Credit Scoring Whitepaper

Abstract

This document highlights the data, models and process used to develop a new Credit Scoring Model.

Eric Frayer
info@ericfrayer.com

September 2019

Contents

Introduction

This document provides an outline of the data and model construction used in building a new Credit Scoring model. It follows a framework outlined by Naeem Siddiqi in the second edition of his book "Intelligent Credit Scoring". The data comes from two sources: actual Experian Retro Data and a "bank loan" dataset provided by IBM. The goal of this paper is to highlight the process undertaken to build and deliver an actual credit scoring model – providing examples from both the "real" and "sample" datasets.

A version of this document was prepared for an actual client. This document presents how the results of the model were analyzed and shared. A formal presentation was provided to a credit/risk group at a large financial institution. The audience for the whitepaper included both technical and non-technical professionals. A Logistic Regression model was developed to provide better performance than using FICO alone. The FICO plus model out performed FICO alone allowing lenders to approve more borrowers (FICO scores of 680 to 720) with a positive economic benefit.

A new credit scoring model was built using Logistic Regression. Logistic Regression was chosen because of its straightforward approach and relative transparency. Binomial logistic regression attempts to predict the probability that an observation falls into one of two categories of a dichotomous dependent variable based on one or more independent variables. Those variables can be either continuous (e.g. Income, balance, FICO) or categorical (e.g. Presence of Co-signer, Level of Education). The ability to predict "good and bad" loans allows the underwriter greater flexibility in approving more originations.

Executive Summary

A credit scoring model provides Lenders with the option of increasing the number of approved loans applicants with relatively small increases in expected level of default. Model performance shows significant lift in approved applications with within expected default tolerance.

Customer Credit Model

A binomial logistic regression was used to predict the likelihood that borrowers will have a loss/default using the following variables: Borrower FICO, Co-Borrower FICO, Co-Borrower FICO = 0 (either Missing or no Cosigner), Borrower FICO = 0, Total Balance Open Personal Finance/Student Loan Accounts w/Update w/in 3 Months, Number Retail Accounts Satisfactory w/in 3 Months, Total Past Due Amount Revolving Accounts, Number Open Auto Finance Accounts w/Update w/in 3 Months w/ Balance >= 75% Loan Amount, Number Open Bankcard Accounts w/Update w/in 3 Months w/ Balance >= 75% Credit Limit/High Credit, Current Days Delinquent on Open Student Loan Trades, Worst Delinquency on Student Loan Trades, and presence of Co-Signer.

The logistic regression model was statistically significant. The model explained 21.9% (Nagelkerke R2) of the variance in loss/default and correctly classified 84.1% of cases.

Example Credit Model

A binomial logistic regression was run using the following 4 variables: Years with current employer, Years at current address, Household income in thousands, Debt to income ratio (x100).

The logistic regression model was statistically significant. The model explained 41.7% (Nagelkerke R2) of the variance in loss/default and correctly classified 82% of cases.

Data

The data used to build the **Customer Credit Model** came from a dataset provided by Experian. The model was built and evaluated using two datasets training and testing with the model being trained against 70% of the data with holdouts used to test and validate the model.

The variables used for the **Customer Scoring Model**:
Borrower FICO
Invalid or Missing Borrower FICO (Dummy)
Co-Borrower FICO
Invalid, Missing or No Co-Borrower FICO (Dummy)
Delinquency on Specific Tradelines
Current Days Delinquent on Specific Tradelines
Total Balance Open Personal Finance Loan Accounts w/Update w/in 3 Months
Number Retail Accounts Satisfactory w/in 3 Months
Total Past Due Amount Revolving Accounts
Number Open Auto Finance Accounts w/Update w/in 3 Months w/ Balance >= 75% Loan Amount
Number Open Bankcard Accounts w/Update w/in 3 Months w/ Balance >= 75% Credit Limit/High Credit
Co-Signer on loan (Distinguishes between invalid cofico0 and sole borrower)

The data used to build the model **Example Credit Model** come an IBM "bank loan" dataset. This is a hypothetical data file that concerns a bank's efforts to reduce the rate of loan defaults. The file contains financial and demographic information on 850 past and prospective customers. The first 700 cases are customers who were previously given loans. The last 150 cases are prospective customers that the bank needs to classify as good or bad credit risks.

The variables used for the **Example Scoring Model:**
Age in years
Level of education
Years with current employer
Years at current address
Household income in thousands
Debt to income ratio (x100)
Credit card debt in thousands
Other debt in thousands
Previously defaulted

Appendix A: provides additional details on the variables used in the two Credit models.

Model Construction

The Credit Underwriting Model was built using SAS JMP, R and SPSS. The Appendix provides the SAS, SPSS and R output and code needed for this project.

Logistic Regression

In many ways, binomial logistic regression is like linear regression, except for the measurement type of the dependent variable (i.e., linear regression uses a continuous dependent variable rather than a dichotomous one). However, unlike linear regression, you are not attempting to determine the predicted value of the dependent variable, but the probability of being in a “specific category” of the dependent variable given the independent variables. An observation is assigned to whichever category is predicted as most likely. As with other types of regression, binomial logistic regression can also use interactions between independent variables to predict the dependent variable.

A credit scoring model is generally used in the decision-making process of accepting or rejecting a loan. The credit scoring model is the result of a statistical model which, based on information about a borrower (e.g. age, number of previous loans, etc.), allows one to distinguish between "good" and "bad" loans and give an estimate of the probability of default.

One of the most common, successful and transparent ways to do the required binary classification to “good” and “bad” is via a logistic function. This is a function that takes as input the borrower characteristics and outputs the probability of default or loss.

$$p = \frac{\exp(\beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_n \cdot x_n)}{1 + \exp(\beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_n \cdot x_n)}$$

where in the above:

- p is the probability of default
- xi is the explanatory factor i
- βi is the regression coefficient of the explanatory factor i
- n is the number of explanatory variables

For each of the existing data points it is known whether the loan is a good or bad loan (i.e. p=1 or p=0). The aim in the here is to find the coefficients β0,… , βn such that the model’s probability of loss equals the observed probability of loss. The above logistic function contains the borrower characteristics in a linear way (i.e. as $\beta 0 + \beta 1 \cdot x1 + \cdots + \beta n \cdot xn$). Note: while Linear Regression uses Sum of Least Square estimates Logit Regression uses Maximum Likelihood.

The assumptions of a binomial logistic regression will allow you to: (a) provide information on the accuracy of your predictions; (b) test how well the regression model fits your data; (c) determine the variation in your dependent variable explained by your independent variables; and (d) test hypotheses on your regression equation. If these assumptions are violated, you need to make corrections and re-test these assumptions. If they still do not pass, you must find alternative statistical tests. For the Credit Model no assumptions were violated.

Model Performance

To evaluate the performance of the model several methods were used. Many of these are likely familiar to Credit and Risk officers at financial institutions. In addition to the statistical measures described below the models were executed in SAS JMP, SPSS and R.

C-Statistic

The C-statistic, which is also called the AUC or area under the ROC curve, is an R-square-like measure used in logistic regression. The C-statistic can range from 0.50 to 1.00, with higher values indicating better predictive models. A rough rule for interpretation is that C-statistics above 0.80 indicate very good models, between 0.70 and 0.80 good models, and between 0.50 and 0.70 weak models.[i]

This is the most powerful nonparametric two-sample test, and the measure is equivalent to the area under the ROC curve, Gini coefficient, and the Wilcoxon-Mann-Whitney test. It measures classifier performance across all score ranges and is a strong measure of overall scorecard strength.

For the Customer **Credit Model,** the **c-stat** observed was **76.9** outperforming the FICO alone **C-stat** of **71.7**[ii] For the **Example Credit Model** the c-stat was 87.0

KS -Kolmogorov-Smirnov (KS) statistic

This measures the maximum vertical separation (deviation) between the cumulative distributions of goods and bads and is a very widely used measure of divergence/separation. One issue with KS is that the separation is measured only at the one point (which may not be around the expected cutoff point), and not on the entire score range. If the intended scorecard cutoff is at the upper or lower range of scores, this measure may not provide a good method of scorecard comparison, since the statistic would be irrelevant to the decision at hand. In such cases, it might be better to compare the deviation at the intended cutoff, since that is where maximum separation is most required, and indeed, if divergence is a priority.[iii]

Confusion Matrix

An Error or Confusion Matrix can be used to evaluate the performance of a Credit Scoring Model. This simple and relatively straightforward method presents the Accuracy and Precision. The cutoff (50%) is being used.

		Model Prediction	
		No Loss (0)	Loss (1)
Actual Loan Status	No Loss (0)	True Negative (TN)	False Positive (FP)
	Loss (1)	False Negative (FN)	True Positive (TP)

True Positives (TP): These are cases in which the model predicted loss correctly
True Negatives (TN): The model predicts loss and there wasn't a loss
False Positives (FP): Predicted Loss, but not a loss. (Also known as a "Type I error.)
False negatives (FN): The model predicts no loss, but case was actually a loss.
(Also known as a "Type II error.")[iv]

This is a list of rates that are often computed from a confusion matrix for a binary classifier:

		Model Prediction	
Credit Model	n=4198	No Loss (0)	Loss (1)
Actual Loan Status	No Loss (0)	3459 or 82.4%	51 or 1.2%
	Loss (1)	620 or 14.8%	68 or 1.6%

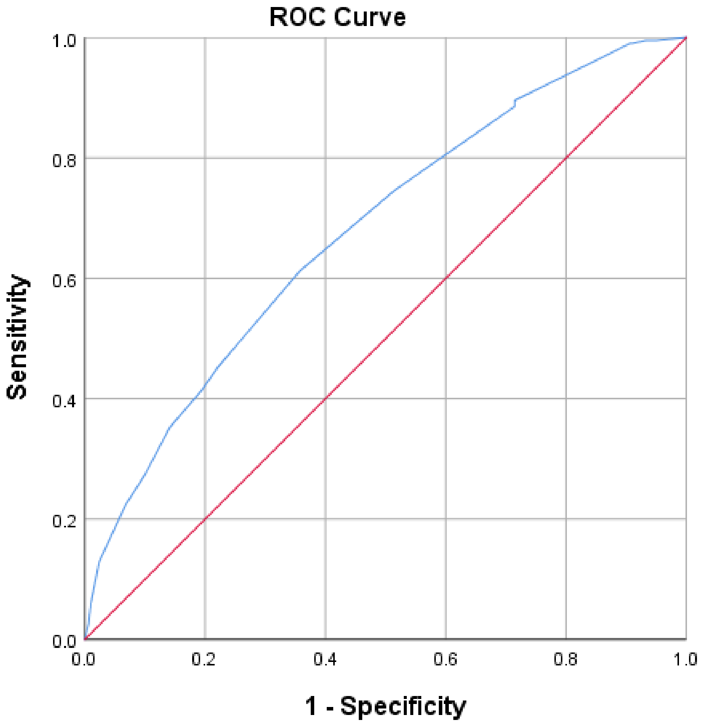
Accuracy: How often is the classifier correct? (TP+TN)/total = (3459+68)/4198 = **84.8%**
Error Rate: How often is it wrong? (FP+FN)/total = (51+620)/4198 = **15.2%**

		Model Prediction	
Example Credit Model	n=700	No Loss (0)	Loss (1)
Actual Loan Status	No Loss (0)	478 or 68.3%	39 or 5.6%
	Loss (1)	91 or 13.0%	92 or 13.1%

Accuracy: How often is the classifier correct? (TP+TN)/total = (478+91)/700 = **81.4%**
Error Rate: How often is it wrong? (FP+FN)/total = (51+620)/700 = **18.6%**

ROC Curve – Area Under Curve

The following ROC and AUC curves were reproduced both in SPSS and R. The graph below is for the Credit **Model**. The **AUC value .769** is consistently reported in all three applications (SAS, SPSS and R).

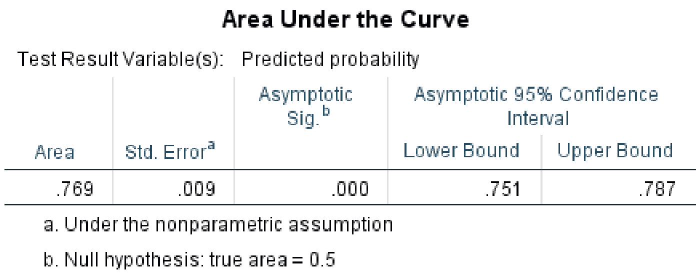


The sub-note (a.) shows the positive actual state is "1.00 Yes", indicating that we have correctly stated the event (i.e., the event of interest in this example is having loss/default, which was coded as "1").

The further the blue line is above the straight line, the better the discrimination. The area under the ROC curve is equivalent to the concordance probability (Gönen, 2007). The concordance (c) statistic is the most common measure of the ability binomial logistic regression model to discriminate.

It is equivalent to the area under the ROC curve for a dichotomous dependent variable (i.e., for binomial (or binary) logistic regressions) (Gönen, 2007; Steyerberg, 2009).

The AUC (Area Under Curve) is the probability that a randomly chosen positive case receives a score higher than a randomly chosen negative case.



Area Under the Curve

Test Result Variable(s): Predicted probability

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.769	.009	.000	.751	.787

a. Under the nonparametric assumption

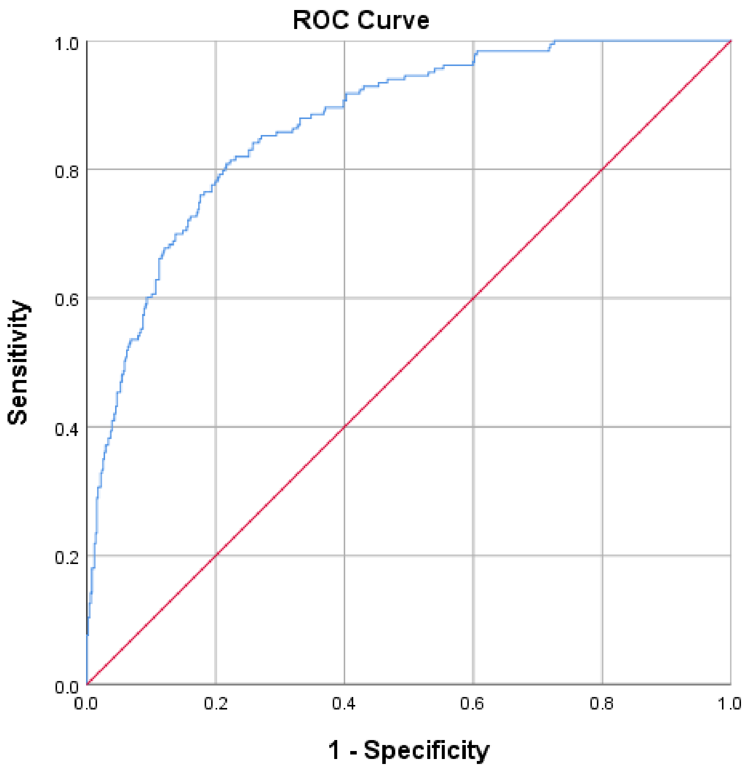
b. Null hypothesis: true area = 0.5

The **Customer Credit Model** area under the ROC curve is **.769**. The area can range from 0.5 to 1.0 with higher values representing better discrimination. According to Hosmer et al. (2013) a value of .769 puts the discrimination of this model at the middle range of acceptable discrimination.

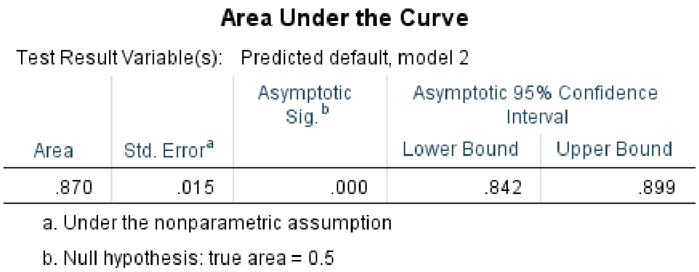
The general rules of thumb of Hosmer et al. (2003) are presented below:

AUC	Classification
0.5	This suggests no discrimination, so we might as well flip a coin.
0.5 < AUC < 0.7	We consider this poor discrimination, not much better than a coin toss.
0.7 ≤ AUC < 0.8	We consider this acceptable discrimination.
0.8 ≤ AUC < 0.9	We consider this excellent discrimination.
AUC ≥ 0.9	We consider this outstanding discrimination.

The **Example Credit Model** area under the ROC curve is **.870**.



The **Example Credit Model** area under the ROC curve is **.870**.



Area Under the Curve

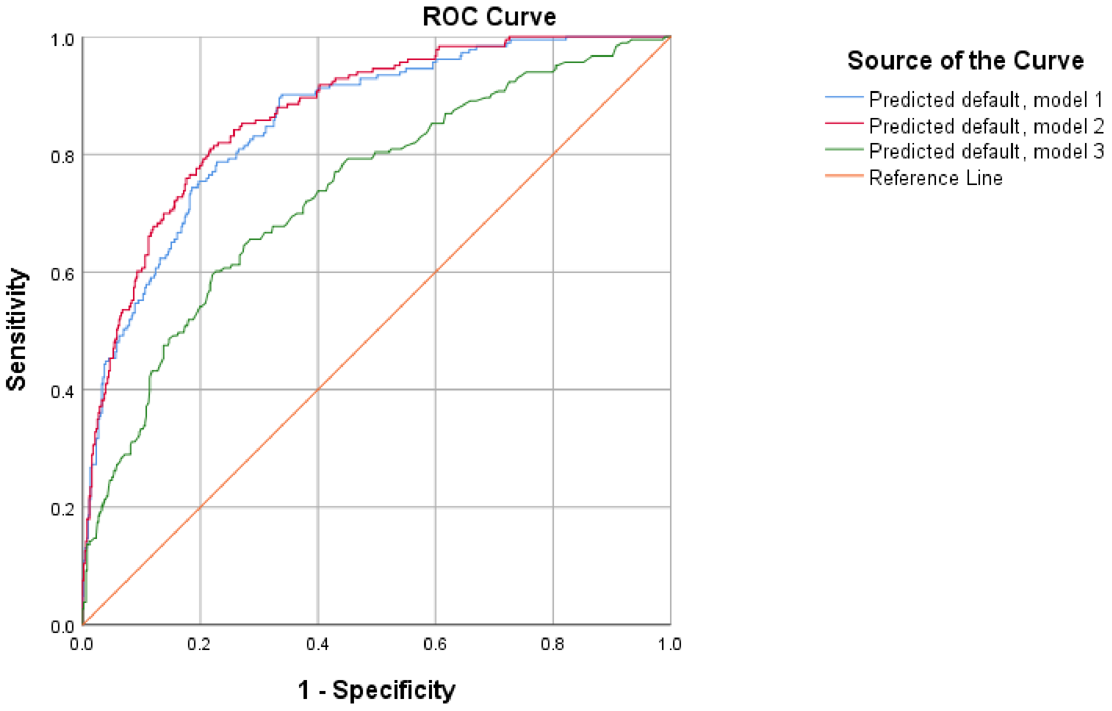
Test Result Variable(s): Predicted default, model 2

		Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
Area	Std. Error[a]		Lower Bound	Upper Bound
.870	.015	.000	.842	.899

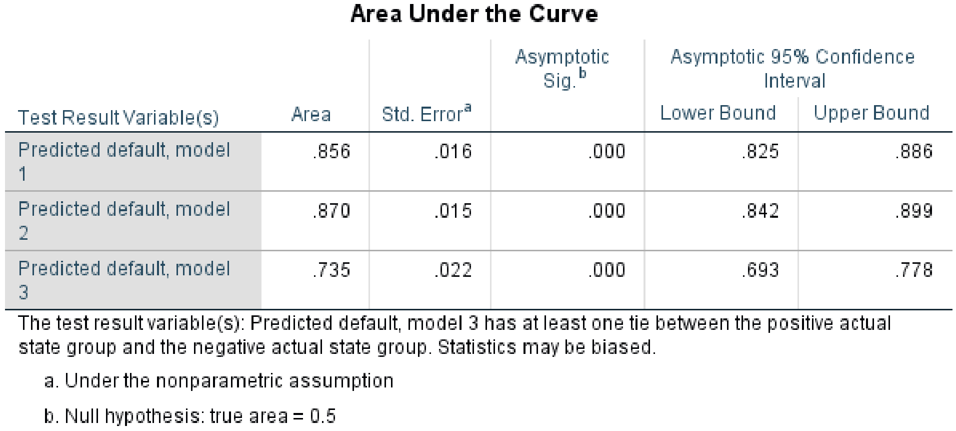
a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Note: the SPSS example included three models each slightly different.



Diagonal segments are produced by ties.



Area Under the Curve

Test Result Variable(s)	Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
Predicted default, model 1	.856	.016	.000	.825	.886
Predicted default, model 2	.870	.015	.000	.842	.899
Predicted default, model 3	.735	.022	.000	.693	.778

The test result variable(s): Predicted default, model 3 has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Appendix A: Data Characteristics

The data used for the models contain the following fields:

Experian Data

Variable Name	**Variable Type**	**Measure**	**Model Variable Name**	**Transform**
Loss	Dependent	Nominal	Loss	Intercept
Total Balance Open Personal Finance/Student Loan Accounts w/Update w/in 3 Months	Independent	Scale	l_cma3166_br_lk	Log
Number Retail Accounts Satisfactory w/in 3 Months	Independent	Scale	l_cma3335_br_lk	Log
Total Past Due Amount Revolving Accounts	Independent	Scale	l_cma3246_co_lk	Log
Number Open Auto Finance Accounts w/Update w/in 3 Months w/ Balance >= 75% Loan Amount	Independent	Nominal	o_cma3725_co_lk	Dummy
Number Open Bankcard Accounts w/Update w/in 3 Months w/ Balance >= 75% Credit Limit/High Credit	Independent	Nominal	o_cma3726_br_lk	Dummy
Co-Borrower FICO	Independent	Scale	lastcofico	Linear
Borrower FICO	Independent	Scale	lastknownborrfico_t	Linear
Co-Borrower FICO = 0 (either Missing or no Cosigner)	Independent	Nominal	cofico0	Dummy
Borrower FICO = 0	Independent	Nominal	bfico0	Dummy
Cosigner Exist	Independent	Nominal	cosigner_exist	Dummy
Current Days Delinquent on Open Student Loan Trades	Independent	Scale	worstdel_t	Log
Worst Delinquency on Student Loan Trades	Independent	Scale	worstdel24_t	Log

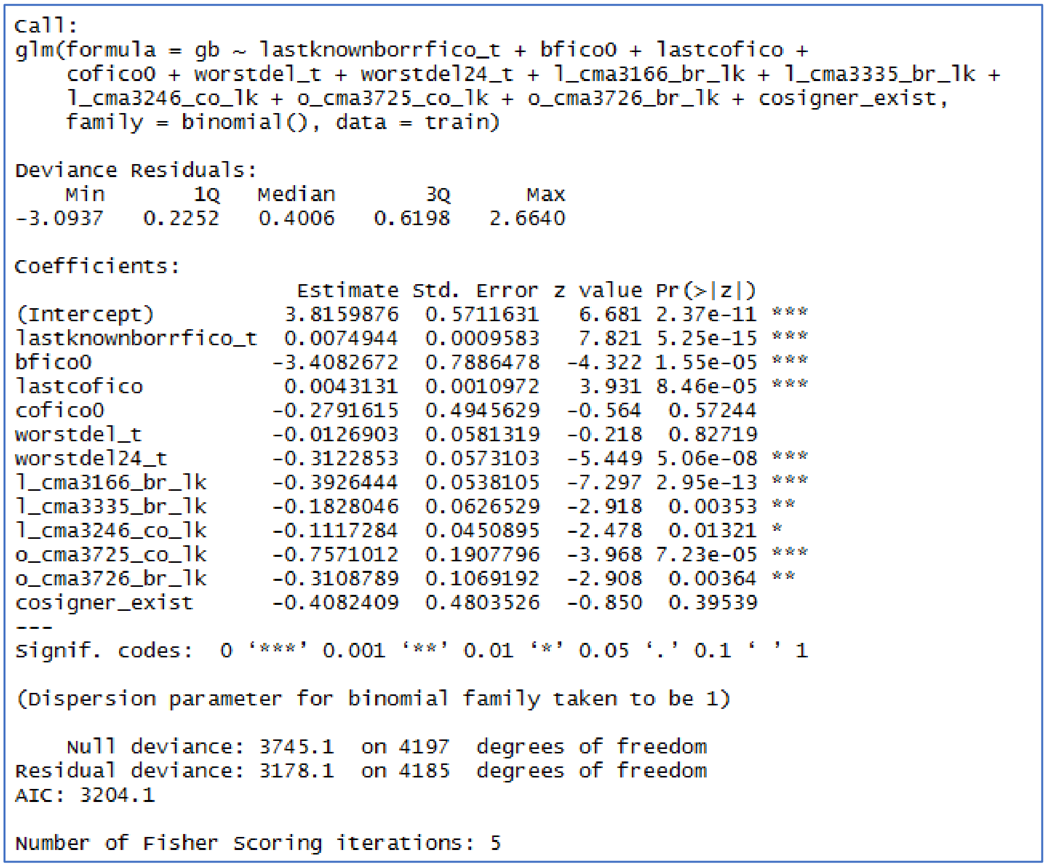
Bank Loan Data

Variable Name	**Variable Type**	**Measure**	**Model Variable Name**	**Transform**
Previously defaulted	Dependent	Nominal	default	Intercept
Age	Independent	Scale	age	Log
Years with current employer	Independent	Scale	employ	Log
Years at current address	Independent	Scale	address	Log
Household income in thousands	Independent	Scale	income	Log
Debt to income ratio (x100)	Independent	Scale	debtinc	Log
Credit card debt in thousands	Independent	Scale	creddebt	Log
Other debt in thousands	Independent	Scale	othdebt	Log

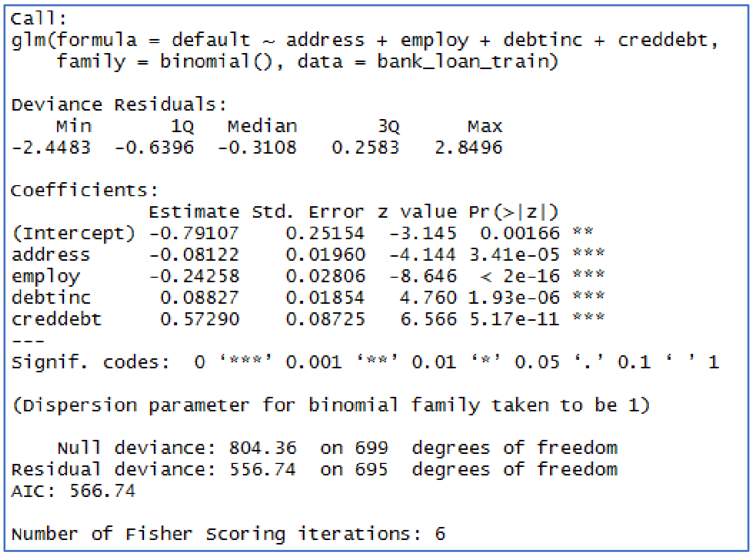
Appendix B: R, SPSS and SAS JMP

To validate the model and conclusions R, Rattle and SPSS were used. The data presented below represents the output in R for the **Customer Scoring** and **Example Scoring** models respectively.

Model Performance - The figure below shows the output in R of the GLM model. The variables used in the model are presented at the top. The coefficients, std error and p-value all align.



```
Call:
glm(formula = gb ~ lastknownborrfico_t + bfico0 + lastcofico +
    cofico0 + worstdel_t + worstdel24_t + l_cma3166_br_lk + l_cma3335_br_lk +
    l_cma3246_co_lk + o_cma3725_co_lk + o_cma3726_br_lk + cosigner_exist,
    family = binomial(), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0937   0.2252   0.4006   0.6198   2.6640

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          3.8159876  0.5711631   6.681 2.37e-11 ***
lastknownborrfico_t  0.0074944  0.0009583   7.821 5.25e-15 ***
bfico0              -3.4082672  0.7886478  -4.322 1.55e-05 ***
lastcofico           0.0043131  0.0010972   3.931 8.46e-05 ***
cofico0             -0.2791615  0.4945629  -0.564  0.57244
worstdel_t          -0.0126903  0.0581319  -0.218  0.82719
worstdel24_t        -0.3122853  0.0573103  -5.449 5.06e-08 ***
l_cma3166_br_lk     -0.3926444  0.0538105  -7.297 2.95e-13 ***
l_cma3335_br_lk     -0.1828046  0.0626529  -2.918  0.00353 **
l_cma3246_co_lk     -0.1117284  0.0450895  -2.478  0.01321 *
o_cma3725_co_lk     -0.7571012  0.1907796  -3.968 7.23e-05 ***
o_cma3726_br_lk     -0.3108789  0.1069192  -2.908  0.00364 **
cosigner_exist      -0.4082409  0.4803526  -0.850  0.39539
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3745.1  on 4197  degrees of freedom
Residual deviance: 3178.1  on 4185  degrees of freedom
AIC: 3204.1

Number of Fisher Scoring iterations: 5
```



```
Call:
glm(formula = default ~ address + employ + debtinc + creddebt,
    family = binomial(), data = bank_loan_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4483  -0.6396  -0.3108   0.2583   2.8496

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.79107    0.25154  -3.145  0.00166 **
address     -0.08122    0.01960  -4.144 3.41e-05 ***
employ      -0.24258    0.02806  -8.646  < 2e-16 ***
debtinc      0.08827    0.01854   4.760 1.93e-06 ***
creddebt     0.57290    0.08725   6.566 5.17e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 804.36  on 699  degrees of freedom
Residual deviance: 556.74  on 695  degrees of freedom
AIC: 566.74

Number of Fisher Scoring iterations: 6
```

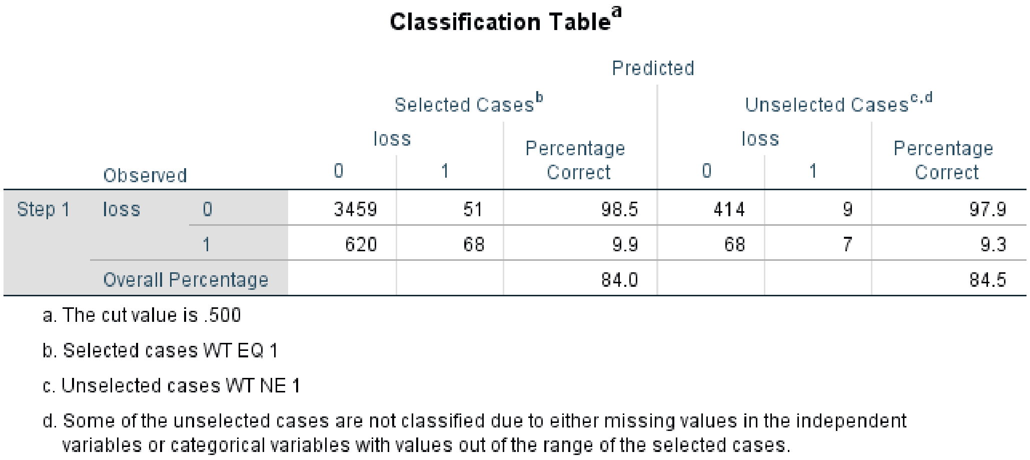
Analysis in SPSS

A copy of the data was also used in a model built with SPSS Statistics 25.

Category prediction

Binomial logistic regression estimates the probability of an event occurring. If the estimated probability of the event occurring is greater than or equal to 0.5 (better than even chance), SPSS Statistics classifies the event as occurring. If the probability is less than 0.5, SPSS Statistics classifies the event as not occurring. It is very common to use logistic regression to predict whether cases can be correctly classified (i.e., predicted) from the independent variables.

Therefore, it becomes necessary to have a method to assess the effectiveness of the predicted classification against the actual classification. There are many methods to assess this with their usefulness often depending on the nature of the study conducted. However, all methods revolve around the observed and predicted classifications, which are presented in the Classification Table, as shown below.



Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c,d]		
			loss		Percentage Correct	loss		Percentage Correct
	Observed		0	1		0	1	
Step 1	loss	0	3459	51	98.5	414	9	97.9
		1	620	68	9.9	68	7	9.3
	Overall Percentage				84.0			84.5

a. The cut value is .500

b. Selected cases WT EQ 1

c. Unselected cases WT NE 1

d. Some of the unselected cases are not classified due to either missing values in the independent variables or categorical variables with values out of the range of the selected cases.

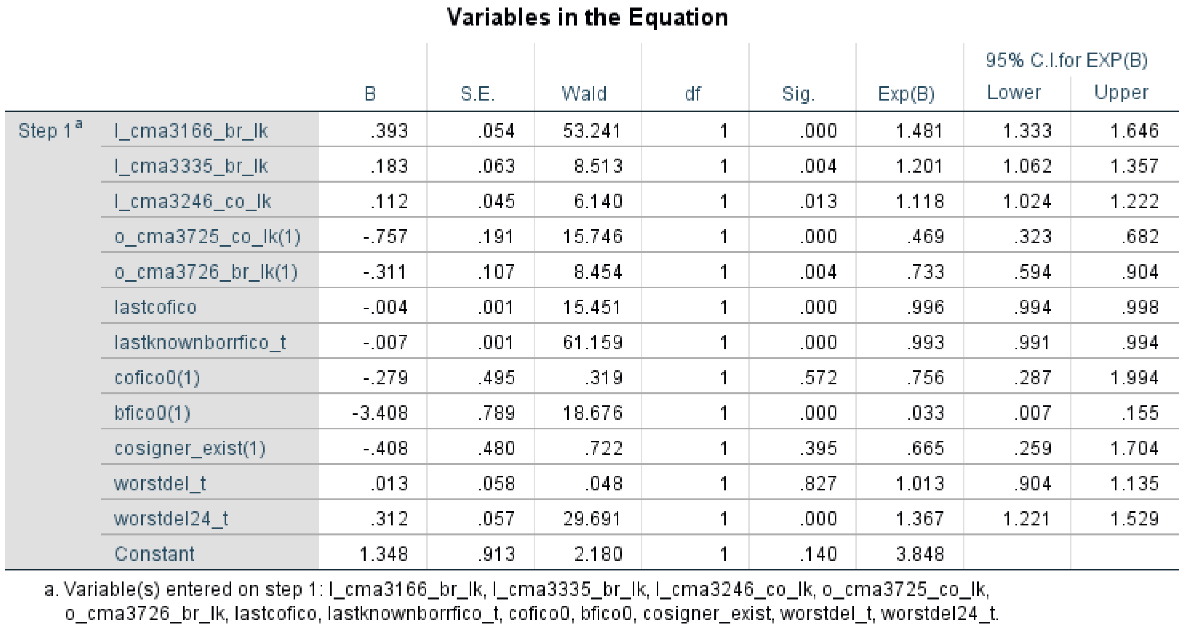
The first table subscript (a.) states "The cut value is .500". This means that if the probability of a case being classified into the "yes" category is greater than .500, then that particular case is classified into the "yes" category. Otherwise, the case is classified as in the "no" category.

The model correctly classifies 84.02% of cases overall (see "Overall Percentage" row). The addition of the independent variables improves the overall prediction of cases into their observed categories of the dependent variable. This measure is referred to as the percentage accuracy in classification (PAC).

Another measure is the sensitivity, which is the percentage of cases that had the observed characteristic (e.g., "yes" for loss/default) which were correctly predicted by the model (i.e., true positives). In this case, 15.98% of participants who were loss/default were also predicted by the model to fall into the loss/default category (see the "Percentage Correct" column in the "Yes" row of the observed categories).

Specificity is the percentage of cases that did not have the observed characteristic (e.g., "no" for loss/default) and were also correctly predicted as not having the observed characteristic (i.e., true negatives). In this case, 84.8% of “good” borrowers were correctly predicted by the model not to fall into the loss/default category (see the "Percentage Correct" column in the "No" row of the observed categories).

The **Variables in the Equation** table shows the contribution of each independent variable to the model and its statistical significance.



Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	l_cma3166_br_lk	.393	.054	53.241	1	.000	1.481	1.333	1.646
	l_cma3335_br_lk	.183	.063	8.513	1	.004	1.201	1.062	1.357
	l_cma3246_co_lk	.112	.045	6.140	1	.013	1.118	1.024	1.222
	o_cma3725_co_lk(1)	-.757	.191	15.746	1	.000	.469	.323	.682
	o_cma3726_br_lk(1)	-.311	.107	8.454	1	.004	.733	.594	.904
	lastcofico	-.004	.001	15.451	1	.000	.996	.994	.998
	lastknownborrfico_t	-.007	.001	61.159	1	.000	.993	.991	.994
	cofico0(1)	-.279	.495	.319	1	.572	.756	.287	1.994
	bfico0(1)	-3.408	.789	18.676	1	.000	.033	.007	.155
	cosigner_exist(1)	-.408	.480	.722	1	.395	.665	.259	1.704
	worstdel_t	.013	.058	.048	1	.827	1.013	.904	1.135
	worstdel24_t	.312	.057	29.691	1	.000	1.367	1.221	1.529
	Constant	1.348	.913	2.180	1	.140	3.848		

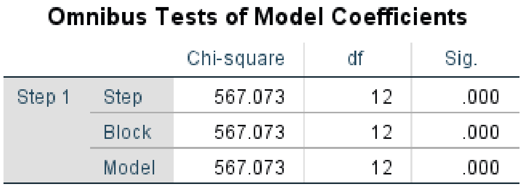
a. Variable(s) entered on step 1: l_cma3166_br_lk, l_cma3335_br_lk, l_cma3246_co_lk, o_cma3725_co_lk, o_cma3726_br_lk, lastcofico, lastknownborrfico_t, cofico0, bfico0, cosigner_exist, worstdel_t, worstdel24_t.

The Wald test ("Wald" column) is used to determine statistical significance for each of the independent variables. The statistical significance of the test is found in the "Sig." column.

The B coefficients ("B" column) are used in the equation to predict the probability of an event occurring, but not in an immediately intuitive manner. The coefficients do, in fact, show the change in the log odds that occur for a one-unit change in an independent variable when all other independent variables are kept constant.

SPSS Statistics also includes the odds ratios of each of the independent variables in the "Exp(B)" column along with their confidence intervals ("95% C.I. for EXP(B)" column). This informs of the change in the odds for each increase in one unit of the independent variable.

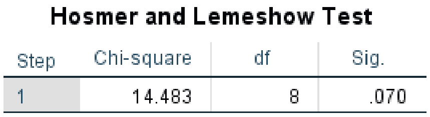
The next section examines model fit. The following tables represent the results of the main logistic regression analysis with all independent variables added to the equation.



Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	567.073	12	.000
	Block	567.073	12	.000
	Model	567.073	12	.000

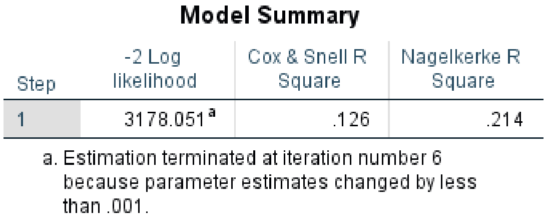
For this type of binomial logistic regression, the important reference the "Model" row. This table above shows the model is statistically significant (p < .0005; "Sig." column). Another way of assessing the adequacy of the model is to analyze how poor the model is at predicting the categorical outcomes. This is tested using the Hosmer and Lemeshow goodness of fit test. Hosmer and Lemeshow test is not statistically significant (p = .009; "Sig." column), indicating that the model is not a poor fit.



Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	14.483	8	.070

Finally, the Model Summary table contains the Cox & Snell R Square and Nagelkerke R Square values. Both are methods of calculating "explained variation". These measures are not as straightforward as R2 in multiple regression but are sometimes referred to as pseudo R2. Note: the values and will have lower values than in multiple regression. However, they are interpreted in the same manner, but with more caution.



Model Summary

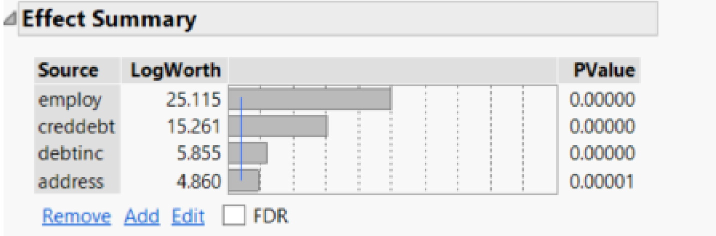
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	3178.051[a]	.126	.214

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

The explained variation in the dependent variable based on our model ranges from 12.6% (Cox & Snell R2) to 21.4% (Nagelkerke R2). Nagelkerke R2 is a modification of Cox & Snell R2, the latter of which cannot achieve a value of 1. For this reason, it is preferable to look at the Nagelkerke R2 value.

Analysis in SAS JMP

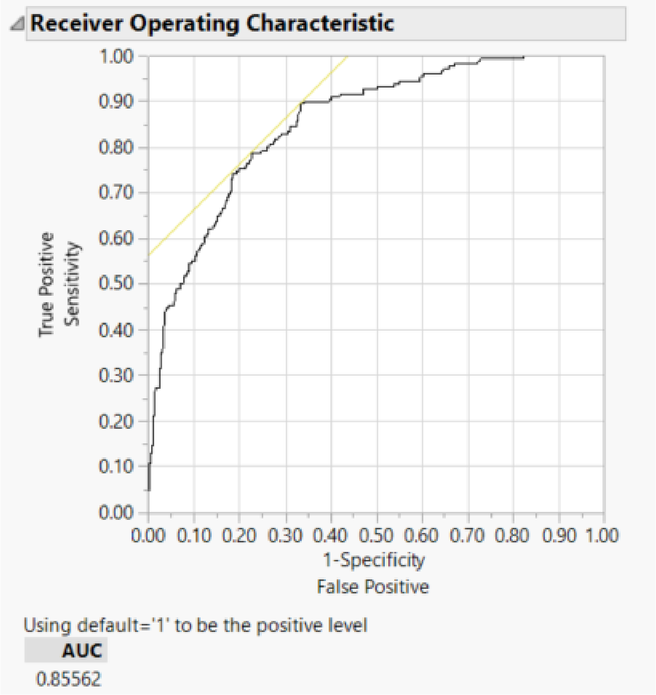
The **Example Credit Model** analysis aligned with the Logistic Regression models run in R and SPSS. While the number of observations varied based on the Training/Test sizes the results are comparable. The coefficients strength in the Example Credit model are shown below. The order is Years at Current Employer, Credit card debt in thousands, Debt to income ratio (x100) and Years at current address.



Effect Summary

Source	LogWorth	PValue
employ	25.115	0.00000
creddebt	15.261	0.00000
debtinc	5.855	0.00000
address	4.860	0.00001

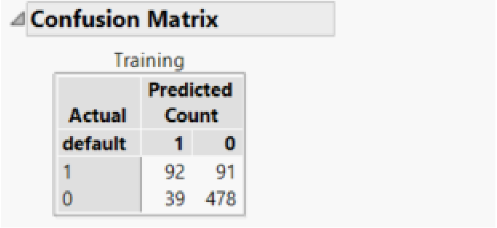
The AUC and ROC curve is the same in SAS, SPSS and R.



Using default='1' to be the positive level

AUC
0.85562

The Area Under the Curve can also be called the C-Stat. Using SAS and 700 records the C-Stat was **85.5**



Confusion Matrix

Training

Actual	Predicted Count	
default	1	0
1	92	91
0	39	478

Confusion Matrix - SAS

N = 700	Predicted			
Actual		0	1	
	0	68.3%	5.6%	
	1	13.0%	13.1%	
				100.0%
			Correct	81.4%
			Wrong	18.6%

Notes

[i] Siddiqi, Naeem. *Intelligent credit scoring: building and implementing better credit risk scorecards*. Hoboken, New Jersey: Wiley, 2017. Print.
[ii] "(ibid. p. 57)"
[iii] “(ibid. p. 230)”
[iv] https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/