Credit Scorecard Modelling

Alexander James Waudby

August 11, 2020

Abstract

Abstract Placeholder

Contents

List of Figures

Chapter 1

Credit Scoring

1.1 Introduction

Credit scoring is a method used by financial institution globally to assess whether a customer should be taken on. This can be for a variety of services such as credit cards, loans, mortgages, etc. It's development originated from the need of risk vs rewards. Lenders needed a way of determining if a potential customer would be able to pay back their credit and as such not costing the lender money by taking on creditees which end up being unable to repay the debt. A credit score is usually just a number indicating your quality as a creditee. The scale of the score can vary on which lender is providing the score but usual ones are 0-999 or 0-500 with the lower the score the less likely you would be offered the service.

Although used globally, there is no widely accepted "perfect" model or method. All companies assess their customers differently, a customer could be rejected from one lender and be accepted by another based on what they would define as an acceptable client. Even within companies the models and methods can change due to new circumstances and the changing financial climet. What previously could of been a strong predictor of a bad client could now be insignifcant. A recent example of this is the technology development of mobile phones. Previously, if a client did not have access home phone this could be an indicator of a possible bad client. Now, with the development and wide public access to mobile phones, access to a home phone has become mostly irrelevant with most of the public having no use for them anymore. Changes such as this and others require lenders to be constantly evaluating how they assess customers to prevent the rejection of good clients and the accepting of bad ones.

1.2 Modelling

Credit score modelling is often discrete based with the most common being a logistic regression with the response variable being either a good ($y = 0$) or bad customer ($y = 1$). Predictors can be a variety of variable such as personal characteristics, age, gender or economic status e.g. car owner, home owner/rentor etc, to financial characteristics like amont of current debt and repayment statuses. One thing to note, certain personal characteristics are off limit to company due to discrimination laws. Predictors such as race, may be shown to have some use in scoring but cannot be used as the model would become discriminatory.

1.3 WOE and IV

Weight of evidence (WOE) is a popular method used in score card modelling, often used because the variables used in credit scoring can have a large amount of categories which cause impractiallities when converting these to dummy variables. WOE is an alternative to that, rather than create a large amount of dummy variables, the method produces a numerical value (weight of evidence) for each category which is prdouced by (1.1). with $f()$ being the distribution of category X for goods and bads. These value then replace their respective categorical value when modelling the scorecard.

$$WOE = \ln \frac{f(X = x|y = 0)}{f(X = x|y = 1)} \tag{1.1}$$

IV, the information value. Is a measure of the weight of evidence for categories $IV \geq 0$. A value of 0 indicates the variable has no predictive power i.e. no valueable information in the variable. IV is calculated by (1.2). A guideline produced by Bailey[2] is below for evaluating the IV values.

$$\text{IV} = \sum (\text{\% of Bad} - \text{\% of Good}) \cdot \text{WOE} \tag{1.2}$$

IV	Recommendation
Less than 0.03	Poor Predictor
From 0.03 to less than 0.10	Weak Predictor
From 0.10 to less than 0.30	Average Predictor
From 0.30 to less than 0.50	Strong Predictor
Over 0.50	Very Strong Predictor

Figure 1.1: Information Value Table
[2]

1.4 Performance Evaulation

1.4.1 ROC and AUC

ROC, Receiver Operating Characteristic. Was a method of analysis developed during World War II under "Signal Detection Theory". It was originally used for radar operators and their ability to determine if a blip on screen was an enemy or just noise, hence the name Receiver Operating Characteristics.[3] Since then, the method has been applied into a variety of fields for visuallising the accuary of classification models.

Understanding the ROC Curve is relatively simple, the plot is the false positive rate against true positive rate for different cutoff points. The true positive rate is seen as the sensitivity and the false positive being (1 - specificity) An example figure can be found below, the higher the curve, the more accurate the model can be seen as, with the neutral line going 45 degrees through the plot can be seen as the model being the same as a 50/50 guess on the outcome. In some cases these curves can overlap and cause some ambiguity on which curve is overall the best so the measure used to remove this amiguity is the AUC, Area under the curve (1.3). A higher AUC inidicates a stronger disciminatory power with 0.5 being none and 1 being a "perfect model". As such the model with a higher AUC can be considered "a better model". Generally, an $AUC > 0.8$ is considerd good.

$$A = \int_c F_1(c){F_0}'(c)dc \tag{1.3}$$

A more common representation of the AUC is the gini coefficient (1.4). A linear transformation of the AUC to allow the measure to have a preferred 0 to 1 scale rather than 0.5 to 1.

$$gini = (2 \cdot AUC) - 1 \tag{1.4}$$

1.4.2 K-S Statistic

The K-S Statistic (Kolmogorov-Smirnov Statistic) is a measurement of the scorecards ability to seperate the goods from bads. The K-S Statistic is the maximum distance between the cumulative distributions of both the goods and bads, or alterntively, if $F_g(x)$ is the cumulative distribution of goods and bads is $F_b(x)$ where x is the score then the KS Statistic is (1.5)

$$KS = max(F_g(x) - F_b(x)) \tag{1.5}$$

The statistic can be expressed visually by plotting the cumlative distributions as seen below. An issue of this measurement is that it only provides the score at which the scordcard seperates the goods and bads the most. The cutoff score for the card might not necessarily be this score and a higher K-S score does not imply the scorecard is a better fit.

1.4.3 Divergence

Divergence is a measurement of the distributions of goods and bads. The idea is that the scorecard on average will assign a lower score to bads than goods i.e. $\mu_b < \mu_g$. Divergence is a way to assess this performance.

1.5 Cut-off

A scorecard in simple terms is just a method producing a score for each individual. To put the scorecard into use the difference between the scores needs to be classified this is done by a cut-off score. This score is a point on the scorecard which would seperate accepted applicants from rejected. A simple cut-off method would be to have a single score, any applicants above the score are accepted and anyone below the score is rejected. The benefit of a simple method is the ability to quickly process applicants and move desired applicants onto the next stage faster. The issue with the single cutoff comes with the applicants that are close to the cutoff, having a strict cut-off can cause a company to take on bad applicants or reject good applicants where futher investigation would prove the applicant more likely to be the opposite.

An alternative to this would be a two score cut-off. This would be done by having two scores like Rejected $< S_1 <$ Refer $< S_2 <$ Accepted. Any score above S_2 is automatically accepted and any below S_1 is rejected. Scores which the land in between and moved to a referral stage where a lender can further look into the applicants case by case to decide the outcome. This comes with added benefit of removing the issue of applicants close to the single cutoff. The idea is that with the lenders insight, more good applicants will be accepted and more bads rejected compared to the single cutoff, thus possibly reducing the bad rate of accepted applicants.

The cut-off scores can be determined by varying factors which can change depending on the companies interest. Four of these are specified by Bailey [2]. Acceptance rate, the percentage of all applicants accepted by the cut-off. Overall bad rate, the percentage of all accepted applicants that end up being bads. Marginal bad rate, the percentage of accepted applicancts that are bad close to the cut-off score. Profitability, the possible profit from goods minus the loss from bads. Depending on the situation of the business and its goals would determine the importance of each factor with overal bad rate being the usual priority.

Chapter 2

Data

The data I am using for this project is a collection of observations of 5,960 home equity loans which is provided by Baesens, Bart, Roesch, Daniel and Scheule, Harald [1]. Home equity loans are when an applicant borrows agains the value or 'equity' of their home. **Talk about home equity loans and what they are here.** You can find a full description of each variable in **??**.

Variables used in the Data Set [1]

Variable	Definition
BAD	1 = Applicant defaulted on loan or seriously delinquent; 0 = applicant pain load
LOAN	Amount of requested loan
MORTDUE	Amount due on exisiting mortgage
VALUE	Value of property the loan is to go against
REASON	The reason the applicant is applying for the loan. DebtCon = Debt condsolidation; HomeImp = Home Improvement
JOB	Occupational categories
YOJ	Years at present job
DEROG	Number of major derogatory reports
DELINQ	Number of delinquent credit lines
CLAGE	Age of oldest credit line in months
NINQ	Number of recent credit inquiries
CLNO	Number of credit lines
DEBTINC	Debt-to-income ratio

2.1 Data Cleaning

The data provided needed some initial cleaning. 2596 observations were missing atleast one value with some missing several variables. The biggest culprit of this would be DEBTINC with 1,267 missing values. I decided to handle these missing values on a case by case basis applying different methods. First I decided to exclude observations missing more than a third of their variables, 339 fit this criteria. Next before I went forward with any imputing I considered any possible outliers within my numerical data, using the summary table 2.1. You can see for the quantile ranges that there will most likely be some outliers occuring in the majority of the numerical variables. To solve this I removed the 99th percentile for every numerical variable excluding BAD, this ended up removing 589 rows.

Moving onto imputing variable, for MORTDUE and VALUE I imputed their values using a simple linear regression of the other. This was going on the assumption that the mortgage due on a house has a strong relationship with the value of property. The assumption is further backed up with the correlation between the two being 0.8748 before imputing, far higher than any of the other correlations in the data. So for MORTDUE I used (2.1) and for VALUE I used (2.2). This was applied to any missing value where the other was present and for the remaining I took the mean of each variable from the original data before the imputations.

$$\text{MORTDUE} = \beta_0 + \beta_1\text{VALUE} \tag{2.1}$$

β_0	-2145.6497
β_1	0.7177

$$\text{VALUE} = \beta_0 + \beta_1\text{MORTDUE} \tag{2.2}$$

β_0	21340.4803
β_1	1.1253

For the remaining numerical variables I chose to take the median of the values as there was no highly correlated variables I could take to impute using linear regression. There is an argument that because DEBTINC is missing 991 (19.7 %) that some other method from using the median value should be used but the only reasonable solution with the data available would be to drop the variable all together or drop the missing rows, both would result in a large loss of data. As this will most likely have a significant impact on the importance of the variable I planned to look into the performance of a few datasets with different solutions for this variable.

Last was the two categorical variables REASON (DebtCon, HomeImp) and JOB (Other, Office, Sales, Mgr, ProfExe, Self). REASON's categories were specified in the data dictionary but JOB's categories were not, for simplicity I am going to assume that the missing values are of the categories just specified. Their missing values were 138 and 127 respectively and I decided to impute these values using a weighted random sample with the weights being the counts of the respective category. Below is two table summarizing the imputes, I applied a seed to reproduce the sampling to remian consistent.

REASON

Category	Original	Weight For Sampling	New	Count Imputed
DebtCon	3448	0.7005	3528	80
HomeImp	1474	0.2995	1504	30

JOB

Category	Original	Weight For Sampling	New	Count Imputed
Other	2056	0.4175	2100	44
ProfExe	1123	0.2281	1150	27
Office	862	0.1751	877	15
Mgr	646	0.1312	657	11
Self	144	0.0292	151	7
Sales	93	0.0189	97	4

With these two completed I had no more missing values an no other noticeable issues which needed to be corrected before I could further look into the variables. A summary of actions taken on missing values can be found in table (**??**)

Missing Variables Breakdown

Variable	No. Missing	Solution
BAD	0	N/A
LOAN	0	N/A
MORTDUE	316	Imputed from a linear regression (2.1). Mean taken when VALUE was unavailable
VALUE	59	Imputed from a linear regression (2.2). Mean taken when MORTDUE was unavailable
REASON	110	Random weighted sample taken
JOB	108	Random weighted sample taken
YOJ	311	Median taken
DEROG	362	Median taken
DELINQ	246	Median taken
CLAGE	67	Median taken
NINQ	178	Median taken
CLNO	0	N/A
DEBTINC	991	Median taken

	BAD	LOAN	MORTDUE	VALUE	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
count	5621.0	5621.00	5278.00	5537.00	5294.00	5206.00	5356.00	5549.00	5422.00	5621.00	4447.00
mean	0.2	18846.02	73977.01	103025.40	9.00	0.24	0.45	179.77	1.19	21.45	34.07
std	0.4	11301.47	44813.54	58002.35	7.61	0.80	1.13	85.70	1.73	10.13	8.47
min	0.0	1100.00	2063.00	8000.00	0.00	0.00	0.00	0.00	0.00	0.00	0.52
25%	0.0	11300.00	46385.00	66922.00	3.00	0.00	0.00	115.57	0.00	15.00	29.43
50%	0.0	16500.00	65000.00	90008.00	7.00	0.00	0.00	173.63	1.00	20.00	35.02
75%	0.0	23500.00	91989.25	120724.00	13.00	0.00	0.00	230.72	2.00	26.00	39.14
max	1.0	89900.00	399550.00	855909.00	41.00	10.00	15.00	1168.23	17.00	71.00	203.31

Figure 2.1: Summary Before Outliers Removed

	BAD	LOAN	MORTDUE	VALUE	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
mean	0.17	17944.69	68427.20	98341.35	8.60	0.15	0.32	174.81	1.04	20.65	33.88
std	0.38	9547.25	36656.24	45860.89	6.95	0.49	0.77	76.09	1.35	9.08	6.11
min	0.00	1100.00	2063.00	8000.00	0.00	0.00	0.00	0.51	0.00	0.00	0.52
25%	0.00	11100.00	44317.25	66343.75	3.00	0.00	0.00	114.59	0.00	14.00	30.94
50%	0.00	16200.00	62562.50	89033.00	7.00	0.00	0.00	170.72	1.00	20.00	34.88
75%	0.00	22725.00	87368.50	117696.75	12.00	0.00	0.00	225.11	2.00	26.00	37.97
max	1.00	62700.00	207687.00	271738.00	29.00	3.00	4.00	398.40	7.00	48.00	48.28

Figure 2.2: Summary After Outliers Removed

	BAD	LOAN	MORTDUE	VALUE	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
BAD	1.000	-0.073	-0.031	-0.033	-0.065	0.297	0.354	-0.170	0.173	-0.010	0.155
LOAN	-0.073	1.000	0.228	0.332	0.093	-0.032	-0.037	0.097	0.043	0.069	0.062
MORTDUE	-0.031	0.228	1.000	0.895	-0.064	-0.047	0.002	0.128	0.027	0.331	0.146
VALUE	-0.033	0.332	0.895	1.000	0.003	-0.043	-0.004	0.173	-0.006	0.261	0.099
YOJ	-0.065	0.093	-0.064	0.003	1.000	-0.060	0.047	0.190	-0.066	0.023	-0.066
DEROG	0.297	-0.032	-0.047	-0.043	-0.060	1.000	0.176	-0.073	0.156	0.054	0.028
DELINQ	0.354	-0.037	0.002	-0.004	0.047	0.176	1.000	0.027	0.055	0.157	0.042
CLAGE	-0.170	0.097	0.128	0.173	0.190	-0.073	0.027	1.000	-0.112	0.237	-0.049
NINQ	0.173	0.043	0.027	-0.006	-0.066	0.156	0.055	-0.112	1.000	0.082	0.120
CLNO	-0.010	0.069	0.331	0.261	0.023	0.054	0.157	0.237	0.082	1.000	0.143
DEBTINC	0.155	0.062	0.146	0.099	-0.066	0.028	0.042	-0.049	0.120	0.143	1.000

Figure 2.3: Correlation Table

2.2 Variables

With the data cleaned I then was able to look more into the properties of the variables and their relationship with the default outcome. First looking at the summary data found in Table **??**

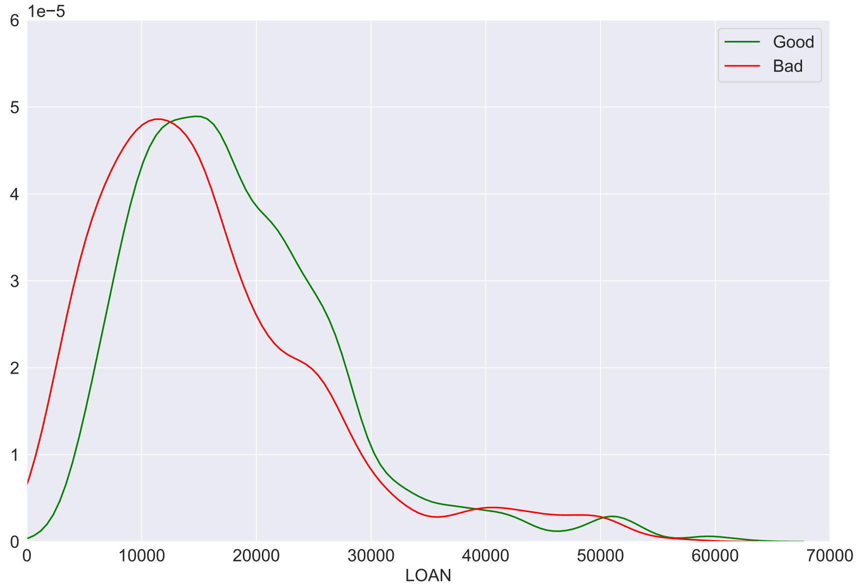

Figure 2.4: Distribution of LOAN by BAD.

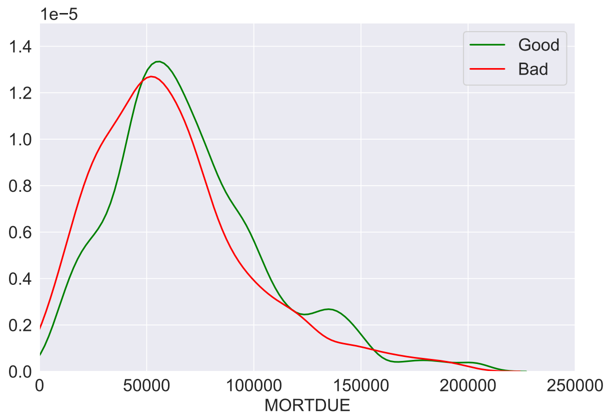

Figure 2.5: Distribution of MORTDUE by BAD.

Breakdown of each variable go here.

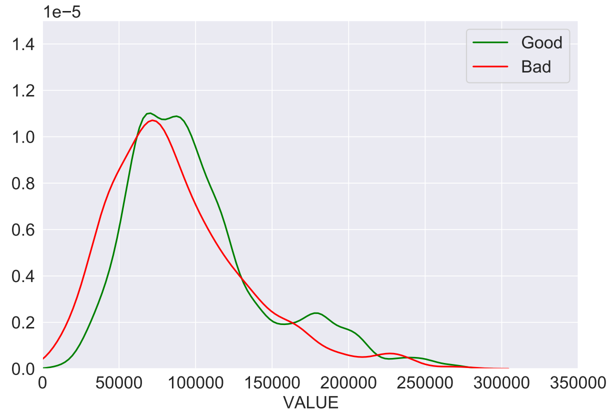

Figure 2.6: Distribution of VALUE by BAD.

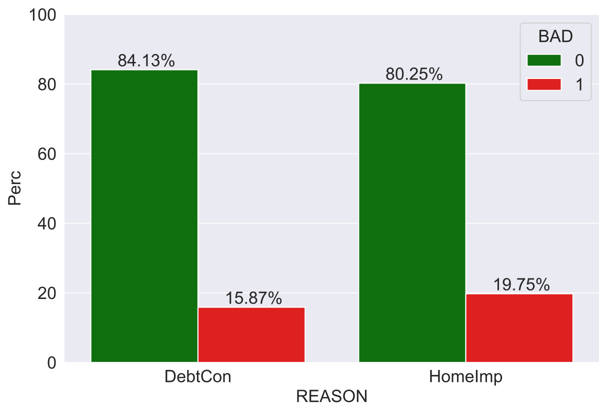

Figure 2.7: Category plot of REASON by BAD.

2.3 WOE and IV

WOE results here.

IV results here.

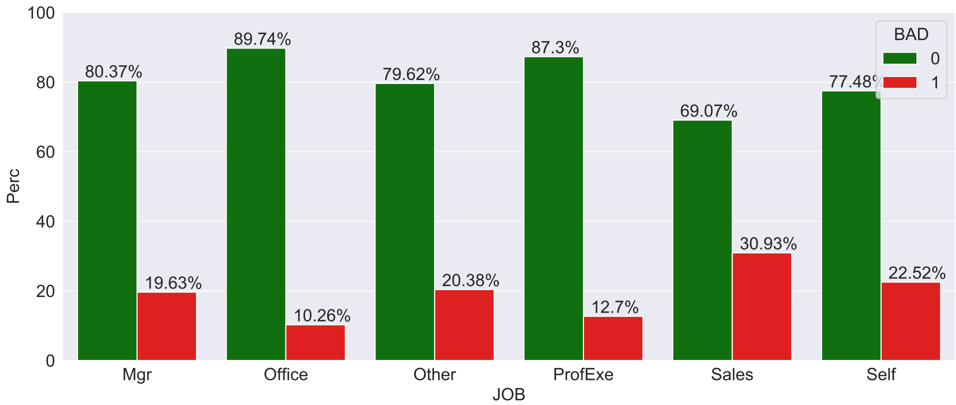

Figure 2.8: Category plot of JOB by BAD.

Chapter 3

Modelling

3.1 Logistic Regression

Chapter 4

Selection

Chapter 5

Alternative Uses - Covid-19

Bibliography

[1] Bart Baesens, Daniel Roesch, and Harald Scheule. *Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS*. John Wiley & Sons, 2016.

[2] Murray Bailey. *Credit scoring: The principles and practicalities*. White Box Publ., 2004.

[3] TG Tape. "Using the Receiver Operating Characteristic (ROC) curve to analyze a classification model". In: *University of Nebraska* (2000), pp. 1–3.