

Data Mining

by Iwan Dyke

Submission date: 13-May-2024 08:23AM (UTC+0100)

Submission ID: 233361226

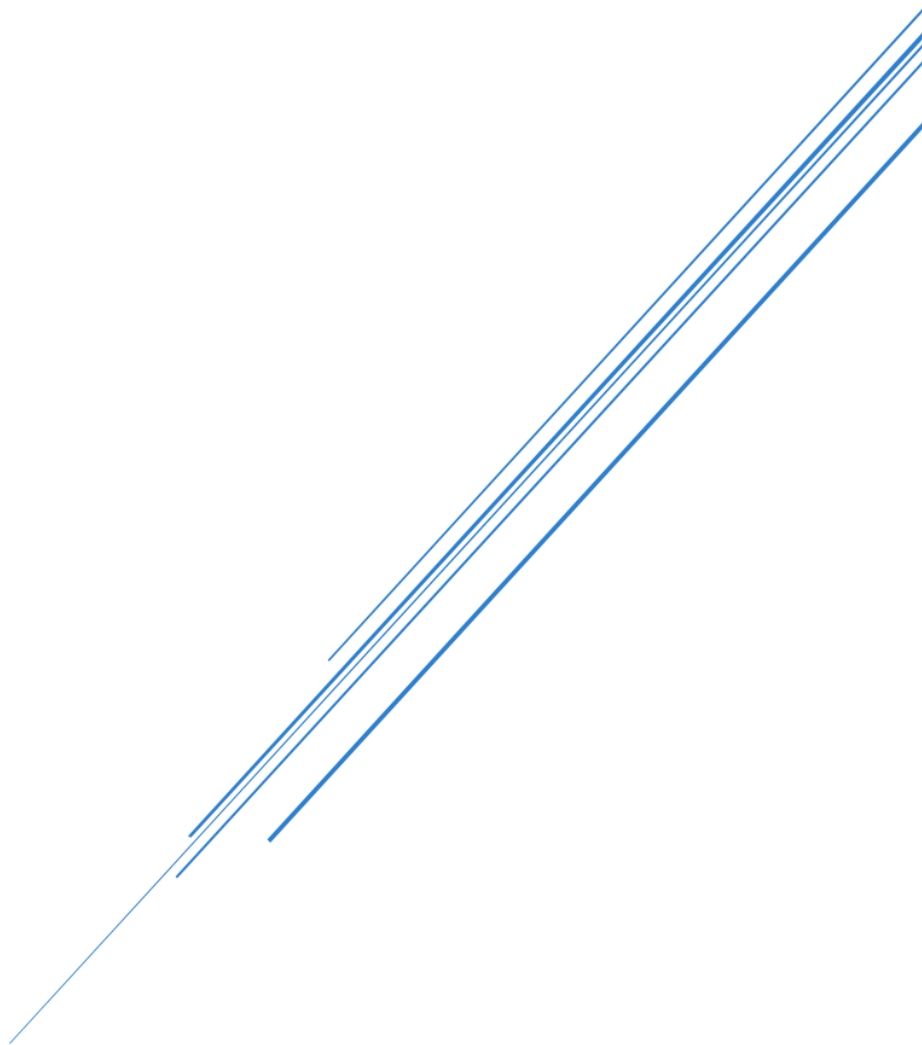
File name: 24865_Iwan_Dyke_Data_Mining_565911_1563544061.docx (1.97M)

Word count: 3608

Character count: 19105

DATA WAREHOUSING AND DATA MINING

ACCA6026



UWTSD
Iwan Dyke

Contents

Introduction	2
Dataset Overview	2
Data Preparation	5
Pre-processing Techniques	5
Data Cleaning.....	5
Data Transformation	5
Final Dataset Description	6
Dataset Snapshot	6
Methodology	7
Random Forest.....	7
Logistic Regression	7
Naïve Bayes	7
Analysis and Results.....	8
Random Forest Analysis	8
Logistic Regression Analysis.....	9
Naïve Bayes Analysis.....	10
Business Insights and Recommendations	10
Car Loans.....	11
Employment and Duration of Credit.....	12
Personal Status against Age.....	13
Saving Status and Age	14
Installment Commitment and Credit Amount.....	15
Purpose and Age	16
Reflective Commentary	16
Assessment of Method Effectiveness (Accuracy, Efficiency)	16
Advantages and Disadvantages of Approaches Used.....	17
Lessons Learned and Possible Improvements.....	17
Conclusion	17
References	17
Figure 1 (Data Column Exploration)	5
Figure 2 (ARFF File Snapshot)	6
Figure 3 (Random Forest Numerical Analysis)	8
Figure 4 (Logistic Regression Numerical Analysis).....	9
Figure 5 (Naïve Bayes Numerical Analysis).....	10
Figure 6 (Loan Purpose Against Credit Amount).....	11

Figure 7 (Employment Duration Against Loan Duration)	12
Figure 8 (Personal Status Against Age)	13
Figure 9 (Saving Status Against Age)	14
Figure 10 (Installment Commitment Against Credit Amount)	15
Figure 11 (Purpose Against Age).....	16

Introduction

Data mining is an analytical process that is used to explore large data sets to discover patterns and explore actionable insights. Various techniques can be applied to enable the extraction of predictive information which can be used to guide decision making across a business. Many common machine learning algorithms can be used such as clustering which groups similar data items. Classification which assigns data to categories which have been predefined, and regression which can predict continuous outcomes. [10]

Data mining can be pivotal for driving informed decision making which can enhance operational efficiency such as streamlining production, boost customer satisfaction such as grouping frequently bought products in a store, and overall increasing profitability. [5] The ability to predict customer behaviours help businesses to tailor their services promoting growth and customer satisfaction.

The main objective of this analysis is to explore the South German Credit Dataset and attempt to identify predictive factors that may influence a customer's ability to repay a loan.

Dataset Overview

The Dataset is obtained from the UCI machine learning repository which is a common source for machine learning datasets. The South Germany Credit Data is a publicly available which allows reproducibility in research. A description of the data can be found below.

Attribute	Data Type	Description	Value(s)	Data Values Count	Missing Data?
Checking Status	Nominal	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
Duration	Numerical	Duration of the loan in months	4 - 72		No
Credit History	Nominal	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	0,1,2,3,4,	0 = 40 1 = 49 2 = 530 3 = 88 4 = 293	No

		A33: delay in paying off in the past A34: critical account/other credits existing (not at this bank)			
Purpose	Nominal	What the load 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 8 = 9 9 = 97 10 = 12	No
Credit Amount	Numerical	Amount in DM's loaned to the applicant	250 – 18,424		No
Savings Status	Nominal	Debtor's savings A61: ... < 100 DM A62: 100 <= ... < 500 DM A63: 500 <= ... < 1000 DM A64: ... > 1000 DM A65: unknow/no savings account	1,2,3,4,5	1 = 603 2 = 103 3 = 63 4 = 48 5 = 183	No
Employment Duration	Nominal	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
Instalment Commitment	Numerical	Instalment rate in percentage of disposable income	1,2,3,4	1 = 136 2 = 231 3 = 157 4 = 476	No
Personal Status and Sex	Nominal	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
Other Parties	Nominal		1,2,3	1 = 907	

		Is there another person who is standing as a guarantor for the debt A101: none A102: co-applicant A103: guarantor		2 = 41 3 = 52	
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
Property Magnitude	Nominal	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
Age	Numerical	Age of the debtor at the time of the loan in years	19 - 75		No
Other Payment Plans	Nominal	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
Housing	Nominal	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
Existing Credits	Numerical	Number of credits including current one that the debtor has	1,2,3,4	1 = 633 2 = 333 3 = 28 4 = 6	
Job	Nominal	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

Number of Dependents	Numerical	The number of people who are financially reliant on the debtor.	1,2	1 = 155 2 = 845	No
Own Telephone	Nominal	Does the debtor have a landline registered in their name A191: none A192: yes	1,2	1 = 596 2 = 404	No
Foreign worker	Nominal	Is the debtor a foreign worker A201: yes A202: no	1,2	1 = 37 2 = 963 No missing values	No
Class	Nominal	Cost Matrix 1 = good 2 = bad	1,2	Good = 700 Bad = 300	No

Figure 1 (Data Column Exploration)

Data Preparation

Pre-processing Techniques

Data usually undergoes pre-processing before any analysis is performed to ensure that it is cleaned, and any outliers would not skew the results. It also ensures that the data is all in the correct format to be processed by mining and analysis tools.

Data Cleaning

Missing Values

One of the main issues with datasets are NULLs (unknown data) in the test data used for this practical there were no missing data values found in the test data set. However, in a real-world scenario it is highly probable that data would have missing values. There are a few different methods of dealing with this. Replacing numerical NULLs with the mean or median of the column, or simply omitting the row from the dataset in pre-processing.

Removing Outliers

It is possible to remove outliers by calculating the interquartile range (IQR) and then removing values which fall outside of 1.5 times the IQR from the first and third quartiles. By doing this it stabilizes the dataset and can improve the accuracy of analysis.

Data Transformation

Normalisation

Once the data has been cleansed transformations can take place to improve the reliability of the analysis. Normalisation is the process of scaling any numeric attribute so that they range between a specific interval, in this case 0 and 1. [7]

Feature Engineering

Feature engineering uses domain knowledge to create a new predictor from raw data which can help to improve the robustness and accuracy of different machine learning models. Feature engineering can leverage human intuition to compensate for limitations in machine learning algorithms. There are a few different techniques including transforming and standardising the variables. Can create new variables such as adding variables together to provide new insights.

Possible examples of feature engineering for the south German credit data:

- Debt-to-income ratio: Creating a feature that combines information from the borrower's debt and income. This is often used in credit scoring models.
- Age Groups: By converting age to a categorical field it can remove the bias that age and credit worthiness is not always linear.
- Interaction Terms: New features which represent interactions between existing features which could highlight new risks.

Final Dataset Description

Preprocessing measures were implemented to optimize data quality and enhance the suitability for predictive modelling. Missing values should have been addressed at this stage although none were found for this test data, categorical variables were encoded to improve analytical compatibility with multiple algorithms.

Dataset Snapshot

```
% Relabeled values in attribute class
% From: 1          To: good
% From: 2          To: bad
%
%Relation german_credit
%Attribute 'checking_status' ( '0', '0<=x<200', '>=200', 'no checking')
%Attribute 'duration' real
%Attribute 'credit_history' ( 'no credits/all paid', 'all paid', 'existing paid', 'delayed previously', 'critical/other existing credit')
%Attribute 'purpose' ( 'new car', 'used car', 'furniture/equipment', 'radio/tv', 'domestic appliance', 'repairs, education, vacation, retraining, business, other')
%Attribute 'credit_amount' real
%Attribute 'savings_status' ( '0', '100<=x<500', '500<=x<1000', '>=1000', 'no known savings')
%Attribute 'employment' ( unemployed, '1', '1<=x<4', '4<=x<7', '>=7')
%Attribute 'installment_commitment' real
%Attribute 'personal_status' ( 'male div/sep', 'female div/dep/mar', 'male single', 'male mar/wid', 'female single')
%Attribute 'other_parties' ( none, 'co applicant', guarantor)
%Attribute 'residence_since' real
%Attribute 'property_magnitude' ( 'real estate', 'life insurance', car, 'no known property')
%Attribute 'age' real
%Attribute 'other_payment_plans' ( bank, stores, none)
%Attribute 'housing' ( rent, own, 'for free')
%Attribute 'existing_credits' real
%Attribute 'job' ( 'unemp/unskilled non res', 'unskilled resident', skilled, 'high qualif/self emp/mgmt')
%Attribute 'num_dependents' real
%Attribute 'own_telephone' ( none, yes)
%Attribute 'foreign_worker' ( yes, no)
%Attribute 'class' ( good, bad)
%Data
'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
'0<=x<200',48,'existing paid',radio/tv,5951,'0',1,'0<=x<4',2,'female div/dep/mar',none,2,'real estate',22,none,own,1,skilled,1,none,yes,bad
'no checking',12,'critical/other existing credit',education,2096,'0',1,'4<=x<7',2,'male single',none,3,'real estate',45,none,own,1,'unskilled resident',2,none,yes,good
'0',42,'existing paid',furniture/equipment,7882,'0',1,'4<=x<7',2,'male single',guarantor,4,'life insurance',45,none,'for free',1,skilled,2,none,yes,good
'0',24,'delayed previously',new car,4870,'0',1,'0<=x<4',3,'male single',none,4,'no known property',53,none,'for free',2,skilled,2,none,yes,bad
'no checking',36,'existing paid',education,9055,'no known savings','0<=x<4',2,'male single',none,4,'no known property',35,none,'for free',1,'unskilled resident',2,yes,yes,good
'no checking',24,'existing paid',furniture/equipment,2835,'500<=x<1000','>=7',3,'male single',none,4,'life insurance',53,none,own,1,skilled,1,none,yes,good
'0<=x<200',36,'existing paid',new car,6948,'0',1,'0<=x<4',2,'male single',none,2,'car,35,none,rent,1,'high qualif/self emp/mgmt',1,yes,yes,good
'no checking',12,'existing paid',radio/tv,3055,'>=1000','4<=x<7',2,'male div/sep',none,4,'real estate',61,none,own,1,'unskilled resident',1,none,yes,good
'0<=x<200',36,'critical/other existing credit',new car,3234,'0',1,'unemployed,4,'male mar/wid',none,2,'car,28,none,own,2,'high qualif/self emp/mgmt',1,none,yes,bad
'0<=x<200',12,'existing paid',new car,1295,'0',1,'0',3,'female div/dep/mar',none,1,'car,25,none,rent,1,skilled,1,none,yes,bad
'0',48,'existing paid',business,4308,'0',1,'0',3,'female div/dep/mar',none,4,'life insurance',24,none,rent,1,skilled,1,none,yes,bad
'0<=x<200',12,'existing paid',radio/tv,1567,'0',1,'0<=x<4',1,'female div/dep/mar',none,1,'car,22,none,own,1,skilled,1,yes,yes,good
'0',24,'critical/other existing credit',new car,1195,'0',1,'>=7',4,'male single',none,4,'car,60,none,own,2,'unskilled resident',1,none,yes,bad
'0',15,'existing paid',new car,1403,'0',1,'0<=x<4',2,'female div/dep/mar',none,4,'car,28,none,rent,1,skilled,1,none,yes,good
'0',24,'existing paid',radio/tv,1282,'100<=x<500','0<=x<4',4,'female div/dep/mar',none,2,'car,32,none,own,1,'unskilled resident',1,none,yes,bad
'no checking',24,'critical/other existing credit',radio/tv,2424,'no known savings','>=7',4,'male single',none,4,'life insurance',53,none,own,2,skilled,1,none,yes,good
'0',30,'no credits/all paid',business,8072,'no known savings','0',2,'male single',none,3,'car,25,bank,own,3,skilled,1,none,yes,good
'0<=x<200',24,'existing paid',used car,12579,'0',1,'>=7',4,'female div/dep/mar',none,2,'no known property',44,none,'for free',1,'high qualif/self emp/mgmt',1,yes,yes,bad
'no checking',24,'existing paid',radio/tv,3430,'500<=x<1000','>=7',3,'male single',none,2,'car,31,none,own,1,skilled,2,yes,yes,good
'no checking',9,'critical/other existing credit',new car,2134,'0',1,'0<=x<4',4,'male single',none,4,'car,48,none,own,3,skilled,1,yes,yes,good
'0',6,'existing paid',radio/tv,2647,'500<=x<1000','0<=x<4',2,'male single',none,3,'real estate',44,none,rent,1,skilled,2,none,yes,good
'0',15,'critical/other existing credit',new car,2241,'0',1,'0',1,'male single',none,3,'real estate',48,none,rent,2,'unskilled resident',2,none,own,good
'0<=x<200',12,'critical/other existing credit',used car,1804,'100<=x<500','0',1,'3,'male single',none,4,'life insurance',44,none,own,1,skilled,1,none,yes,good
'no checking',10,'critical/other existing credit',furniture/equipment,2049,'no known savings','0<=x<4',2,'male mar/wid',none,1,'car,26,none,own,2,skilled,1,none,no,good
```

Figure 2 (ARFF File Snapshot)

Methodology

Three different classification techniques have been employed to analyse the south German credit data. Random forests, logistic regression, and support vector machines. These three analyses have different strengths in analysis capabilities. [11]

Random Forest

The random forest builds multiple distinct decision trees using different subsets of data, and then combines the predictions to make a final decision. [2] Using this algorithm reduces errors because each decision tree is based on smaller subsections of data. However, this algorithm has a known tendency to predict more false positives. Fine tuning the processing options such as the number of trees and the complexity could reduce the likelihood of false positives.

Multiple decisions trees are used for classification accuracy and control over fitting. Overfitting is when an algorithm is extremely accurate when used on training data but not for production or new data. Forests can handle a large number of input variables easily.

Logistic Regression

Logistic regression estimates probabilities using a logistic function which is good for binary classification tasks. Logistic functions input features to produce outcomes between 0 and 1. The probabilities represent the outcome of belonging to a particular class. In this case having a good or bad credit. This is typically split with below 0.5 being one class and above as another. Logistic regression directly shows the different attributes influence on the probability which is where the high level of interpretability is shown.

Logistic regression assumes a linear relationship and can struggle with datasets which are non-linear. Much like random forests this algorithm can be prone to overfitting on training datasets. With a smaller dataset such as the German credit, specific feature selection is crucial to omit this risk. [3]

Naïve Bayes

Naïve bayes is named as such because it assumes that all features included in the algorithm are equally as important. The class with the highest probability using bayes theorem is then selected as the prediction. It has a straightforward probabilistic framework. This is particularly useful for credit scoring classification. It also is used more for interpretability in cases such as financial data where the impact of these variables is crucial. While not the most accurate its computational efficiency is unmatched. However, this algorithm can struggle with limited data as the test dataset used only has 1000 rows this could fit into this category.

Naïve Bayes even though can be considered a simple model can in fact outperform more complex algorithms. It is a fast and reliable algorithm which provides better results when features are conditionally independent of one another. [4] The equation for bayes theorem is as follows:

$$P(C_k | x) = P(x | C_k) * P(C_k) / P(x)$$

Analysis and Results

Random Forest Analysis

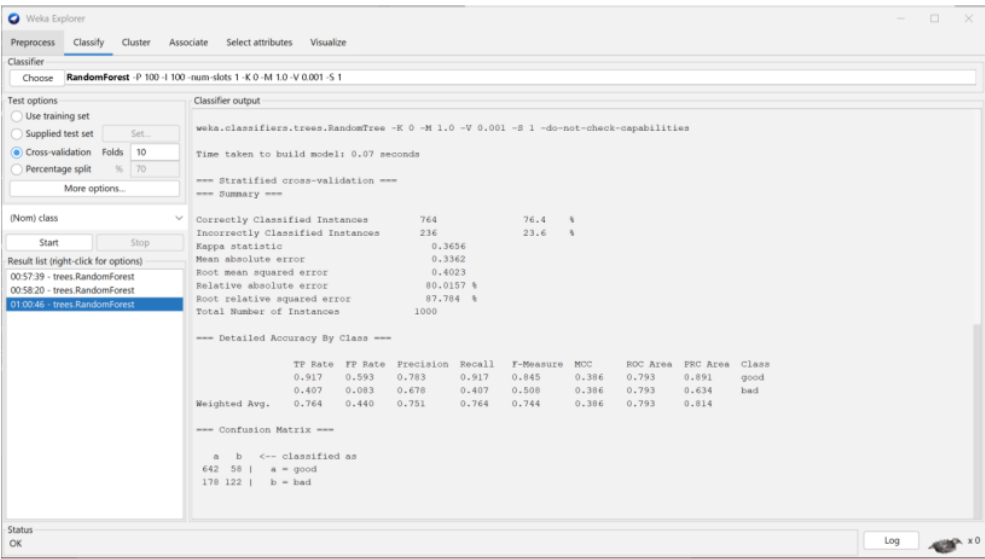


Figure 3 (Random Forest Numerical Analysis)

The classification accuracy of the random forest algorithm reached 76.4% identifying 764 instances while 236 were misclassified causing an error rate of 23.6%. This classification technique demonstrated a strong true positive rate of 91.7% when classifying good credit outcomes, however this method of analysis encountered a rather high false positive rate of nearly 60%. [6] The confusion matrix indicates a tendency of the model to predict more false positives than negatives potentially indicating an optimism bias when predicting creditworthiness. While this algorithm has correctly identified around 75% of the good credit scores it has a high error margin and leaves room for improvement.

Logistic Regression Analysis

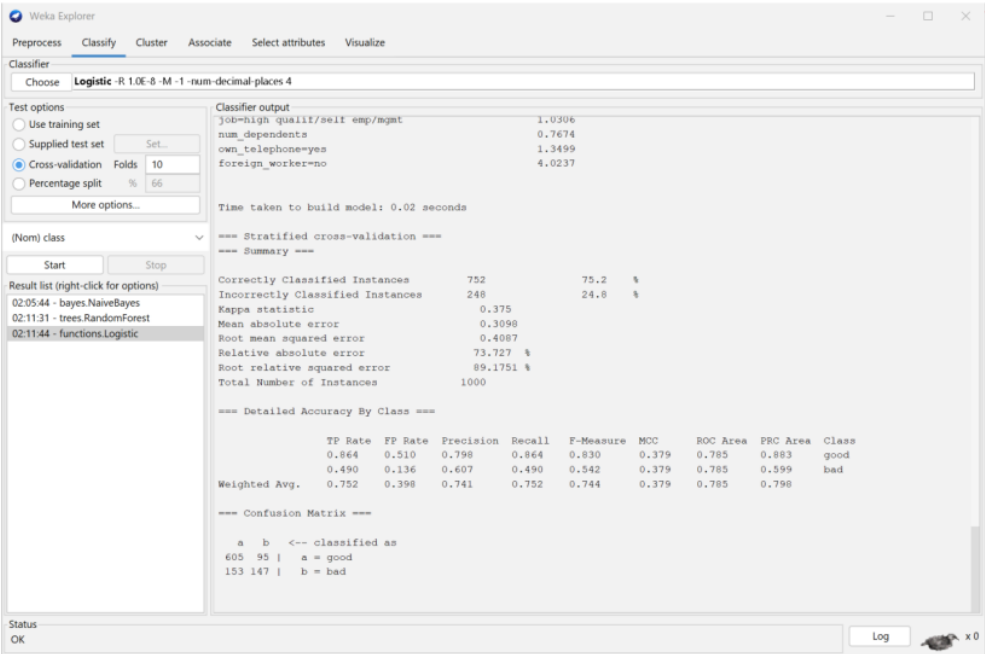


Figure 4 (Logistic Regression Numerical Analysis)

A slightly lower classification accuracy has been shown when using a logistic regression analysis down to 75%. The false positive rate reached around 50% for the logistic regression indicating it is potentially not the best method for analysing good or bad credit scores based on the data available. The model was able to easily discriminate between classes with a with an ROC area stood as 0.785. If this model was to be used in a production environment, then refinement would need to be conducted to improve accuracy.

Naïve Bayes Analysis

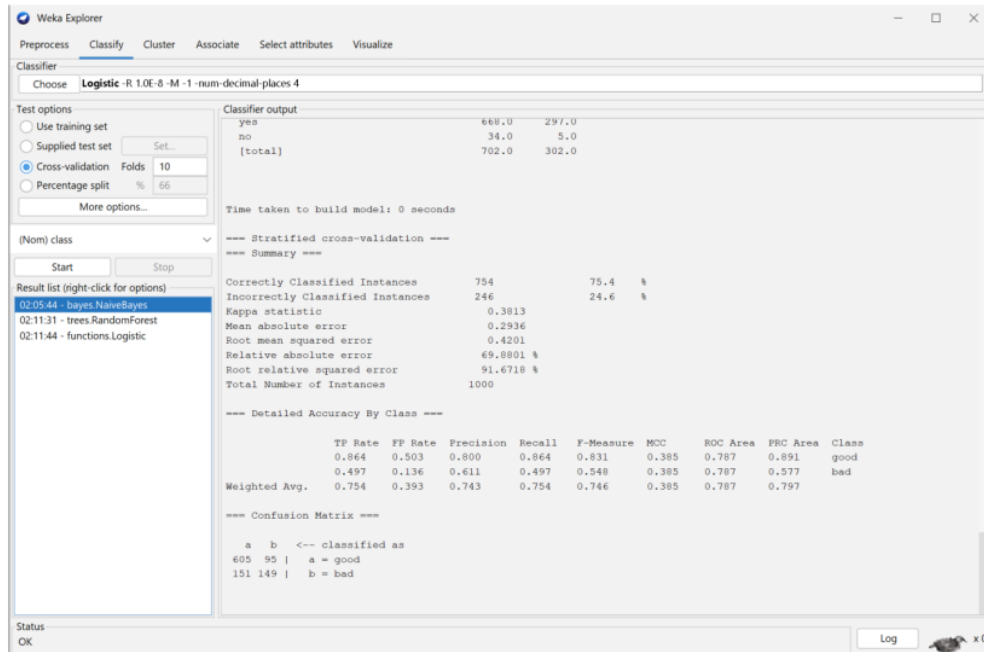


Figure 5 (Naive Bayes Numerical Analysis)

The last algorithm employed has a similar accuracy to logistic regression. However, this algorithm has the highest level of reliable predictions based on the kappa statistic compared to the others employed. Once again, the false positive rate for this algorithm was around 50% which in a financial setting can be quite risky. However, another positive from this algorithm is an f-measure of 83% which shows a balanced classification performance. This algorithm can be further refined to improve accuracy.

Business Insights and Recommendations

Due to the higher percentage of true positives found by the random forest algorithm that will be used for further analysis and looking into business insights which can be gained from the credit data.

Car Loans



Figure 6 (Loan Purpose Against Credit Amount)

As can be seen from figure 6, when specifically looking at car related loans between new and used. The trends shown here specifically related to new and used cars show a significant insight to the amount of the loan the purpose and if they are a good or bad credit risk. The visualisation clearly shows that smaller loans used for new cars have a high credit risk. While these cars are typically more reliable, they also have higher depreciation rates. It is possible the cars market value depreciates before the loan value. If the debtor's financial situation worsens then there's a higher chance of a default.

Employment and Duration of Credit

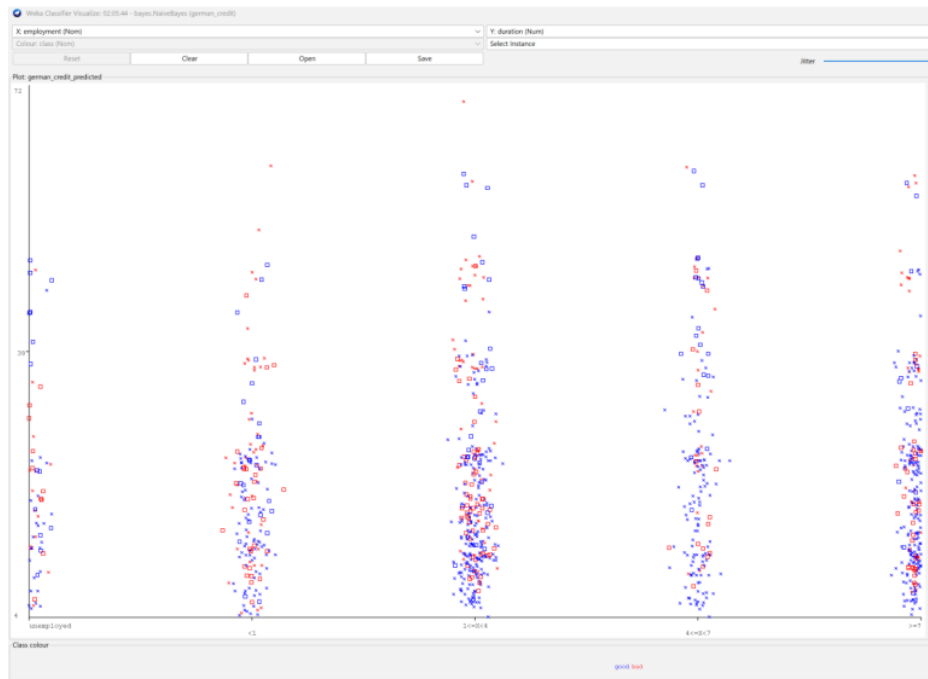


Figure 7 (Employment Duration Against Loan Duration)

Figure 7 shows that a lower employment duration implies a higher chance of a default and bad credit. Whereas those with a longer employment history have a lower chance of defaulting. It can be deduced that those with shorter employment history may have less financial stability. Employment between 1 – 4 years seems to be the highest density of people and has a similar distribution of good and bad credits.

Personal Status against Age

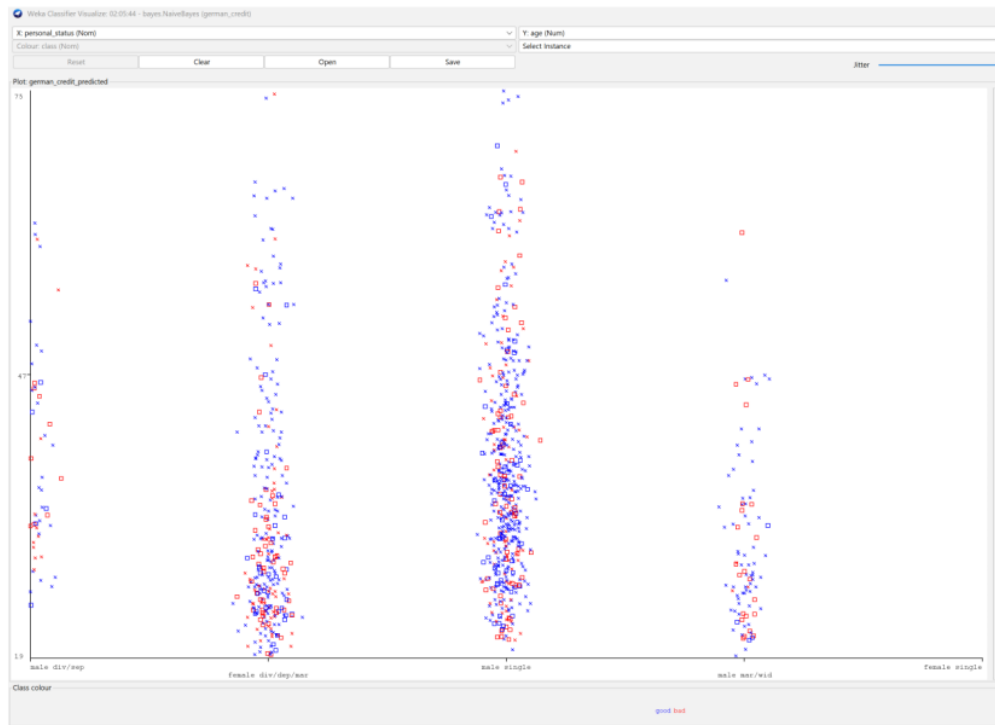


Figure 8 (Personal Status Against Age)

It is shown in Figure 8 that younger individuals when divorced or separated are more likely to have a bad credit, which correlates with lower financial stability or potential more financial obligations. These could be divorcing settlements or childcare payments. Opposite to that older individuals seem to have a better credit rating across the board. It should be noted that there is no data for single females. All other female categories are grouped together. Indicating that there are more males looking for loans than females. With younger females having a distinct separation of bad credit compared to older females.

Saving Status and Age

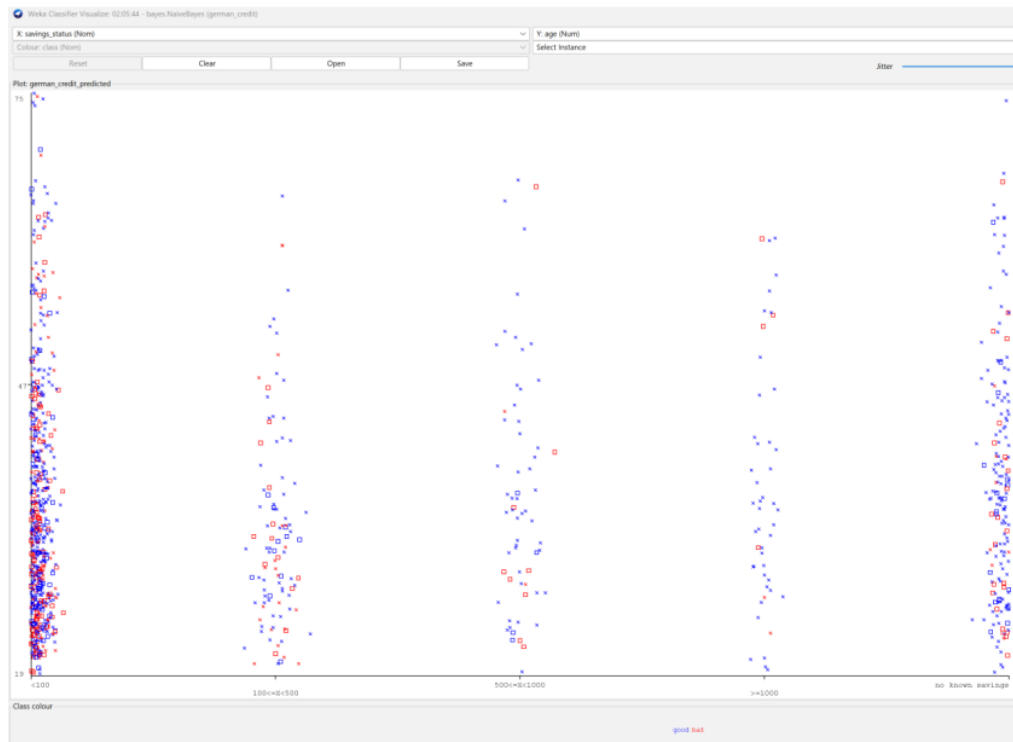


Figure 9 (Saving Status Against Age)

There is a clear distinction of younger debtors with no savings or savings under 100 having bad credit. However, there is a vastly low amount of debtors with savings who have bad credit potentially due to having more collateral. It should also be noted the significant number of debtors in the dataset which have little to no savings. A potential insight which could be gleaned from this is that those who have savings may not necessarily need to take out loans and would be in a stable financial position.

Installment Commitment and Credit Amount

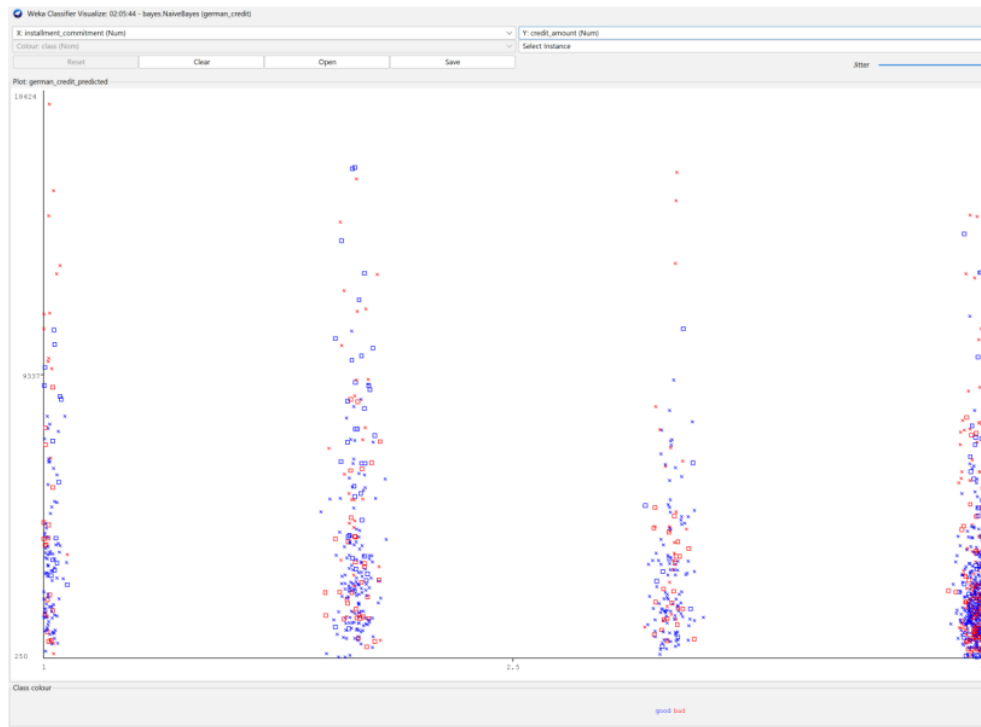


Figure 10 (Installment Commitment Against Credit Amount)

When plotting installment commitments and credit amount against each other, the smaller the loan amount and the shorter the duration of the loan the better chance of a good credit score. It clearly shows the longer the duration of the loan the higher chance of a default and a bad credit prediction. While a longer credit installment duration doesn't necessarily correlate with a default or bad credit outcome. There is a significant distribution of bad credit points in the 4 year and over bracket. Potentially showing that loans limited to under 4 years could be better for at risk debtors.

Purpose and Age

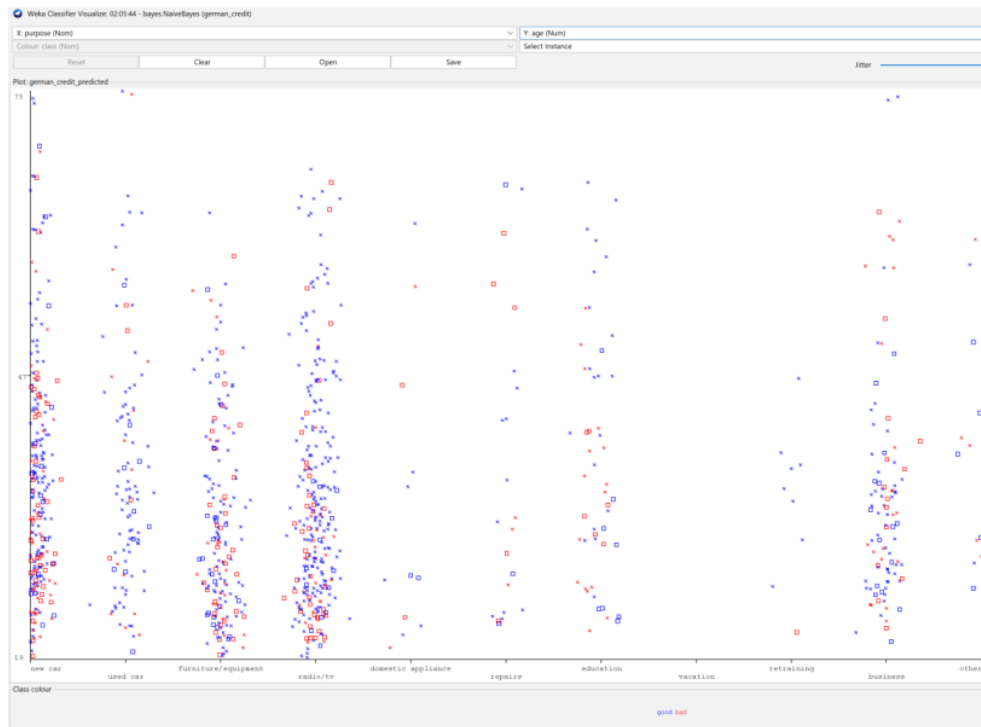


Figure 11 (Purpose Against Age)

It is shown in Figure 11 that younger debtors make up the vast majority of the statistics in the dataset. With a high proportion of the debtors under 30 defaulting with bad credits. With a greater number of bad credits falling under the new car, education and radio/tv categories. This indicates a potential financial instability or risk prone behaviour from this demographic and should be factored into loan agreements. Loaners could also investigate flexible loan arrangements for purposes such as education to stop people trying to improve their earning potential from defaulting.

Reflective Commentary

Various machine learning algorithms were used to analyse the south German credit dataset such as Random Forest, Logistic Regression and Naïve Bayes. These all provided some interesting insights for understanding financial risk factors when dealing with certain demographics and loan purposes. The effectiveness of the models varied with Random Forest being the most accurate even though it had a relatively high false positive rate. However, this can be fine tuned with more data and altering the algorithm tuning. [8]

Assessment of Method Effectiveness (Accuracy, Efficiency)

The employed methods demonstrated different degrees of accuracy and efficiency. While the models were generally reliable, they could be fine tuned to try and reduce false positives and enhancing the precision of the predictions. The efficiency of the models did not cause any concerns as the credit

dataset is a smaller dataset. However Naïve Bayes was the quickest and could be used for initial screening before a person fills out a complete application form.

Advantages and Disadvantages of Approaches Used

The three different algorithms used had different strengths when handling the dataset. Random Forest had an ensemble approach which avoids overfitting by reducing the variance in the data. The Logistic Regression algorithm offered a high degree of interpretability which is crucial for understanding credit decisions. While lastly Naïve Bayes provided a quick baseline model however its main weakness assumes the features are independent of one another, which is not always the case when dealing with credit ratings.

Lessons Learned and Possible Improvements

The analysis conducted using Weka highlighted the importance of further preprocessing of the data. For example, feature engineering and potentially removing some of the attributes which have a low IQR (inter-quartile range). Creating more complex features like debt-to-income ratio could enhance model accuracy. There is also a class imbalance between good and bad credits which may skew the results.

Conclusion

Several machine learning algorithms were used: Random Forest, Logistic Regression, and Naïve Bayes—to the South German Credit Dataset to elucidate patterns of credit risk that correlate with demographic and financial variables. The analysis confirmed that factors such as loan type, borrower's age, employment stability, and financial behavior significantly influence credit risk outcomes. Notably, it was found that younger borrowers, particularly those under the age of 30, exhibit a higher propensity for defaults, predominantly in loan categories associated with new cars, educational pursuits, and lifestyle-related expenditures such as radios and televisions.

Data mining revealed that loans of shorter durations and smaller amounts are generally associated with improved credit scores, suggesting a lower risk profile. However, the occurrence of false positives, especially prevalent in the Random Forest model, indicates an imperative for further algorithmic refinement to bolster both accuracy and applicability in real-world settings.

The investigation yielded insights that are instrumental for financial institutions aiming to develop more nuanced risk management strategies and financial products tailored specifically for younger demographics and categories within high-risk loan segments. By continuing to refine the analytical methods and data management practices, it should augment the accuracy of the predictions and contribute to the development of more stable and equitable lending practices. [9]

References

- [1] "Explore gpts," ChatGPT, <https://chatgpt.com/gpts?oai-dm=1> (accessed May 13, 2024). (Used for turning the credit data into an ARFF file.)
- [2] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.
- [3] D. W. Hosmer Jr., S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed., Wiley, 2013.

- [4] G. H. John and P. Langley, "Estimating Continuous Distributions in Bayesian Classifiers," in *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, 1995*, pp. 338-345.
- [5] F. Provost and T. Fawcett, *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*, O'Reilly Media, Inc., 2013.
- [6] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.
- [7] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms*, 2nd ed. IEEE Press, 2011.
- [8] S. Raschka, *Python Machine Learning*, 2nd ed., Packt Publishing, 2017.
- [9] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*, 3rd ed., Morgan Kaufmann, 2011.
- [10] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 1, pp. 16–23, 2010.
- [11] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, 2002.

Data Mining

GRADEMARK REPORT

FINAL GRADE

GENERAL COMMENTS

72/100

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

