpose | Furtniture/  equipment | New car | Radio/Tv | Radio/Tv | Furniture/Equipment | Furniture/Equipment |
| Credit\_amount | 3071 | 2933 | 3513 | 3291 | 3032 | 3516 |
| Saving\_status | <100 | <100 | <100 | <100 | No known savings | <100 |
| Employment | 4<=x<7 | 1<=x<4 | 1<=x<4 | >=7 | <1 | 1<x<4 |
| Personal\_status | Female div/sep/mar | Male single | Male single | Male single | Female div/sep/mar | Female div/sep/mar |
| Age | 31 | 37 | 34 | 41 | 30 | 31 |
| Job | Skilled | Unskilled resident | Skilled | Skilled | Skilled | Skilled |
| Class | good | good | good | good | good | bad |

On the [Cluster list](#clusterlist) we can see on cluster 5 the one which has less chance to receive the loan, so the clients whose checking status is less than 0 and has existing paid credits and the purpose is to get a Furniture/Equipment, and wants a credit amount of 3516 but has less than 100 credits in savings, as well has been working between 1 and 4 years, and is a female divorced, separated or marriage, with an age of 31 and skilled on a job, they are not going to get any loan.

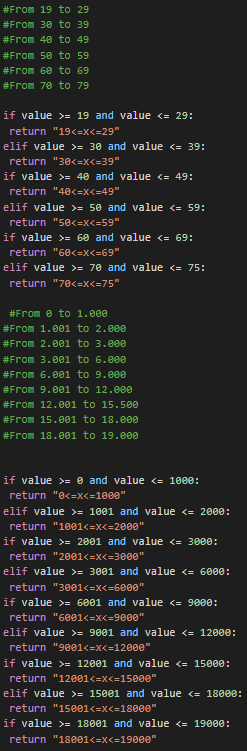
Another of the details we can notice from the list is that in every cluster except in one, exactly on the ‘Cluster 4’, their saving account is <100 except on the cluster 4 which is no known savings.

***4.- Conclusion***

By doing this coursework I learned how to clean a data which comes with some data which is incorrect using OpenRefine, which is helpful with the tools that are implemented on that software.

Then on do a data analysis, where I had to use 3 different kinds of algorithms for each kind of data analysis, the one that I found more useful was association since it gives you the rules straight using Weka and even tells you the confidence of the rule.

It was an interesting module, thanks for both coursework, they were fun.



**FIG1**