

Assignment2_Kw

by Karen Windsor

Submission date: 13-May-2024 11:44AM (UTC+0100)

Submission ID: 233397587

File name: 24482_Karen_Windsor_Assignment2_Kw_565911_413263478.docx (7.62M)

Word count: 6719

Character count: 36930

Data Mining

ACCA6026

KAREN WINDSOR (1918576)

Table of Contents

1	Introduction	3
2	Data Mining	3
2.1	Definition	3
2.2	Data Mining vs Data Analysis	3
2.3	Data Mining vs Data Visualisation	4
2.4	Application(s) for Business	4
3	Description of South German Dataset	5
3.1	Initial patterns	7
3.1.1	Unbalanced Dataset.....	7
3.1.2	Dataset Patterns.....	8
4	Pre and Post Processing.....	9
4.1	Pre-processing.....	9
4.1.1	Primary Pre-processing.....	9
4.1.2	Secondary Pre-processing	9
4.1.3	Human readability pre-processing.....	9
4.1.4	WEKA specific pre-processing	10
4.2	Post-processing.....	10
4.2.1	Visualisation	10
4.2.2	Weightings.....	22
4.2.3	Normalisation.....	24
4.2.4	Training and testing split.....	25
5	Analysis Methods	26
5.1	Linear regression.....	26
5.2	Density-Based Spatial Clustering of Applications with Noise (DBSCAN)	27
5.3	K-Nearest Neighbour (KNN)	30
5.4	Random Trees	34
6	Business Recommendations	37
7	Reflections	40
8	Bibliography.....	41
9	Appendix	43
9.1	South German Credit CSV file	43
9.2	Data uploaded onto WEKA.....	43

Table 1 Data Mining Software [4]	3
Table 2 Attributes	5
Table 3 Imbalanced Data [16]	7
Figure 1 Python script	9
Figure 2 ARFF Header	10
Figure 3 Visual all attributes	10
Figure 4 Good vs Bad	11
Figure 5 Duration in months	12
Figure 6 Credit history	12
Figure 7 Purpose	12
Figure 8 Credit amount	13
Figure 9 Saving status	13
Figure 10 Employment	14
Figure 11 Instalment commitment	14
Figure 12 Personal status	15
Figure 13 Other parties	16
Figure 14 Residence since	16
Figure 15 Property magnitude	17
Figure 16 Age	17
Figure 17 Other payment plans	18
Figure 18 Housing	18
Figure 19 Existing credits	19
Figure 20 Job	19
Figure 21 Number of dependents	20
Figure 22 Own telephone	20
Figure 23 Foreign worker	21
Figure 24 Information gain	22
Figure 25 Attributes post discarding	23
Figure 26 Normalisation feature in WEKA	24
Figure 27 Randomising the dataset	25
Figure 28 Linear Regression	26
Figure 29 Credit amount vs predicted credit amount	27
Figure 30 DBSCAN with high noise level	28
Figure 31 Second run of DBSCAN	28
Figure 32 Remaining noise points	29
Figure 33 Outliers Personal status vs credit amount	29
Figure 34 KNN parameter 1	30
Figure 35 KNN parameter 2	31
Figure 36 Credit amount vs predicted credit amount	32
Figure 37 Purpose vs predicted credit amount	32
Figure 38 Outlier 1	33
Figure 39 Outlier 2	33
Figure 40 Random tree	34
Figure 41 Cost benefit Analysis	35
Figure 42 Training and test random tree	36
Figure 43 Employment vs credit amount	37
Figure 44 Credit history	38
Figure 45 Married	39
Figure 46 Purpose	39

1 Introduction

This report aims to explain what 'Data Mining' is, how it differs from Data Analysis and how it can benefit businesses. It highlights the importance of pre-processing primary and secondary data, researching the selection of appropriate analysis in WEKA and Tableau, and how results can be interpreted into useful banking statistics using the South German Credit Dataset [1]. The dataset contains 20 variables, and the classification depicts an applicant's creditworthiness. It holds information on 1000 loan applicants and this report endeavours to create a model that predicts evaluations of whether to approve a prospective applicant based on profile data [2].

2 Data Mining

2.1 Definition

Data mining is described as a method of discovering patterns in raw data [3], this process can utilise Supervised or unsupervised machine learning. It is vital to enable organisations/businesses to identify trends within their data. Patterns or trends must have a meaningful business objective that will give the organisation an advantage over its competitors. There are multitudes of different data mining tools available to data scientists, the top 10 include:

Table 1 Data Mining Software [4]

Data Mining Software	Description
MonkeyLearn	No code text mining tools
RapidMiner	Drag and drop workflows or data mining in Python
Oracle Data Miner	Predictive data mining models
IBM SPSS Modeler	A predictive analytics platform for data scientists
WEKA	Open-source software used for data mining
KNIME	Includes pre-built components used for data miner and other applications
H2O	Open-sourced library which includes data mining in Python
Orange	Open-sourced data mining toolbox
Apache Mahout	Used for large complex data mining
SAS Enterprise Miner	Data mining tool used to solve business problems

Data Mining Tool WEKA [5] will be used to explore the South German Dataset and evaluate results. WEKA is an open-sourced data mining tool, developed by the Machine Learning Group, University of Waikato, New Zealand. It is named after the flightless indigenous bird, but also the abbreviation stands for Waikato Environment for Knowledge Analysis [6]

2.2 Data Mining vs Data Analysis

The main difference between data mining and data analysis is that data mining is that data mining comprises the process(es) needed to extract patterns from large (big data) datasets. Data analysis focuses on how trends within data can be interpreted. Both are essential to enable organisations to gain insights when evaluating their data [5] [6].

2.3 Data Mining vs Data Visualisation

The focus of Data Mining as explained above is the recognition and extraction of patterns from big data, data visualisation presents these patterns in a visual format to provide organisations with a graphical tool to aid decision-making [7].

2.4 Application(s) for Business

There are myriad of different applications that data mining can be used for in business:

- i. Target marketing for retail businesses by constructing algorithms to predict customer responses by identifying past and current purchasing trends. A good example is how Amazon knows what products to recommend to their customer, data mining algorithms are used to make customer-specific recommendations to tempt the customer to make additional purchases [8].
- ii. Insurance firms underwrite risk by using algorithms to predict and detect insurance fraud [9] by utilising a framework called predictive profiling that can predict future purchasing behaviour given historical data to calculate the probability of a fraudulent claim.
- iii. Healthcare companies like Bupa can use data mining to provide clients with improved levels of care and tailor treatment options for patients. Insights gained can result in patients receiving more affordable care and higher-quality care [10].
- iv. Social media data mining is the process of extracting patterns and trends from vast amounts of unstructured data available on social platforms like Facebook, TikTok and X. This involves deep dives into data to analyse user behaviours to uncover patterns. This can be used to identify customer pain points, and trending topics or gauge how particular brands are viewed [11].

3 Description of South German Dataset

The dataset consists of 1,000 instances with 20 attributes and a classification as to whether an applicant is a good or bad credit risk. An instance is a single line of line/observation of the data, and each instance is characterised by values of attributes that measure different values within instances. There are several different types of attributes used in data mining, but typical data mining schemes usually only deal with two:

- i. Numeric
- ii. attributes are measurable and quantitative and as the name implies are numbers, represented integers or real values. There are two types of numeric attributes interval and ratio scaled. The interval-scaled attribute has values where the difference between them is meaningful. And ratio-scaled is an attribute with a fixed zero-point.
- iii. Nominal (categorical) is also known as qualitative data, and it represents categorical information. That is it cannot be quantified or ranked using numbers. Values in nominal data denote a category and there is no intrinsic order or hierarchical structure. Examples of nominal data include gender, ethnicity, religion and job title. Nominal data is used in data mining for classification and grouping purposes [12].

The dataset consists of 7 numerical and 13 categorical attributes. It also contains a binary attribute 'credit risk' which denotes whether a debtor is a good or bad credit risk. The dataset was downloaded in an ASC format and needs to be opened using a text editor. To enable familiarisation of the data the file was re-formatted into a [comma delimiter CSV](#) file and meaningful headings were given to each attribute. The table below details each attribute.

Table 2 Attributes

Attribute	Type	Description	Value(s)	Breakdown
Status	Categorical	Status of the applicants/debtor's bank account A11: < 0 DM A12: 0<=...<200 DM A13: ... >=200 DM/ salary assignments for at least 1 year A14: no checking account	1,2,3,4	1 = 274 2 = 269 3 = 63 4 = 394 No missing values
Duration	Numerical	Duration of the loan in months	4 - 72	No missing values
Credit History	Categorical	Debtors' payment history and current credit contracts A30: no credits taken/ all credits paid back duly A31: all credits at this bank paid back duly A32: existing credits paid back duly till now A33: delay in paying off in the past A34: critical account/other credits existing (not at this bank)	0,1,2,3,4,	0 = 40 1 = 49 2 = 530 3 = 88 4 = 293 No missing values
Purpose	Categorical	What the loan is needed for A40: car(new) A41: car(old) A42:furniture/equipment A43:radio/television A44:domestic appliances A45:repairs A46:education A47:vacation A48:retraining A49:business A410:others	0,1,2,3,4, 5,6,8,9,10	0 = 234 1 = 103 2 = 181 3 = 280 4 = 12 5 = 22 6 = 50 No number 7 – no loans for vacations 8 = 9 9 = 97 10 = 12 No missing values
Credit Amount	Numerical	Amount in DM's loaned to the applicant	250 – 18,424	No missing values
Savings	Categorical	Debtor's savings A61:...< 100 DM A62: 100<=...<500 DM A63: 500<=...<1000 DM	1,2,3,4,5	1 = 603 2 = 103 3 = 63

		A64:...>1000 DM A65: unknow/no savings account		4 = 48 5 = 183 No missing values
Employment duration	Categorical	Duration of the debtor's employment with current employer A71: unemployed A72: ...< 1 year A73: 1 <=...4 years A74: 4 <=...< 7 years A75:...>= 7 years	1,2,3,4,5	1 = 62 2 = 172 3 = 339 4 = 174 5 = 253 No missing values
Instalment rate	Numerical	Instalment rate in percentage of disposable income	1,2,3,4	1 = 136 2 = 231 3 = 157 4 = 476 No missing values
Personal status and sex	Categorical	Combined information about the debtor's gender, marital status A91: male divorced/separated A92: female divorced/separated/married A93: male single A94: male married/widowed A95: female single	1,2,3,4	1 = 50 2 = 310 3 = 548 4 = 92 No missing values but no value for widowed female
Other debtors	Categorical	Is there another person who is standing as a guarantor for the debt A101: none A102: co-applicant A103: guarantor	1,2,3	1 = 907 2 = 41 3 = 52 No missing values
Present residence since	Numerical	The length of time in years that the debtor has resided in their current property	1,2,3,4	1 = 130 2 = 308 3 = 149 4 = 413 No missing values
Property	Categorical	The debtor's most valuable assets A121: real estate A122: building society savings agreement/life insurance A123: car or another asset A124: unknown/no property	1,2,3,4	1 = 282 2 = 232 3 = 332 4 = 154 No missing values
Age	Numerical	Age of the debtor at the time of the loan in years	19 - 75	No missing values
Other instalment plans	Categorical	Instalment plans from other credit plans that the debtor may have A141: bank A142: stores A143: none	1,2,3	1 = 139 2 = 47 3 = 814 No missing values
Housing	Categorical	Type of housing that the debtor lives in at the time of the loan A151: rent A152: own A153: for free	1,2,3	1 = 179 2 = 714 3 = 107 No missing values
Number of credits	Numerical	Number of credits including current one that the debtor has	1,2,3,4	1 = 633 2 = 333 3 = 28 4 = 6

				No missing values
Job	Categorical	Quality of the debtor's job A171: unemployed/unskilled - non-resident A172: unskilled resident A173: skilled employee/official A174: management/self-employed/highly qualified employee/officer	1,2,3,4	1 = 22 2 = 200 3 = 630 4 = 148 No missing values
People liable	Numerical	The number of people who are financially reliant on the debtor	1,2	1 = 155 2 = 845 No missing values
Telephone	Categorical	Does the debtor have a landline registered in their name A191: none A192: yes	1,2	1 = 596 2 = 404 No missing values
Foreign worker	Categorical	Is the debtor a foreign worker A201: yes A202: no	1,2	1 = 37 2 = 963 No missing values
Credit risk	Cost matrix	1 = good 2 = bad	1,2	Good = 700 Bad = 300

3.1 Initial patterns

3.1.1 Unbalanced Dataset

This dataset is said to be imbalanced because it contains a disparate number of examples from each class [15]:

- i. 700 instances labelled as 'good' credits (majority class)
- ii. 300 instances labelled as 'bad' credits (minority class)

This results in the dataset having a class imbalance ratio of around 7:3 or 2.33:1 between majority and minority classes.

Table 3 Imbalanced Data [16]

Degree of imbalance	Proportion of Minority Class
Mild	20-40 of the datasets
Moderate	1-20% of the datasets
Extreme	<1% of the dataset

Class imbalance is a common issue in financial credit scoring and loan approval datasets, where it is usual for the number of 'bad' credits to be lower than the number of good credits [17]. Since many typical machine learning methods are biased towards the majority class during training, imbalanced datasets might provide difficulties. Especially when it comes to the minority class examples, which are sometimes the most crucial to identify, this might result in poor predicting performances. To manage this imbalance data scientists use methods such as class weights, (oversampling minority class, under-sampling majority class) or particular algorithms that have the functionality to cope with imbalanced classification during modelling [18].

3.1.2 Dataset Patterns.

When viewed in a CSV format it was clear that there were no missing data values. On further analysis of the formulae for ensuring a successful loan application several attributes have distinct differences in the ranges:

- i. Only 63 of the applicants had bank accounts status of more than 200 Deutsche marks and the majority of applicants had no checking account (394).
- ii. Over half (530) applicants had previous credit paid back duly.
- iii. Single males made up over half of the applicants (548), and there were no single female applicants, and no category for widowed females. It is unclear if this is a social-economic reflection of Germany at the time of the dataset's creation, or if there is insufficient data to include these.
- iv. Just over 90% of the applicants (907) are sole applicants.
- v. Loans appear to be for new cars and radio/televisions, but no loans were for vacations, is this another reflection of the social-economic environment in Germany or a deliberate/accidental representation of the data?
- vi. Over 96% of the loan applicants were not foreign workers, is this financial bias or just missing data due to the small dataset?
- vii. Most applicants have their own home 71% and have a skilled/official job 63%

Other attributes have ranges that are more evenly distributed:

- i. Telephone
- ii. Property/assets

4 Pre and Post Processing

4.1 Pre-processing

4.1.1 Primary Pre-processing

A fully pre-prepared dataset was available that had been thoroughly cleansed, but if the dataset was not available the following steps would have been completed:

- i. Handling missing values
Should instances containing missing values be removed or should the missing values be filled using techniques like mean imputation or regression imputation?
- ii. Removal of duplicates based on specific criteria, to ensure that each of the instances is unique.
- iii. Identification of outliers, using statistical techniques such as z-scores or the interquartile range, to evaluate and consider either removal or transformation of elements.

4.1.2 Secondary Pre-processing

Secondary pre-processing entails additional data transformation to enhance the quality and suitability of the data for analysis. This phase encompasses dimensionality reduction, feature engineering, and feature scaling. Standardising a feature's range to make it comparable is known as feature scaling. Z-score normalisation and min-max scaling are popular methods for feature scaling. Feature engineering generates new features from pre-existing ones to enhance forecast performance or collect more data [19]. This can involve polynomial expansion, binning, or generating interaction terms. Principal component analysis (PCA) and feature selection algorithms are two dimensionality reduction techniques used to reduce the number of features while preserving the most crucial information [20].

4.1.3 Human readability pre-processing

Unlike algorithms and software, humans are notoriously bad at reading numbers [21], to make the file more readable python scripts can be written to transform the numbers into the corresponding text.

Figure 1 Python script

```
In [ ]:
# Open the input file for reading
with open("German.txt", "r") as file:
    lines = file.readlines()

# Transform the elements in the "OtherInstallments" column
for i in range(len(lines)):
    # Split each line by tab to separate columns
    columns = lines[i].split("\t")

    # Get the value in the "OtherInstallments" column
    other_installments = int(columns[3])

    # Transform the value according to the conditions
    if other_installments == 1:
        columns[3] = "Bank"
    elif other_installments == 2:
        columns[3] = "Stores"
    elif other_installments == 3:
        columns[3] = "None"

    # Join the columns back into a line and replace the original line
    lines[i] = "\t".join(columns)

# Open the output file for writing
with open("German_transformed.txt", "w") as file:
    file.writelines(lines)

#This script reads the contents of the "German.txt" file, transforms the elements in the
#"OtherInstallments" column based on the provided conditions, and writes the updated lines
#to a new file named "German_transformed.txt". Make sure to replace "German.txt" with the
#actual file path if it's located in a different directory.
```

4.1.4 WEKA specific pre-processing

Files uploaded into WEKA need to be in an Attribute Relation File Format (ARFF) [22]. The ARFF file is an ASCII text file which describes a list of instances (data rows) sharing a set of attributes (columns) and was developed for analysis in WEKA. The first part of the file is the ARFF header which contains information about the type of data held within the attribute, followed by the Data.

Figure 2 ARFF Header

```
281
282 @relation german_credit
283 @attribute 'checking_status' { '<0', '0<=X<200', '>=200', 'no checking'}
284 @attribute 'duration' real
285 @attribute 'credit_history' { 'no credits/all paid', 'all paid', 'existing paid', 'delayed previously', 'critical/other existing
credit'}
286 @attribute 'purpose' { 'new car', 'used car', 'furniture/equipment', 'radio/tv', 'domestic appliance', 'repairs', 'education', 'vacation',
retraining, 'business', 'other'}
287 @attribute 'credit_amount' real
288 @attribute 'savings_status' { '<100', '100<=X<500', '500<=X<1000', '>=1000', 'no known savings'}
289 @attribute 'employment' { 'unemployed', '<1', '1<=X<4', '4<=X<7', '>=7'}
290 @attribute 'installment_commitment' real
291 @attribute 'personal_status' { 'male div/sep', 'female div/dep/mar', 'male single', 'male mar/wid', 'female single'}
292 @attribute 'other_parties' { 'none', 'co applicant', 'guarantor'}
293 @attribute 'residence_since' real
294 @attribute 'property_magnitude' { 'real estate', 'life insurance', 'car', 'no known property'}
295 @attribute 'age' real
296 @attribute 'other_payment_plans' { 'bank', 'stores', 'none'}
297 @attribute 'housing' { 'rent', 'own', 'for free'}
298 @attribute 'existing_credits' real
299 @attribute 'job' { 'unemp/unskilled non res', 'unskilled resident', 'skilled', 'high qualif/self emp/mgmt'}
300 @attribute 'num_dependents' real
301 @attribute 'own_telephone' { 'none', 'yes'}
302 @attribute 'foreign_worker' { 'yes', 'no'}
303 @attribute 'class' { 'good', 'bad'}
304 @data
305 '<0',6,'critical/other existing credit',radio/tv,1169,'no known savings','>=7',4,'male single',none,4,'real
estate',67,none,own,2,skilled,1,yes,yes,good
306 '0<=X<200',48,'existing paid',radio/tv,5951,'<100', '1<=X<4',2,'female div/dep/mar',none,2,'real
estate',22,none,own,1,skilled,1,none,yes,bad
307 'no checking',12,'critical/other existing credit',education,2096,'<100', '4<=X<7',2,'male single',none,3,'real
estate',49,none,own,1,'unskilled resident',2,none,yes,good
308 '<0',42,'existing paid',furniture/equipment,7882,'<100', '4<=X<7',2,'male single',guarantor,4,'life insurance',45,none,'for
free',1,skilled,2,none,yes,good
309 '<0',24,'delayed previously', 'new car',4870,'<100', '1<=X<4',3,'male single',none,4,'no known property',53,none,'for
free',2,skilled,2,none,yes,bad
310 'no checking',36,'existing paid',education,9055,'no known savings', '1<=X<4',2,'male single',none,4,'no known
property',35,none,'for free',1,'unskilled resident',2,yes,yes,good
311 'no checking',24,'existing paid',furniture/equipment,2835,'500<=X<1000', '>=7',3,'male single',none,4,'life
```

@Relation – the dataset name

@Attribute – data type which can be numeric, nominal, string, or date. The above attributes that are real are types of numeric attributes. Nominal values are defined by possible values they can include, in the South German dataset the attribute class can contain either 'good' or 'bad'.

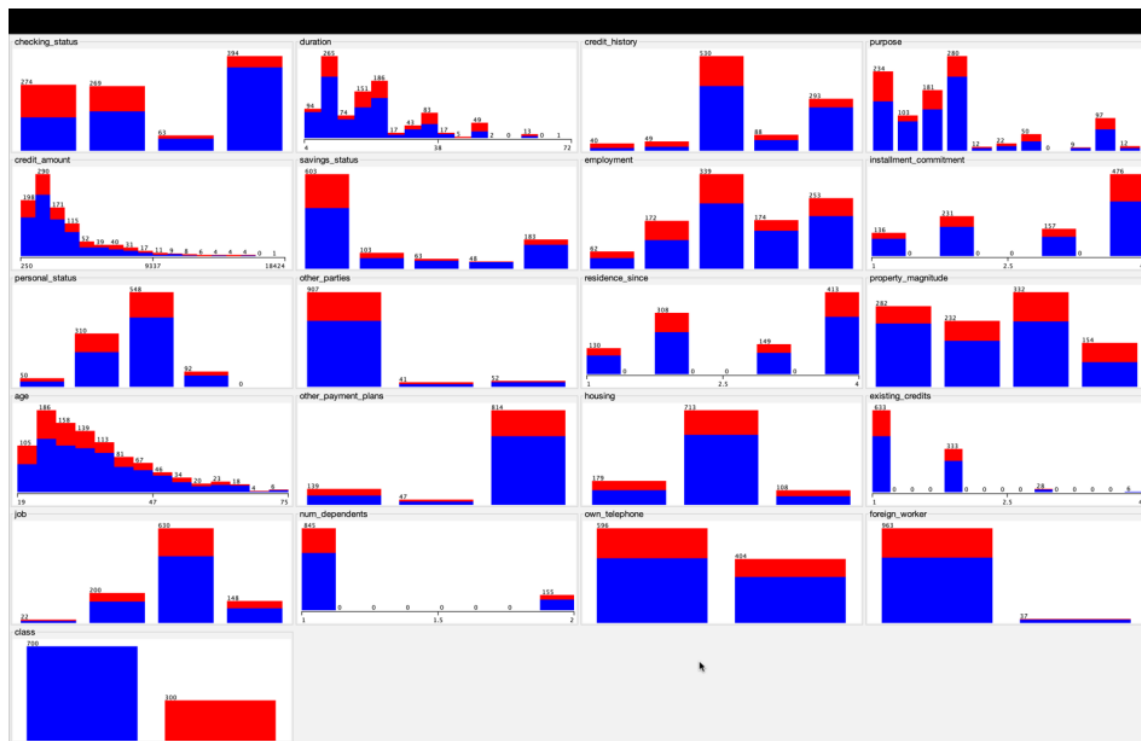
4.2 Post-processing

Post-processing in data mining refers to the techniques and procedures applied to the results obtained from a machine learning or data mining algorithm to refine, filter, or transform them into a more understandable, interpretable, or actionable form. The post-processing is performed on the data once it is [uploaded into WEKA](#).

4.2.1 Visualisation

The 'visual all' tab in WEKA can visualise the attributes in the dataset graphically. It provides insights into the data by creating various charts to allow the data scientist to understand the relationships and patterns within the data. There are options to apply filters and adjust settings to meet analysis requirements.

Figure 3 Visual all attributes



To understand the charts it is important to make sense of the colour schematic. The below figure illustrates that 'good credit' is depicted in blue whilst, and that 70% of the bank's loans are considered 'good' credit risks.

Figure 4 Good vs Bad

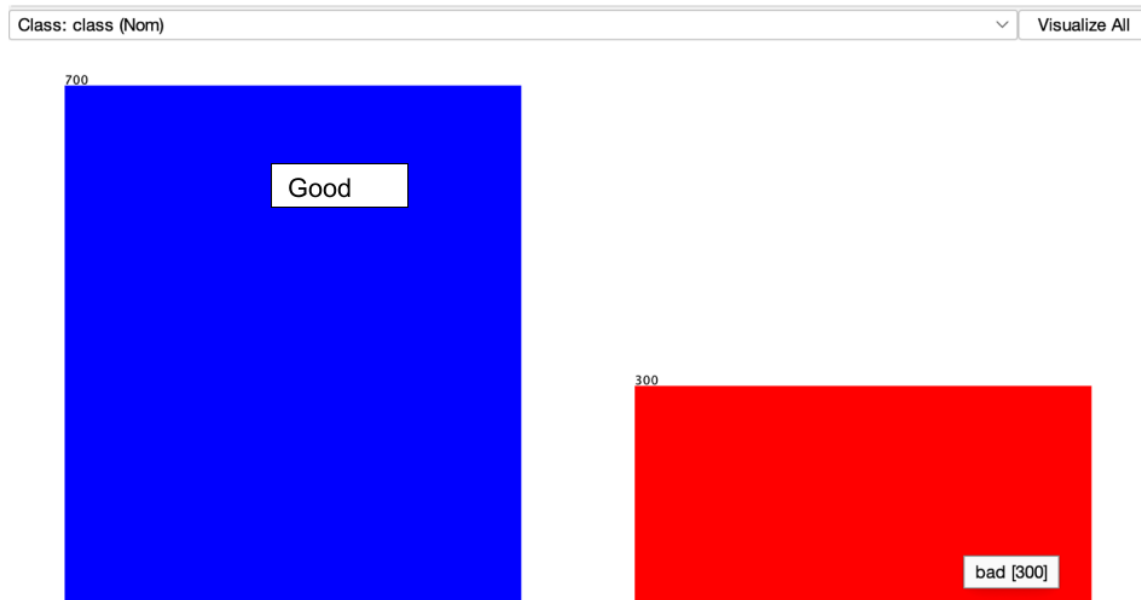
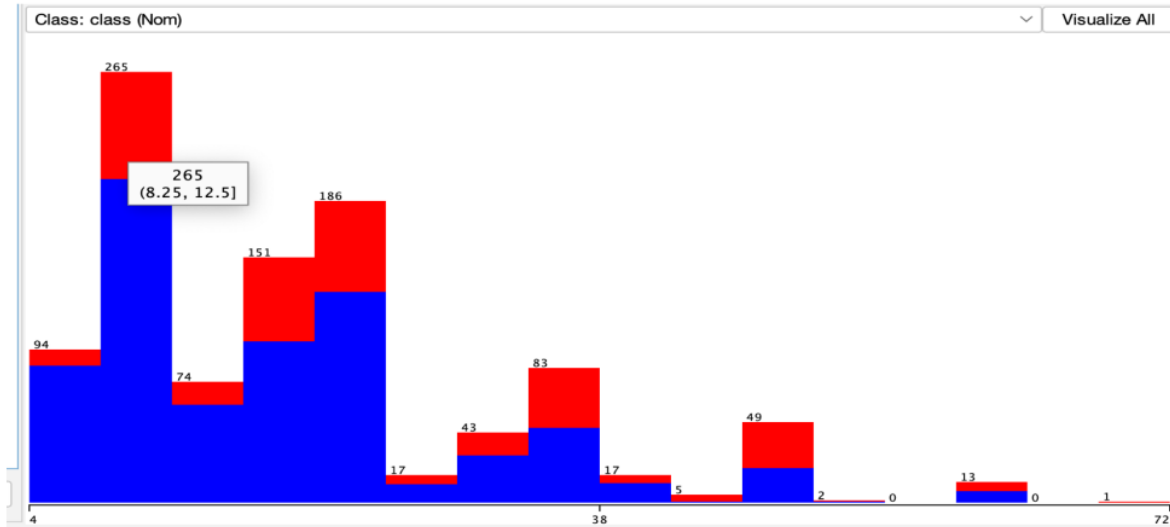
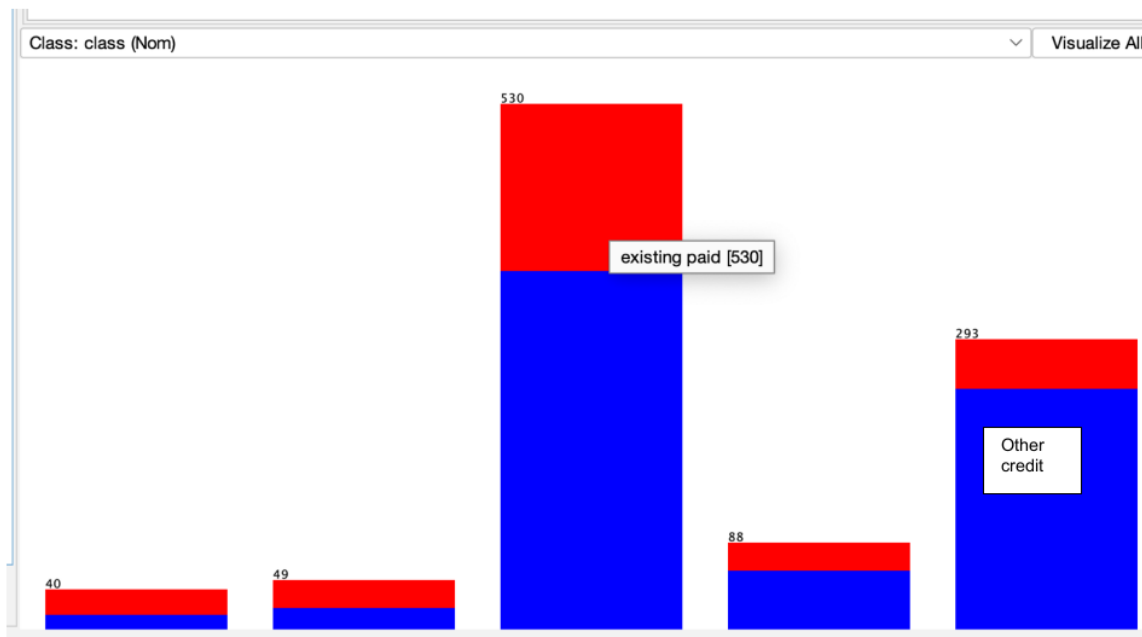


Figure 5 Duration in months



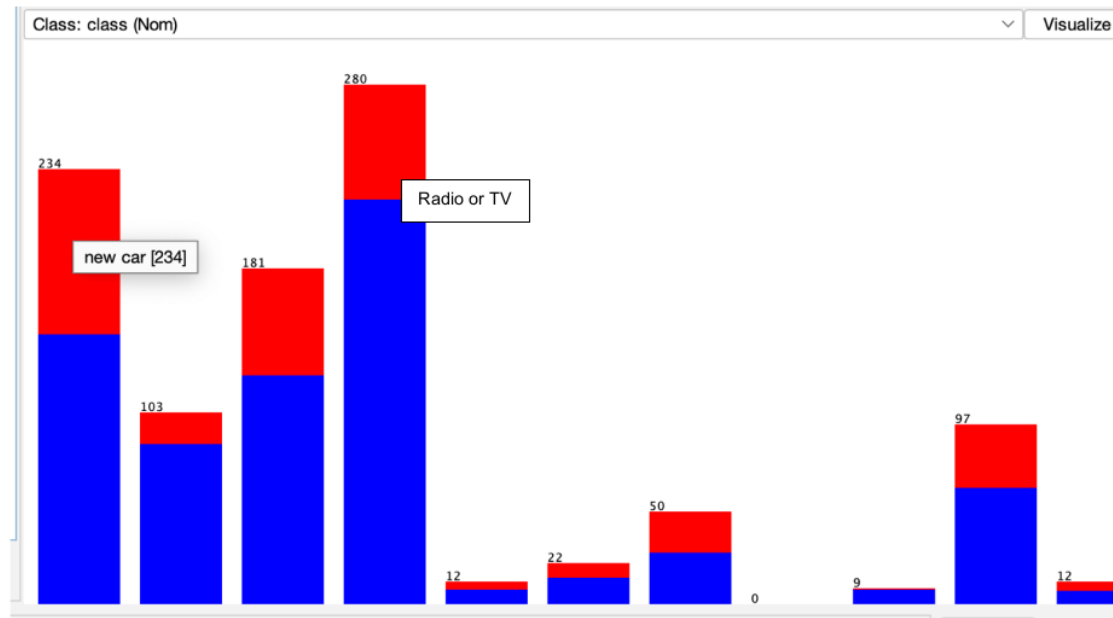
The chart shows that the majority of loans have a duration between 8.25 and 12.5 months and there is an interesting second peak (hump) to investigate between 21-25.25-months duration.

Figure 6 Credit history



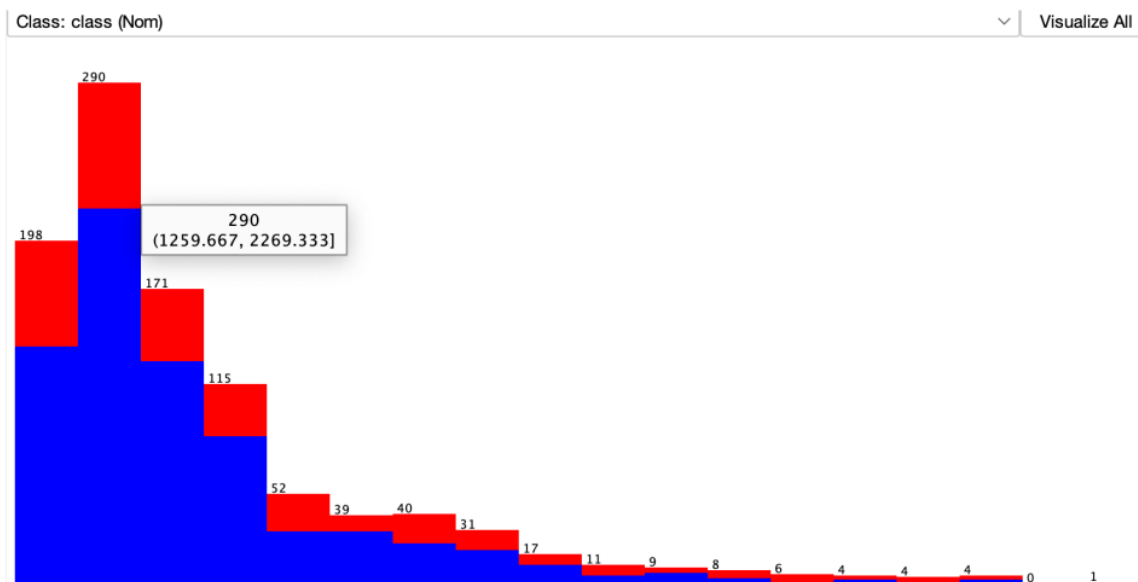
The credit history attribute shows a 'good credit' majority in two clear humps for those applicants that had previously paid back or have existing credit agreements.

Figure 7 Purpose



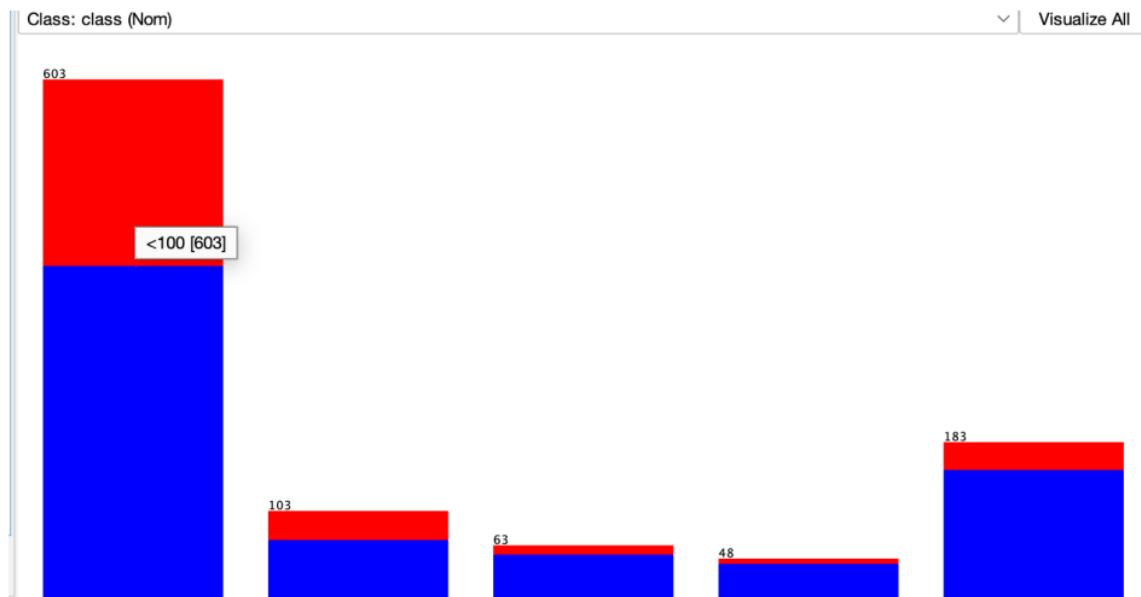
The above figure shows the different purposes of loans and the distribution of 'good' and 'bad' credit risks.

Figure 8 Credit amount



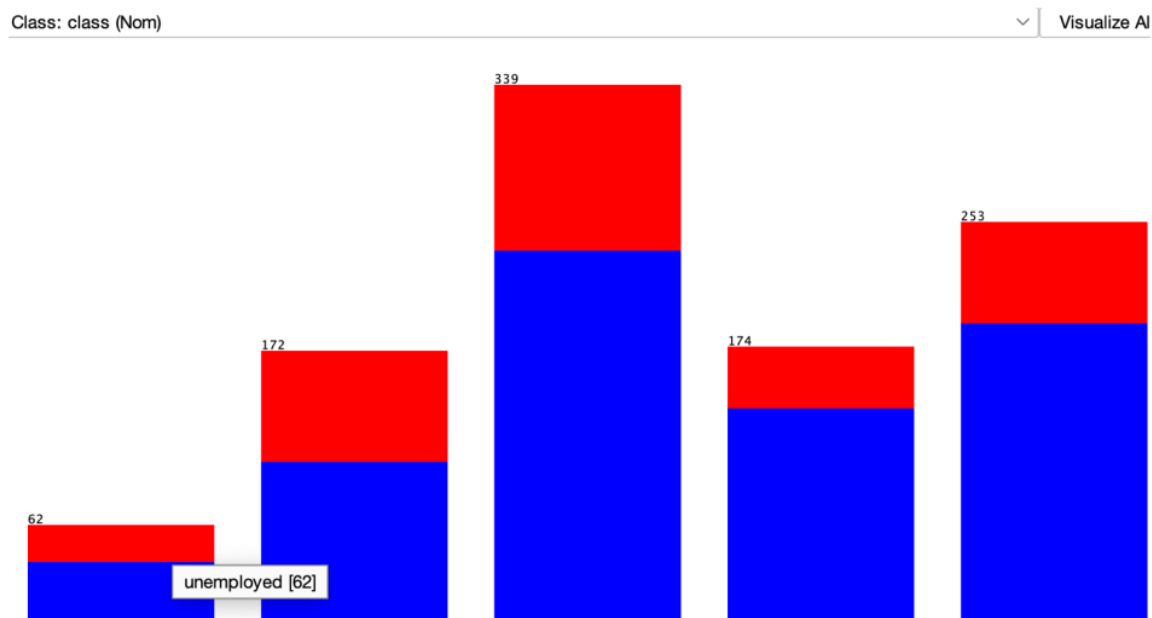
The amount loaned peaks between 1259-2269 DMs(290 applicants) and then evenly decreases to only 1 applicant for loans between 17,414 – 18,424 DMs. This chart clearly shows where outliers, those applicants with the highest loan amounts could affect the results of any analysis. A decision will have to be made as to what action will need to be taken with these outliers, remove them or leave them and note that they are part of the dataset.

Figure 9 Saving status



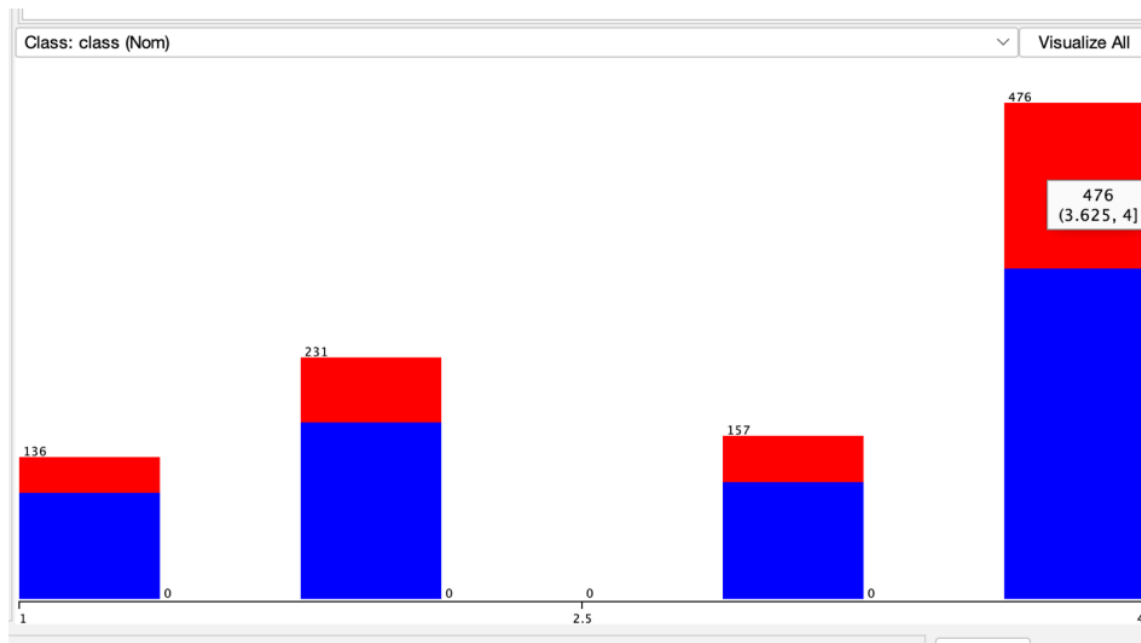
Applicants with less than 100 DMs in savings make up the majority of the applicants, and the least loans were given to those with 1000 DMs or more in savings. The second highest number was given to applicants with no savings.

Figure 10 Employment



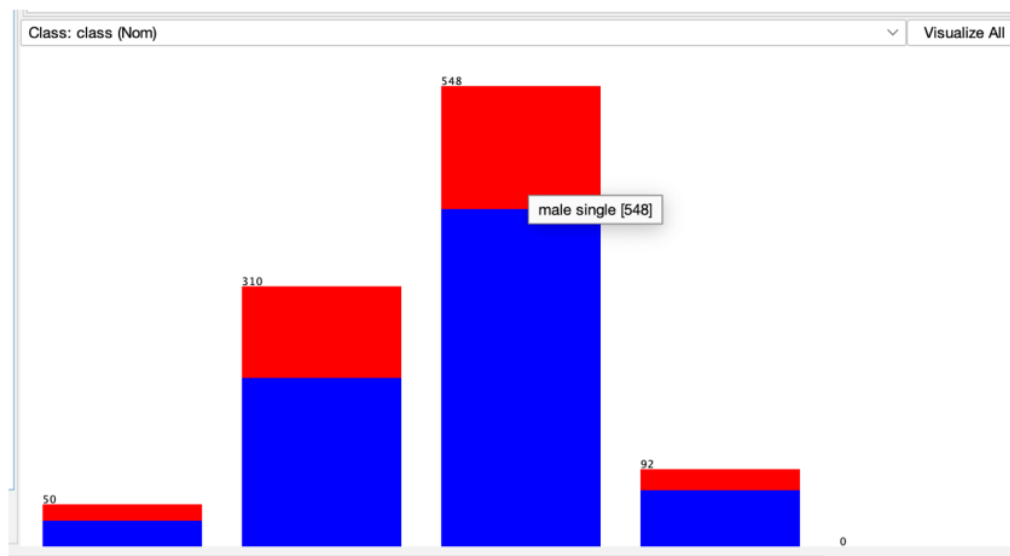
There were 62 unemployed applicants, an interesting analysis would be to compare applicants with an unemployed status with those with higher savings.

Figure 11 Instalment commitment



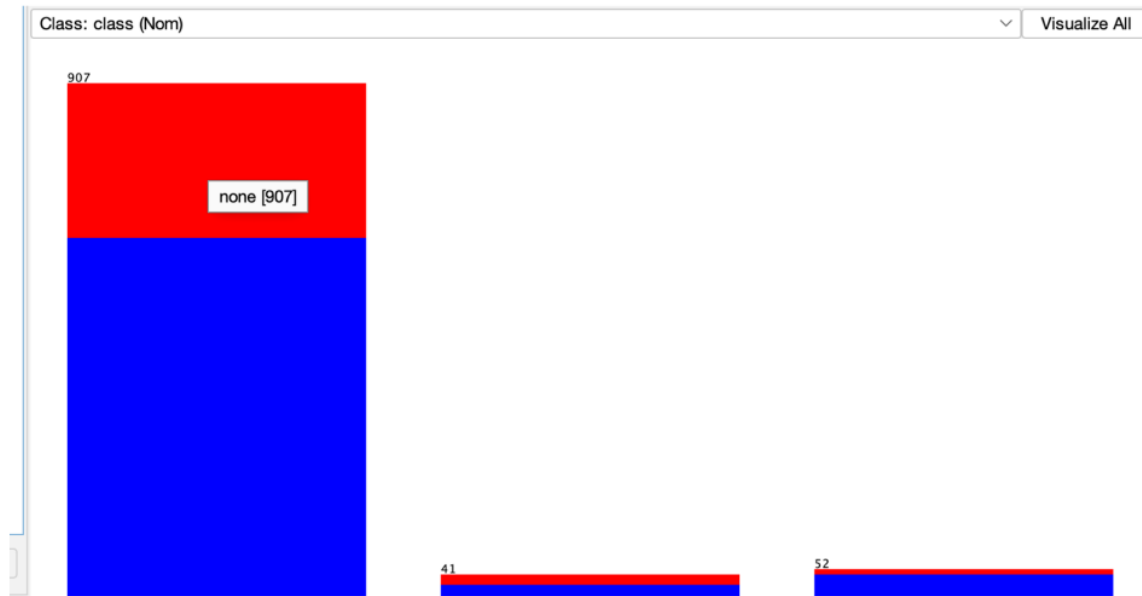
This attribute contains information on what percentage the monthly instalment is of the applicant's gross incomings, the majority of the applicants have instalments that equate to 4% of their income. On Examination the visual shows that the lowest percentage of bad risk is for those applicants for whom the monthly commitment is only 1% of their income.

Figure 12 Personal status



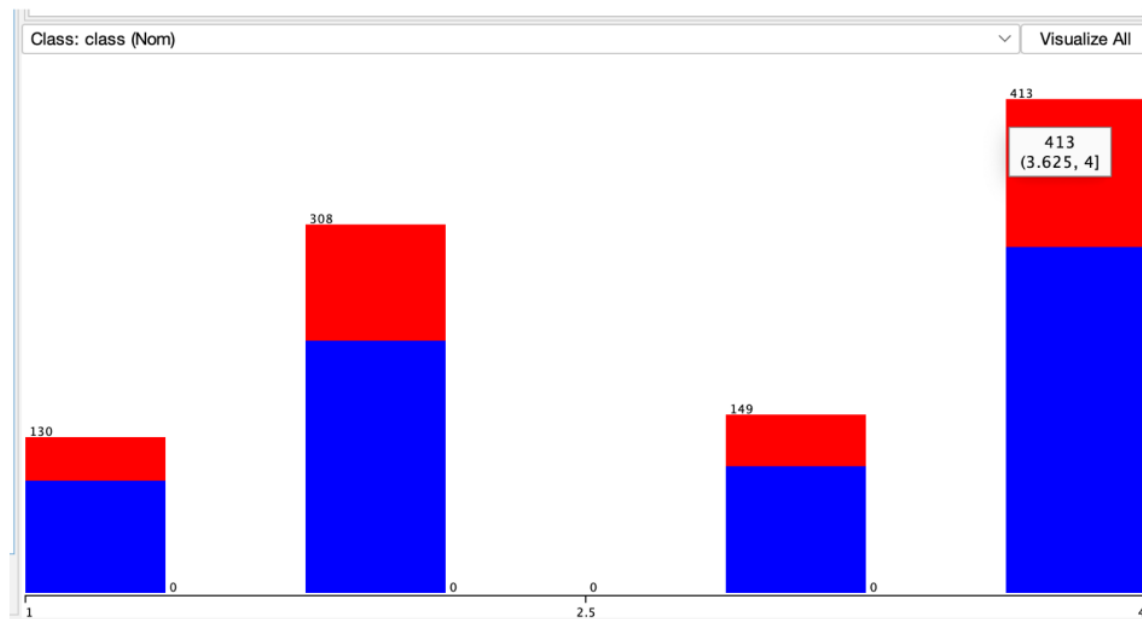
As expected single males make up the majority of applicants, followed by divorced, married, dependent females. Scanning the visual it appears that the 'bad' credit risk is equally distributed between the 4 categories.

Figure 13 Other parties



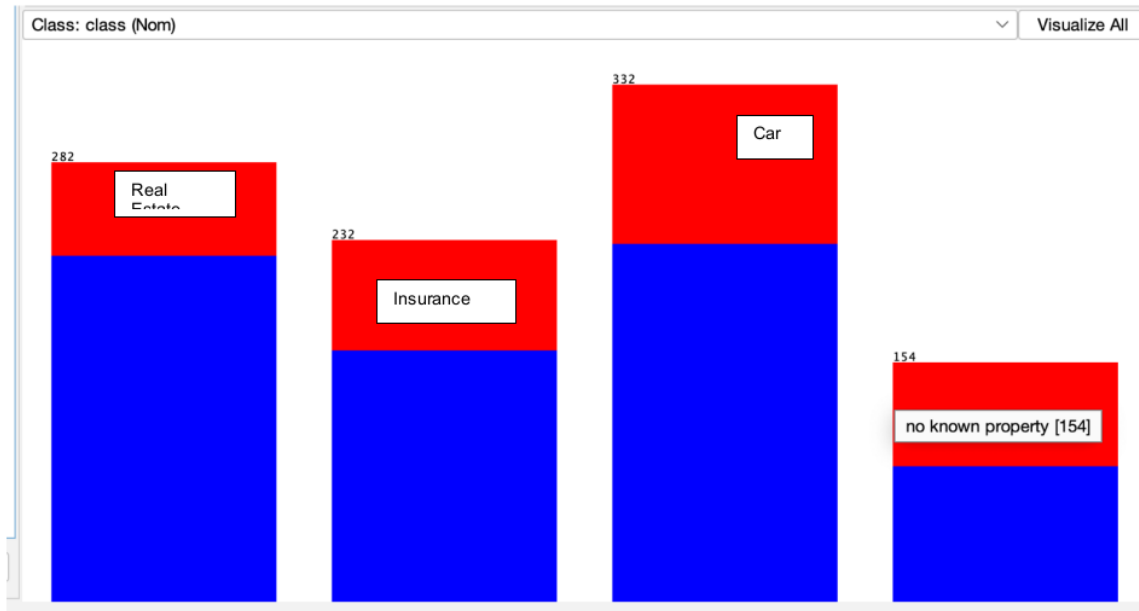
This attribute does not add depth to the dataset as over 90% of the applicants are sole applications.

Figure 14 Residence since



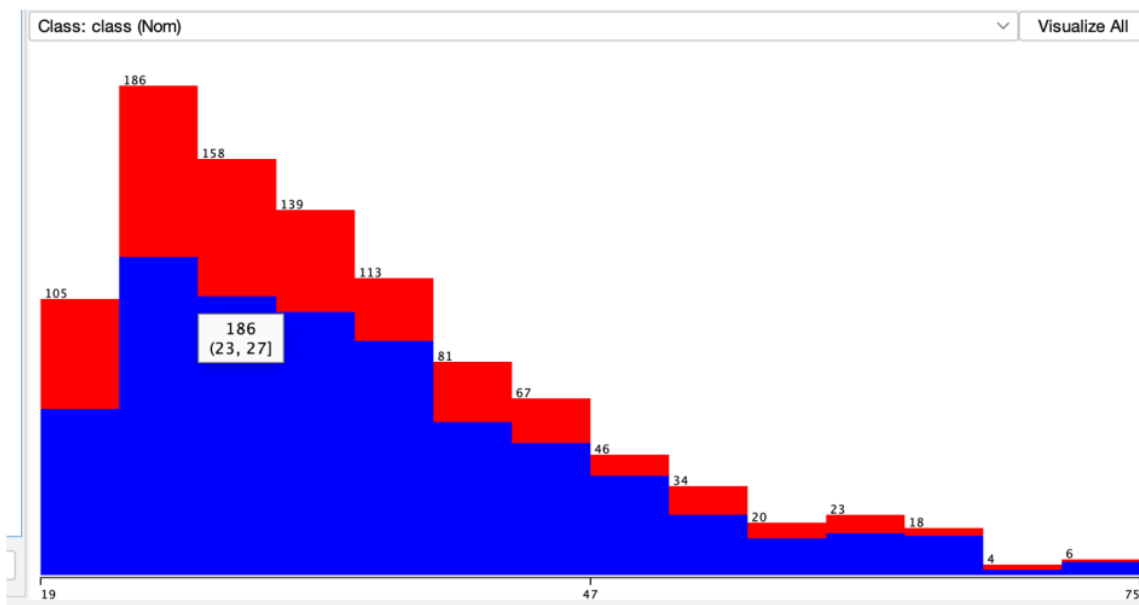
There is not much variation in this attribute with a nearly equal split between those applicants in their current residence under 2.5 years and those living there longer. And there does not seem to be any interesting split in the 'good' or 'bad' credit to explore.

Figure 15 Property magnitude



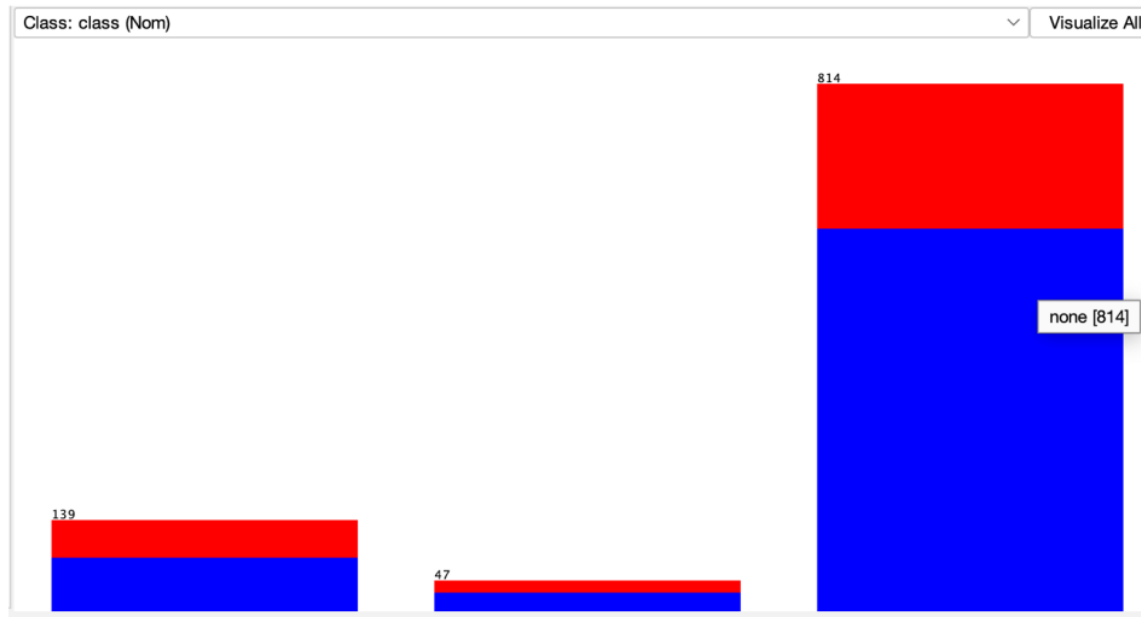
The applicants' assets seem equally distributed with no outstanding category within the attribute.

Figure 16 Age



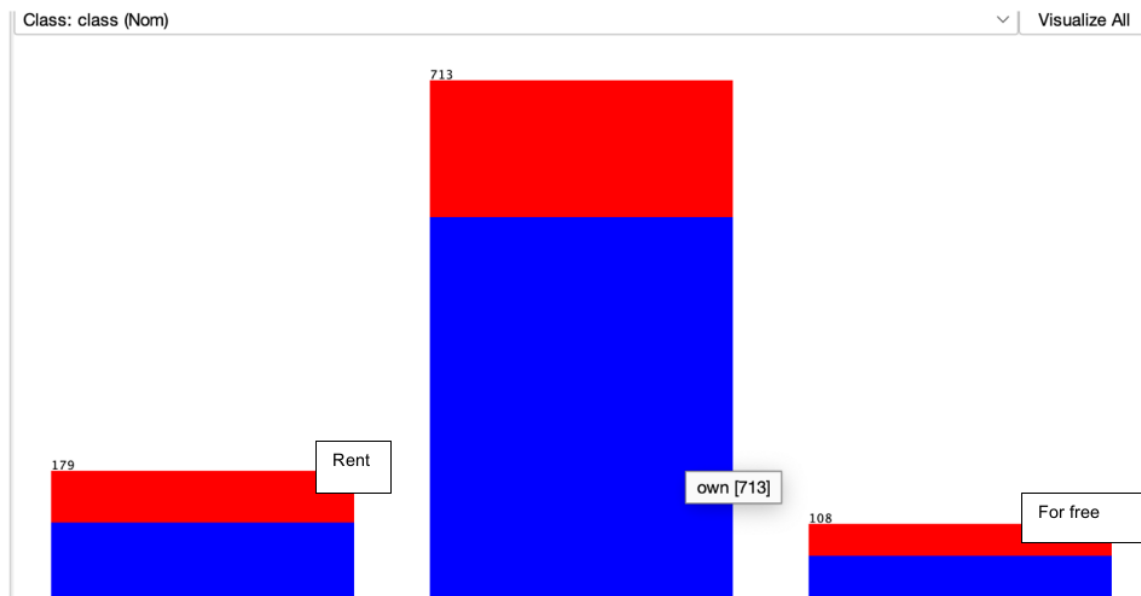
The number of applicants peaks between the ages of 23 - 27, then steadily decreases until the applicant is 75.

Figure 17 Other payment plans



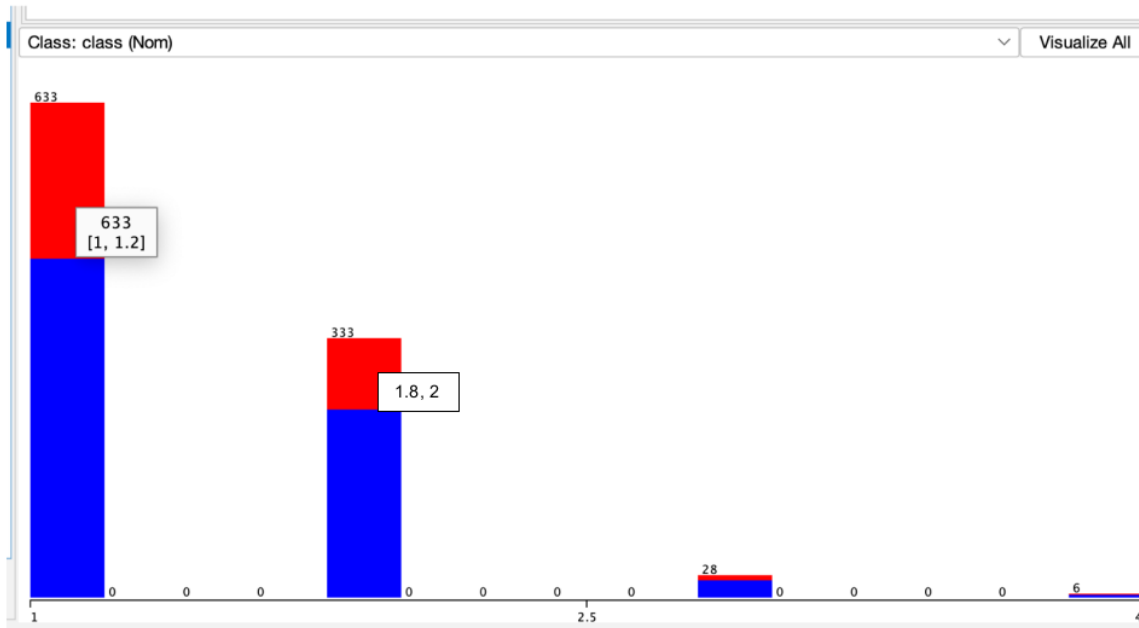
This attribute does not demonstrate any provoking information with 81% of applicants having no other payment plans.

Figure 18 Housing



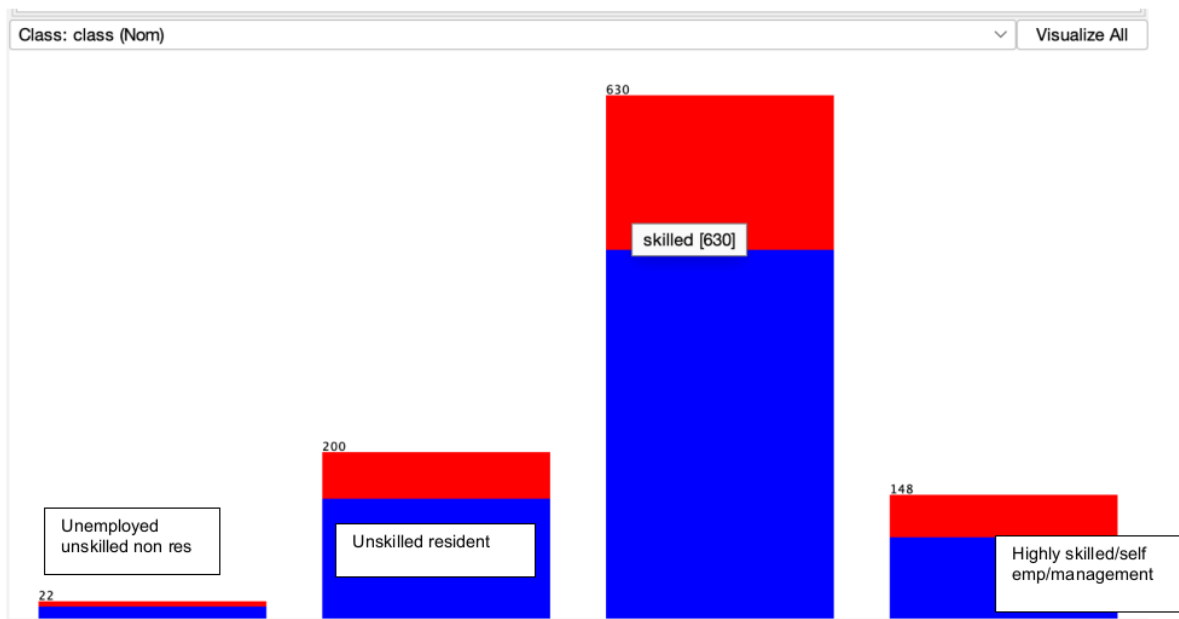
This attribute contains no attention-grabbing elements, except that 10% of applicants live in their homes for free. is this because they have paid off their mortgage or reside with family?

Figure 19 Existing credits



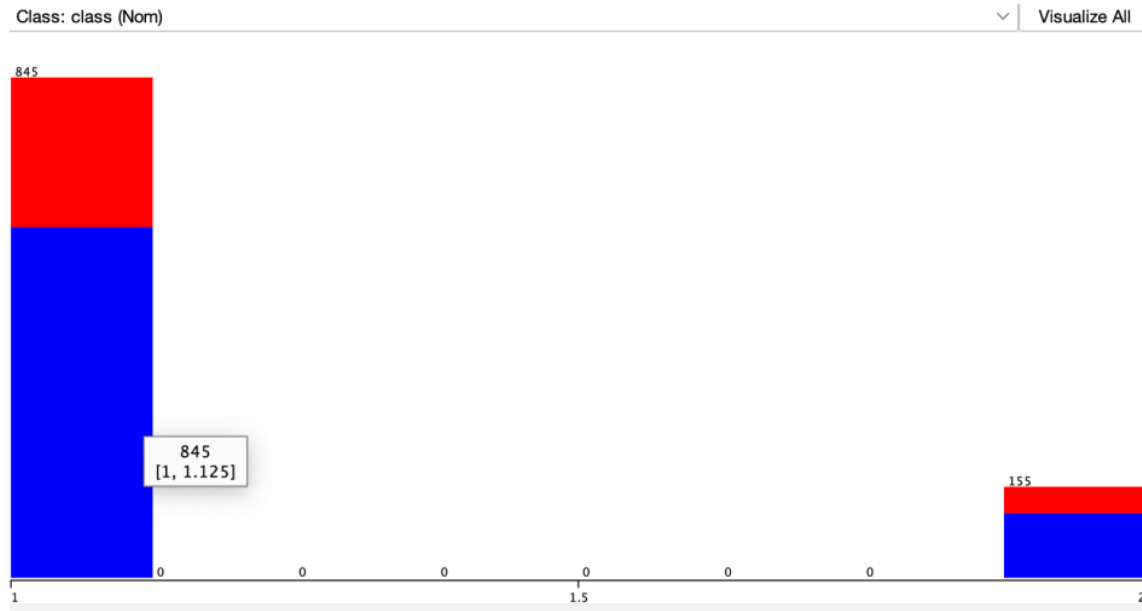
The 2 main types with the attribute hold between them 966 instances with 34 outliers that need to be weighed up before analysis.

Figure 20 Job



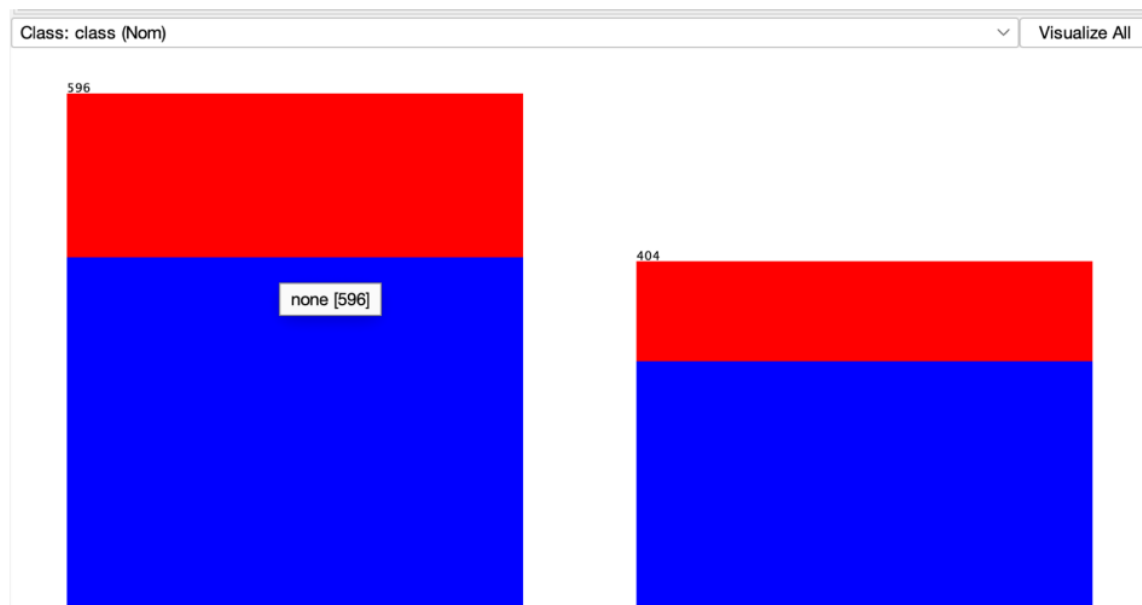
Outliers of 22 instances where the applicant is either unemployed or an unskilled non-resident.

Figure 21 Number of dependents



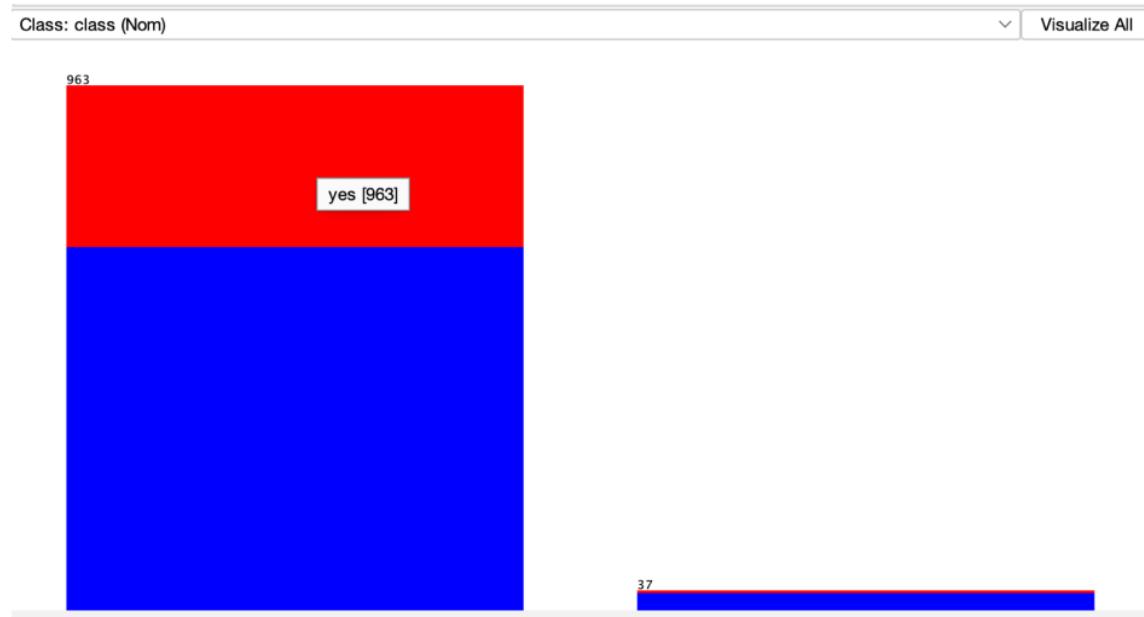
This attribute shows that all the applicants either have 1 or 2 dependents which would seem contradictory to the personal status attribute why the biggest group are applications from single males.

Figure 22 Own telephone



The attribute dates the dataset, because in 2023 when mobile phone ownership was reported at 143.6% of the population [23]. This attribute has a nearly even split between no phone and phone owners and does not provide the analyst with any informative data.

Figure 23 Foreign worker



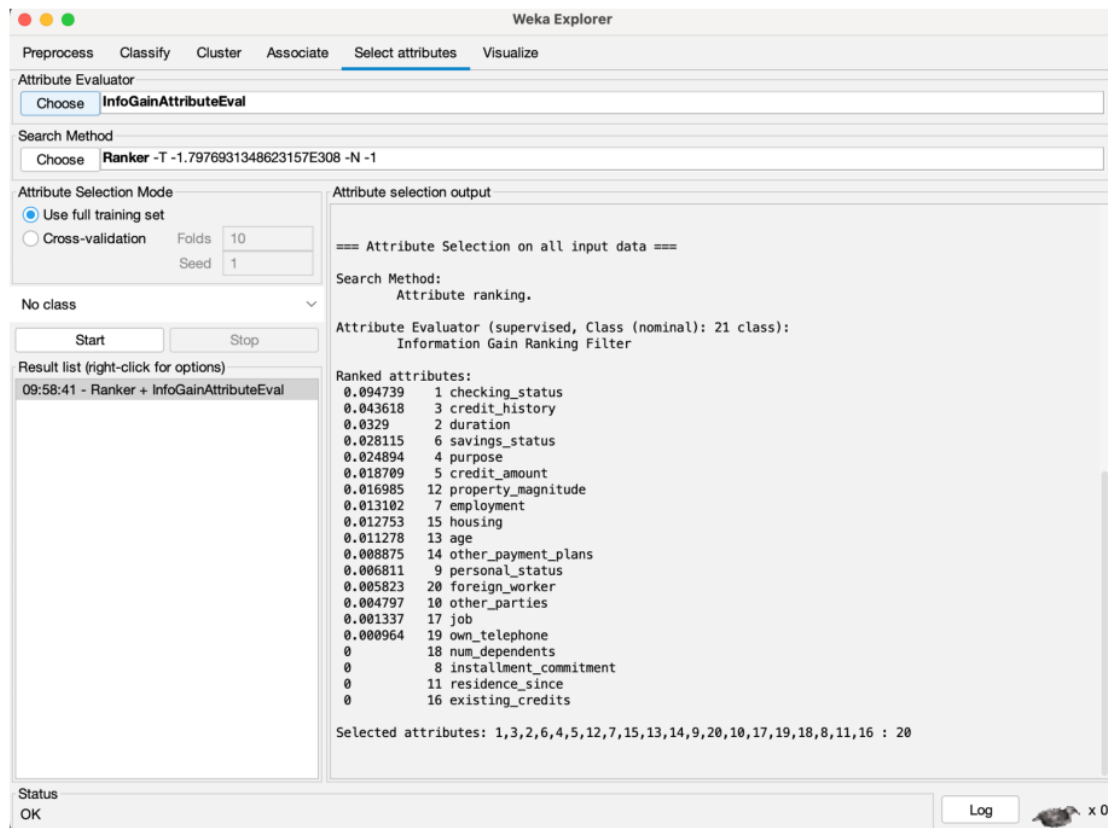
In the original dataset, 37 applicants were classed as foreign workers, however, in WEKA this is showing as 963 foreign workers. This discrepancy will need to be investigated if the attribute is used in any analysis to ensure that the analysis is not biased.

4.2.2 Weightings

4.2.2.1 Information gain

A measure concerning the amount of information an attribute provides about the class variable. It is used in feature selection to determine the relevance of each attribute in a classification task. In WEKA the 'InfoGainAttributeEval' calculates the 'information gain' for each attribute by weighing up the entropy of the class variable before and after splitting the data based on the attribute. The attribute with the highest 'information gain' is considered the most informative and relevant for the classification task. This helps to reduce the noise and dimensionality of the dataset and improves the efficiency and accuracy of the classification task [24] .

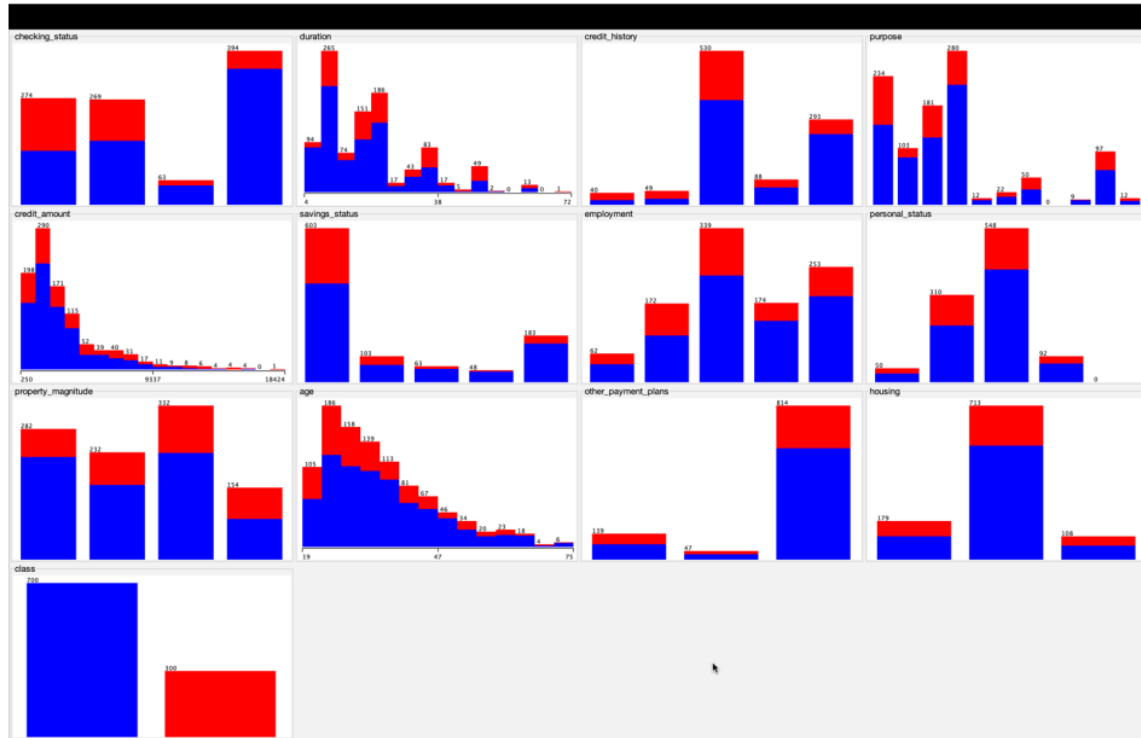
Figure 24 Information gain



Based on the ranked attributes and their weightings, the ones with the lowest can be considered for discarding. In this case, the attributes with weightings of 0 or close to 0 are:

- i. residence since
- ii. existing credits
- iii. nondependent
- iv. installment_commitment
- v. own telephone
- vi. job
- vii. foreign_worker
- viii. other_parties

Figure 25 Attributes post discarding

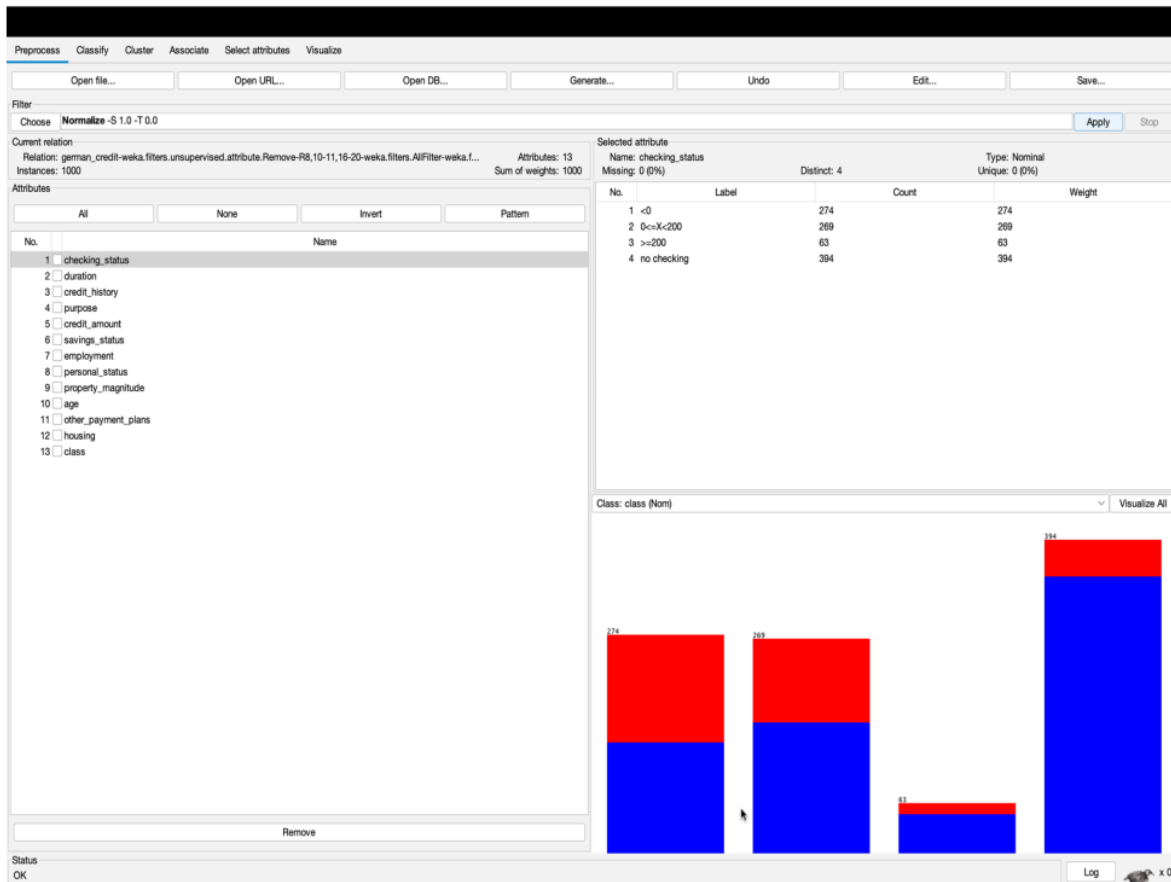


4.2.3 Normalisation

This is the process of rescaling or transforming the attributes of a dataset to a common scale or range and it ensures that:

- No single feature dominates the analysis
- Prevent numerical instabilities and ensure that the optimisation process is not biased towards features with larger scales.
- Reduce the impact of outliers
- Improve performance and convergence of algorithms.

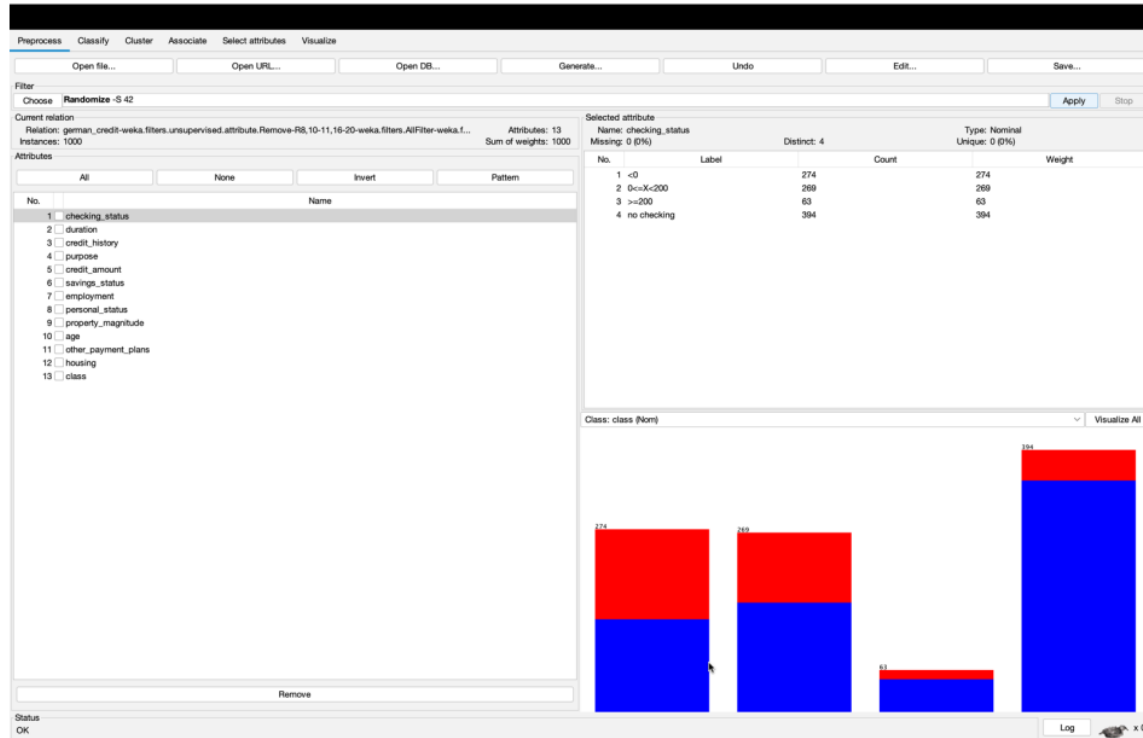
Figure 26 Normalisation feature in WEKA



4.2.4 Training and testing split

A vital step which can be overlooked when splitting data into training and test sets is the randomisation of the dataset. Randomising the dataset ensures that the training and test sets represent the whole dataset. This is essential for obtaining unbiased, accurate, repeatable and reliable analysis models [25]. WEKA provides the analyst within the 'Filter' label see below.

Figure 27 Randomising the dataset



Once the dataset has been randomised, it can be split into training and test sets. The usual ratio for splitting a dataset into training and test sets is 80:20 [26], the Pareto principle [27] vindicates this. The purpose of separating the dataset is twofold:

- The testing data offers an objective assessment of how effectively the model has learnt those patterns by generating predictions on data it has not previously seen. In contrast, the training data teaches the model the underlying patterns and relationships.
- Reduces overfitting, which is where the model performs well on the training examples but is unable to generalise to new data, which is likely to happen when utilising the training data for both training and testing.

The size of the dataset might have a detrimental effect on the results is the size of the dataset (only 1000 instances), as Geron [28] states that even simple machine learning algorithms need thousands of instances to optimise the algorithms. Also imbalanced classes, with more 'good' than 'bad ratings, could pose a challenge for modelling.

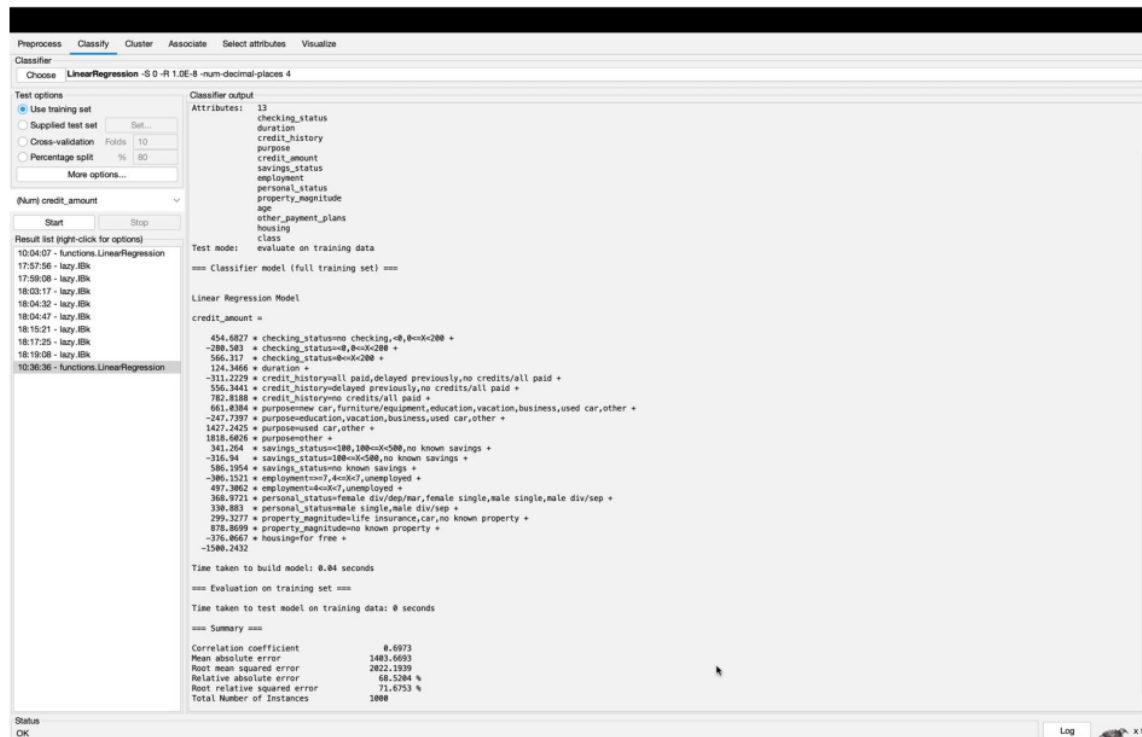
The attributes purpose of the loan, savings levels, and employment history provide useful signals about creditworthiness. Housing, and job stability also likely correlate. The relationship between the attributes of income vs loan amount could aid predictions.

5 Analysis Methods

5.1 Linear regression

One of the best algorithms to use to help analyse the relationship between the age of the applicants and the loan amount would be linear regression. The model can analyse the relationship between a scalar dependent variable (loan amount) and a more explanatory variable (age). The attributes age and loan amount are both continuous numeric attributes, which are attributes that linear regressions are designed to use. The objective is to understand if the loan amount changes as age changes. Linear regression provides a simple linear equation that models the relationship which will make the detection of the strength and direction that age has over the loan amount. Determining how much of a variation in the loan amount can be explained by age alone via the r-squared coefficient of determination.

Figure 28 Linear Regression



The coefficients represent the influence that each of the attributes has on the predicted credit amount, after accounting for other attributes [29]:

- Correlation coefficient 0.6973
- Mean absolute error 1403.6693
- Root mean squared error 2022.1939
- Relative absolute error 68.5204%
- Root relative squared error 71.6753%

The correlation coefficient indicates a moderate positive correlation between the predicted and actual credit amounts. The error metrics show relatively high errors, which implies that the model may not have a high degree of accuracy when predicting credit amounts.

Figure 29 Credit amount vs predicted credit amount



5.2 Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

A density-based clustering algorithm which works to discover clusters of arbitrary shape and size from a large amount of data containing noise and outliers.

When running DBSCAN in WEKA the analyst can specify the epsilon, MinPoints, and distance type to adjust the algorithm to build the best model. The epsilon parameter determines the maximum distance between two points for them to be considered neighbours. The advantages of applying the DBSCAN algorithm to the dataset are that:

- i. It can handle arbitrary-shaped clusters and is robust to noise and outliers
- ii. It does not require the number of clusters to be specified in advance
- iii. It handles datasets with high noise levels and contains clusters of varying densities and shapes

The results from the first run of the DBSCAN algorithm produced a high level of noise and used the default parameters:

- i. Epsilon 0.5
- ii. MinPoints 6

Noise in DBSCAN refers to data points that do not belong to any cluster because they are not dense enough or too far away from other data points.

Figure 30 DBSCAN with high noise level

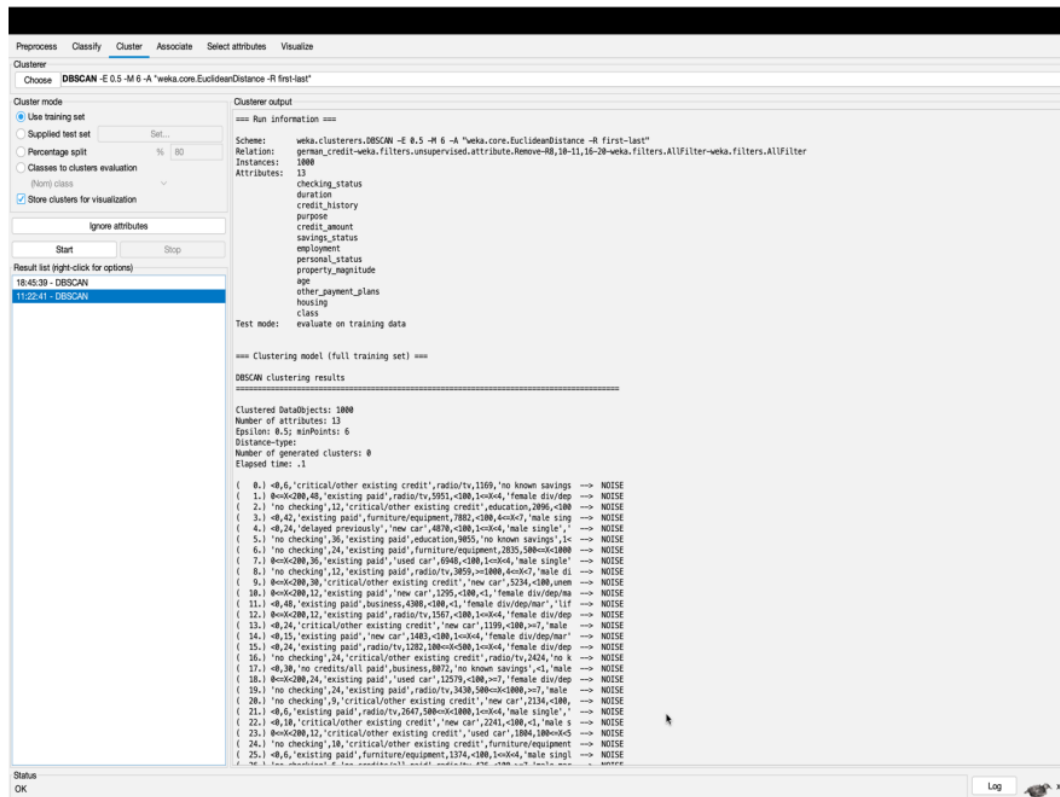
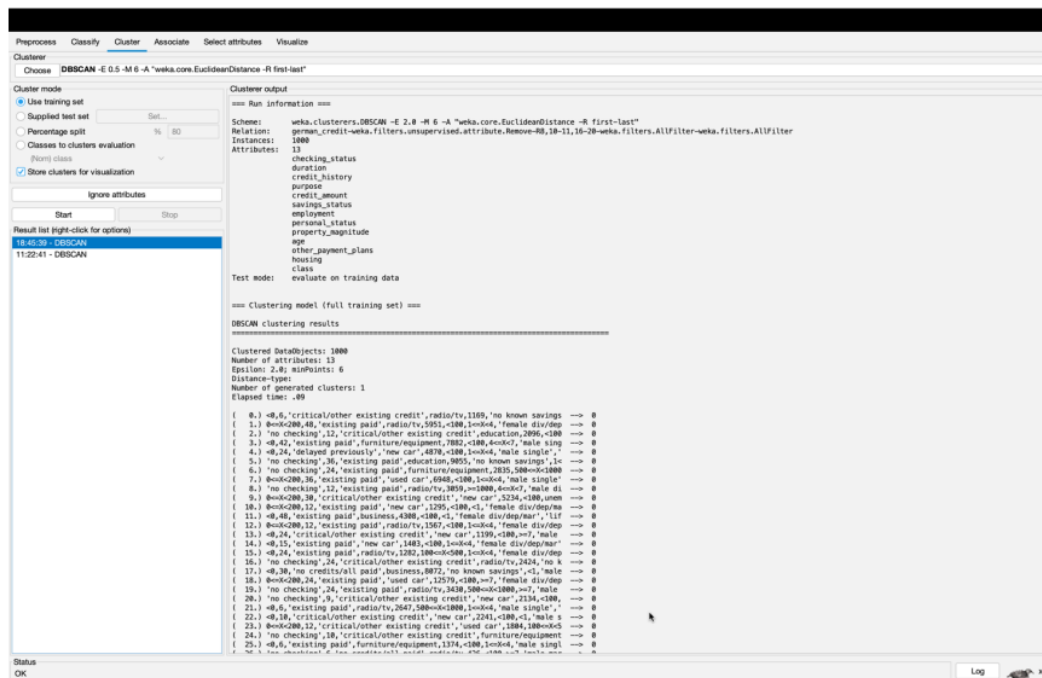


Figure 31 Second run of DBSCAN



In the second run of the DBSCAN algorithm, the epsilon was changed to 2.0 and the MinPoint was left at 6. This results in a reduction in the noise level and the few points that are still creating noise can be investigated using the visualisation option.

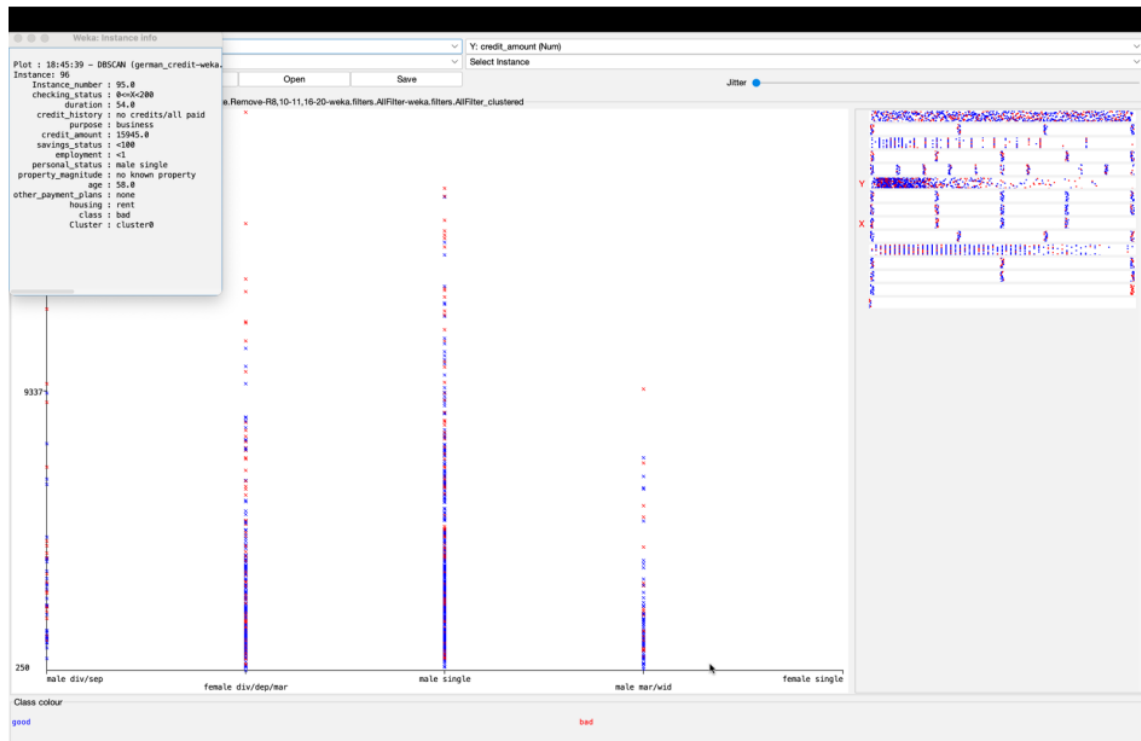
Figure 32 Remaining noise points

```

169.) 0<=X<200,24,'critical/other existing credit',business,1933,<100,>=7, --> 0
170.) <0,15,'no credits/all paid','new car',950,<100,>=7,'male single',car, --> 0
171.) 'no checking',12,'existing paid',furniture/equipment,763,<100,1<=X<4, --> 0
172.) 0<=X<200,24,'delayed previously',furniture/equipment,2064,<100,unempl --> 0
173.) 0<=X<200,8,'existing paid',radio/tv,1414,<100,1<=X<4,'male single','r --> 0
174.) <0,21,'delayed previously',education,3414,<100,<1,'male single','life --> 0
175.) 'no checking',30,'all paid','used car',7485,'no known savings',unempl --> NOISE
176.) <0,12,'existing paid',furniture/equipment,2577,<100,1<=X<4,'male div/ --> 0
177.) <0,6,'critical/other existing credit',radio/tv,338,500<=X<1000,>=7,'m --> 0
178.) 'no checking',12,'existing paid',radio/tv,1963,<100,4<=X<7,'male sing --> 0
179.) <0,21,'critical/other existing credit','new car',571,<100,>=7,'male s --> 0
180.) 'no checking',36,'delayed previously',business,9572,<100,<1,'male div --> 0

```

Figure 33 Outliers Personal status vs credit amount



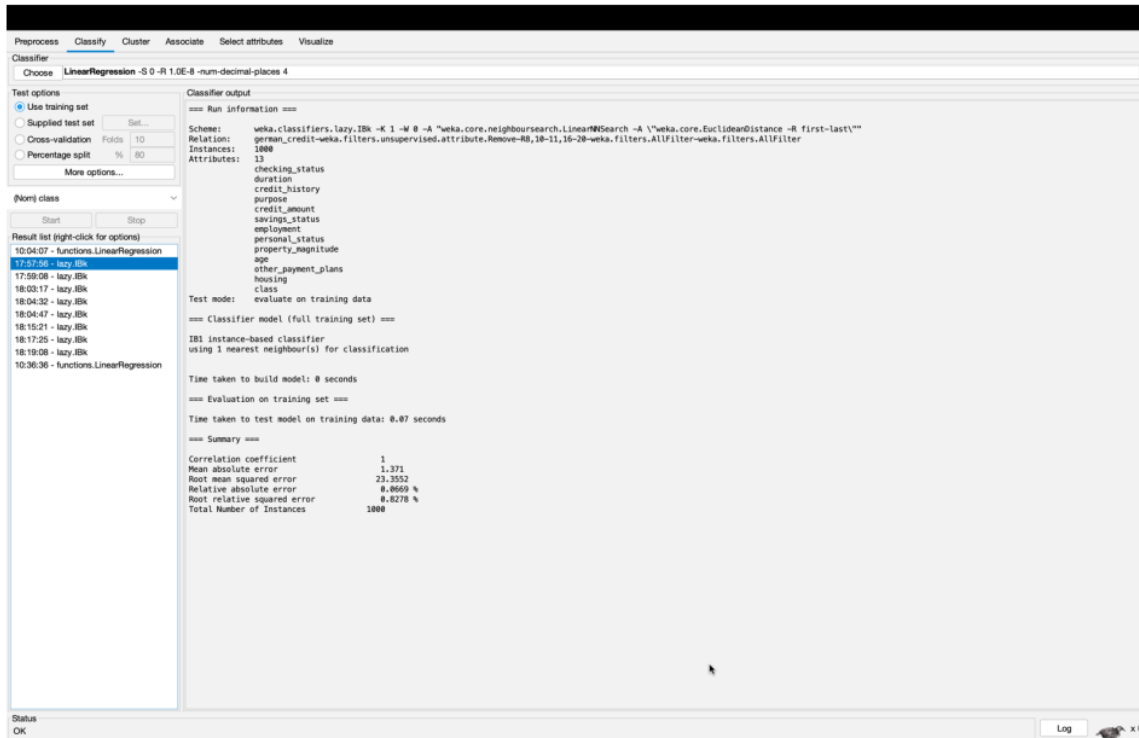
In the above figure, 'bad credit' outliers can be seen for female divorced/dependent/married applicants with loan amounts over 10,00 DMs but no single females have received any loan amounts from the bank.

5.3 K-Nearest Neighbour (KNN)

KNN is a non-parametric, lazy learning algorithm used for classification and regression tasks in machine learning. It makes predictions based on the similarity or distance between the new data point and pre-existing ones in the training set. KNN looks to find the 'k' closest data points (neighbours) within the training set to the new data point, then assign the new data point to the class or value that is most common among those 'k' neighbours [28]. This 'k' value is the parameter that needs to be tuned based on the results gained from running the algorithm. The advantages of using this algorithm for the analysis are:

- i. KNN is easy to understand and implement
- ii. It does not make any assumptions about the underlying distribution of the data
- iii. It is a 'lazy' learning algorithm, meaning that it does not build a model until a new data point needs to be predicted
- iv. Captures the relevance of features by evaluating the local neighbourhood data points, making it suitable for complex problems.

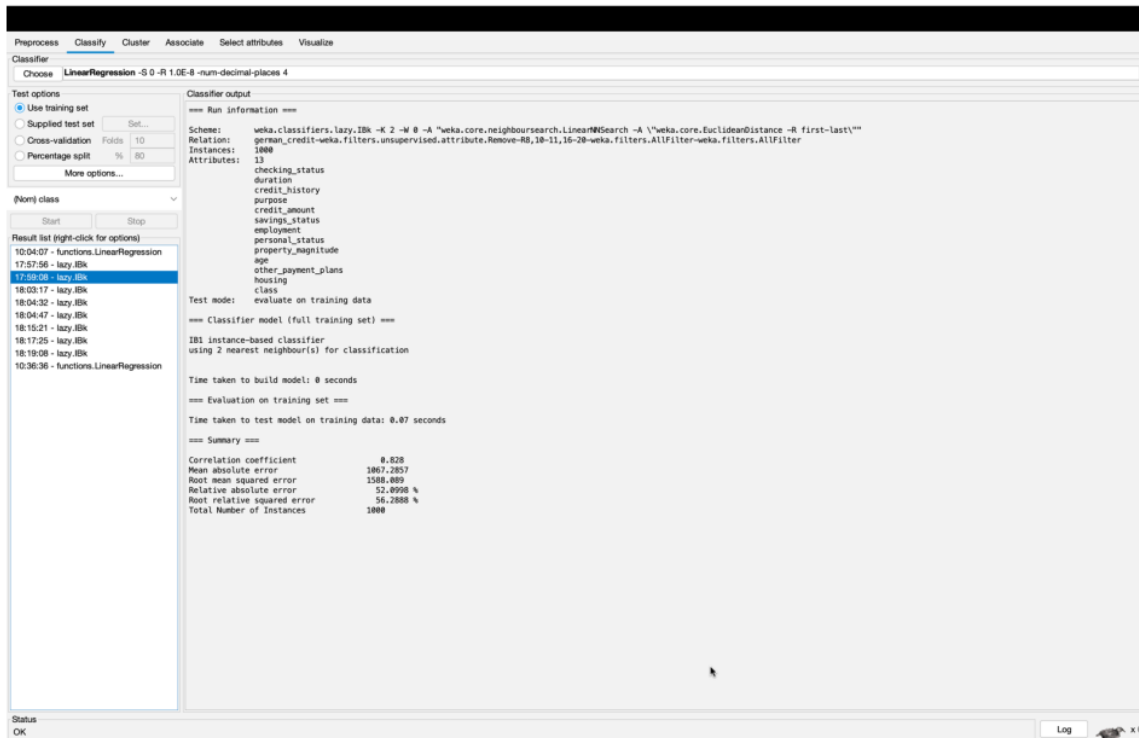
Figure 34 KNN parameter 1



The results from running with the KNN parameter set to the default 1:

- i. Correlation coefficient: 1.0 (perfect correlation)
- ii. Mean absolute error: 1.371
- iii. Root mean squared error: 23.3552
- iv. Relative absolute error: 0.0669%
- v. Root relative squared error: 0.08278%

Figure 35 KNN parameter 2



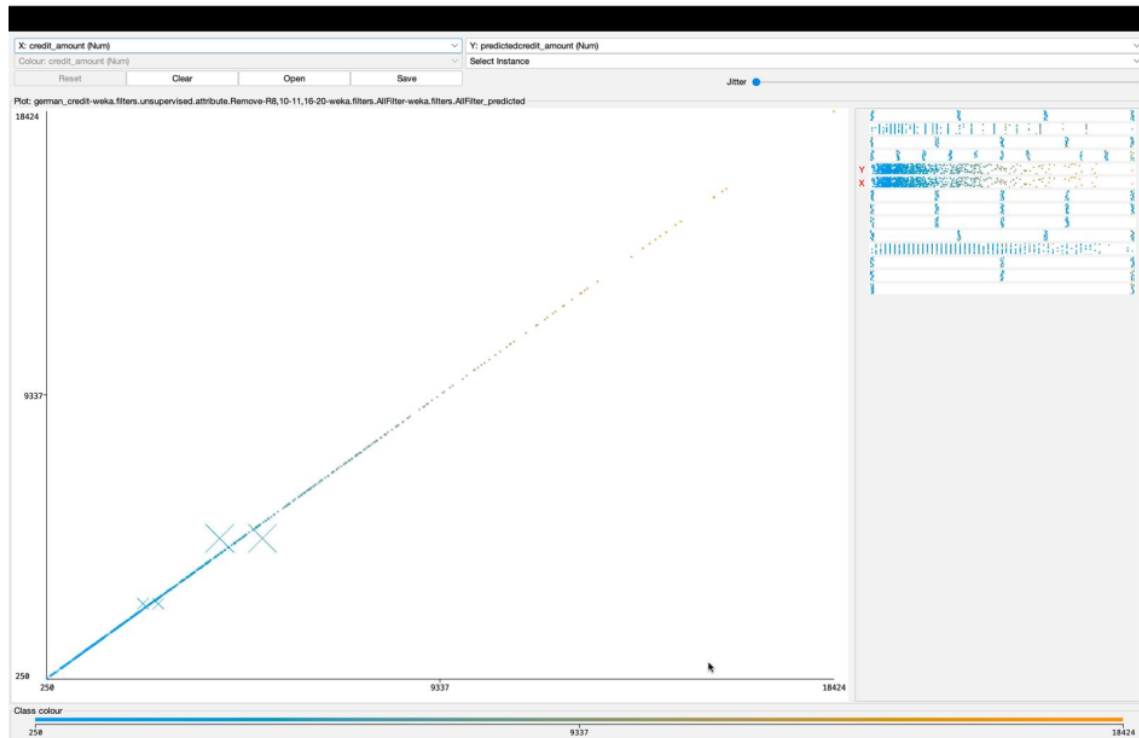
The results from running with the KNN parameter set to the default 2:

- i. Correlation coefficient: 0.828
- ii. Mean absolute error: 1067.2857
- iii. Root mean squared error: 1588.089
- iv. Relative absolute error: 52.0998%
- v. Root relative squared error: 56.2888%

The KNN 1 model has a perfect correlation coefficient of 1.0, indicating a strong positive correlation between the predicted and actual values. It also has significantly lower error rates across all metrics (mean absolute error, root mean squared error, relative absolute error, and root relative squared error) when compared to the KNN 2 model.

The lower error rates and a higher correlation coefficient suggest that the KNN 1 model is more accurate in predicting the target variable (class) based on the given features. Therefore, the KNN model with 1 nearest neighbour (KNN 1) has better accuracy than the KNN model with 2 nearest neighbours (KNN 2) on the training dataset.

Figure 36 Credit amount vs predicted credit amount



The above figure shows that the majority of predictions are on the linear regression line with 4 outliers, the 2 outliers shown in the figures below, both show loans to unemployed, males under 30 years of age, for new cars. These would need further analytics to discover why loans would be granted to these outliers.

Figure 37 Purpose vs predicted credit amount

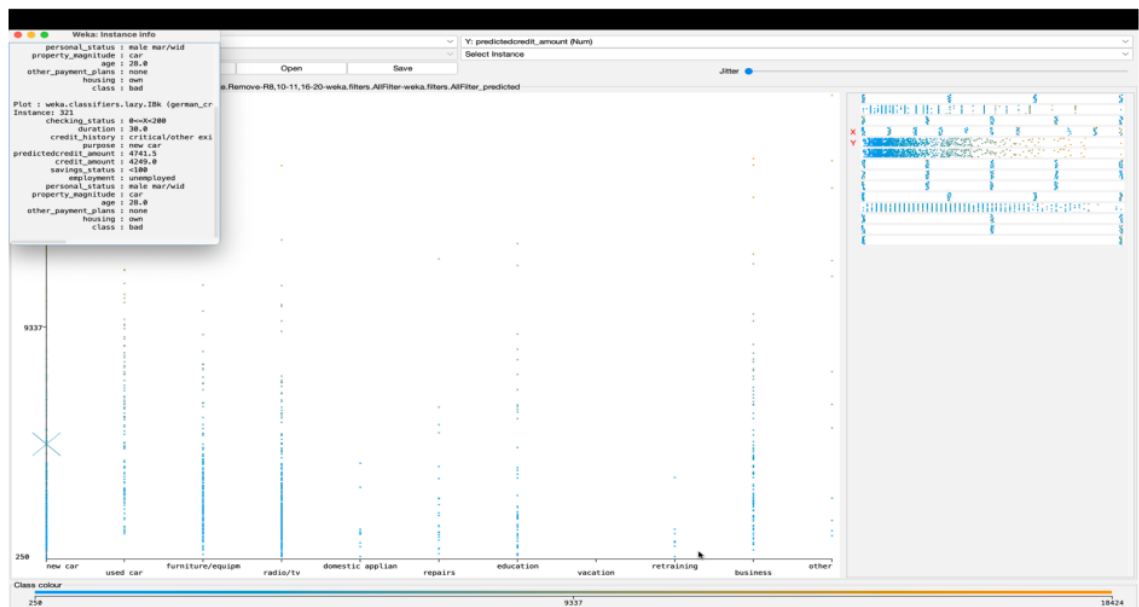


Figure 38 Outlier 1

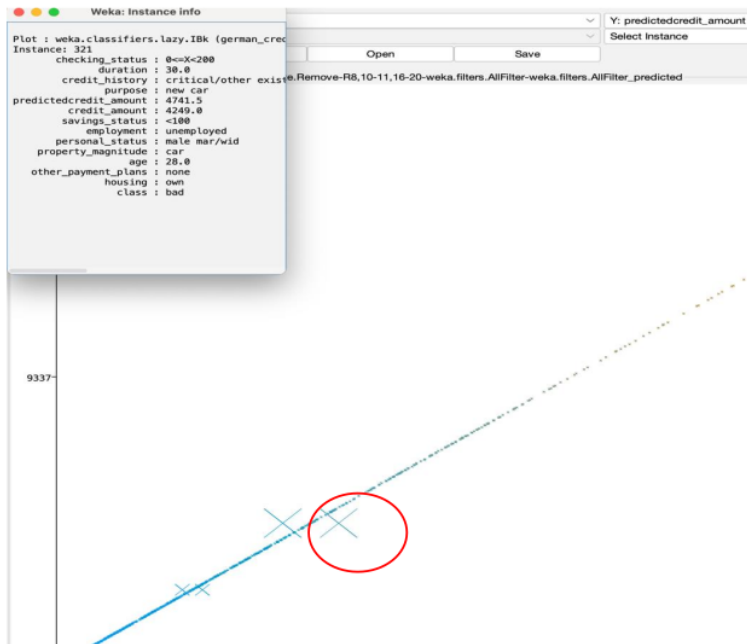
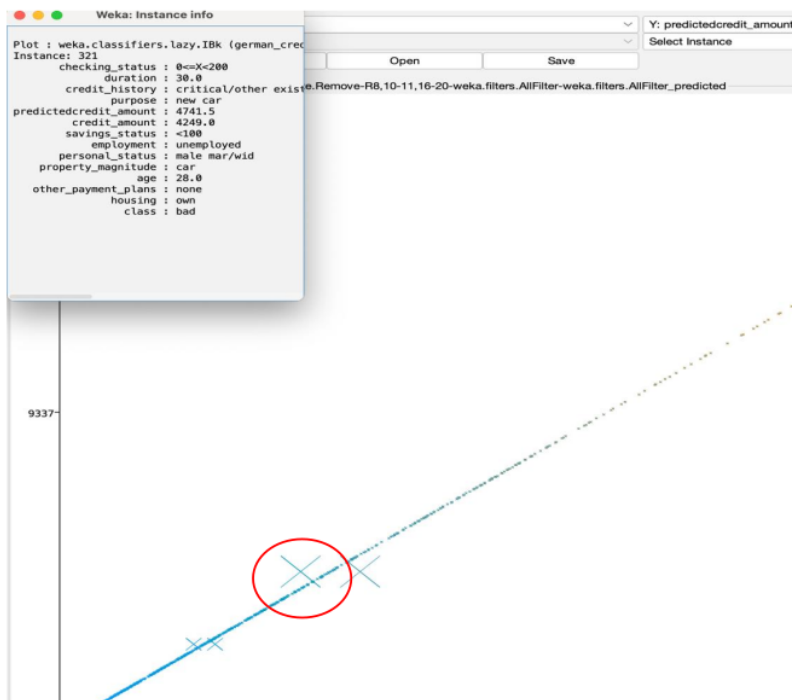


Figure 39 Outlier 2



5.4 Random Trees

Create decision trees by introducing randomness into the construction process. Each tree is built using a bootstrap sample of the training data and a random subset of features is considered for splitting at each node [30]. The advantages of using a random tree are:

- It helps to reduce the correlation between individual trees for a more robust and accurate model
- Resistant to overfitting
- Able to manage missing values and capture non-linear relationships.

Figure 40 Random tree



The above figure shows the results from the random tree algorithm:

- Correctly Classified Instances: 1000 (100%)
- Incorrectly Classified Instances: 0 (0%)

This indicates that the random tree model correctly classified all 1000 instances in the dataset, achieving 100% accuracy. The Kappa statistic measures the agreement between predicted and actual classes, correcting for random chance. A value of 1 indicates perfect agreement. (figure 41)

- Mean absolute error: 0
- Root mean squared error: 0
- Relative absolute error: 0%
- Root relative squared error: 0%

These error metrics all have a value of 0, indicating that the model made no errors in its predictions.

The results change when the data is split into training and test sets. (figure 42)

Figure 41 Cost benefit Analysis

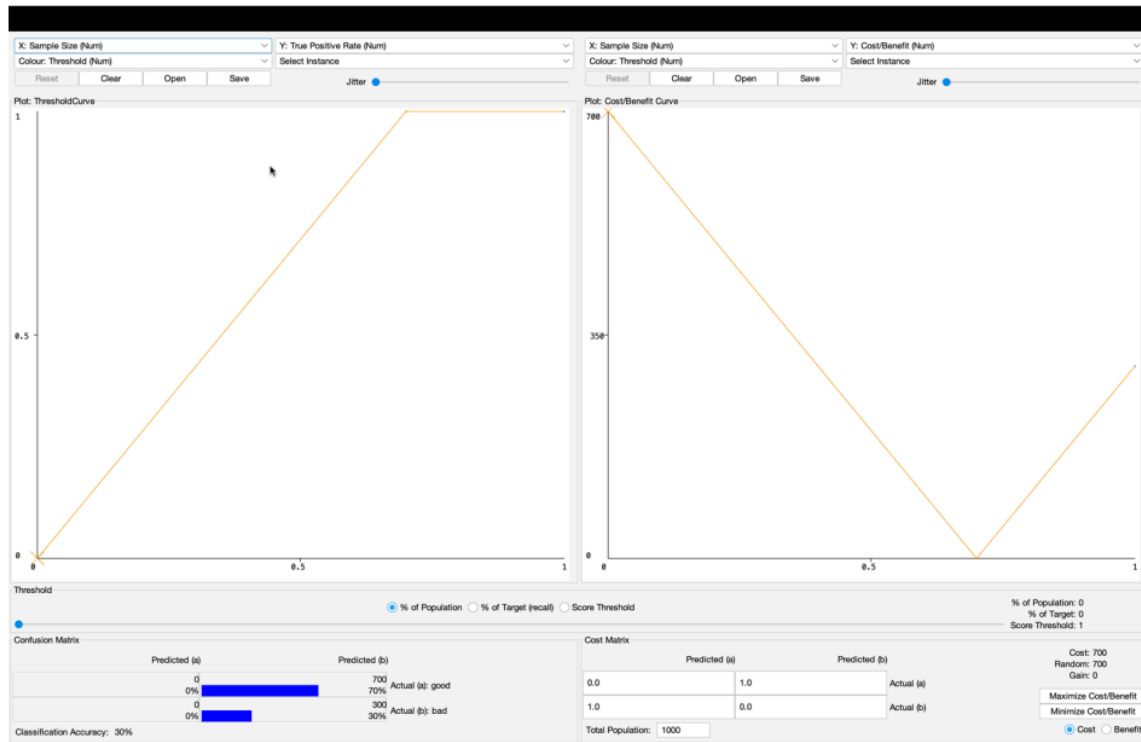


Figure 42 Training and test random tree



The key differences between the two results are:

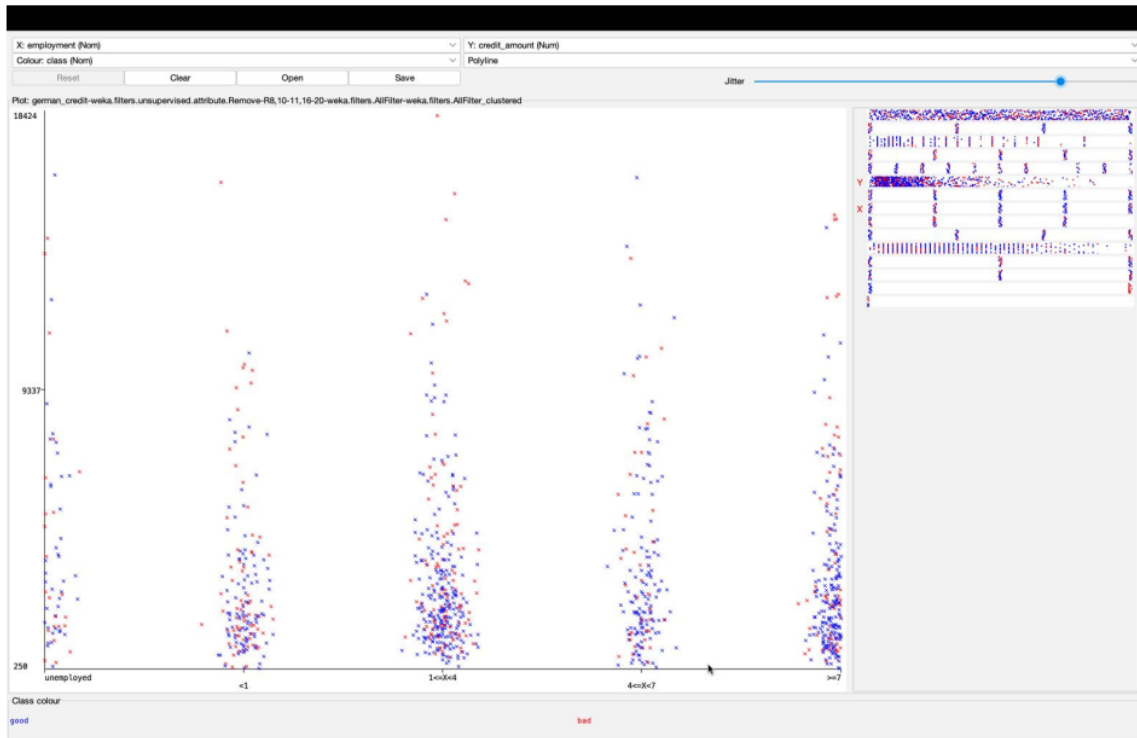
- This model uses only on 68% of the instances compared 100% previously
- The Kappa statistic is now 0.3517 which still indicates a fair agreement between the predicted and actual classes
- This model does highlight non-zero error values
- Accuracy by class shows it correctly identifies 64% of 'good' credit with some false positives. For 'bad' credit the new model has a true positive of 0.773 and a precision of 0.479 which means it correctly identifies 77.3% of 'bad' credit.

The overall accuracy of the model is 68% with better performance on identifying bad credit than it does good credit.

6 Business Recommendations

One notable observation is that unemployed applicants or those employed for under a year have a much higher risk of default. Almost 75% of applicants in these categories defaulted on their loans. Considering someone's employment stability and history is important, as it directly correlates with their ability to repay debt. The credit institution should avoid or carefully evaluate loan applications from unemployed individuals or those with short time in employment.

Figure 43 Employment vs credit amount



Past credit history is also a strong predictor of future credit repayment. Applicants with delays in paying off previous credits or existing credits at other institutions defaulted over 60% of the time. Thoroughly checking an applicant's past credit performance and making lending decisions accordingly could significantly reduce defaults. Customers demonstrating responsible credit use in the past are more likely to repay future loans. (figure 44)

Married applicants, regardless of gender, seem to have much lower default rates than divorced, separated, or single applicants. This makes intuitive sense, as married individuals likely have a partner sharing living expenses and resources to repay debt. However, it should be stated that the data is imbalanced as there are no single female loan applicants this hypothesis will need to be further investigated with additional data to cover all categories. (figure 45)

The purpose of the loan also shows those applicants seeking funds for cars, home repairs, or education typically repay, whilst those seeking business loans or consolidating debts have high default probabilities. Considering the loan's intended use and whether it supports necessities or discretionary spending allows better risk assessment. (figure 46)

Another important factor is the small size of the dataset, future analysis should ensure that the dataset should include at least 20 thousand instances to ensure that the data is not imbalanced. By regularly re-analysis the data the company will stay on top of any fluctuating patterns. The company needs to improve the risk model basing it on real-world performance (i.e. ownership of telephone), allowing them to make informed credit decisions, and minimise losses due to non-payment.

Figure 44 Credit history

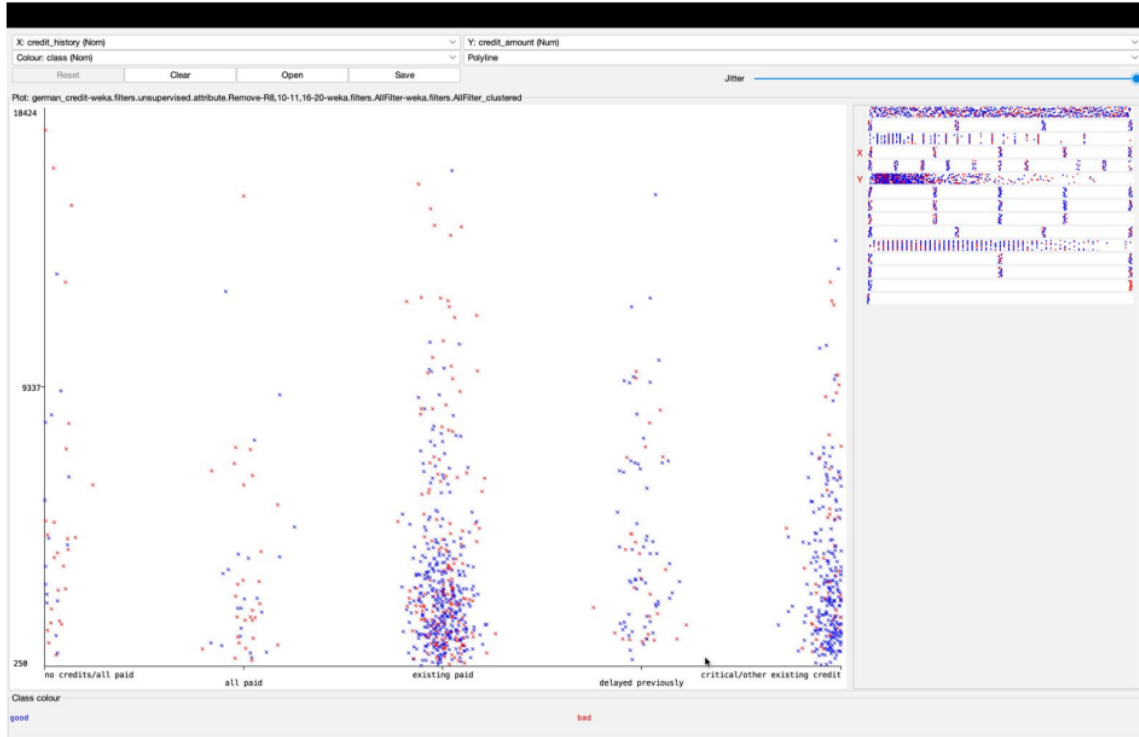


Figure 45 Married

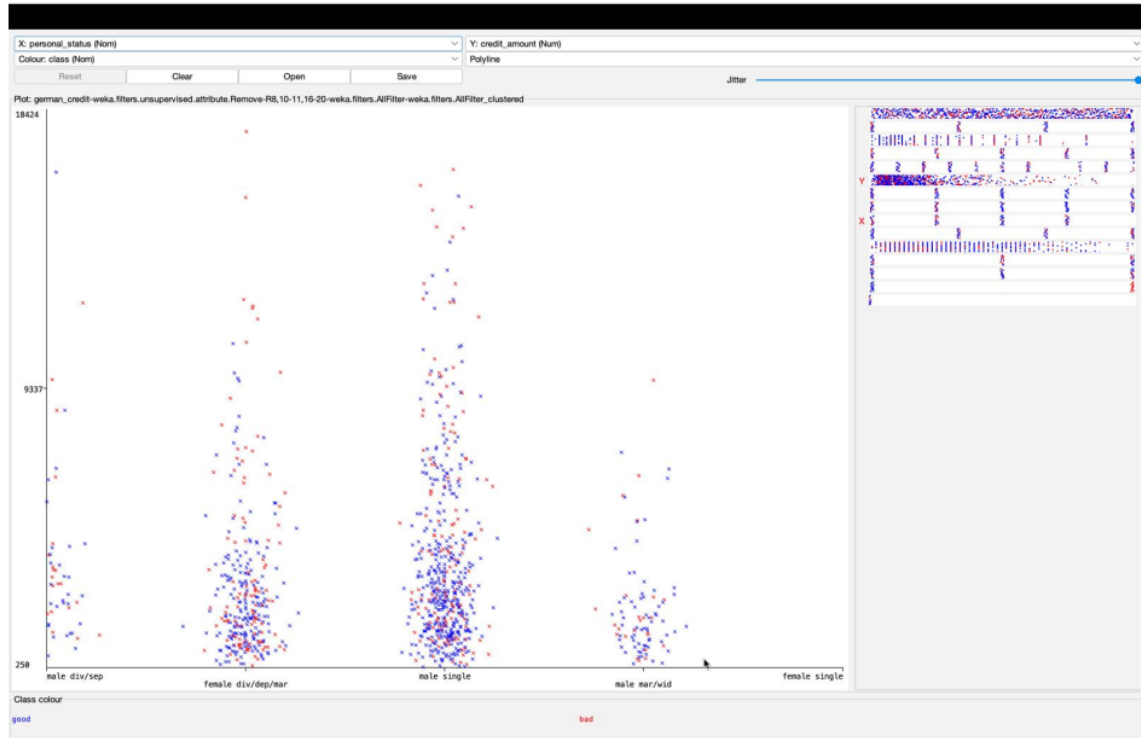
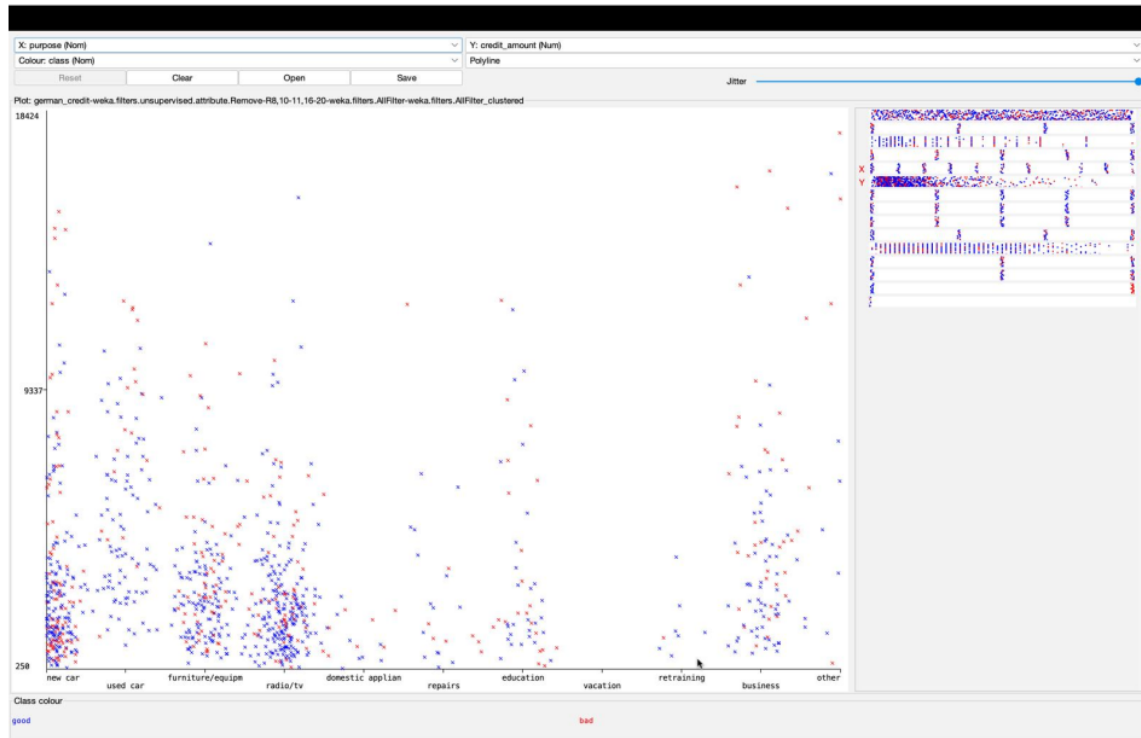


Figure 46 Purpose



7 Reflections

The text analysis provided some useful insights into different machine learning algorithms that could be applied to a loan approval prediction problem. Linear regression, KNN, decision trees and random forests were explored which are commonly used supervised learning methods. Key variables like age, employment, credit history and loan purpose were identified as important factors to analyse. The analysis evaluated different parameter settings and provided error and accuracy metrics to assess model performance. I enjoyed using WEKA and finding my way around the software, never having used software for data mining it opened my eyes to what can be accomplished.

However, there are also some limitations:

- i. The size of the dataset used was small and this might have caused overfitting and imbalanced results
- ii. There was only time to use a small percentage of all the algorithms available in WEKA and I would have liked to be able to explore a wider range of models like Naïve Bayes or neural networks.
- iii. Testing the models on a separate holdout dataset was also only briefly mentioned. More rigorous testing is needed to validate the findings.

In the future, I would like to see a larger and more realistic dataset obtained to perform a more comprehensive analysis. Ensemble and deep learning methods could leverage more predictive signals. Proper training, validation and testing protocol would strengthen confidence in any business recommendations. While some valuable starting points were identified, more advanced analytical methods applied to better-quality data would likely produce even stronger insights.

8 Bibliography

- [1] "South German Credit," UC Irvine Machine Learning Repository, 1st August 2019. [Online]. Available: <https://archive.ics.uci.edu/dataset/522/south+german+credit>. [Accessed 27th April 2024].
- [2] "Analysis of German Credit Data," Online Stat Psu Edu, [Online]. Available: <https://online.stat.psu.edu/stat508/book/export/html/796>. [Accessed 27th April 2024].
- [3] I. Witten, E. Frank and M. Hall, *Data Mining Practical Machine Learning Tools and Techniques*, San Francisco: Elsevier, 2011.
- [4] R. Wolff, "10 Best Data Mining Tools in 2022," MonkeyLearn, 22 December 2020. [Online]. Available: <https://monkeylearn.com/blog/data-mining-tools/>. [Accessed 22nd April 2024].
- [5] "The Data Platform for the Cloud and AI Era," Weka.io, 2024. [Online]. Available: https://www.weka.io/?utm_medium=cpc&utm_source=google&utm_term=weka&utm_campaign=&_bt=656194683085&_bk=weka&_bm=p&_bn=g&_bg=149262287918&qgad=656194683085&qgterm=weka&qgad_source=1&gclid=CjwKCAjwrcKxBhBMEiwAIVF8rFr5wmNKrH1gKnOJ1BmOKuNPH_OuPPXk9fZDCYlpxBBqG. [Accessed 30th April 2024].
- [6] P. Bhatia, *Data Mining and Data Warehousing*, Cambridge: Cambridge University Press, 2019.
- [7] R. Chen, "The application of data mining in data analysis," 2023.
- [8] M. Mulin, "The research and analysis of data mining technology," *Academic journal of science and technology*, vol. 3, no. 3, pp. 127 - 129, 2022.
- [9] H. Alshareef, A. Mairashi, M. Helal and M. Tahir, "Knowledge Extraction and Data Visualisation: A Proposed Framework for Secure Decision Making using Data Mining," *Internation Journal of Advanced Computer Science and Applications*, vol. 12, no. 8, 2021.
- [10] "How is Data mining used in marketing," CompTIA, [Online]. Available: <https://www.comptia.org/content/articles/how-is-data-mining-used-in-marketing>. [Accessed 27th April 2024].
- [11] IBM, "Using Data Mining to detect insurance fraud," 2020. [Online]. Available: <https://www.ibm.com/downloads/cas/7ZY8N46D>. [Accessed 27th April 2024].
- [12] Snowflake, "Data mining for business analytics," 2022. [Online]. Available: <https://www.snowflake.com/trending/data-mining-for-business-analytics/>. [Accessed 27th April 2024].
- [13] D. Mironov, "What is Social Media," Improvado, 2nd February 2024. [Online]. Available: <https://improvado.io/blog/what-is-social-media-data-mining#:~:text=Social%20media%20data%20mining%20involves,user%20behavior%20to%20uncover%20patterns..> [Accessed 27th April 2024].
- [14] geekforgeeks, "Data Mining: Data attributes and quality," geek for geeks, 6th May 2023. [Online]. Available: <https://www.geeksforgeeks.org/data-mining-data-attributes-and-quality/>. [Accessed 28th April 2024].
- [15] J. Brownlee, "Develope a Model for the Imbalanced Classification of Good and Bad Credit," *Machine Learning Mastery*, 5th January 2021. [Online]. Available: <https://machinelearningmastery.com/imbalanced-classification-of-good-and-bad-credit/>. [Accessed 30th April 2024].
- [16] "Imbalanced Data," *Machine Learning*, 9th June 2023. [Online]. Available: <https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data>. [Accessed 30th April 2024].
- [17] U. Lal, "Mastering Loan default prediction tackling imbalanced dataset for effective risk assessment," *Medium.com*, 14th April 2023. [Online]. Available: <https://medium.com/geekculture/mastering-loan-default-prediction-tackling-imbalanced-datasets-for-effective-risk-assessment->

8e8dfb2084d0#:~:text=The%20data%20received%20by%20the,balanced%20datasets%20f or%20accurate%20predictions.. [Accessed 30th April 2024].

- [18] "Classification on imbalanced data," Tensorflow.org, 23rd March 2024. [Online]. Available: https://www.tensorflow.org/tutorials/structured_data/imbalanced_data. [Accessed 30th April 2024].
- [19] M. Bilal, G. Iqbal, M. Anwar and K. R. Malik, "Auto-Prep: Efficient and Automated Data Preprocessing Pipeline," *IEEE Access*, vol. 10, pp. 107764-107784, 2022.
- [20] P. Gemperline, "Principle component analysis," *Practical guide to chemometrics*, vol. 2, pp. 69-104, 2006.
- [21] S. Kamath, "A small problem with big numbers," Towards science, 5th October 2020. [Online]. Available: <https://towardsdatascience.com/the-small-problem-with-big-numbers-4f3dad23ce01>. [Accessed 5th May 2024].
- [22] P. Bhatia, Data Mining and Data Warehousing, Cambridge: Cambridge University Press, 2019.
- [23] "Mobile connections in Germany 2023," DataReportal, 2023. [Online]. Available: <https://datareportal.com/reports/digital-2023-germany#:~:text=Data%20from%20GSMA%20Intelligence%20shows,total%20population%20in%20January%202023..> [Accessed 6th May 2023].
- [24] L.-H. Chen, "Research on Distributed Software of College Online Education Based on Information Gain Optimization Algorithm," *Internation Journal of Science and Engineering Applications*, 2023.
- [25] D. Rabideau and R. Wang, "Randomization based confidence intervals for cluster randomization trials," *Biostatistics*, pp. 1 - 28, 2021.
- [26] R. Joseph, "Optimal ratio for data splitting," *Statistical analysis and data mining*, vol. 15, no. 4, pp. 531-538, 2022.
- [27] The Investiopedia Team, Investiopedia, 19th March 2024. [Online]. Available: <https://www.investopedia.com/terms/p/paretoprinciple.asp>. [Accessed 9th May 2024].
- [28] A. Geron, Hands on Machine Learning with Scikit-learn, Keras and TensorFlow, Sebastopol: O Reilly, 2019.
- [29] P. Schober, "Correlation Coefficients: Appropriate Use and Interpretation," *Anesthesia & Analgesia*, vol. 126, no. 5, pp. 1763 - 1768, 2018.
- [30] Y. Religia, G. Pranto and E. Santosa, "South German Credit Data Classification Using Random Forest Algorithm to Preditc Bank Credit Receipts," *J. Inf. Appl*, vol. 3, no. 2, p. 837, 2020.

9 Appendix

9.1 South German Credit CSV file



South_german.csv

9.2 Data uploaded onto WEKA

WEKA interface showing the 'Preprocess' tab with the 'Randomize-S 42' filter applied. The 'Current relation' is 'german_credit' with 1000 instances and 21 attributes. The 'Selected attribute' is 'checking_status' with 4 distinct values and a nominal type.

No.	Label	Count	Weight
1	<0	274	274
2	0<=X<200	269	269
3	>=200	63	63
4	no checking	394	394

The 'Class: class (Nom)' is visualized as a stacked bar chart with blue and red segments. The counts for each class are: 274 (blue), 269 (blue), 63 (blue), and 394 (blue).

Status: OK

Assignment2_Kw

GRADEMARK REPORT

FINAL GRADE

GENERAL COMMENTS

73/100

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17

PAGE 18

PAGE 19

PAGE 20

PAGE 21

PAGE 22

PAGE 23

PAGE 24

PAGE 25

PAGE 26

PAGE 27

PAGE 28

PAGE 29

PAGE 30

PAGE 31

PAGE 32

PAGE 33

PAGE 34

PAGE 35

PAGE 36

PAGE 37

PAGE 38

PAGE 39

PAGE 40

PAGE 41

PAGE 42

PAGE 43

