SET09120 Coursework 2

40205163

1. Introduction

This coursework was designed to test our ability to explore and understand a dataset, our ability to clean the data and then draw useful conclusions based on what we discovered. I was confident in my ability to complete a high-quality report as I have prior experience with ML algorithms and extracting information from datasets.

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

Generally, when cleaning data you only remove samples from the data if necessary. Especially since the dataset is relatively small, I've tried to ensure that all data could be kept, and inconsistent or incorrect values were updated instead of removed.

Below are the initial changes I made to clean the dataset:

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 (Figure. 1). 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 ran the data through a Feedforward Perceptron with a single hidden layer and checked the weightings of the Network. This was to try and get a feel for the weight of each attribute (positive and negative), to understand which attributes could be merged, which attributes generally correlated with the positive class and which attributes correlated with the negative class.

I then tackled the issue of the unbalanced split between the Positive and Negative class. I decided to use SMOTE as a Pre-processing method, 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.

Below are the secondary changes I made to clean the dataset:

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

2.4 Data Conversion

After I was happy with the Numerical dataset that had been created, I created a Nominal dataset which could be used for Apriori, the Association algorithm. I didn't want to discard the data in the 'Credit Amount' and 'Age' columns, so I updated the columns based on value ranges.
I updated the range of 'Credit Amount' to 1000, 2000 and 2500 intervals and tested each on a standard J48 tree, finding the results were best for the 2000 range.
I also updated the range of 'Age' to (10-19, 20-29 etc) and (18-27, 28-37 etc) and tested both on a standard J48 tree. I found that the secondary schema generated better results.

Below are the changes I made to create the Nominal dataset:

Column	Change from	Change to	Change Count	Why
Credit Amount	Numerical data	Credit<2000, 2000<=Credit<4000…., 18000<=Credit<20000	1300	1000 split left holes in the dataset and produced too many groupings 2500 split left a good number of groups, but most of the data was in a small range (79% of the data is <5000, while only 72% of the data < 4000).
Age	Numerical data	18<=X<28, 28<=X<38…., 68<=X<78	1300	You can't get a loan if you're under 18. I therefore chose to start my age split at 18

I have included the code which I used to 'transform' the columns alongside the report.

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 feature associations, and an underlying understanding of feature relationships seemed very useful.

3.2 Classification

Classification is a supervised machine learning technique which looks to predict the class of a single feature, given 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). For Classification, I used my numerical dataset.

On the default J48 settings, I was already achieving an accuracy of 77.2% (Figure. 2), 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. I used a weighting of 1.25 for the False Positive penalty, with the standard weighting of 1 for the False Negative and increased the J48 confidence factor to 0.3. This had a slightly higher accuracy (78.2%) (Figure. 3), 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 taken 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. 1) 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

Association is an unsupervised machine learning technique which looks to uncover associations and relationships between seemingly unrelated data. A common use of Association methods is to try and understand the buying patterns of customers.

For generating the relationships, I used the Apriori algorithm, where I designated the minimum confidence as 0.8 and the rules to be generated as 100, as I wanted to find a greater variety of rules. I used my Nominal dataset for the Association, as Apriori requires a Nominal dataset.

Apriori takes a 'bottom-up' approach to generate rules, where it iteratively tests associations against each other until no more rules can be derived from the underlying data.

The following rules have been taken from the Apriori algorithm:

3.31 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 there's a negative current balance on the current account, they're single in a skilled occupation and have been rejected for a loan, it's highly likely that they have little money in their savings account.
With such high confidence and conviction, it suggests that generally people who are single and skilled are generally a good lend provided they have savings (since if they're in the Negative class and a savings number can be generated with such confidence, it implies that almost every single skilled loan decline must have low balance on their saving and current accounts).

3.32 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)

If you have no current account with the bank and have existing or critical credit, you're considered a good lend.
It suggests the experience of previously having taken a loan combined with approaching the bank as a new customer is a good sign for lending. This is along similar lines to the credit scoring: to get a high credit score you need to have taken on debt previously and proved you can pay it off.

3.33 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)

If you have no current account with the bank, are single and work in a skilled job, you're considered a good lend.
Like rule 3.31, non-customers who are single, skilled employees are generally a good lend, unless they have no money in the account (suggests a high cash outflow).

3.34 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)

If there's a negative current balance on the current account, they have less than €100 in savings and are single working in a skilled occupation then it's considered a bad lend.
The most confident indication of a bad lend from the Apriori rules is being single, having a skilled job and having little cash available. The skilled job suggests a higher level of income, being single suggests a younger lend and the little money in the account suggests bad spending habits, indicating a possible inability to repay the loan.

3.35 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)

If there's a negative current balance on the current account, they have less than €100 in savings, have worked less than four years for their current employer in a skilled job then they're considered a bad lend.
Alongside a status of 'single', a relatively short current employment term also indicates a bad lend.

3.36 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)

If you have no current account with the bank and have paid off existing credit, then you're considered a good lend.
Like rule 3.33, approaching the bank with a new lending proposition with an existing history of paying off debt suggests a good lend.

3.4 Clustering

Clustering is an unsupervised machine learning technique which looks to uncover 'clusters' in the dataset. Within K-Means Clustering, k clusters are placed and then data points are associated with the closest cluster. This is iterated until a suitable outcome is found. I used Simple K-Means Clustering on my Numerical dataset, with a Canopy initialisation and k = 6. I got an SSE of 3206.27.

The following clusters have been generated by the K-Means Clustering algorithm:

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**
Checking Status	0<=X<200	<0	no checking	<0	<0	no checking
Credit History	existing paid	existing paid	critical/other existing credit	existing paid	no credits/ all paid	existing paid
Purpose	radio/tv	furniture/equipment	radio/tv	new car	new car	new car
Credit Requested	2997	2604	3397	2319	6020	2987
Saving Status	<100	<100	unknown	<100	<100	<100
Employment	<4	<4	>=7	>=7	<4	<4
Personal Status	div/sep/mar/wid	div/sep/mar/wid	single	div/sep/mar/wid	single	single
Age	31	29	42	38	37	37
Job	skilled	skilled	skilled	skilled	skilled	unskilled
Class	bad	good	good	bad	bad	good
% of total dataset	21% (268)	18% (233)	17% (225)	10% (125)	19% (248)	15% (201)

4 Conclusion

Of the algorithms used, I found that Classification was the most useful, followed by Association and then Clustering. A large part of this is to do with the Problem Domain – helping a bank further understand their target market and assist them with the creation of rules to help a bank lend safely.

- Classification created a decision tree, allowing us to generate loan decision rules from the existing lending decision labelled data
- Association created various rules allowing us to understand the relationship between difference customer attributes and how strongly they affected lending decisions
- Clustering created six separate clusters, allowing us to further explore the individual customer groupings within the lending data

I spent a lot more time on the Classification technique than the other two as it had a plethora of Machine Learning algorithms and customisation options.

In the dataset, I found there was too much variance in the 'Purpose' column, however I couldn't merge the data any further without decreasing the performance of the Classifier.
Ideally in a dataset, there is more numerical data and less nominal data, allowing Regression to be used in a more effective manner. It also allows Feature Extraction from the data provided.

I found Classification useful, however I struggled to get a high accuracy without overfitting (e.g. I got an 87% accuracy on a tree with around 900 leaves). Ideally, more data would improve the accuracy of the prediction. In addition, the imbalance of data in the individual classes made it so that the training model underfit to the smaller classes (e.g. 358 radio/tv and only 12 other).

I also found Association useful, however it kept giving very similar suggestions for rules. The highest confidence rules were mostly a variation of each other. This made it hard to draw different conclusions and I had to decrease the lower confidence bound to allow additional rules to start appearing. Association however did give more of an insight into why combinations came up.

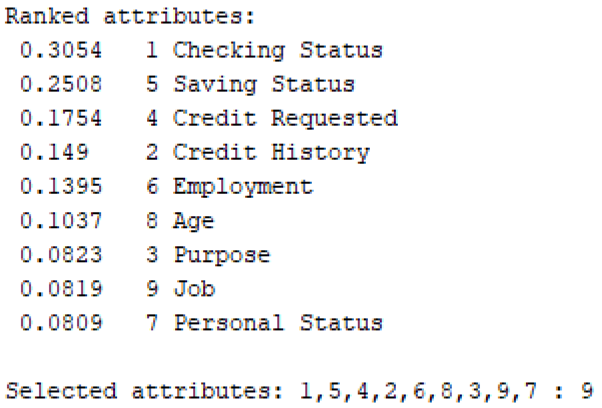
I found Clustering the least useful. I kept re-running the clustering algorithm trying to improve the accuracy and try and get more variation within the classes, however a lot of the clusters seemed very similar. It suggests that more clusters are required to get an accurate representation of the dataset. When there were only six clusters with the original settings, it gave 5 'Checking Status = no checking', and with the settings changed it's given five 'skilled' Jobs and three 'new cars' in Purpose.

Normally I would use scikit-learn on Supervised and Unsupervised learning techniques for a dataset of this size. I found Weka initially a bit unintuitive, however I quickly became comfortable using it. In future I'd consider using Weka if I'm looking to put an ML model together quickly, but like the flexibility and customisability offered from scikit-learn.

References:

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

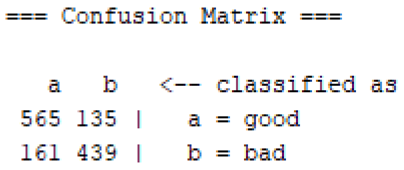
Figure 1 - Correlation Attribute Evaluation:



```
Ranked attributes:
 0.3054    1 Checking Status
 0.2508    5 Saving Status
 0.1754    4 Credit Requested
 0.149     2 Credit History
 0.1395    6 Employment
 0.1037    8 Age
 0.0823    3 Purpose
 0.0819    9 Job
 0.0809    7 Personal Status

Selected attributes: 1,5,4,2,6,8,3,9,7 : 9
```

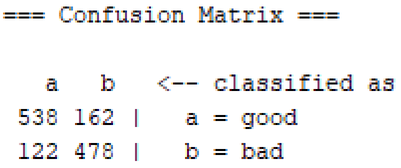
Figure 2 - Standard J48:



```
=== Confusion Matrix ===

   a   b   <-- classified as
 565 135 |   a = good
 161 439 |   b = bad
```

Figure 3 - J48 with Cost Sensitive Classifier and Bagging:



```
=== Confusion Matrix ===

   a   b   <-- classified as
 538 162 |   a = good
 122 478 |   b = bad
```

Bibliography

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

Apriori Algorithm:

https://www.codeproject.com/Articles/70371/Apriori-Algorithm
https://www.geeksforgeeks.org/apriori-algorithm/

Association Rules:

https://towardsdatascience.com/association-rules-2-aa9a77241654
https://towardsdatascience.com/complete-guide-to-association-rules-2-2-c92072b56c84
https://www.quora.com/What-are-association-rules-in-data-mining