QUANTITATIVE RISK MANAGEMENT

ASSIGNMENT 3: CREDIT RISK MODELLING

Pan Pan Wu
Student Number: 2662780
pan-pan.wu@student.uva.nl

Anna-Maria Angelova
Student Number: 2683090
angelova8668@gmail.com

June 16, 2020

ABSTRACT

In this assignment we classify bank loans as performing or defaults using a Kaggle dataset. We apply different classification methods, which achieve robust accuracy of the predictions of around 77% - 79%.

Contents

1 Data and Motivation

For the current assignment, we use a Bank Loan Status Dataset from Kaggle [1]. We chose a dataset, which has interesting and sufficient data, but not too many submitted entries on the Kaggle board so that we could perform interesting new analysis. There are 9 existing submissions in the Kaggle kernel board as of beginning of June 2020. They contain exploratory data analysis, data cleaning and classification methods. We noticed that none of the analyses contain extensive fine and coarse classing of the features based on default rate. In addition, whenever classification was performed, Null (reference) groups were assigned with automatic "drop first" method (i.e. by assigning Null group to the first class of the variable) rather than by manually choosing the most suitable Null class with average default rate and highest number of observations. In addition, diagnostics of the logistic methods were rarely present, so we could not assess the appropriateness of the models based on variable significance and logic of the coefficients. The fact that the dataset is imbalanced is not taken into consideration. We include these analyses the current report.

We use the training dataset, which we split into training (70%) and validation (30%) sample. The initial data contains 100,000 instances and 19 attributes, shown in the Table 9 in the Appendix.

2 Data Preprocessing

2.1 Missing Treatment

The first step of the data cleaning process is treatment of missing values. The number of missing observations per variable are shown in Table 9 in the Appendix. There are many missing values for Credit Score, Annual Income and Months since last delinquency. The following treatments are applied:

- Annual Income: impute the missing values with the mean
- Credit Score: impute the missing values with 2 different means based on short and long-term debt
- Months since last delinquency: drop the variable as it is used for calculating other key variables
- Drop missing observations for the other variables

There are 77,271 observations after applying treatment for missing values.

2.2 Correlations

Figure 1 shows the correlations among the variables. As it can be observed, the highest correlations are less than 0.5, which by rule of thumb means that it is not necessary to exclude any variables at this stage.

2.3 Variable Transformations

Monthly Debt to Income is created as the ratio between Monthly Debt to Annual Income (divided by 12 months) and is used as a proxy for the ability of the borrower to repay their debt.

$$Monthly\_Income = \frac{Annual\_Income}{12} \tag{1}$$

$$Monthly\_Debt\_to\_Income = \frac{Monthly\_Debt}{Monthly\_Income} \tag{2}$$

[1] Bank Loan Status Dataset: https://www.kaggle.com/zaurbegiev/my-dataset

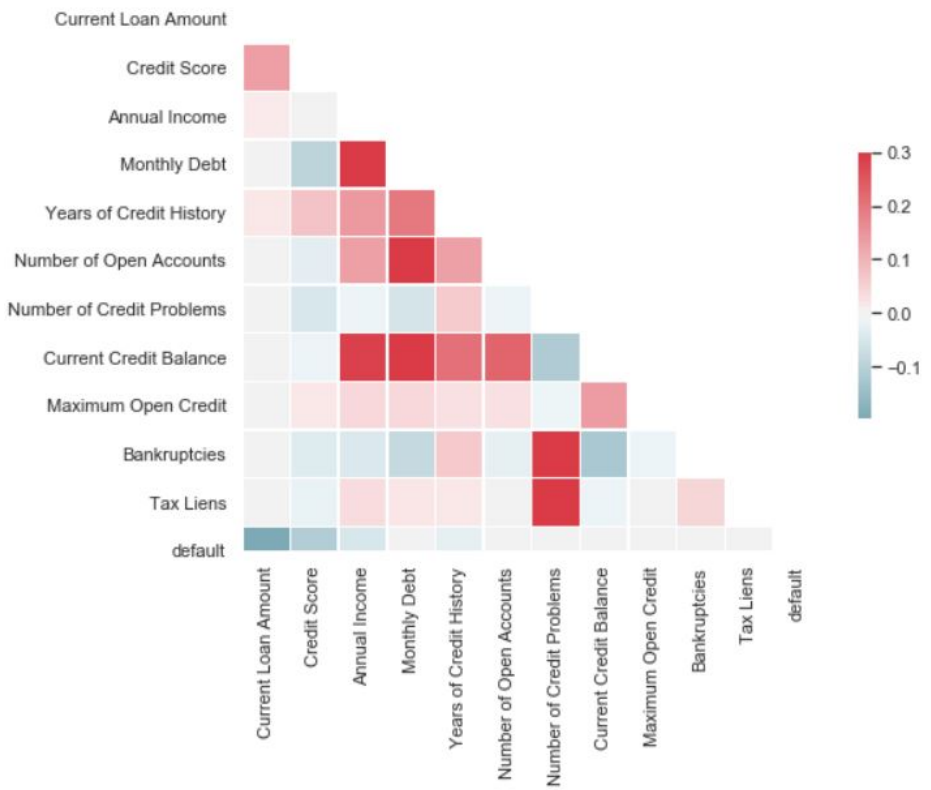

Figure 1. Correlations Heatmap

Credit Scores above 850 are scaled by 10, which seems to be a data issue, which we address by rescaling them.

2.4 Variable Classing

Variable classing is an iterative process alongside modelling. Different versions of bucketing were tested in the model. In this section we presented the classing for the variables that enter the final model with their final buckets alongside with the ratio of default to performing status for each class. Bad (default) rates and Good/Bad Odds for the individual buckets are available in Tables 3 - 5.

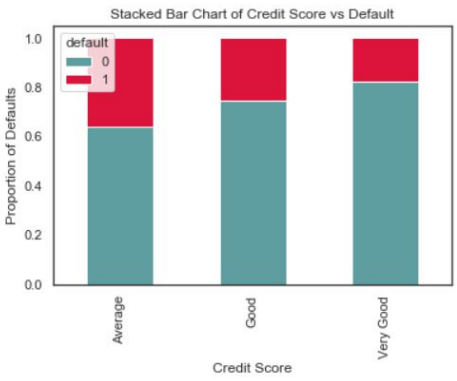

Figure 2. Credit Score

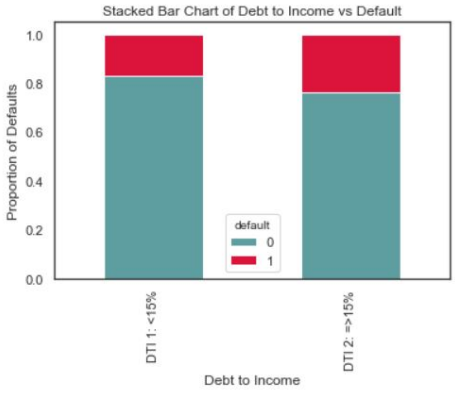

Figure 3. Debt to Income

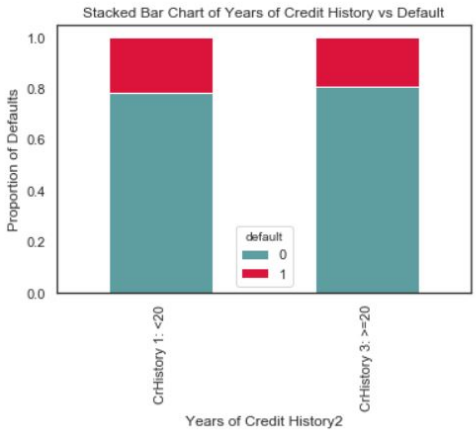

Figure 4. Years of Credit History

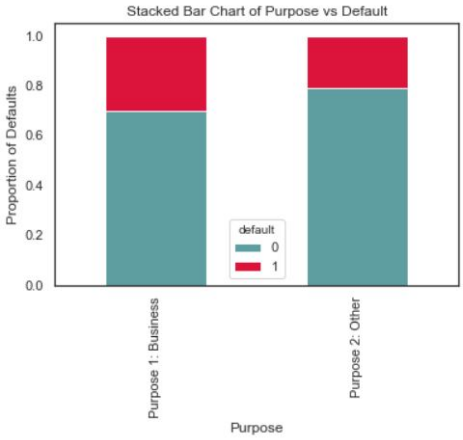

Figure 5. Loan Purpose

The standard rules for creating buckets are as follows:

- The buckets should follow a monotonic (and reasonable) trend with the default rate/GB odds
- The buckets should contain minimum 50 performing and 50 non-performing defaults
- The Null group is the bucket with biggest number of observations and/or closest to the average default rate

2.4.1 Credit Score

We use the Experian credit score ranges to class the Credit Score [2]. Table 1 provides a summary:

Credit Score	Range
Poor	$<$579
Average	580-669
Good	670-739
Very Good	740-799
Exceptional	800-850

Table 1. Experian Credit Score Range

In our dataset we only have Average, Good and Very Good credit scores, which show correct trend in the data as the proportion of defaults is highest for Average credit scores, medium for Good credit scores and lowest for Very good credit scores. The Null group is Good credit scores.

[2]Experian Credit Score Ranges: https://www.experian.com/blogs/ask-experian/infographic-what-are-the-different-scoring-ranges/

Credit Score	% Total	Performing Obs.	Performing %	Default Obs.	Default %	GB Odds
Average	7.46%	3861	66.9%	1907	33.1%	2.02
Good	69.97%	41930	77.9%	11908	22.1%	3.52
Very Good	22.86%	15266	86.4%	2399	13.6%	6.36

Table 2. Credit Score Buckets

2.4.2 Debt to Income ratio

Low Debt-to-income ratios are less likely to default than high ones, which is the intuitive expectation. The Null group is Debt to Income $>= 15\%$.

Debt to Income	% Total	Performing Obs.	Performing %	Default Obs.	Default %	GB Odds
<15%	41.5%	26551	82.9%	5474	17.1%	4.85
>=15%	58.5%	34506	76.3%	10740	23.8%	3.21

Table 3. Debt to Income Buckets

2.4.3 Credit History

Longer credit history is less likely to default than short credit history. The variable does not discriminate strongly between the two classes in terms of default percentage, but is still significant in the regression analysis as the next section will show. The Null group is < 20 years of credit history.

Credit History	% Total	Performing Obs.	Performing %	Default Obs.	Default %	GB Odds
<20 years	68.1%	41230	78.3%	11416	21.7 %	3.61
>= 20 years	31.9%	19827	80.5%	4798	19.5%	2.38

Table 4. Credit History Buckets

2.4.4 Purpose of Loan

Business lending is associated with highest risk. More granular classing did not improve the performance of the model. In addition, the logic behind the trend of the default rates could not be justified without specific knowledge of the portfolio. Business activities in general could be more risky category since especially small and medium enterprises fail often. The Null group is Others.

Purpose	% Total	Performing Obs.	Performing %	Default Obs.	Default %	GB Odds
Business	1.89%	1021	70.0%	437	30.0%	2.34
Others	98.11%	60036	79.2%	15777	20.8%	3.80

Table 5. Purpose of Loan Buckets

2.5 Undersampling

Since the dataset is unbalanced, we perform undersampling. The number of observations in the training sample before and after balancing is 54,089 and 34,014, respectively. Figures 6 and 7 show the performing and defaulted observations in each.

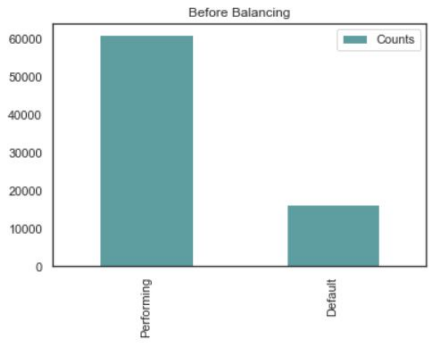

Figure 6. Before Balancing

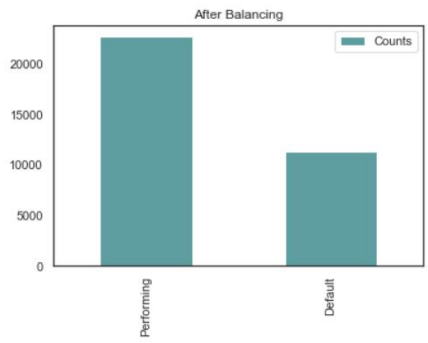

Figure 7. After Balancing

3 Data Modelling

Modelling was done in an iterative fashion, starting with all dummies (except for the Null groups) and step by step excluding insignificant buckets (with p-value > 0.05) and buckets, which exhibited illogical coefficients in terms of observed sign. After each iteration, regression coefficients were reestimated. Different classing of the dummies was also tested. We first exclude Tax Liens which has *p-value* 0.6504. Using the new model without Tax Liens, we next exclude Long Term feature which has *p-value* 0.6292. After excluding both features with stepwise selection, we acquire the result which every remaining features are important (*p-value* $<<$ 0.05). In the following section, we fit the preprocessed data with logistic regression and observe if there is any outlier with residuals histogram. Lastly, we compare the resulting accuracy of logistic regression with four other classification algorithms.

3.1 Logistic Regression

The logistic regression results are shown in Table 6. All variables are significant (p-values < 0.05). In terms of coefficients, higher values and positive coefficients mean that the dummy contributes to the default status. The results are logical as we expect change in the sign of the coefficients around the Null groups. Very good credit scores are less likely to default (coef. -) than Good scores, which is the assigned Null group. Average credit scores are more likely to default (coef. +) than Good scores . Low Debt-to-income ratios are less likely to default (coef. -) than high ones, which are the Null group. Business lending is associated with higher risk, as observed during the bucket creation process,

where business was linked to highest default rate. Hence, we observe no inverse trend as it enters the regression (coef. +)

	Coef.	Std. Err.	z	p-value	[0.025	0.975]
Credit Score: Average	0.3025	0.0398	7.6083	0.0000	0.2245	0.3804
Credit Score: Very Good	-0.7851	0.0300	-26.1799	0.0000	-0.8439	-0.7263
Debt to Income: $<15\%$	-0.7024	0.0211	-33.2642	0.0000	-0.7438	-0.6610
Credit History: $>=$ 20 years	-0.3805	0.0229	-16.6075	0.0000	-0.4254	-0.3356
Purpose of Loan: Business	0.3037	0.0807	3.7636	0.0002	0.1455	0.4618

Table 6. Logistic Regression Results

The coefficients are the logarithmic value of the odds ratio (OR), we can hence obtain the odds ratio from the above table by taking the exponential value of the coefficients. We have OR of 1.353 average 1.353 which indicates if having average credit score, the odds of being defaulted is higher than non-average group. If having the purpose of loan being business, the odds of being defaulted is also higher than other purposes (OR = 1.355). On the other hand, if having very good credit score, low debt to income ratio and high credit history, the odds of being defaulted is mostly half less than other groups (OR: 0.456, 0.495 and 0.684 respectively). These outcomes are reasonable: the people having lower debit, higher income and higher credit score have lower odds of being defaulted.

If the fitted logistic regression model is true, we would expect to see that the residuals are falling between 3 and are normally distributed, meaning that the model is wrong in the same way in both directions. Positive values for the residual mean the prediction was too low (optimistic), and negative values mean the prediction was too high (conservative). For our logistic regression we observe that the residuals satisfy the condition for the boundaries, but they are not normally distributed. Since they are more skewed on the negative side (between -1.0 and -0.5), we conclude the model might be consistently slightly conservative. This is still a good outcome for regulatory purposes on capital determination, but might hurt profitability.

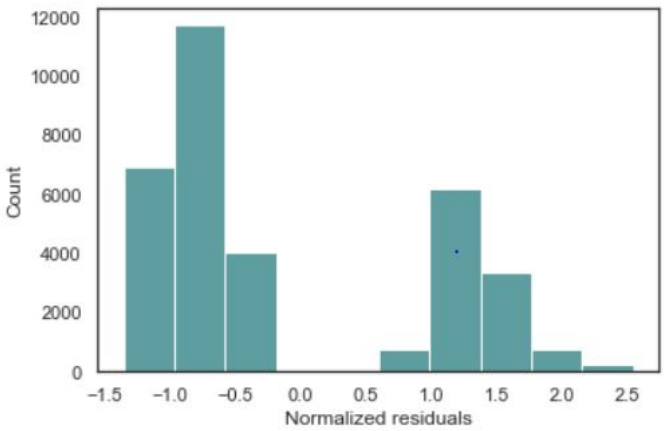

Figure 8. Residuals Logistic Regression

3.2 Model Performance and Other Classification Methods

In this section we discuss the accuracy of the logistic regression model on the validation sample and compare it to other classification methods. We show that the chosen set of final variables leads to robust accuracy regardless of the classification method choice (Table 8). On the other hand, Figure 9 shows that the area under the curve (AUC) is not ideal which is close to random (0.52), when examing the confusion matrix, we found out there are large number of false positive cases, this indicates that our model easily classify the non-default cases into default class. This can be interpreted as a conservative result.

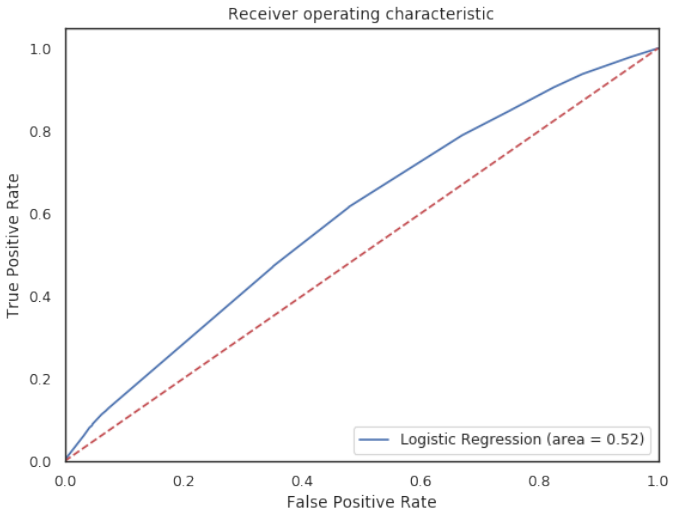

Figure 9. Receiver operation curve (ROC) of the result using logistic regression

	Non-default	Default
Non-default	17534	772
Default	4477	399

Table 7. Confusion matrix of the logistic regression result

	Accuracy
Logistic Regression	77.36%
K Nearest Neighbors (n=5)	78.57%
Support Vector Classification	78.97%
Decision Tree Classifier	77.38%
Random Forest Classifier (250 trees)	77.38%

Table 8. Accuracy Classification Methods.

4 Appendix

#	Variables List	Missing Obs.
1	Loan ID	514
2	Customer ID	514
3	Loan Status	514
4	Current Loan Amount	514
5	Term	514
6	Credit Score	19,668
7	Annual Income	19,668
8	Years in Current Job	4,736
9	Home Ownership	514
10	Purpose	514
11	Monthly Debt	514
12	Years of Credit History	514
13	Months since Last Delinquency	53,655
14	Number of Open Accounts	514
15	Number of Credit Problems	514
16	Current Credit Balance	514
17	Maximum Open Credit	516
18	Bankruptcies	718
19	Tax Liens	524

Table 9. Variables List