Credit Scorecard Modelling

Alexander James Waudby

August 25, 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 Home Equity Loans

.

1.3 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.

Chapter 2

Literature Review

2.1 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 **bailey2004credit**. 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.

2.2 Weight of evidence and Information value

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 Equation (2.1). with $f()$ being the distribution of category X for goods and bads. These value then replace their respective categorical value when modelling the scorecard.

$$WOE = \ln \frac{f(X = x|y = 0)}{f(X = x|y = 1)} \tag{2.1}$$

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 Equation (2.1). A guideline produced by Bailey**bailey2004credit** is below for evaluating the IV values.

$$\text{IV} = \sum(\% \text{ of Bad} - \% \text{ of Good}) \cdot \text{WOE} \tag{2.2}$$

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 2.1: Information Value Table
bailey2004credit

2.3 Performance Evaulation

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.**tape2000using** 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 (2.3). 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.

$$A = \int_c F_1(c){F_0}'(c)dc \tag{2.3}$$

A more common representation of the AUC is the gini coefficient (2.4). A linear transformation of the AUC to allow the measure to have a preferred 0 to 1 scale rather than 0.5 to 1.

$$gini = (2 \cdot AUC) - 1 \tag{2.4}$$

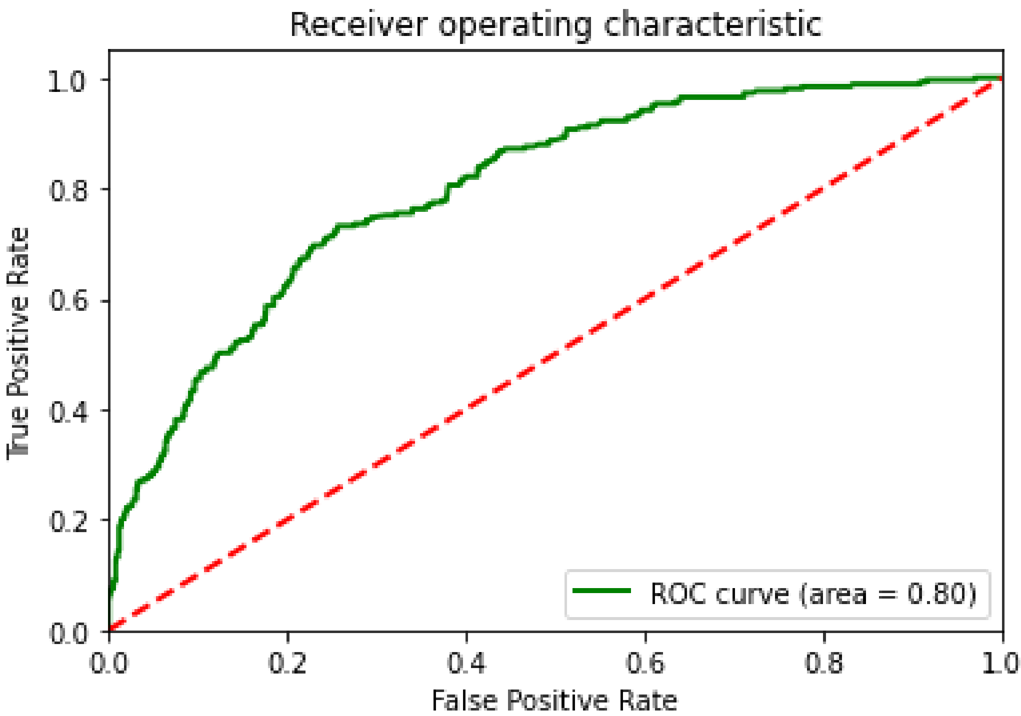

Figure 2.2: ROC Example

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 K-S Statistic is (2.5)

$$KS = max(F_g(x) - F_b(x)) \tag{2.5}$$

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.

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.

Population Stability Index

.

2.4 Data Cleaning Methods

2.5 Chimerge Discretization

For the application of the woe methods a python package called scorecardpy will be used to help automate the process by finding the optimal bins for the numerical variables. The package has the two options for optimizing, tree based and chimerge. For this project I will be using the chimerge method and will explain its application.

The chimerge methods uses the χ^2 statistic to bin numerical variables. It can be seen in detail in **kerber1992chimerge**. The intial step is for the variables to be sorted and then each observation will be split into it's own bin. Each bin will then be compared to its adjacent and calculate the χ^2 value, the bin is then merged into the adjcent bin with the lowest χ^2 value. This step is repeated until all pairs have χ^2 values exceeding a threshold. The formula for computing χ^2 can be seen in Equation (2.6).

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}} \tag{2.6}$$

where, m = 2, the two intervals being compared. k is the number of classes, in our case 2 (Good and Bad). A_{ij} is the number of observations in ith interval and jth class. E_{ij} is the expected frequency of A_{ij} which is calculated by Equation (2.7).

$$A_{ij} = \frac{R_i - C_j}{N} \tag{2.7}$$

where, R_i is the number of observations in ith interval. C_j is the number of observations in jth class and N is the total number of observations.

Chapter 3

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 **baesens2016credit**. Home equity loans are when an applicant borrows agains the value or 'equity' of their home. You can find a full description of each variable in Table (3.1).

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

Figure 3.1: Variables used in the Data Set **baesens2016credit**

3.1 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 B.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 Equation (3.1) and for VALUE I used Equation (3.2). 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.

$$\text{MORTDUE} = \beta_0 + \beta_1\text{VALUE} \tag{3.1}$$

β_0	-2145.6497
β_1	0.7177

$$\text{VALUE} = \beta_0 + \beta_1\text{MORTDUE} \tag{3.2}$$

β_0	21340.4803
β_1	1.1253

For the remaining numerical variables excluding DEBTINC I chose to take the median of the values as there were only a small amount missing from each with no highly correlated variables to take a regression from. There is an argument that because DEBTINC is missing 991 (19.7 %) that some other method from using the median value should be used. After some further analysis the decsion to drop the variable was made, imputing did not appear to be an option as of the 857 bad applicants, 585 (68.2%) of them were missing DEBTINC compared to 4175 and 405 (9.7%) for good applicants. Dropping every row with DEBTINC did not appear to be practical either as it would result in the loss of almost 70% of the bad applicants and their data whilst also taking the bad rate down to 6%. Dropping DEBTINC would be the lower loss of information agains the alternative of dropping missing rows (5032 values lost vs 11892).

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 (3.2)

3.2 Variables

With data cleaned the remaining observations was 5032 of 12 independ variables with no missing values. A summary of the numerical values can be found in Table (B.2). There you can see the remaining data has a bad rate of 17% which equates to 857 defaulted applicants.

LOAN

LOAN, the amount requested for the home equity loan by the applicant can be seen in Figure (3.3). The initial assumption was that higher loan values would have a higher bad rate due to the larger amount to pay back, increasing the length and difficulty for the applicant to pay back the loan. Looking at the figure this does not appear to be the case, although small, larger loans tend to be payed off more often. The reasons for this are unclear, a couple suggested could be that since we do not know exactly how this data was gathered the case could be that for larger loans the bank/company offering these loans had higher cutoffs on their applicant scoring to prevent higher risks in higher potential loss causing a shift down in the bad rate. Another could be the arguement that larger requested loans are coming from owners of higher valued properties, which could be the case when looking at the correlation table B.3. A higher property value indicates a better economic status and less likely to default on a loan. It would be reasonable to expect this variable to have a significant effect on the credit score.

Variable	No. Missing	Solution
BAD	0	N/A
LOAN	0	N/A
MORTDUE	316	Imputed from a linear regression (3.1). Mean taken when VALUE was unavailable
VALUE	59	Imputed from a linear regression (3.2). 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	Variable dropped as too much information missing

Figure 3.2: Missing Variables Breakdown

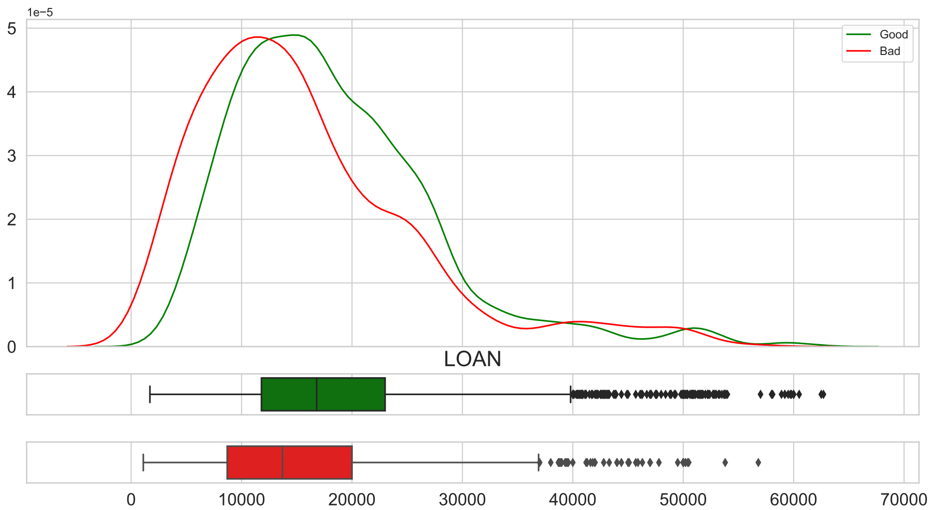

Figure 3.3: Distribution of LOAN by BAD.

MORTDUE

MORTDUE, the outstanding balance on the applicants existing mortgage. Assumption here would be similar to LOAN, a higher outstanding balance on their mortgage mean a large amount of debt and an increased risk of defaulting due to the larger payments. Looking at figure (3.4) you can see it does not follow this assumption, again a small but clear difference in the distribution shows that applicants with a higher outstanding balance on their mortgage are less likely to default. Whatever the reason behind this is would most likely be the same as the reason behind LOAN.

VALUE

VALUE, the property of the applicants and the equity the loan is being put against. The same initial assumption being made with MORTDUE is also here, a value of an applicants property is an indication of their economic status. An owner of a higher valued property should be able to payback a larger loan and less likely to default on smaller ones. From figure (3.5) there is a small visible effect of loan on their probability of defaulting.

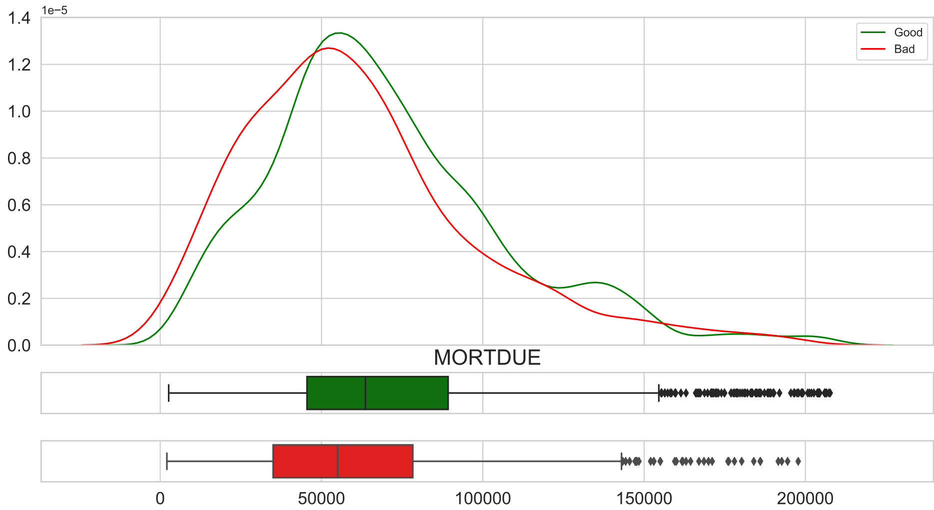

Figure 3.4: Distribution of MORTDUE by BAD.

REASON

REASON, the reason for the applicant's request. There are only two categories as seen in 3.7, DebtCon, the loan would be used for a debt consolidation. HomeImp, the loan is being used for a home improvement. A breakdown of their splits between good and bad can be found in table (3.6). The two categories continue to have similar splits when they do not default but when they do default, HomeImp's split increases by 5%. Looking at this further in figure (3.7) the bad rate for DebtCon is 15.87% and HomeImp 19.75%. Although there does appear to be a difference, it is small and compared to other variables in the group I would expect this to be on the lower end of signifiance.

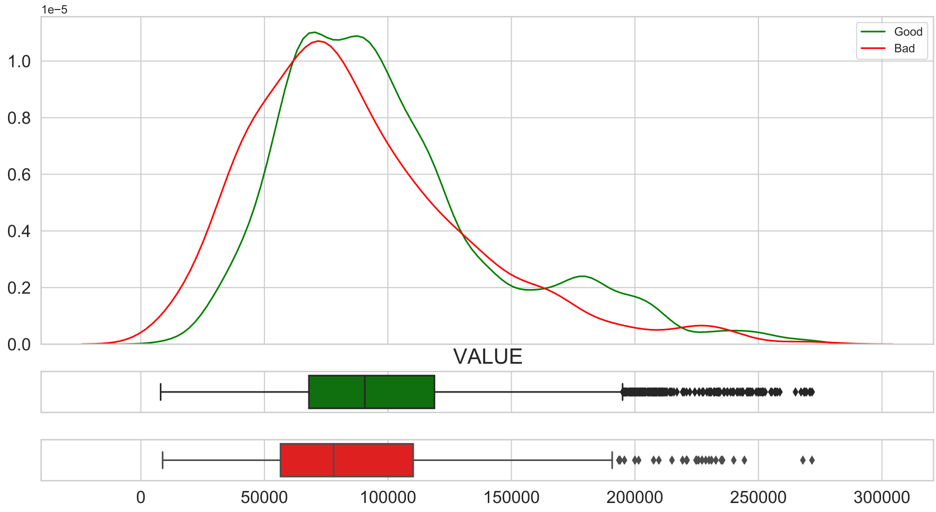

Figure 3.5: Distribution of VALUE by BAD.

Category	% of Total (N = 5032)	% of Good (N = 4175)	% of Bad (N = 857)
DebtCon	70.0% (3528)	71.1% (2968)	65.3% (560)
HomeImp	30.0% (1504)	28.9% (1207)	34.7% (297)

Figure 3.6: REASON breakdown

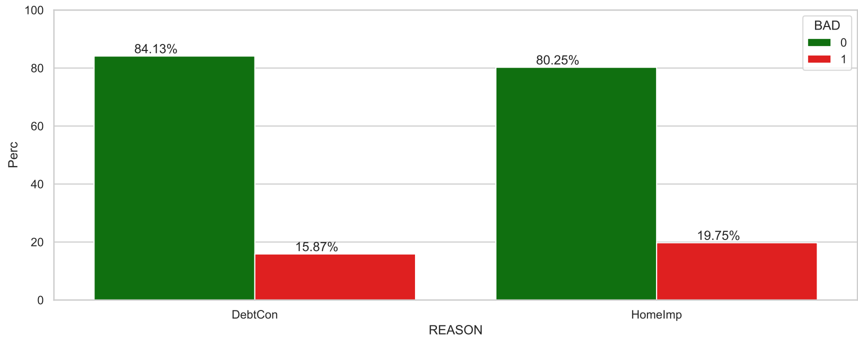

Figure 3.7: Category plot of REASON by BAD.

JOB

JOB, categorical job occupation. Categories can be seen in 3.9. Occupation could be used as an indicator for the applicants economic status e.g. a ProfExe, proffesional executive is more likely to have a higher income than office staff or someone who is self employed. It could also be used to see how volatile their employement status is, someone who is self employed can be seen as a possible risk due to their income being potentially unstable. Although we do not have a way of looking at their job security, we do have YOJ, years at present job, as an indicator of Job security. Comparing figure (3.10) and figure (3.9), Sales, the job with the lowest average YOJ has the highest bad rate at 30.93% followed by Self at 22.52%.

Category	% of Total (N = 5032)	% of Good (N = 4175)	% of Bad (N = 857)
Other	41.7% (2100)	40.0% (1672)	49.9% (428)
ProfExe	22.9% (1150)	24.0% (1004)	17.0% (146)
Office	17.4% (877)	18.9% (787)	10.5% (90)
Mgr	13.1% (657)	12.6% (528)	15.1% (129)
Self	3.0% (151)	2.8% (117)	4.0% (34)
Sales	1.9% (97)	1.6% (67)	3.5% (30)

Figure 3.8: JOB breakdown

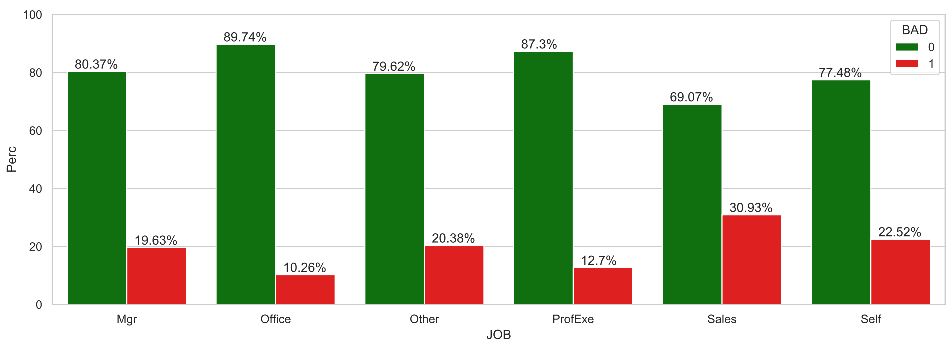

Figure 3.9: Category plot of JOB by BAD.

YOJ

YOJ, number of years the applicant has been at present job. Can be an indicator of job security, an applicant losing their job can be a high risk of defaulting on their loan. Figure (3.11 shows that the majority of applicants are between 3 to 14 years at their current job with the mean for bads and goods being relatively the same. The difference between them starts to become noticable at higher values where an applicant who has been at their present job for more than 20 years starts to become less likely to default.

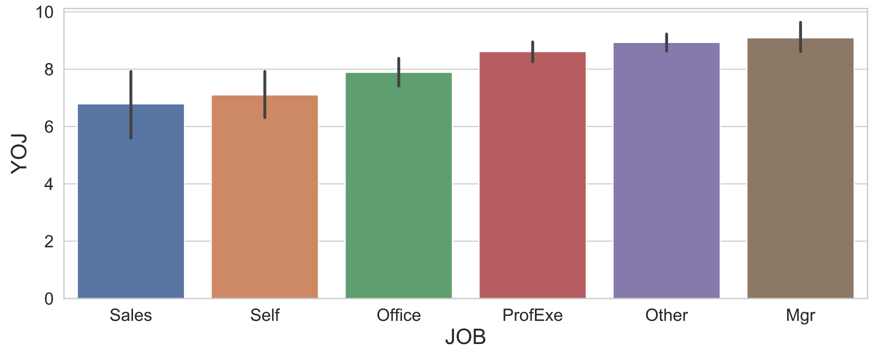

Figure 3.10: Category plot of JOB by mean of YOJ.

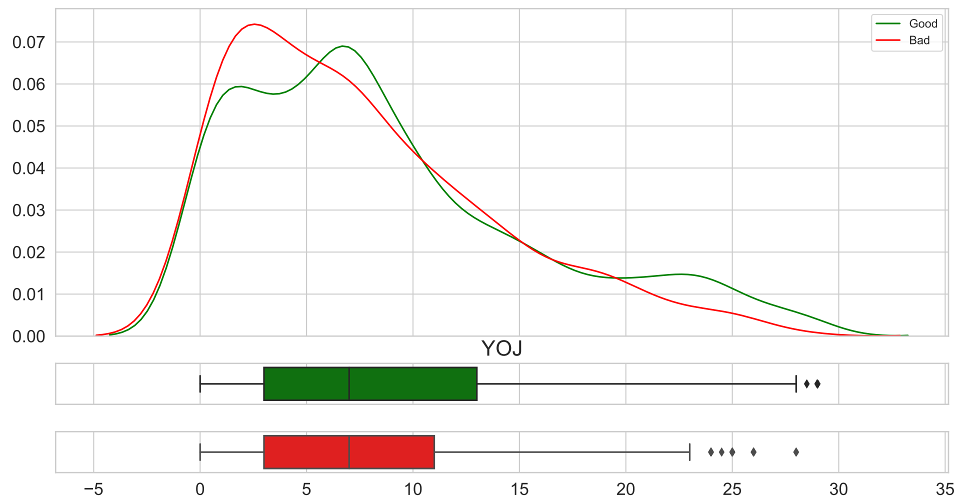

Figure 3.11: Distribution of YOJ by BAD.

DEROG

DEROG, number of dergoatory marks on aginast the applicant. A derogatory mark can have a large impact on your chances of being accepted for a loan. In this dataset, only 12.3% of applicants have 1 or more dergoatory mark. An applicant can recieve a derogatory mark for various reasons such as, missing payments, bankruptcy, repossession, etc. The severity of the reason for the derogatory mark is often used in credit scoring but for this dataset we only have the numbers of marks against the applicant. Figure (3.12) shows the impact having a derogatory mark can have on your chance of defaulting with 0 having a bad rate of 14.3% and 3 having a bad rate of 65.91%.

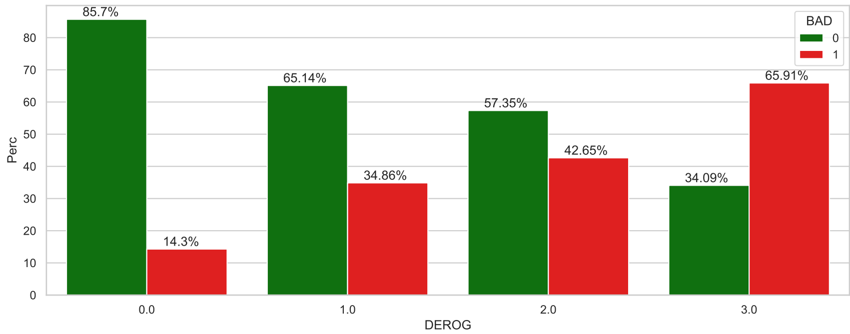

Figure 3.12: Category plot of DEROG by BAD.

DELINQ

DELINQ, Number of delinquent credit lines. Delinquency is used to describe when a borrower has missed a payment, the borrower is referred to as delinquent. Delinquency preceeds defaulting, once a borrower has been deqlinquent for a time it comes apparent that the borrower is unable to pay back the loan. The data we are provided is the number of deqlinquent credit lines, which is the number of credit lines the borrower has missed payments on.

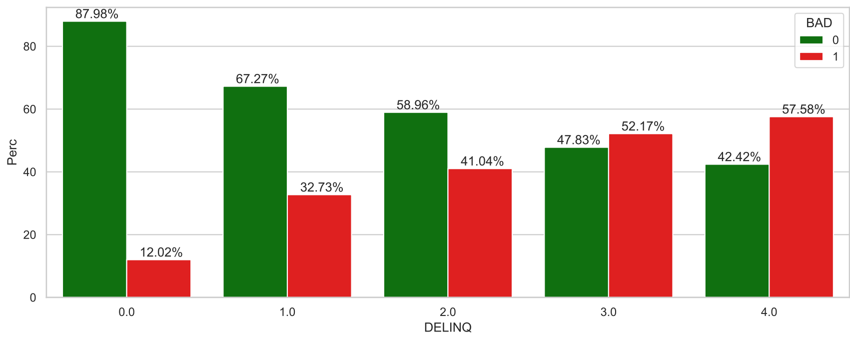

Figure 3.13: Category plot of DELINQ by BAD.

CLAGE

CLAGE, Age of oldest credit line in months. Assumption here is the longer an applicant has held credit lines the more experienced they are in repaying payments. A applicant who has experience in credit repayments for 20 years is going to have a better time in avoiding delinqency and defaulting that someone who is new to credit. Figure (3.14) reinforces this assumption, there is a clear difference in the distributions of goods and bads. Applicants between 100 and 150 months on thier oldest credit line have a bad rate of 21.8% and applicants between 200 and 300 have a bad rate of 9.5%. Expecting this value to have a strong signifance for the credit score.

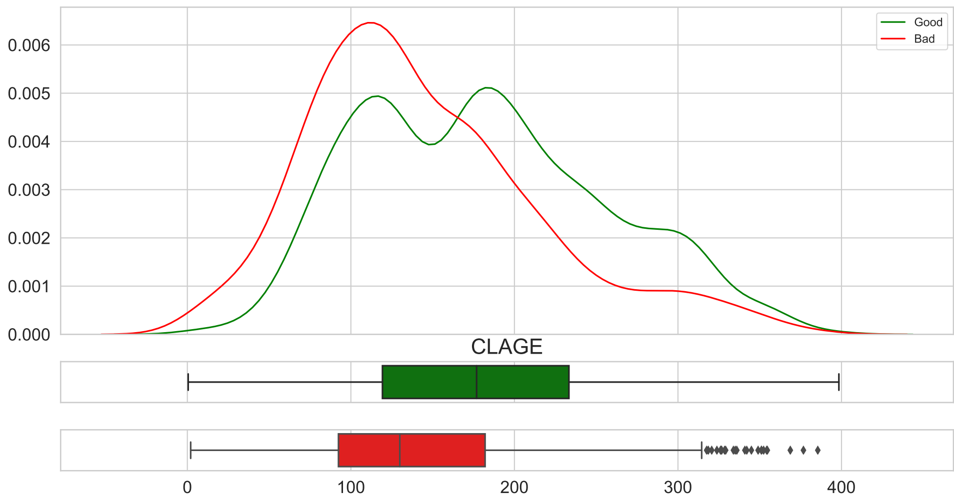

Figure 3.14: Distribution of CLAGE by BAD.

NINQ

NINQ, number of recent credit inquiries. Figure () indicates that the more recent credit equiries an applicant makes the more likely they are to default, this trend is followed until the higher values are reached but this deviation could be due to the low sample size at larger values. 73.8% of applicants have made no more than 1 recent credit equiry where as only 3% of applicants made 5 or more. This could create an issue depending on how the WOE method bins the values.

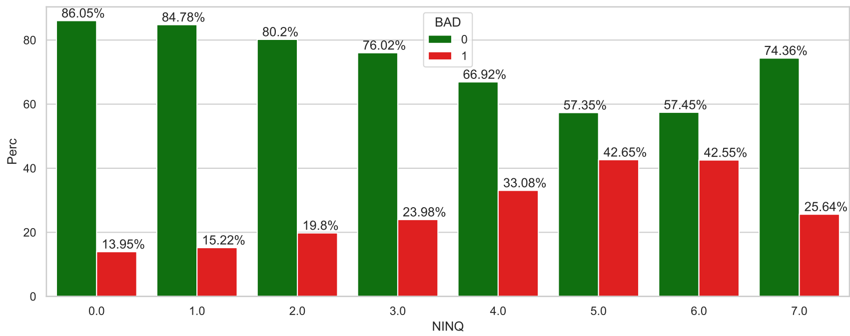

Figure 3.15: Category plot of NINQ by BAD.

CLNO

CLNO, number of credit lines. A credit line can be any method in which someone can recieve credit such as an overdraft, credit card, etc. Figure (3.16) suggests that applicants with low or high values for credit lines have a higher bad rate that ones in the centre. A reason behind this could be that a applicant with a low number of credit lines could be inexperience with debt management or don't have access to other credit to ensure payment on loans is on time. On the other end it could be that applicants with a large number of credit lines become incapable of managing the potential debt from the numerous sources. Where as the centre is seen to be where applicants have a good control over their credit lines. Applicants with a CLNO value between 20 and 30 have a bad rate of 15.1% where applicants outside of this group have a bad rate of 23.1%. A differnce of 8%.

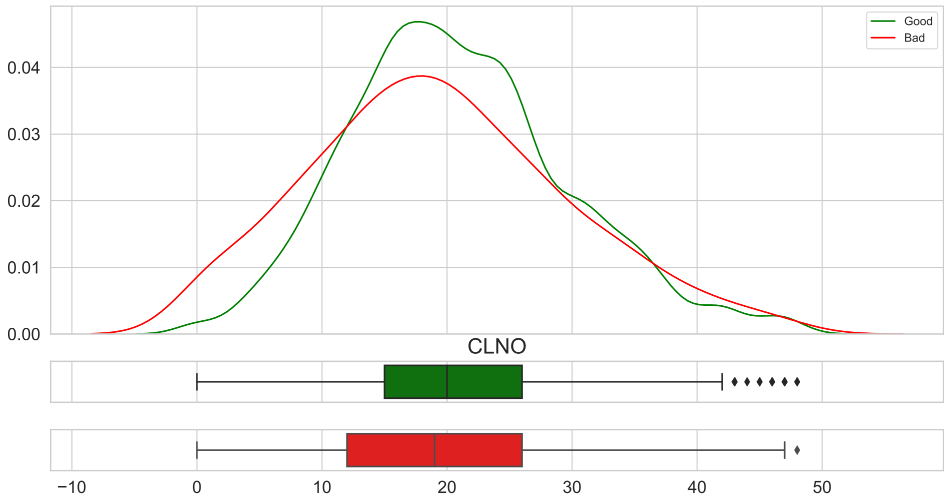

Figure 3.16: Distribution of CLNO by BAD.

3.3 WOE and IV

WOE binning and calculation was done using the scorecardpy package provided by ShichenXie **scorecardpy**, results can be found in the two tables (3.19) & (3.20). I allowed the function to determine optimal bins itself using the chimerge method described in Section 2.5. In the table we have a breakdown for each bin providing their woe value and individual information values. We can also see the total information value of each variable in the last column (total_iv). Based on IV, our strongest predictor by a large margin is DELINQ with a IV value of 0.47 and our weakest is Reason with an IV of 0.02. Referring back to Bailey's guideline (2.1)**bailey2004credit**, we can categorise each variable on their strength from the IV.

Looking back on our data exploration we can see that our assumption about some of the variables appear to be reflecting in the WOE bins. Reason, a variable I expected to be rather insignifcant due to the bad rate being similar in both categories has the lowest IV. NINQ appears to have been binned appropriately with the upper bin being [4.0, inf) meaning the drop in bad rate at 7.0, assumed to be from a small sample, should not have a siginicant effect on the woe value of the group. The assumption on CLAGE, the variable with the second highest IV, is also reflected in the WOE values with a clear trend appearing in the bad probabilities of each bin seen in figure (3.18).

Variables Prediction Strength

Variable	Strength	IV
LOAN	Average	0.25
MORTDUE	Average	0.10
VALUE	Average	0.20
REASON	Poor	0.02
JOB	Average	0.12
YOJ	Weak	0.07
DEROG	Average	0.23
DELINQ	Strong	0.47
CLAGE	Average	0.28
NINQ	Average	0.12
CLNO	Average	0.12

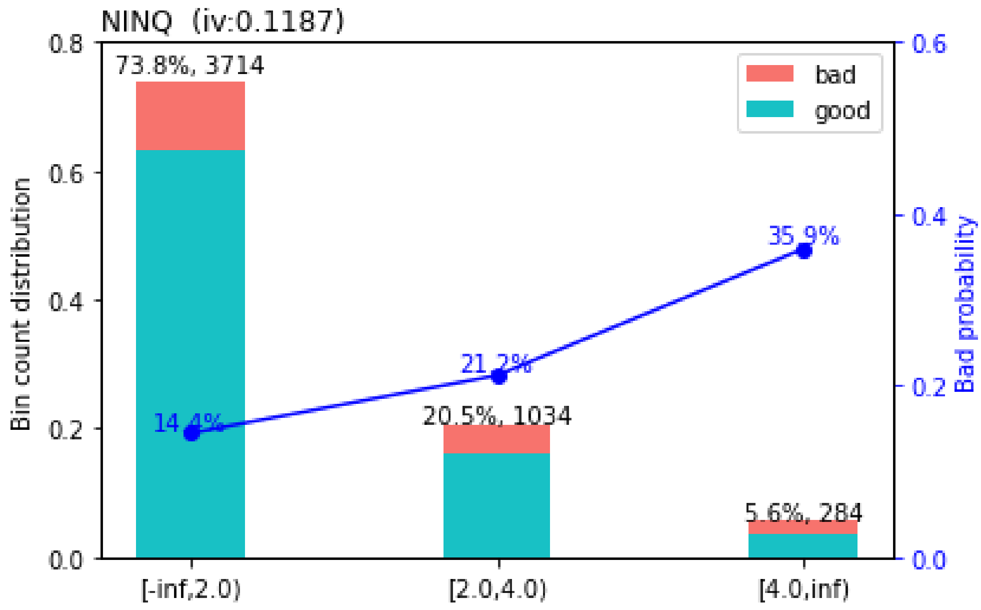

Figure 3.17: NINQ woe plot

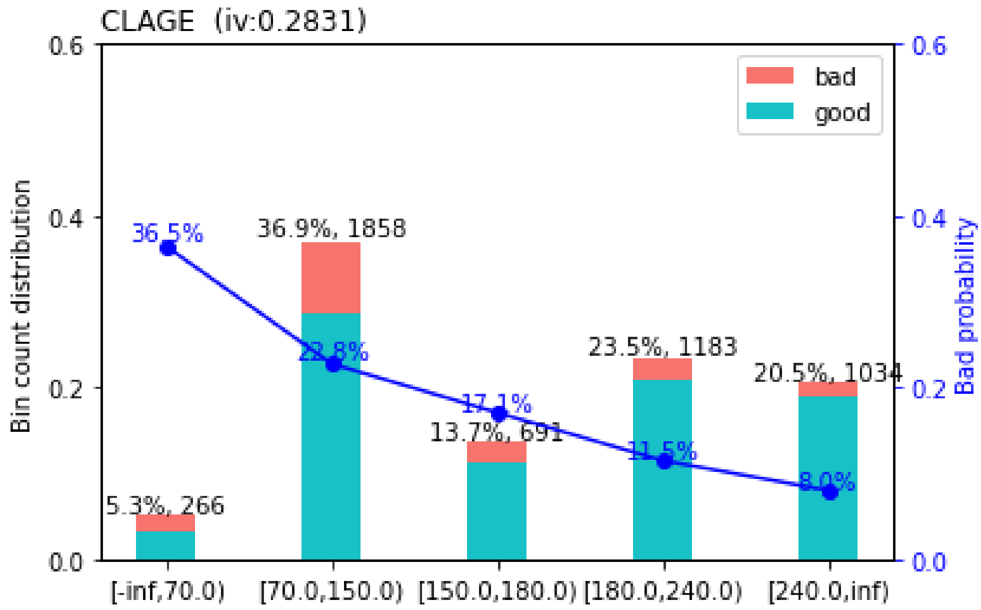

Figure 3.18: Clage woe plot

variable	bin	count	count_distr	good	bad	badprob	woe	bin_iv	total_iv
LOAN	[-inf, 6000.0)	254	0.05	136	118	0.46	1.44	0.15	0.25
	[6000.0, 8000.0)	270	0.05	204	66	0.24	0.45	0.01	0.25
	[8000.0, 11000.0)	678	0.13	546	132	0.19	0.16	0.00	0.25
	[11000.0, 15000.0)	936	0.19	790	146	0.16	-0.10	0.00	0.25
	[15000.0, 16000.0)	312	0.06	245	67	0.21	0.29	0.01	0.25
	[16000.0, 21000.0)	1006	0.20	879	127	0.13	-0.35	0.02	0.25
	[21000.0, 25000.0)	597	0.12	541	56	0.09	-0.68	0.04	0.25
	[25000.0, inf)	979	0.19	834	145	0.15	-0.17	0.01	0.25
MORTDUE	[-inf, 35000.0)	835	0.17	621	214	0.26	0.52	0.05	0.1
	[35000.0, 55000.0)	1172	0.23	960	212	0.18	0.07	0.00	0.1
	[55000.0, 60000.0)	355	0.07	309	46	0.13	-0.32	0.01	0.1
	[60000.0, 75000.0)	892	0.18	742	150	0.17	-0.02	0.00	0.1
	[75000.0, 105000.0)	1048	0.21	917	131	0.12	-0.36	0.02	0.1
	[105000.0, 130000.0)	348	0.07	289	59	0.17	-0.01	0.00	0.1
	[130000.0, inf)	382	0.08	337	45	0.12	-0.43	0.01	0.1
VALUE	[-inf, 50000.0)	488	0.10	318	170	0.35	0.96	0.12	0.2
	[50000.0, 70000.0)	998	0.20	835	163	0.16	-0.05	0.00	0.2
	[70000.0, 80000.0)	529	0.11	422	107	0.20	0.21	0.01	0.2
	[80000.0, 125000.0)	1952	0.39	1694	258	0.13	-0.30	0.03	0.2
	[125000.0, 175000.0)	619	0.12	499	120	0.19	0.16	0.00	0.2
	[175000.0, inf)	446	0.09	407	39	0.09	-0.76	0.04	0.2
REASON	DebtCon	3528	0.7	2968	560	0.16	-0.08	0.00	0.02
	HomeImp	1504	0.3	1207	297	0.20	0.18	0.01	0.02
JOB	Other	2100	0.42	1672	428	0.20	0.22	0.02	0.12
	Office	877	0.17	787	90	0.10	-0.58	0.05	0.12
	Sales	97	0.02	67	30	0.31	0.78	0.01	0.12
	Mgr	657	0.13	528	129	0.20	0.17	0.00	0.12
	ProfExe	1150	0.23	1004	146	0.13	-0.34	0.02	0.12
	Self	151	0.03	117	34	0.23	0.35	0.00	0.12

Figure 3.19: WOE results table.

variable	bin	count	count_distr	good	bad	badprob	woe	bin_iv	total_iv
YOJ	[-inf, 2.0)	750	0.15	615	135	0.18	0.07	0.00	0.07
	[2.0, 6.0)	1190	0.24	935	255	0.21	0.28	0.02	0.07
	[6.0, 12.0)	1729	0.34	1467	262	0.15	-0.14	0.01	0.07
	[12.0, 21.0)	918	0.18	750	168	0.18	0.09	0.00	0.07
	[21.0, inf)	445	0.09	408	37	0.08	-0.82	0.04	0.07
DEROG	[-inf, 1.0)	4482	0.89	3841	641	0.14	-0.21	0.04	0.23
	[1.0, inf)	550	0.11	334	216	0.39	1.15	0.20	0.23
DELINQ	[-inf, 1.0)	4086	0.81	3595	491	0.12	-0.41	0.12	0.47
	[1.0, 2.0)	553	0.11	372	181	0.33	0.86	0.11	0.47
	[2.0, inf)	393	0.08	208	185	0.47	1.47	0.24	0.47
CLAGE	[-inf, 70.0)	266	0.05	169	97	0.36	1.03	0.07	0.28
	[70.0, 150.0)	1858	0.37	1435	423	0.23	0.36	0.05	0.28
	[150.0, 180.0)	691	0.14	573	118	0.17	0.00	0.00	0.28
	[180.0, 240.0)	1183	0.24	1047	136	0.11	-0.46	0.04	0.28
	[240.0, inf)	1034	0.21	951	83	0.08	-0.86	0.11	0.28
NINQ	[-inf, 2.0)	3714	0.74	3178	536	0.14	-0.20	0.03	0.12
	[2.0, 4.0)	1034	0.21	815	219	0.21	0.27	0.02	0.12
	[4.0, inf)	284	0.06	182	102	0.36	1.00	0.08	0.12
CLNO	[-inf, 10.0)	496	0.10	339	157	0.32	0.81	0.08	0.12
	[10.0, 24.0)	2763	0.55	2332	431	0.16	-0.10	0.01	0.12
	[24.0, 27.0)	600	0.12	537	63	0.10	-0.56	0.03	0.12
	[27.0, inf)	1173	0.23	967	206	0.18	0.04	0.00	0.12

Figure 3.20: WOE results table.

Chapter 4

Modelling

4.1 Results

For the modelling, the data was split into train and test sets with proportions 0.7 and 0.3 respectively. The bad rate was maintained in each data at 17% to ensure fairness and the variables were converted to their woe values for their respective bin from tables (3.19) & (3.20). Once split the train data was passed through a glm model with logit link, for this the python package statsmodels was used **statsmodels**. Three models were created, first was the base model with every variable included with no changes. For the second model I looked at applying log transformations to the variables which had a heavy right skew, LOAN, MORTDUE, VALUE and YOJ. These values were then passed through the woe binning and the values converted. The IV of LOAN and YOJ improved by 0.01 and 0.03 respectively whilst MORTDUE and VALUE's IV decreased, based on this only the log transformations of LOAN and YOJ were kept and used for the second model. Finally, a third model was made from dropping variables with high p-values, these were MORTDUE and REASON.

The results for each of these models can be found in the appendix (), () and (). The performance of each model is shown and compared in Table (4.2). It can be seen from this table that based on the performance evaulation mentioned in Section 2.3. Model 2 and 3 appear to out perform Model 1 but when comparing the two the difference between performance indicators becomes smaller. In the case of Model 3 the AIC is lower but the GINI coefficient is also lower than Model 2. Based on Table (4.2) either Model 2 or 3 would be an appropriate choice but I decided to go with Model 2. This is because the difference in AIC is only 3.26 and when looking at the Log-Likelihood for each model the difference is 0.4. This is implying that the performance based on AIC is only better because of the removal of two variables making it a smaller and less complicated model. Where as the GINI coefficient doesn't consider the complication of the model.

Dep. Variable:	BAD	**No. Observations:**	3522
Model:	GLM	**Df Residuals:**	3510
Model Family:	Binomial	**Df Model:**	11
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1236.8
Date:	Sat, 15 Aug 2020	**Deviance:**	2473.6
Time:	13:16:04	**Pearson chi2:**	3.55e+03
No. Iterations:	6		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	-1.6002	0.054	-29.683	0.000	-1.706	-1.495
CLNO_woe	0.8862	0.154	5.755	0.000	0.584	1.188
REASON_woe	-0.1580	0.455	-0.347	0.729	-1.051	0.735
VALUE_woe	0.8818	0.159	5.558	0.000	0.571	1.193
YOJ_woe	1.1226	0.214	5.238	0.000	0.703	1.543
DELINQ_woe	1.0389	0.072	14.443	0.000	0.898	1.180
LOAN_woe	0.7584	0.115	6.581	0.000	0.533	0.984
DEROG_woe	0.8037	0.100	8.062	0.000	0.608	0.999
CLAGE_woe	0.8006	0.102	7.861	0.000	0.601	1.000
NINQ_woe	1.0246	0.139	7.397	0.000	0.753	1.296
JOB_woe	0.7379	0.155	4.764	0.000	0.434	1.041
MORTDUE_woe	0.0044	0.234	0.019	0.985	-0.455	0.464

Figure 4.1: Generalized Linear Model Regression Results

Model	AIC	KS	GINI	Potential Loss
Model 1 (Default)	2491.73	0.4892	0.6156	$0
Model 2 (Log transformations)	2460.77	0.4890	0.6186	$0
Model 3 (REASON and MORTDUE dropped)	2457.51	0.4858	0.6180	$0

Figure 4.2: Performance Evaluation Results On Test

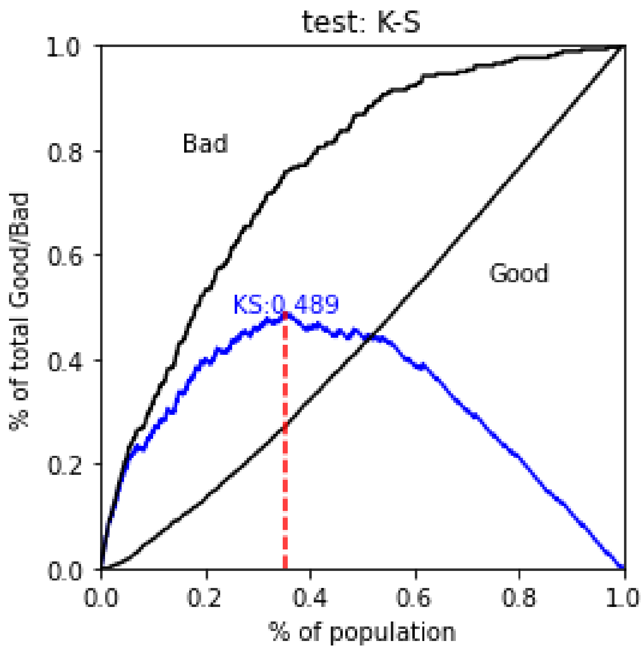

Figure 4.3: KS Plot

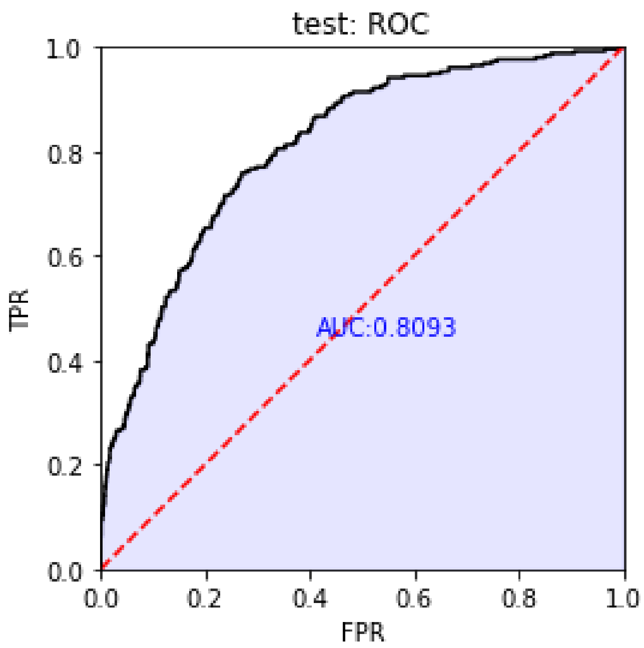

Figure 4.4: ROC Plot

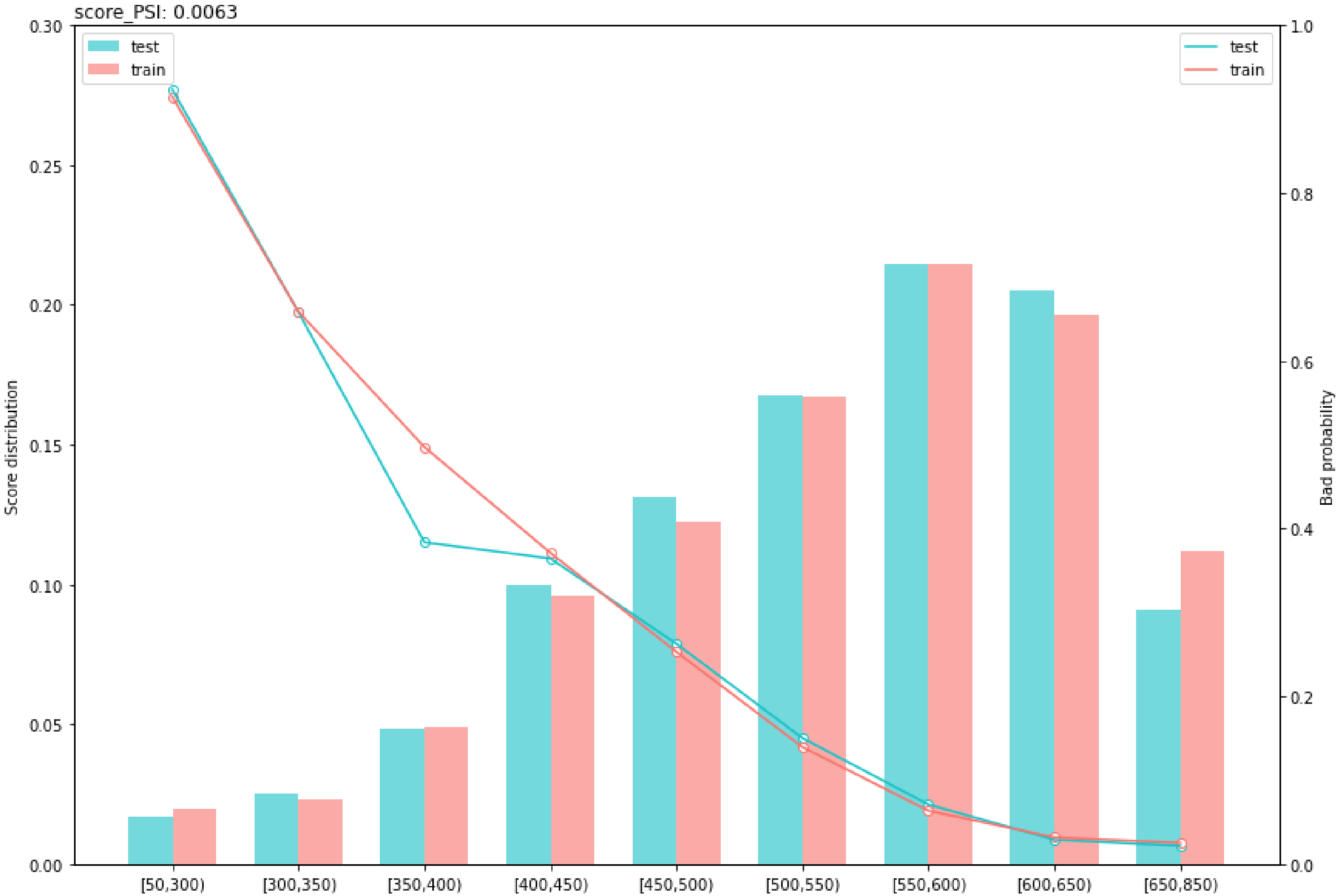

Figure 4.5: Scorecard Plot

Chapter 5

Alternative Uses: Covid-19

Appendices

Appendix A

Definintions

Definition 1. Goods. The term to define a good client, most often meaning a client which does not default on a loan

Definition 2. Bads. The term to define a bad client, most often meaning a client which ends up defaulting on a loan

Definition 3. Bad Rate. The percentage of clients that have defaulted.

Appendix B

Tables

	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 B.1: Summary Before Outliers Removed

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

Figure B.2: Summary After Cleaning

	LOAN	MORTDUE	VALUE	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO
BAD	-0.105	-0.0842	-0.1049	-0.0620	0.2175	0.2858	-0.1820	0.1369	-0.0632
LOAN		0.1730	0.3045	0.0453	0.0036	-0.0946	0.1172	0.0661	0.1117
MORTDUE			0.8994	-0.0693	-0.0432	-0.0424	0.1065	-0.0061	0.3389
VALUE				-0.0160	-0.0570	-0.0518	0.1775	-0.0267	0.3107
YOJ					-0.0464	0.0341	0.1669	-0.0488	0.0307
DEROG						0.1680	-0.0614	0.1249	0.0060
DELINQ							-0.0108	0.0298	0.1101
CLAGE								-0.0906	0.2202
NINQ									0.1046

Figure B.3: Correlation Table

Appendix C

Logistic Regression Results

Model:	GLM	AIC:	2491.7333
Link Function:	logit	BIC:	-26197.6795
Dependent Variable:	BAD	Log-Likelihood:	-1233.9
Date:	2020-08-25 15:28	LL-Null:	-1607.6
No. Observations:	3522	Deviance:	2467.7
Df Model:	11	Pearson chi2:	3.62e+03
Df Residuals:	3510	Scale:	1.0000
Method:	IRLS		

	Coef.	Std.Err.	z	P> |z|	[0.025	0.975]
const	-1.5952	0.0539	-29.5904	0.0000	-1.7009	-1.4896
DELINQ_woe	1.0533	0.0724	14.5417	0.0000	0.9114	1.1953
CLAGE_woe	0.7701	0.0988	7.7933	0.0000	0.5764	0.9638
CLNO_woe	0.9026	0.1435	6.2917	0.0000	0.6214	1.1838
REASON_woe	-0.1519	0.4517	-0.3364	0.7366	-1.0373	0.7334
MORTDUE_woe	0.0866	0.2208	0.3922	0.6949	-0.3462	0.5194
DEROG_woe	0.7937	0.1000	7.9387	0.0000	0.5978	0.9897
JOB_woe	0.6984	0.1539	4.5390	0.0000	0.3968	1.0000
NINQ_woe	1.0464	0.1420	7.3674	0.0000	0.7680	1.3247
VALUE_woe	0.7590	0.1518	4.9991	0.0000	0.4614	1.0565
LOAN_woe	0.7126	0.1104	6.4559	0.0000	0.4963	0.9289
YOJ_woe	1.0608	0.2156	4.9212	0.0000	0.6383	1.4833

Table C.1: Results: Model 1

Model:	GLM	AIC:	2460.7691
Link Function:	logit	BIC:	-26228.6438
Dependent Variable:	BAD	Log-Likelihood:	-1218.4
Date:	2020-08-25 15:33	LL-Null:	-1607.6
No. Observations:	3522	Deviance:	2436.8
Df Model:	11	Pearson chi2:	3.62e+03
Df Residuals:	3510	Scale:	1.0000
Method:	IRLS		

	Coef.	Std.Err.	z	P> |z|	[0.025	0.975]
const	-1.5975	0.0544	-29.3812	0.0000	-1.7041	-1.4909
DELINQ_woe	1.0724	0.0733	14.6302	0.0000	0.9287	1.2160
CLAGE_woe	0.7804	0.0996	7.8389	0.0000	0.5853	0.9755
CLNO_woe	0.9590	0.1443	6.6452	0.0000	0.6761	1.2418
REASON_woe	-0.3821	0.4607	-0.8295	0.4068	-1.2850	0.5208
MORTDUE_woe	0.0538	0.2224	0.2417	0.8090	-0.3821	0.4896
DEROG_woe	0.8055	0.1009	7.9863	0.0000	0.6078	1.0032
JOB_woe	0.7069	0.1547	4.5689	0.0000	0.4036	1.0101
NINQ_woe	0.9993	0.1421	7.0337	0.0000	0.7208	1.2777
VALUE_woe	0.7692	0.1530	5.0288	0.0000	0.4694	1.0691
LOAN_woe	0.7547	0.1082	6.9759	0.0000	0.5427	0.9668
YOJ_woe	1.1681	0.1790	6.5241	0.0000	0.8171	1.5190

Table C.2: Results: Model 2

Model:	GLM	AIC:	2457.5109
Link Function:	logit	BIC:	-26244.2355
Dependent Variable:	BAD	Log-Likelihood:	-1218.8
Date:	2020-08-25 15:36	LL-Null:	-1607.6
No. Observations:	3522	Deviance:	2437.5
Df Model:	9	Pearson chi2:	3.63e+03
Df Residuals:	3512	Scale:	1.0000
Method:	IRLS		

	Coef.	Std.Err.	z	P> |z|	[0.025	0.975]
const	-1.5969	0.0543	-29.4100	0.0000	-1.7033	-1.4905
DELINQ_woe	1.0703	0.0732	14.6192	0.0000	0.9268	1.2138
CLAGE_woe	0.7837	0.0992	7.9029	0.0000	0.5893	0.9780
CLNO_woe	0.9558	0.1423	6.7157	0.0000	0.6769	1.2348
JOB_woe	0.7140	0.1546	4.6197	0.0000	0.4111	1.0169
NINQ_woe	1.0089	0.1414	7.1329	0.0000	0.7317	1.2862
VALUE_woe	0.7878	0.1188	6.6292	0.0000	0.5549	1.0208
LOAN_woe	0.7237	0.1008	7.1777	0.0000	0.5261	0.9213
DEROG_woe	0.8076	0.1004	8.0444	0.0000	0.6109	1.0044
YOJ_woe	1.1716	0.1787	6.5550	0.0000	0.8213	1.5220

Table C.3: Results: Model 3

Appendix D

Python Code