# 1. Introduction

# 2. Data Preparation

## 2.1 Dataset Observations

I analysed the dataset by briefly scanning through the data in Excel and then OpenRefine alongside the report specification, to gain an understanding of the data provided and the data cleaning required. I was looking to understand the underlying data, the positive and negative class split, and the possible changes required. I have included a few key observations:

* There are no samples of single females in the initial dataset. Data will need to be adapted to ensure that predictions can be made for single females. There are also no samples of credit for Vacations in the dataset, however this is less of an issue as it can be added to ‘Other’ or evaluated separately.
* The problem is an unbalanced classification problem – there is a 70:30 split between the Positive and Negative class. Data will need to be adapted to ensure that good predictions can be made with the Negative class.
* A lot of data required cleaned in the dataset. I have explained the cleaning below.

## 2.2 Dataset Cleaning

Normally in data science, will only lose data if necessary. (IBM?) Steps have taken in cases to write ML algorithms to try and update data rather than lose it

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Column** | **Change from** | **Change to** | **Change Count** | **Why** |
| N/A | No headers | Added headers | All columns | Makes data clearer – data requires headers |
| ID | Column present | Removed Column | 1000 | Redundant data |
| Job | yes | skilled | 2 | Assume that yes means the applicant replied “yes” to skilled |
| Job / Employment / Saving Status / Purpose/ Credit History /  Checking Status | Removed ‘’s | - | 998 / 938 /  1000 / 349 /  1000 /  1000 | Cleans up the data |
| Personal Status | ‘female div/dep/mar’ | female mar/wid/div/sep | 310 | Cleaning data and removing ‘dep’ spelling error |
| Personal Status | ‘male single’ | male single | 548 | Cleans up data |
| Personal Status | 'male mar/wid’ and 'male div/sep' | male mar/wid/div/sep | 92 mar/wid  50 div/sep  (142 total) | Merging checking data to resemble female group |
| Saving Status | no known savings | unknown | 183 | More specific / clearer |
| Credit Requested | 111328000 | 13280 | 1 | Looks like duplicated ‘1s’ and ‘0s’ |
| Credit Requested | 19280000 | 19280 | 1 | Looks like duplicated ‘0s’ |
| Credit Requested | 13580000 | 13580 | 1 | Looks like duplicated ‘1s’ and ‘0s’ |
| Credit Requested | 13860000 | 13860 | 1 | Looks like duplicated ‘1s’ and ‘0s’ |
| Credit Requested | 5180000 | 5180 | 1 | Looks like duplicated ‘1s’ and ‘0s’ |
| Credit Requested | 5850000 | 5850 | 1 | Looks like duplicated ‘1s’ and ‘0s’ |
| Credit Requested | 7190000 | 7190 | 1 | Looks like duplicated ‘1s’ and ‘0s’ |
| Credit Requested | 63610000 | 6361 | 1 | Looks like duplicated ‘0s’ (£63,610 doesn’t make sense) |
| Purpose | ather / busines / Radio/Tv / Eduction / busness / radio/Tv | other / business / radio/tv / education / business / radio/tv | 1 / 3 / 2 /  1 / 3 /  2 | Wrongly typed |
| Age | -29 / -34 / -35 | 29 / 34 / 35 | 3 | Updating negatives |
| Age | 0.44 / 0.24 / 0.35 | 44 / 24 / 35 | 3 | Updating error |
| Age | 6 / 222 / 1 / 333 | 60 / 22 / 25 / 33 | 4 | Updating error |

## 2.3 Data Merging

For the process of merging data, I firstly ran the Correlation Attribution Evaluation. This Attribute Selector evaluates each Feature in a dataset and ranks them based on the correlation between it and the class (for nominal data, each feature is considered separately). This allowed me to see which attributes correlated more highly with the class prediction being made.

I then tackled the issue of the unbalanced split between the Positive and Negative class. I decided to use SMOTE to rebalance the class distribution. SMOTE is an oversampling technique relying on k-NN to produce synthetic data samples. I decided that SMOTE was best on the dataset for nearest neighbours = 4, and although it didn’t give me as many True Negatives as I was hoping for (less than 3 or 5), it gave me a higher number of True Positives, and I could increase True Negatives later.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Column** | **Change from** | **Change to** | **Change Count** | **Why** |
| Purpose | Merging domestic appliance and furniture/equipment | furniture/equipment | 12 (new field – size 193) | Domestic appliance items belong in the furniture/equipment category |
| Credit History | Merging no credits/all paid and all paid | no credits/all paid | 49 (new field – size 89) | ‘All paid’ can be merged into the no credits/all paid category |
| Personal Status | female mar/wid/div/sep & male mar/wid/div/sep | mar/wid/div/sep | 452 | Merging all married/widowed/divorced and separated applications generally improved model accuracy |
| Personal Status | male single | single | 548 | Allows predictions to be made for single females as well as male singles |
| Employment | <1 & 1<=X<4 | <4 | 172 & 339 (new field – size 511) | Merging the employment periods generally improved model accuracy |
| Job | unemp/unskilled non res & unskilled resident | unskilled | 22 & 200 (new field – size 222) | Merging the job skill group generally improved model accuracy |

**TALK ABOUT NOMINAL DATA SPLIT**

# 3. Data Analytics

## 3.1 Introduction

One of the most important aspects of financial lending from the company perspective is not exposing yourself to undue risk. Money is made between the interest rate provided on accounts and the lending interest rate. Typically, APR is between about 2-5% for a large financial institution (currently RBS offers 3.4%, TSB offers 3.2%, and Lloyds offers 3.9%), meaning they make a small amount off each lend.

Lenders also have the potential to lose a lot of money if the borrower can’t repay, as the amount is usually not enough to chase the money through legal action or court orders (especially in the volume of lending in this example). The money lent to 10 people can easily be lost through a bad lend. This means that reducing the number of False Positives is more important than False Negatives, as it’s more important to reduce risky lending (I’m suddenly reminded of the 2008 financial crash).

I chose to use Classification, Association and Clustering algorithms for the Data Analytics. I could have used Regression, in generating the confidence of a model’s class prediction (which could be useful alongside a classification model) however Association generates the confidence of various rule groupings, and an underlying understanding of how features interact seemed very useful.

## 3.2 Classification

Classification is a supervised learning technique which looks to predict the class of a single feature, based on labelled data. Classification can only predict nominal data in a class. The German bank is an example of a Binary Classification problem (only two classes – a positive and a negative), as opposed to a Multi-class Classification problem (more than two classes).

On the default J48 settings, I was already achieving an accuracy of 77.2% **REF HERE**, with an F1 score of 0.772 (Figure. 2). The recall was only 73% on the Negative class however, and I wanted to decrease the False Positive rate for the reasons in the introduction.

A Cost Sensitive Classifier allowed me to increase the penalty weighting for False Positives, Bagging reduced the variance within the Classifier with random sampling of the J48 decision tree.

I used Bagging on the J48 Classifier, within the Cost Sensitive Classifier. This had a slightly higher accuracy (78.2%) **REF HERE**, with an F1 score of 0.782 and an ROC of 0.850 but increased the precision to 80% for the Negative class.

*The following rules have been generated from the decision tree generated:*

### 3.21 IF (Checking Status = no checking) & (Employment < 4) & (Job = skilled) & (Credit Requested < 1544) THEN (Class = good) – [42 & 0]

If the person currently has no current account with the bank, have worked for less than four years for their current employer in a skilled occupation and requested less than €1544, then lend. This rule is 100% accurate and applies to 42 current cases.

Lending to new customers is always a great way for banks to grow their portfolio. With such a high success rate, the bank can start looking to target lending towards similar skilled professionals.

### 3.22 IF (Checking Status = no checking) & (Employment >= 7) THEN (Class = good) – [116 & 11]

If the person currently has no current account with the bank but has worked for more than seven years, then lend. This rule is 91.3% accurate and applies to 127 current cases.

In conjunction with the first rule, a high success rate means that the bank can start targeting lending towards people who have worked under one employer for a long period of time.

### 3.23 IF (Checking Status < 0) & (Saving Status < 100) & (Job = skilled) & (Purpose = radio/tv) THEN (Class = bad) - [84 & 3]

If the customer has less than €0 in their current account and less than €100 in savings, they work in a skilled occupation and are using the money for a radio/tv then don’t lend. This rule is 96.5% accurate and applies to 87 current cases.

It suggests generally that if someone is requesting a loan for a luxury item and they have little money available, that lending is a bad idea.

### 3.24 IF (Checking Status < 0) & (Saving Status < 100) & (Job = unskilled) & (Personal Status = mar/wid/div/sep) & (Employment <4) & (Age <= 34) THEN (Class = bad) - [38 & 6]

If the customer has less than €0 in their current account and less than €100 in savings, they’ve been working in an unskilled role for less than four years, they’re 34 years old or younger and have been married then don’t lend. This rule is 86.3% accurate and applies to 44 current cases.

It suggests that young couples can be a risky lend, especially if they don’t have a financial buffer. Money is shown to be a leading cause of stress in relationships **REF HERE** and potentially a loan puts too much financial pressure on a young family with little money already.

### 3.25 IF (Checking Status = 0<=X<200) & (Credit Requested <= 11998) & (Saving Status >=500) THEN (Class = good) - [27 & 4]

If the customer has between €0 and €200 in their current account, more than or equal to €500 in their saving account and have requested less than €11,998 then lend. The rule is 87.1% accurate and applies to 31 current cases.

It suggests that current customers requesting a small-to-medium size loan are a good lend, as they have savings at the bank to act as collateral in case they can’t pay off the loan.

### 3.26 IF (Checking status = no checking) & (Employment = 4<=X<7) & (Age >= 22) THEN (Class = good) - [67 & 0]

If the person has no current account with the bank, has been employed for between 4 and 7 years and is 22 or older then lend. This rule is 100% accurate and applies to 67 current cases.

Lending to a younger customer can be a great way to incentivise young customers to join the bank. There also seems to be a sizeable number of customers without a checking account which seem a safe lend. This could convey information about additional checks that people without accounts go through or that if a person approaches a new bank with a lending proposition, they’re more serious about receiving lending.

## 3.3 Association

Checking Status=<0 Personal Status=single Job=skilled Class=bad 140 ==> Saving Status=<100 135 <conf:(0.96)> lift:(1.42) lev:(0.03) [39] conv:(7.48)

* If rejection high for bad single skilled, suggests that higher levels of savings than €100 very easily give an acceptance (given all information, must almost certainly have low savings, because higher savings makes them attractive)

Checking Status=no checking Credit History=critical/other existing credit 154 ==> Class=good 143 <conf:(0.93)> lift:(1.72) lev:(0.05) [60] conv:(5.92)

Checking Status=no checking Personal Status=single Job=skilled 156 ==> Class=good 139 <conf:(0.89)> lift:(1.65) lev:(0.04) [55] conv:(4)

Checking Status=<0 Saving Status=<100 Personal Status=single Job=skilled 160 ==> Class=bad 135 <conf:(0.84)> lift:(1.83) lev:(0.05) [61] conv:(3.31)

Checking Status=<0 Saving Status=<100 Employment=<4 Job=skilled 198 ==> Class=bad 167 <conf:(0.84)> lift:(1.83) lev:(0.06) [75] conv:(3.33)

Checking Status=<0 Credit History=existing paid Saving Status=<100 Job=skilled 193 ==> Class=bad 162 <conf:(0.84)> lift:(1.82) lev:(0.06) [72] conv:(3.25)

Checking Status=no checking Credit History=existing paid 195 ==> Class=good 164 <conf:(0.84)> lift:(1.56) lev:(0.05) [59] conv:(2.81)

These calculations can also be based on other factors e.g. applicant, loan volume etc.

Didn’t mix education with retraining due to high volume of “good” in retraining and middle volume of “good” in education (28 good/22 bad) vs (8 good/1 bad) (56% good vs 89% good)

Haven’t merged new and used car either

Used car (86/17| 103) = 83%

New car (145/89| 234) = 62%

**Merging the following**

Domestic appliance (8/4 | 12) = 66%

Furniture/equipment (123/58 | 181) = 68%

radio/tv (218/62 | 280) = 78%

repairs (14/8 | 22) = 64%

retraining (8/1 | 9) = 89%

education (28/22 | 50) = 56%

Stage 1 data: original

Stage 2 data: after first 2 changes (headers and job ‘’s)

Stage 3 data: after updates made to Credit Requested

Stage 4 data: after updates made to rest of table

Stage 5 data: ages updated

Talk about how I merged loads of fields and didn’t keep data backups, which started leading to sub-par training and a real difficulty of backing up data – meaning I had to redo

**Nominal:** Set lending criteria = 18-28, etc. because earliest age that lending can legally occur is 18

Try Increments of 5

Try 20-30 etc too

Mention that gradient boosting is super popular (AVAILABLE?)

Legally, when declining an application for credit, you are required to give a reason why the application was declined. Hence why NNs aren’t used. Meant I had to use a model which could show why the application was declined.

If a user defaults on a loan, it can cost the bank a lot of money. I placed more importance on trying to ensure enough applications were being declined. In some cases, there are more applications being accepted (True positives), but there was a large drop in declines. **Explain how bank makes money on loan**

Also talk about being competitive etc and opening themselves to risk (TN more important than TP

Use large sets for evaluation of rules to ensure max number of customers included

**ROC SUMMARY (**[**http://gim.unmc.edu/dxtests/roc3.htm**](http://gim.unmc.edu/dxtests/roc3.htm)**):**

* .90-1 = excellent (A)
* .80-.90 = good (B)
* .70-.80 = fair (C)
* .60-.70 = poor (D)
* .50-.60 = fail (F)

**Class imbalance:**

<https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/>

<https://content.pivotal.io/blog/how-to-deal-with-class-imbalance-and-machine-learning-on-big-data>

<https://www.einfochips.com/blog/addressing-challenges-associated-with-imbalanced-datasets-in-machine-learning/>

<https://www.researchgate.net/post/What_should_be_the_proportion_of_positive_and_negative_examples_to_make_a_training_set_result_in_an_unskewed_classifier>

<https://www.quora.com/I-have-an-imbalanced-dataset-with-two-classes-Would-it-be-considered-OK-if-I-oversample-the-minority-class-and-also-change-the-costs-of-misclassification-on-the-training-set-to-create-the-model/answer/Shehroz-Khan-2>

SMOTE:

<https://www.dataminingapps.com/2016/11/what-is-smote-in-an-imbalanced-class-setting-e-g-fraud-detection/>

Money is leading cause of stress in a relationship:

<https://www.cnbc.com/2015/02/04/money-is-the-leading-cause-of-stress-in-relationships.html>

**Cluster – Initialised through Canopy as the standard initialisation left 4 radio, 1 furniture and 1 car**

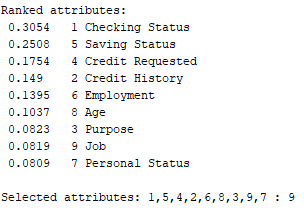
Can be validated through Internal & External indices. Can’t validate through external indices

Clusters:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 4** | **Cluster 5** | **Cluster 6** |
| **Checking Status** | no checking | 0<=X<200 | no checking | 0<=X<200 | <0 | no checking |
| **Credit History** | existing paid | existing paid | existing paid | existing paid | no credits/ all paid | existing paid |
| **Purpose** | furniture/equipment | radio/tv | radio/tv | new car | business | new car |
| **Credit Requested** | 2736 | 3105 | 2998 | 5318 | 5828 | 2982 |
| **Saving Status** | <100 | <100 | <100 | <100 | <100 | <100 |
| **Employment** | 0<X<4 | 0<X<4 | >=7 | >=7 | 0<X<4 | 0<X<4 |
| **Personal Status** | div/sep/mar/wid | div/sep/mar/wid | single | div/sep/mar/wid | single | single |
| **Age** | 31 | 29 | 42 | 42 | 36 | 37 |
| **Job** | skilled | skilled | skilled | skilled | skilled | unskilled |
| **Class** | good | bad | good | bad | bad | good |

# References

## Correlation Attribute Evaluation + Ranker Figure 1



## Standard J48 Test on the k-NN = 4 SMOTE Figure 2

## 

## Cost Sensitive Classified J48:

