UNIVERSITY OF STIRLING

CSCU9T6 - Data Mining Assignment 2019

Student ID: 2520796

Table of Contents

1.0 Introduction

The project refers to the banks that are having some trouble with debt today and they also like to avoid lending money to people who are unlikely to repay their loans in the future. We are about to use some data from a bank that describes 2000 of their previous loan customers and with appropriate data mining techniques and predictions we will be able to tell how likely it is that a new customer would pay back a loan. Predictions will be determined by the outcomes and the given attributes from the dataset for everyone.

There are 2000 customers (instances) with 15 attributes each and a unique 6-digit ID for each customer.

1.1 Terminology

- **Mean:** refers to the average. The sum of the values divided by the number of values.
- **Standard Deviation:** refers to the measure that is used to quantify the amount of variation of a set of data values.
- **Outliers:** refers to a value that considerable differs from the norm
- **Instances:** refers to an individual set of values. A person/customer/individual has 15 attributes and a unique ID (Customer ID) to identify them.
- **Attribute:** refers to variable/field
- **Record(s):** refers to the variables from each row of the data
- **Noise:** refers to a value it is different from the original value or it is not part of the data and should be removed.

1.2 Attributes

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Customer ID	Forename	Surname	Age	Gender	Years at address	Employment status	Country	Current debt	Postcode	Income	Own home	CCJs	Loan amount	Outcome

1. Customer ID
2. Forename
3. Surname
4. Age
5. Gender
6. Years at Address
7. Employment Status
8. Country
9. Current debt
10. Postcode
11. Income
12. Own home
13. CCJs
14. Loan amount
15. Outcome

2.0 Data Summary

2.1 Customer ID

- **Datatype:** Numeric

After some investigation to the database it was found that the Customer ID is a unique value to identify a customer. The datatype of the Customer ID attribute is Numeric, and the values are Discrete. However, 14 values are either duplicated or incorrectly entered into the system and that leads to 99% of unique rate (1986 columns).

CUSTOMER ID	Unique	Distinct	Missing
Value	1986	1993	0

According to the above table, 7 customers use the same Customer ID twice each (after deducting the Distinct value from the Unique). These customers use different named but share the same ID number which that leads to a mistaken input. Customers should be unique otherwise it will not reference the right solution of repaying a loan problem.

CUSTOMER ID	Forename	Surname	Age	Gender
707817	Marcell	Avery	75	M
707817	Toby	Holloway	18	M

2.2 Forename/Surname

- **Datatype:** Nominal

Forename, surname are not very useful attributes for the problem of the customers repaying future loans. However, it may contain invalid data '.', 'A', '@', 'Arber/ Diggins', or mismatched gender attribute as Zoe Bagley has be entered as Male. Some of this data can be removed without producing a serious problem due to the size of the dataset. These attributes are not providing and real usable statistical information, but it is important to avoid these kinds of errors.

CUSTOMER ID	Forename	Surname	Gender
970114	**.**	Azimiaa	M
778207	**A**	Gibbons	F
973657	Zoe	Bagley	**M**
1040020	David	**@**	M
596827	Nick	**Arber/ Diggins**	**F**

2.3 Age

- **Datatype:** Numeric

Age is an attribute that has only 2 mistakes from the whole dataset. However due to the size of the dataset there is not causing a serious change to accuracy of the statistical information. There are total of 73 distinct values which from those the age values range is 17 to 89 ascending.

CUSTOMER ID	Forename	Surname	Age
748920	Victoria	Gogin	**5**
808540	John	Tasker	**174**

- **Mean:** 52.900 (without the above errors), average age 53 rounded
- **Standard Deviation:** 21.996 (without the above errors)

The StDev of 21.996 presenting a decent spread out data that is not grouped, it is unpredictable and close to the value of mean.

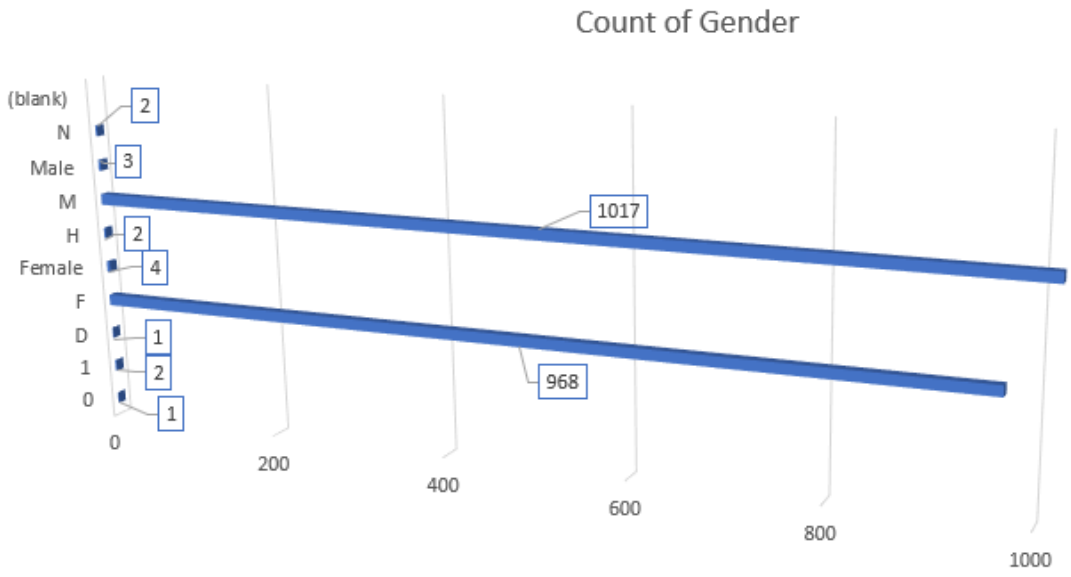
The youth age group (17 – 21) can be used otherwise a different subset can be used to give the most reliable data, i.e. deviations of 10 year age intervals.

The Age attribute combined with an external factor data attribute such as 'Dept Amount' could result very valuable data and probability statistics that could be obtained to present the probability of customers repaying future loans.

2.4 Gender

- **Datatype**: Nominal

In gender attribute after post-analysis, the dataset shows a total of 9 values that represent male and female. Below the graph shows that most values are represent as ‘M’ for males and ‘F’ for females. The other values ‘N’,’H’,’D’,’1’,’0’ are unknown and will be defines as ‘noise’. It is easy to identify the gender by looking at the forename attribute and the gender attribute can be changed accordingly. The remaining attribute variations ‘Male’ and ‘Female’ could be simply replaced by ‘M’ and ‘F’ accordingly. However, under some investigation, the forenames under ‘Female’ gender are male names (David, Stuart, Dan, Simon), therefore they have been changed to ‘M’.



The percentage of M and F values representing the gender of Males and Females are:

- **F:** 48.45% (969)
- **M:** 51.55% (1031)

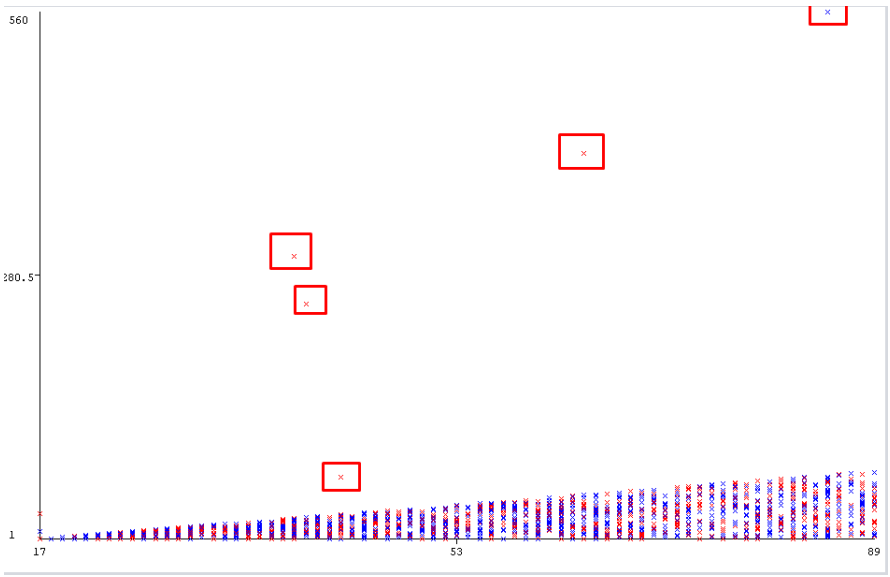
The differentiation of M and F is 3.1%. All things considered, the gender attribute can provide precise statistics to predict the probability of repayment of future loans. However, the usefulness of the gender attribute is very low as the discrimination based on a gender is inappropriate. As mentioned above, the data can be imported inaccurately such as a lot of values(names) in Forename attribute are males but instead they represented as females in the Gender attribute. Therefore, the usefulness of the gender is minimal it is not worthwhile to clean or correct the data.

2.5 Years at Address

- **Datatype:** Numeric

Most of the values for the Years at Address attribute have been entered correct. After scanning the dataset 5 mistakes/anomalies have been detected. The five values presented at the table below, shows some unrealistic period of years that a customer have been living at the address stated.

CUSTOMER ID	Forename	Surname	Age	Years at Address
1039485	Philip	Nurse	43	**66**
722046	Chris	Greenbank	40	**250**
678376	Steve	Hughes	39	**300**
967482	Brian	Humpreys	64	**410**
1075394	Simon	Wallace	85	**560**



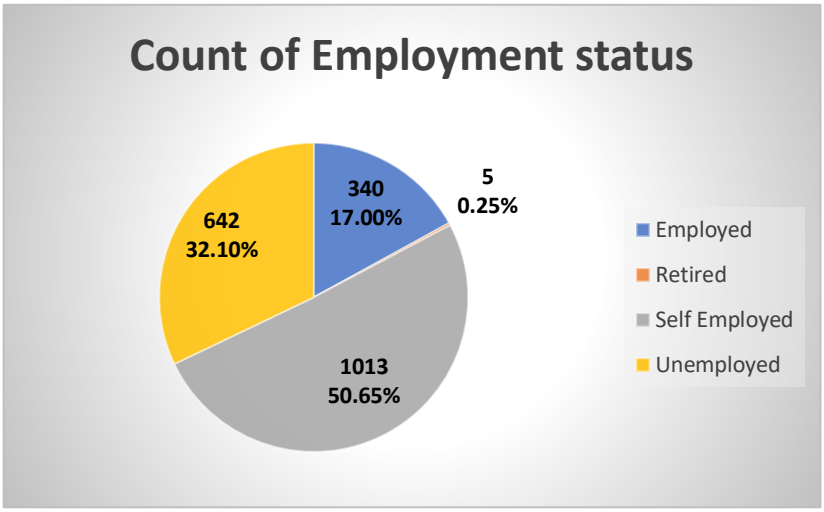
The four customers listed with the unrealistic value of Years at Address could be removed from the dataset as the remaining customers (1996) would provide reliable data. For this attribute, 10 year interval subset could aid a solution to the problem.

Mean – Unedited Dataset:	18.545	**Mean – Edited Dataset:**	17.821
StDev – Unedited Dataset:	23.225	**StDev – Edited Dataset:**	15.800
Min Value – Unedited Dataset:	1	**Min Value – Edited Dataset:**	1
Max Value – Unedited Dataset:	560	**Max Value – Edited Dataset:**	71

2.6 Employment Status

- **Datatype:** Nominal

The greatest value from the Employment Status attribute is the Self-Employed. 1013(50.65%) people out of 2000 claimed to be self-employed which is somehow questionable if this is true, but at the same time there is not a possible way to disprove this figure. The default retirement age is 65, however in this dataset all 5 people that indicates 'Retired' none of them are over the age of 65. These 5 customers will be considered as outliers and should remain in the dataset.



2.7 Country

- **Datatype:** Nominal

There are 4 distinct values in the Country attribute: UK, France, Spain, Germany. As the histogram below shows that the most values are UK (1994) and the remaining 6 values are for the rest of the countries (2 per country). Unfortunately, it is impossible to retrieve a reliable result for the other countries other than UK due to the very small dataset size.

2.8 Current Debt

- **Datatype:** Numeric

This attribute contains 788 distinct values with 294 unique. The higher the amount of dept, the toughest it becomes for a person to repay current loans or loan again. An unproven theory can be hypothesized that a person with low income and high debt will not repay future loans.

Current Debt attribute can be useful for good forecast of the probability that a person will repay future loans, as it provides statistical data from past debts.

- **Minimum Value (Debt):** 0
- **Maximum Value (Debt):** 9980
- **Mean:** 3309.325
- **Standard Deviation:** 2980.629

2.9 Postcode

- **Datatype:** Nominal

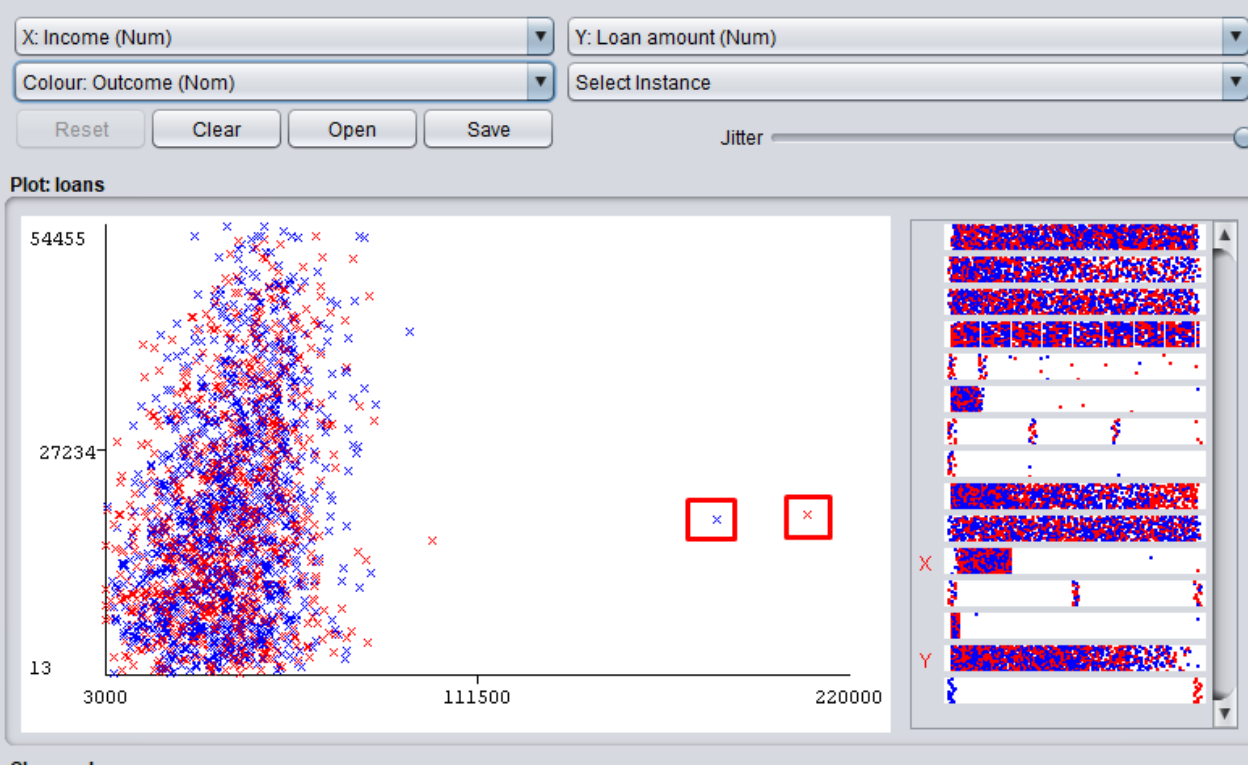
Postcode attribute contains 19171 values and has 0% missing values. Therefore, multiple customers can live in the same postcode and likewise to gender attribute, discrimination based on location is not permitted. This attribute cannot be used to solve the prediction of load repayment but can be used in other outside problems.

2.10 Income

- **Datatype:** Numeric

The Income attribute contains 100 distinct and 7 unique values (0%). The amount of income a person can have is not finite. It can also be hypothesized that the higher the persons income, the less the current debt they have. The scatter graph below represents loan amount against income. The 2 outlier values are shown with a red square with highest incomes and average loan amounts, however one has paid and the other has defaulted.

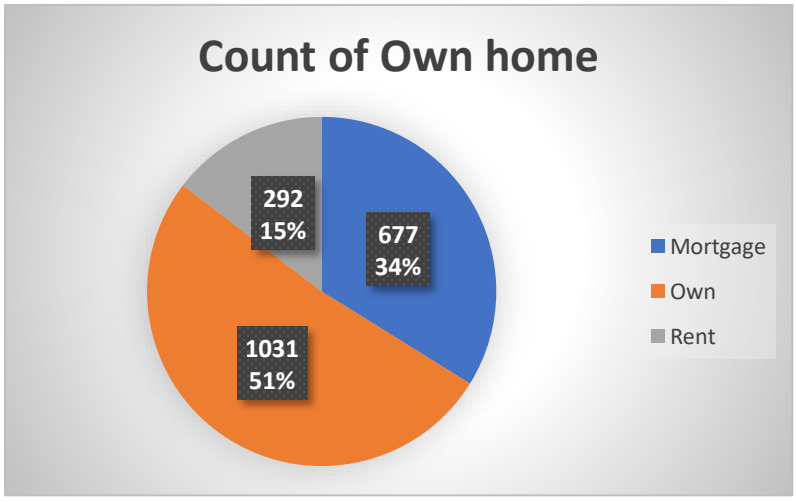
- **Minimum Value (Income):** 3,000
- **Maximum Value (Income):** 220,000
- **Mean:** 38,319
- **StDeviation:** 12,786.506



2.11 Own Home

- **Datatype:** Nominal

The Own Home attribute contains three distinct values and can provide valuable data for future loan repayments. A hypothesis can be stated when people with high income and they are home owners are probable to pay back a loan. There is no noise as there can be only 3 possible values for home ownership and there are 0% missing values in the current dataset.



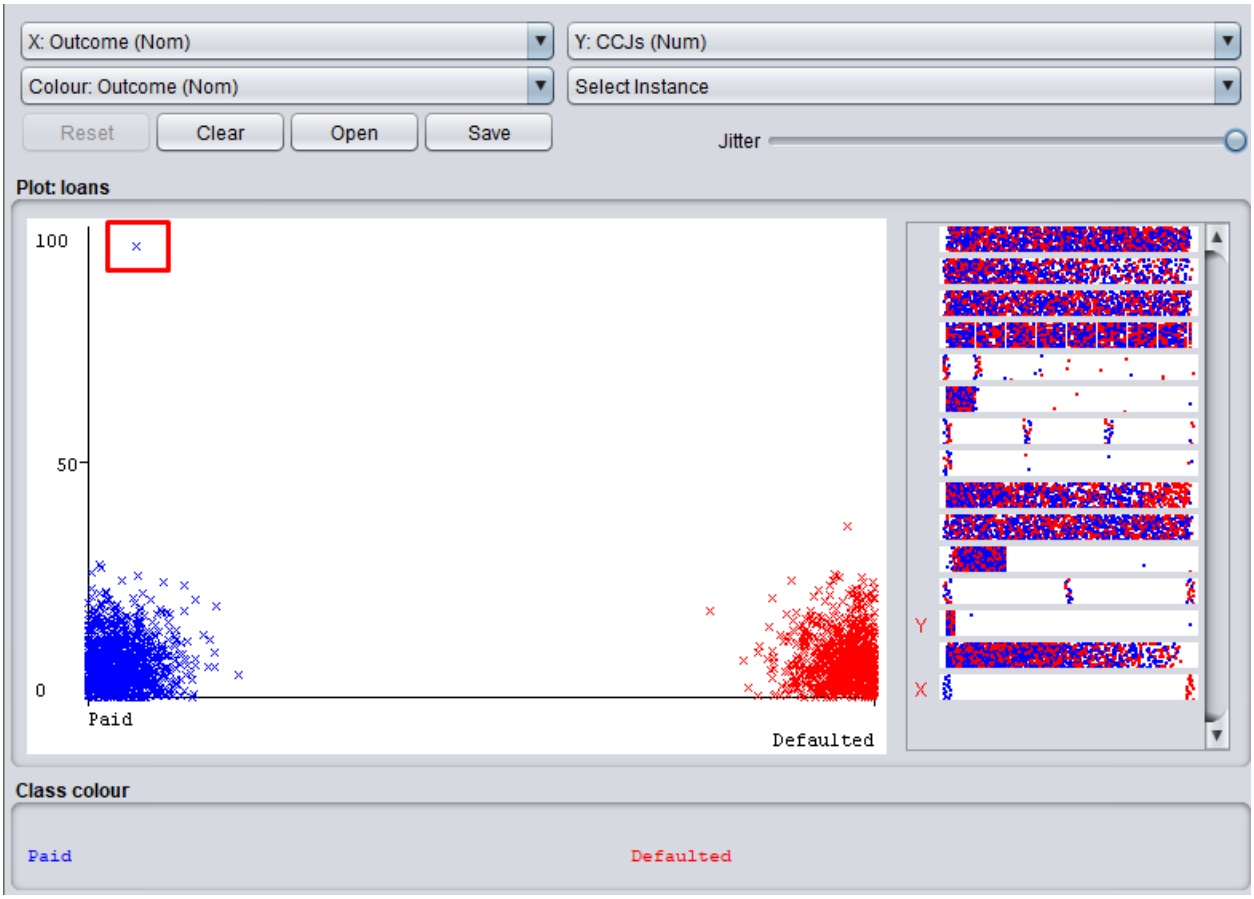
2.12 CCJs

- **Datatype:** Numeric

The CCJs (Country Court Judgement) attribute for this dataset has 6 distinct values, ranging from 0 to 100. If a CCJ issued to a person that means that the person is untrustworthy to repay the money, they owned. The greater the number the higher the risk for loan repayments, making it difficult for a person to secure loans. This is a vital factor within this dataset and could provide a result for the probability of future loans being repaid by customers.

Distinct Values: 0, 1, 2, 3, 10, 100

The record below shows a customer with age 27 and 100 CCJs. This is most likely an error data because it is nearly impossible that a customer who had 100 CCJs paid, therefore this record could be altered to 1, 10, delete it or address it as noise.



CCJs	0	1	2	3	10	100
Total	886	497	347	268	1	1
%	44.3%	24.85%	17.35%	13.4%	0.05%	0.05%

2.13 Loan Amount

- **Datatype:** Numeric

Loan Amount attribute can change continuously as there is no finite amount of loans a person can have.

- **Minimum Value (Loan amount):** 13
- **Maximum Value (Loan amount):** 54455
- **Mean:** 18929.628
- **Standard Deviation:** 12853.189

Above is shows that the lowest amount in the dataset is 13 and can be viewed as an input error (noise) as a bank unlikely would loan small amounts of £13. The real amount could be £1300 or £13000. The same followed for similar values (35, 45, 50…) that are considered to be noise.

2.14 Outcome

- **Datatype:** Nominal

The Outcome attribute contains distinct values of either ‘Paid’ or ‘Defaulted’. These values are based on whether a person paid or defaulted on the loan they had considering their income, current debt and the loan amount.

Below is the table that shows the amount of people who paid and defaulted in both unedited and edited dataset.

Outcome	Paid	Defaulted
Unedited	1118	882
Edited	1115	879

3.0 Data Preparation

The attributes that will be used are:

- Age
- Years at Address
- Employment Status
- Current Debt
- Income
- Own Home
- CCJs
- Loan Amount
- Outcome

The attributes that will not be included in data preparation are:

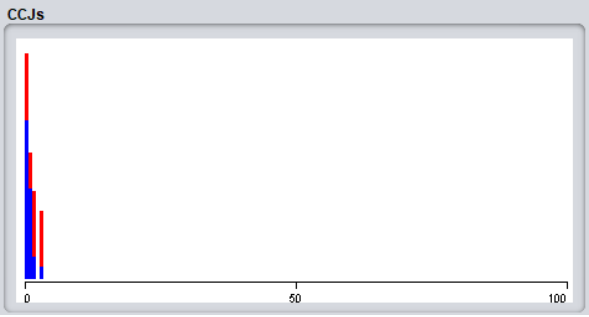
- Customer ID
- Forename
- Surname
- Gender
- Country
- Postcode

3.1 Noise

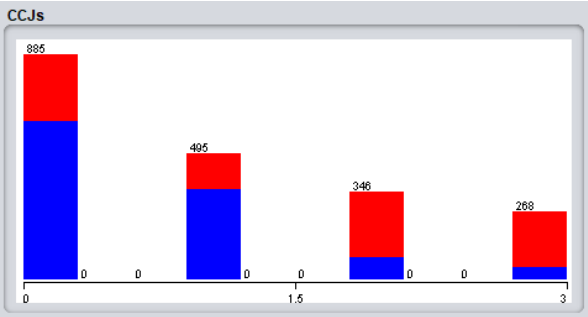
6 Records have been deleted from the dataset. The new total amount of records is 1994.

3.1.1 Years at Address and CCJs

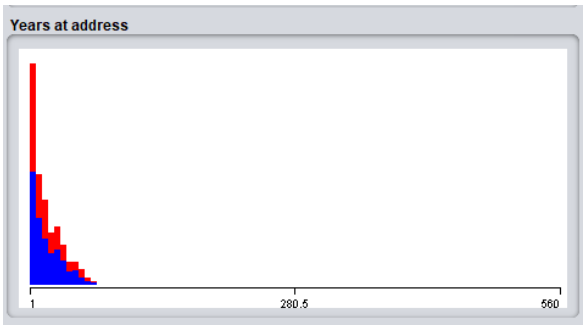
Due to 2 assumed incorrect attributes, 4 records have been removed from Years at Address and CCJs. For instance, 560 for Years at Address and 100 for CCJs. It is easier to label these records as noise and remove them as the big size of the dataset will catch up for very few records being removed. The below histograms represent the unedited and edited dataset before and after the removal of noise.



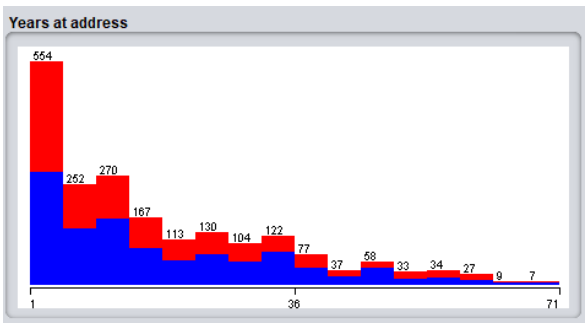
Before



After



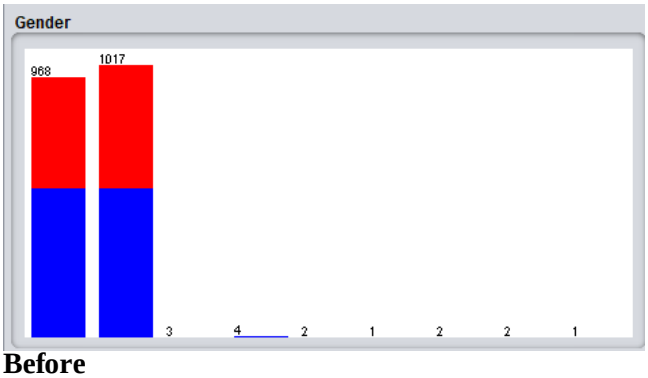
Before



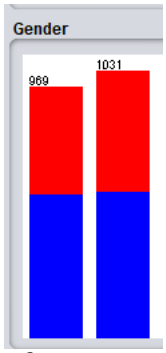
After

3.1.2 Gender

The histograms below represent the unedited and edited dataset before and after the removal of noise. The removed values have been labeled as noise. The 1 and 0 has been categorized as M and F based on the Forename of customers. Gender suggests worthlessness as many of the records are incorrect such as Ben is classed as 'F' (female).



Before



After

4.0 Data Mining Algorithm Selection

4.1 Naïve Bayes

Naïve Bayes can be used to understand how the probability that a theory is true, is affected by another piece of evidence. This algorithm, known as the Bayes theorem of probability, is given the name after Reverend Thomas Bayes in 1702-1761 to forecast a class of an unfamiliar dataset.

Naïve Bayes model has been applied in a real-world scenario i.e. the prediction of late aggression of Alzheimer's disease in 1411 individuals from genome-wide data.

The purpose of this technique is to work out whether a new example is in a class given that it has a certain combination of attribute values.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where A and B are events and $P(B) \neq 0$

- P (A | B) is a conditional probability: the likelihood of event A occurring given that B is true.
- P (B | A) is also a conditional probability: the likelihood of event B occurring given that A is true.
- P(A) and P(B) are the probabilities of observing A and B independently of each other. This is known as the marginal probability.

4.1.1 Pros and Cons of Naïve Bayes

Pros	Cons
Computationally fast	Relies on independence assumption and will perform badly if this assumption is not met
Simple to implement	
Works well with high dimensions	

4.2 Decision Trees

A decision tree is represented by a leaf and node structure, each branch represents the outcome of the test, and each leaf node represents a class label. A common method of producing decision trees is to use TDIDT (Top Down Induction of Decision Trees).

4.2.1 First Phase

The TDIDT method has 2 phases, the first is to build or 'grow' the decision tree. This can be accomplished by determining the 'best' attribute in the dataset, which will convert to the root node of the decision tree. If the nodes are pure will be used and if this option is not available, then the next most pure will be used. Pure nodes do not need to be split further as all fragments of the node have the same class.

4.2.2 Second Phase (The Pruning Phase)

The second phase can be achieved using a bottom-up method. Any data that is allowed to be noise, should be pruned, this will create a leaf by removing the node form the decision tree. Moreover, pruning minimizes the classification error of noisy data.

4.2.3 C.5

Decision trees are created from the C4.5 algorithm and they used for classification problems from training sets, in addition to missing values they can deal with both continuous and discrete attributes. This method it works effectively with noisy data from the dataset. This algorithm can be reinforcing post-competition using the Pruning method to 'train the data' so it can provide changeable outlooks on the dataset. A training set is an important section of a dataset used to evaluate data mining models. Most of the known valued of a dataset are used to discover predictive relationships for future unknown data.

4.2.4 Pros and Cons of Decision Trees

Pros	Cons
Easy to interpret visually when the trees only contain several levels	Prone to overfitting
Can easily handle qualitative features	Possible issued with diagonal decision boundaries
Works well with decision boundaries parallel to the feature axis	

5.0 Modelling

5.1 Experiment

Goal: to investigate whether 50:50 or cross-validation gives the best classification accuracy when using Naïve Bayes and J48.

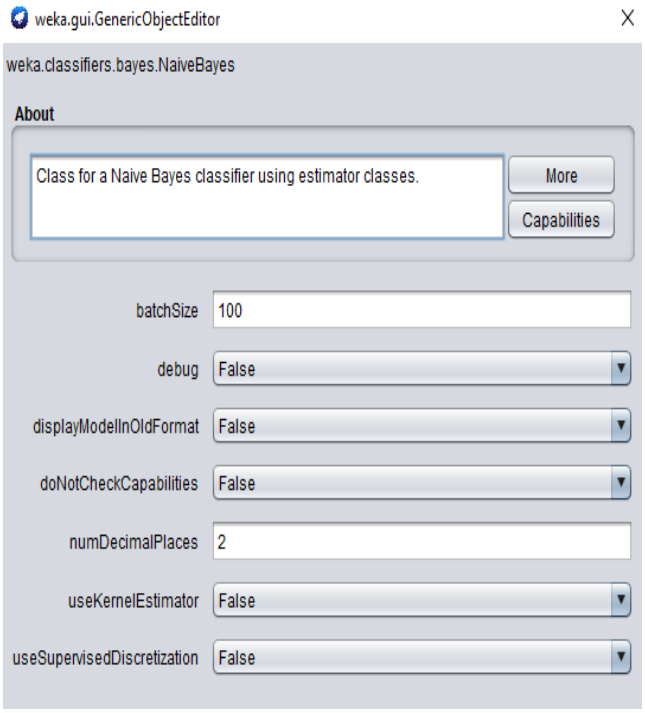
WEKA's C4.5 decision tree algorithm is called J48, which will be tested in addition to Naïve Bayes. The first initial experiment will determine whether a 50:50 split or cross validation training method will produce the highest percentage of correctly classified instances.

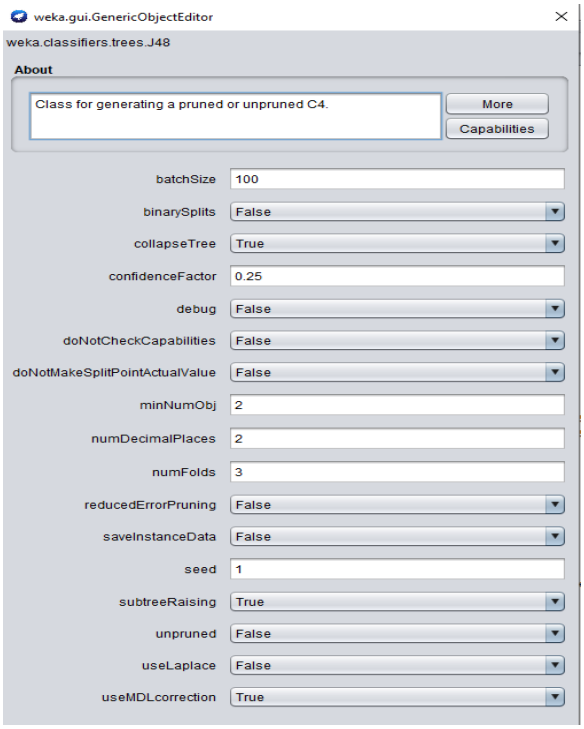
5.2 50:50 Percentage Split

A 50:50 split seems to be the most reliable as the entire dataset is evenly split, without bias. A 90:10 split produces a slightly higher classification accuracy. However, the sample size is too small to form a valid opinion based on reliable data.

5.3 Cross-Validation

This testing method splits the dataset in to 10 equal sections, the algorithm learns from the first section or fold, and applies the 'rules' it learnt to the other remaining folds. The results from the 2 testing options for each methodology will be compared for both J48 and Naive Bayes. The screenshots below indicate the parameter defaults.

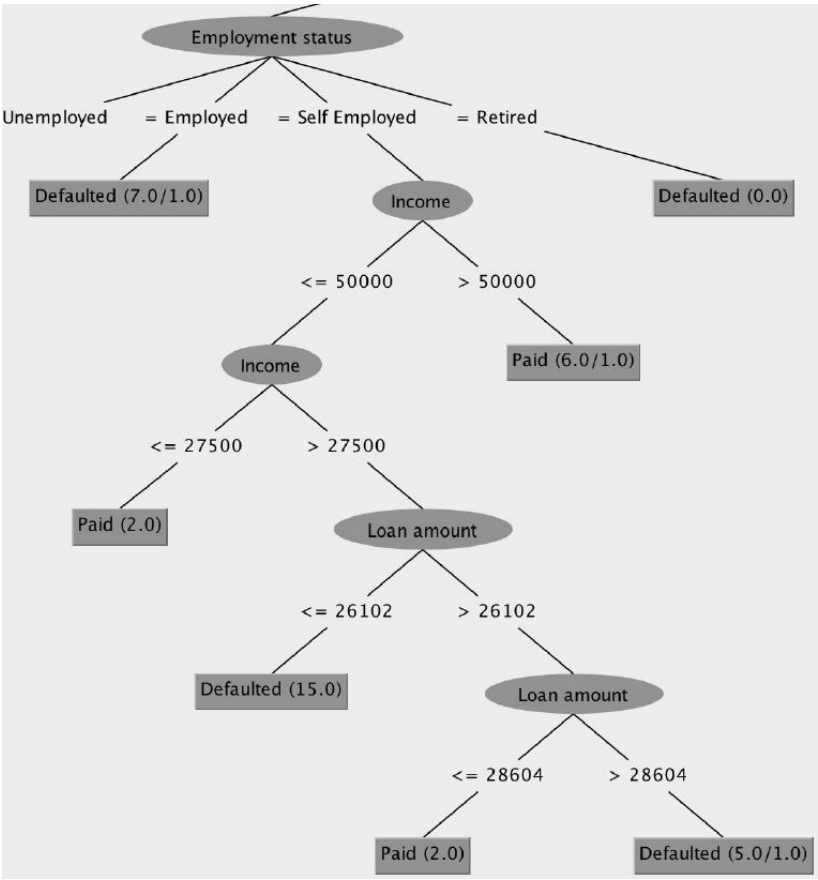




The table below indicated the training and testing methods used on the J48 and Naïve Bayes, and their classification accuracy. The percentage of records for the paid and defaulted attributes are also shown.

Model	Training Testing	Paid (%)	Default (%)	Classification Accuracy (%)
J48	50:50	88.13	70.04	80.44
J48	Cross-Validation	84.39	72.01	78.93
NB	50:50	90.99	64.85	76.62
NB	CV	83.58	63.48	74.72

5.4 Decision Tree



The decision tree shown related to the J48 algorithm, using the percentage split training method.

5.5 Confusion Matrix

The confusion Matrices below are labelled accordingly to the data mining techniques chosen, and training methods used. The confusion Matrices (right) generate the following percentages of correctly classified paid and defaulted records (indicated by P and D respectively). On average J48 (Percentage Split) has the highest average accuracy of correctly classified instances, of paid and defaulted records, with a total figure of 79%. These figures are reliable as it is extremely close to the classification accuracy of the model. In conclusion the J48 classifier with default parameters provides the most reliable overall results.

J48	**PS**	D = 70%	P = 88%
J48	**CV**	D = 72%	P = 84%
NB	**PS**	D = 64%	P = 85%
NB	**CV**	D = 63%	P = 83%

J48 (Percentage Split)

```
=== Confusion Matrix ===

   a   b   <-- classified as
 297 127 |   a = Defaulted
  68 505 |   b = Paid
```

Naive Bayes (Percentage Split)

```
=== Confusion Matrix ===

   a   b   <-- classified as
 275 149 |   a = Defaulted
  84 489 |   b = Paid
```

J48 (Cross-validation)

```
=== Confusion Matrix ===

   a   b   <-- classified as
 633 246 |   a = Defaulted
 174 941 |   b = Paid
```

Naive Bayes (Cross-validation)

```
=== Confusion Matrix ===

   a   b   <-- classified as
 558 321 |   a = Defaulted
 183 932 |   b = Paid
```