Understanding the task and its goals

Data Collection(done by the banks)
Data Preprocessing
Data Visualization
Data Analytics and Application

A given set of data has been collected over the years by banks to access creditworthiness of their customers. Through the data we are able to see if the attributes have a correlation to determine the creditworthiness of a customer. After identifying the attributes linked to creditworthiness, we can help banks understand and solve problems that they face.

The end goal of this data analysis is to classify customers of the bank into risk categories through their past credit scores. This will give an insight to the banks whether the particular customer is reliable enough to receive a financial credit based on their credibility to pay back mo money timely in the past. This helps banks to also give better lending decisions for each customer when they apply for loans or a credit card. By conducting due diligence, the banks can mitigate risks of losing money and have a closer relationship with their customers by providing them with accurate credit limits and caring for their financial health.

This research will greatly help banks in scoring their customers more effectively to minimise credit risks and setting profitable credit limits.

Analysis

For this research, there are 46 attributes in total.

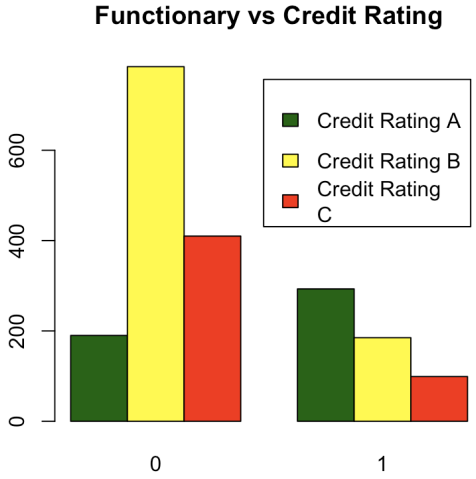
- Checked for missing data
- I divided the data set to find out those who have a credit rating and those who do not have a credit rating. I can only use the data of people who has a credit rating to train my machine learning models as ‘classifiedRows’(1962 entries). I will reserve those without a credit score as my testing set as ‘unclassifiedRows’ (538 entries) to see if my model can accurately predict their creditworthiness.

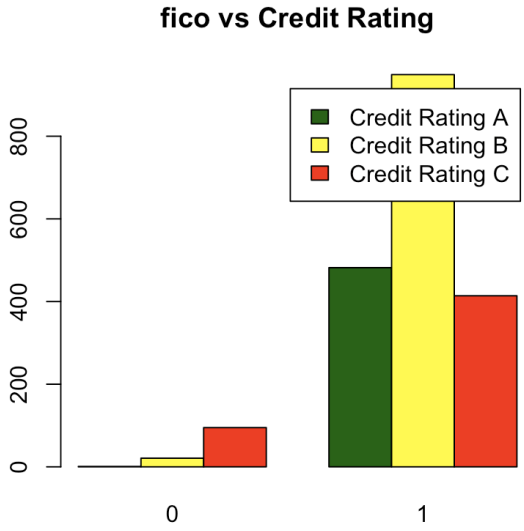
credit.refused.in.past.	0.217838467
max..account.balance.11.months.ago	0.049489339
max..account.balance.3.months.ago	0.038158186
max..account.balance.8.months.ago	0.035423671
avrg..account.balance.9.months.ago	0.029746003
min..account.balance.5.months.ago	0.029634472
min..account.balance.12.months.ago	0.028736469
max..account.balance.12.months.ago	0.027334233
avrg..account.balance.12.months.ago	0.026958574
min..account.balance.8.months.ago	0.025463879
X0..accounts.at.other.banks	0.023139554
max..account.balance.9.months.ago	0.022032572
min..account.balance.2.months.ago	0.021604156
max..account.balance.2.months.ago	0.018010779
min..account.balance.3.months.ago	0.016808148
avrg..account.balance.5.months.ago	0.015258396
avrg..account.balance.8.months.ago	0.013587599
self.employed.	0.012380228
min..account.balance.10.months.ago	0.010292940
years.employed	0.009639419
savings.on.other.accounts	0.007313834
avrg..account.balance.7.months.ago	0.005868788
max..account.balance.5.months.ago	0.004945986
min..account.balance.11.months.ago	0.004944211
avrg..account.balance.10.months.ago	0.004871204
min..account.balance.9.months.ago	0.003433766
max..account.balance.4.months.ago	0.002786613
avrg..account.balance.4.months.ago	-0.004067953
max..account.balance.7.months.ago	-0.007017496
max..account.balance.10.months.ago	-0.007556952

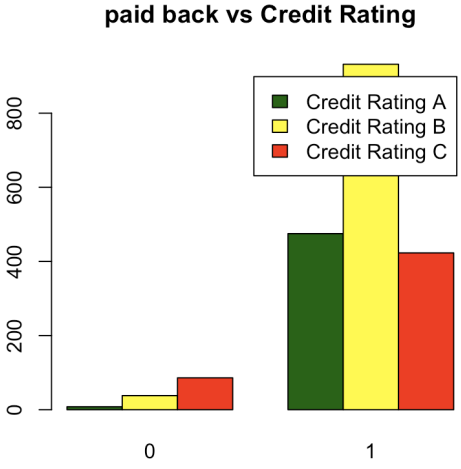
- I then checked every attribute in my ‘classifiedRows’ to see if they have any correlation to the credit.rating.
- I also did some SOM visualisation to help more analysis

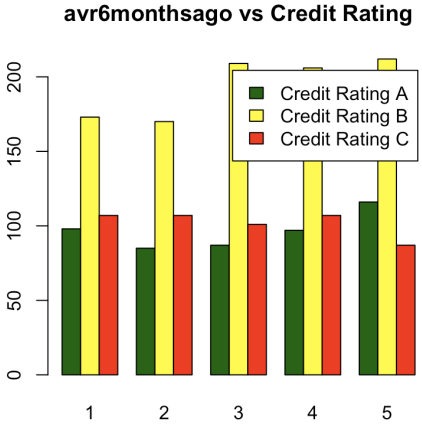
From here I realized that functionary, paid back recently overdrawn, FICO score, gender, and average account balance 6 months ago have the strongest negative correlation with the target.

After exploring the dataset more, I realised the most important features I want to use.

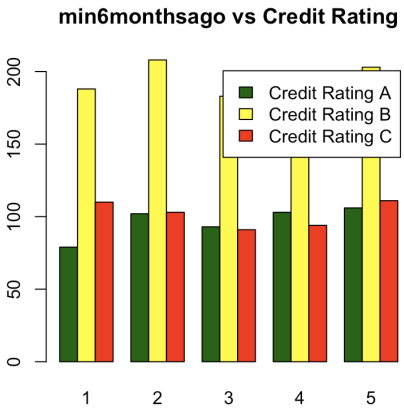








1 = <100, 2 =<1000, 3= <7500 4=<10000, 5=<20000



1 = <100, 2 =<1000, 3= <7500 4=<10000, 5=<20000

Generally with all the graphs shown, there is a strong relationship with a better credit rating in regards to the top 5 attributes i will mention below with the reasons why I think it is such a case.

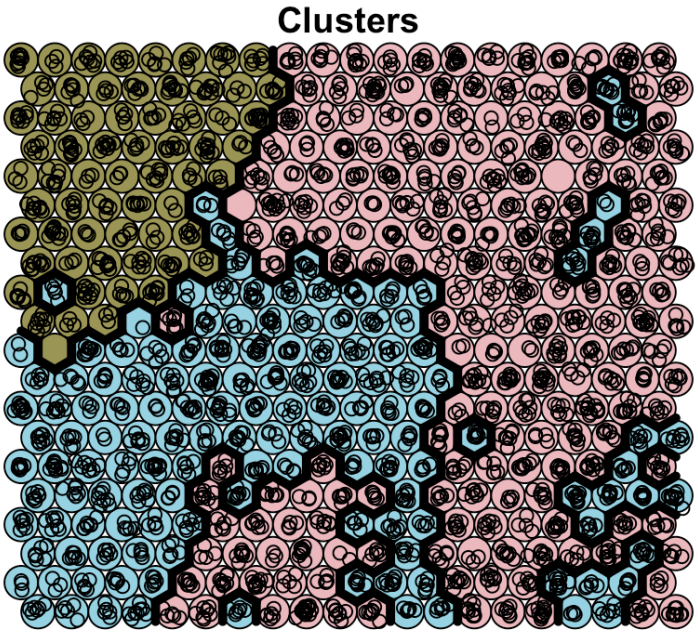
1)functionary. From the functionary vs credit rating graph we can see that there is a relationship of having a better credit score when someone is employed. To me functionary is one of the most interesting attribute as it suggests that people who does not have a job are less likely to have a good credit rating. It is logical to think that if you do not have an income, you might not be able to pay off your loans. In some cases, there might be people who have taken a loan while they had a stable job until one day they lose it suddenly due to recession, layoffs, and accidents. It is also possible that some people might have money management problems despite having a job. Despite the possibilities of unfortunate circumstances, it is not a very common issue that affects the majority of the people. In conclusion, having a job is quite important when considering taking out a loan.

2)re.balanced..paid.back..a.recently.overdrawn.current.acount. This tells me a lot about the customer if they have the capability/habit to pay back their debts on time. If they are able to do so, they are definitely more trustworthy to be able to get a loan as they seem to be able to pay back their loans.

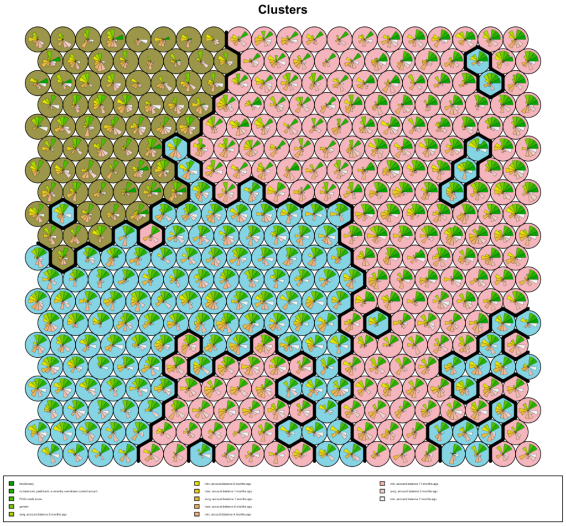
3)FI3O.credit.score. This is a three-digit number that represents the amount of risk a prospective borrower poses to a lender. I think that the banks can avoid potential problems with certain customers by considering this score. Through

4)avrg..account.balance.6.months.ago. This is quite interesting to me as the banks look into the history of the customer's balance. It might sound like a very mundane attribute but in my opinion, people with good financial planning are most likely to have a more stable saving as they are able to maintain their balance/possibly save consistently. It is definitely less risky to loan money to people who seem to be more financially responsible as they seem to have the ability to pay back their loans.

5)min..account.balance.6.months.ago. Similar to point 4, checking the history of the customer's savings could give valuable insights in seeing if they have what it takes to pay back their loans. We can clearly see the amount of remaining money they have in their accounts can actually pay back their loan amounts. Banks definitely need to see if the person is able to pay back their loans in certain circumcisions.



To further support my reasons I used SOM to gain a better understanding of the attributes. These attributes did seem to have a strong relationship to credit rating as seen from the clustering.



Even though i used 13 attributes in the SOM, i can clearly see that the above 5 most interesting attributes to me are the ones causing the clusters. Yays! Now i can perform MLP on these features!

I used one of the most popular library in r for classification models. Its popularity tells me that their models are working great and possibly more accurate compared to others. Along the way i

found out that it was relatively more simple to use caret as they have a wide variety of packages to play around. My strategy will be to use the features that i found to be more useful for training the MLP and then after the first prediction, I will fine tune my model and adjust the hyperparameters to increase my accuracy.

The first prediction results is as follows:

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3
         1  49  37  11
         2  47 141  57
         3   2  14  33

Overall Statistics

               Accuracy : 0.5703
                 95% CI : (0.5196, 0.62)
    No Information Rate : 0.491
    P-Value [Acc > NIR] : 0.001006
```

After fine tuning the results are as follows:

```
Multi-Layer Perceptron, with multiple layers

Pre-processing: centered (13), scaled (13), nearest neighbor imputation (13), principal component signal extraction (13)
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1415, 1413, 1414, 1413, 1413, 1414, ...
Resampling results:

  Accuracy   Kappa
  0.5632031  0.2841003
```

I expected the fine tuning to increase in accuracy but it actually lowered it by about 0.07%.

I had a few concerns regarding how the data was collected, the accuracy of the data, and the duration of the data collection . All these will pose inaccuracy for the training models as the collected data given is possibly also incorrect. No matter how I trained my models or fine tune it, the max accuracy will never pass 60%.

Despite this setback, we are still able to generally catch the trend and make a more informed and educated analysis of the specified issue. It might be that the data collected is not very helpful or that we might need a larger dataset or possibly need more attributes.

Overall, I do feel that maybe I could have chosen better variables. I could have increased the accuracy to 60+%. However, with such a low accuracy regardless, I think that I can argue that the datasets do have problems with it in the first place. The best way to approach this problem again is by having a more robust/revised data collection(done by the banks) and data preprocessing by us(I also realised we did not have to do preprocessing of data…). data collection and data preprocessing are undoubtedly extremely important for data scientist to gain better insight of the population:)