Binary Classification Modeling

Using Logistic Regression to Build Credit Scores

John Michael Croft

Supervised by Jennifer Lewis Priestley, Ph.D.

Kennesaw State University

Submitted April 29, 2016

to fulfill the requirements for STAT8330

Executive Summary

The objective of the analysis is to generate a credit default risk scoring model to maximize expected profitability using SAS. The initial sample data contained over 17 million observations on over 1.25 million customers across more than 300 variables. A binary response is created based on a customer's highest delinquency identification to determine if they are a high credit default risk. Extreme values (greater than four mean standard deviations), and coded values are imputed to the median to minimize distributional impact due to significant right skewness in many variables. Multicollinearity is reduced via variable cluster analysis. Remaining variables undergo two discretization transformations (supervised and unsupervised) allowing for a potential six additional variables for model consideration. Transformations include equal width bins and equal frequency bins as well as the odds and log odds for each. A backward selection logistic regression is implemented. The Wald chi square statistic (higher is better) is used to select the top variable predictors. Model adequacy is measured with the KS statistic and C statistic (higher is better for both). Twelve variables are included in the final model for operational feasibility. Expected profit is evaluated across all possible credit default probabilities using a profit function where each correctly predicted non-default yields a $250 gain and each incorrectly non-default yields a loss of half the credit limit. The final model predicts an average profit per 1,000 customers scored of approx. $110K on validation data. A simplified version of the final model containing the original form of final variables is included for comparison and results in slightly higher expected profit.

Introduction

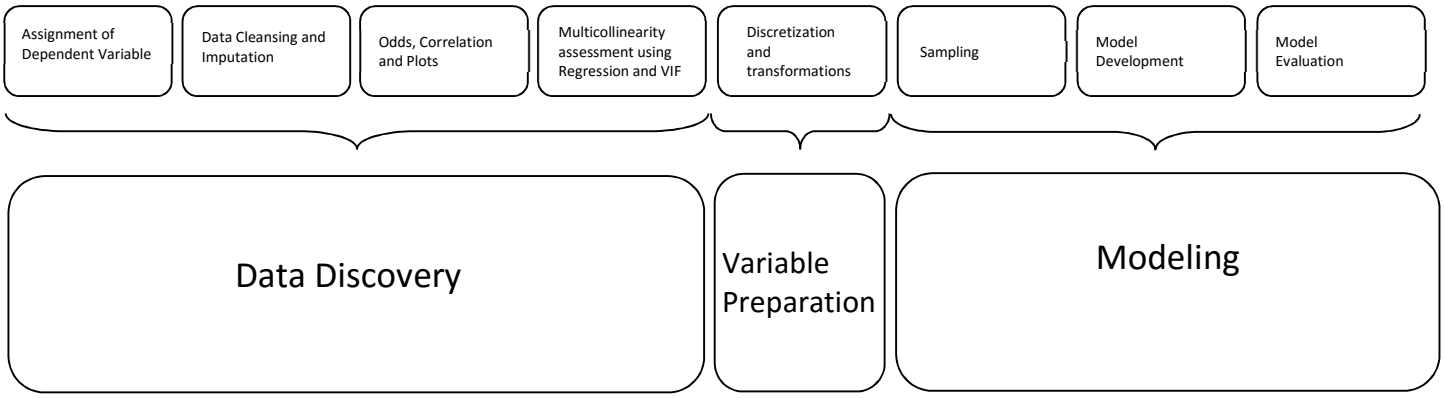
This research paper describes the process and results of developing a binary classification model, using Logistic Regression, to generate Credit Risk Scores. These scores are then used to maximize a profitability function.

The data for this project came from a Sub-Prime lender. Three datasets were provided:

- CPR. 1,444,562 observations and 339 variables. Each observation represents a unique customer. This file contains all of the potential predictors of credit performance. The variables have differing levels of completeness.
- PERF. 17,244,104 observations and 18 variables. This file contains the post hoc performance data for each customer, including the response variable for modeling – DELQID.
- TRAN. 8,536,608 observations and 5 variables. This file contains information on the transaction patterns of each customer.

Each file contains a consistent "matchkey" variable which was used to merge the datasets.

The process for the project included:



Each of these processes will be discussed in turn.

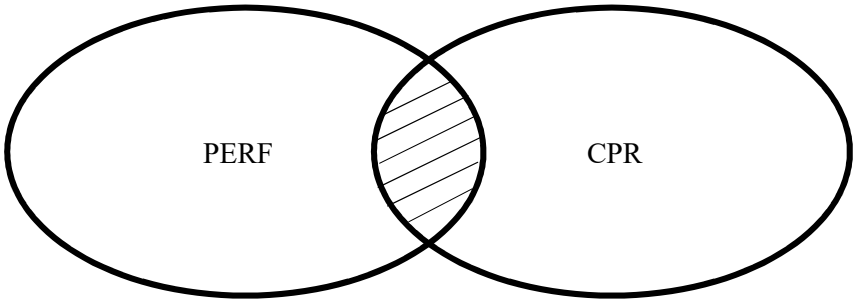
Data Discovery

The MATCHKEY variable is used to merge CPR data with PERF data via an inner join shown in Figure 1. Potential joins considered are listed and discussed below.

- Inner join – Create a new file merging all observations where a single MATCHKEY is found in both the CPR **and** the PERF files.
- Outer join – Create a new file merging all observations where a single MATCHKEY only appears in PERF **or** CPR, but not found in both.
- Right join – Create a new file keeping all CPR and merging in all observations from PERF that matched CPR MATCHKEYs.
- Left join – Keep all PERF and merge in all observations from CPR that matched PERF MATCHKEYs.

Due to the post hoc nature of PERF, only three variables discussed below are kept during the merge. All other PERF variables have been dropped.

Figure 1: Inner Join



- MATCHKEY: ID variable connecting subject across the datasets.
- CRELIM: Credit limit disclosing the amount provided if deciding to issue credit.
- DELQID: Delinquency ID ranging from 0 to 9.
    - 0 – Subject too new to rate
    - 1 – payment is made on time
    - 2 – payment one cycle late
    - Note: the greater the DELQID the later payments are made.

Due to a single MATCHKEY having multiple DELQIDs, shown in Table 1, only the highest DELQID for each MATCHKEY is retained. All lesser DELQIDs for a given MATCHKEY were not merged. The highest DELQID is used to reduce risk of providing someone credit that will default.

TABLE 1: DELQID ISSUE

OBS	DELQID	CRELIM	MATCHKEY
1	0	800	1333324
2	0	800	1333324
3	0	800	1333324
4	1	800	1333324
5	1	800	1333324
6	1	800	1333324
7	2	800	1333324
8	3	800	1333324
9	4	800	1333324
10	5	800	1333324

After the merging process, subjects missing AGE or DELQID are removed. The response variable, GOODBAD, is created: DELQIDs less than three are considered 'good' while DELQIDs greater than two are considered 'bad'. 'Bad' has been coded as '1' since predicting default is the primary objective. The FINALMERGED dataset has 342 variables and 1,255,426 observations.

Table 2 displays the frequency distribution of the response variable, GOODBAD. Overall, 17.57% of our sample is more than two cycles late on payments.

TABLE 2: GOODBAD DISTRIBUTION

GOODBAD	FREQUENCY	PERCENT
0	1034829	82.43
1	220600	17.57

A DELQID score of two (thirty days or one cycle late) is the cutoff for classifying a customer as "good" due to the historically high profits from subjects with this score. Once a customer is two cycles delinquent, there's a historical pattern of falling further behind on payments. Using the highest DELQID per subject to create a GOODBAD variable may inflate our type II error rate (denying a subject credit when they would not have defaulted). For example, a subject with a score of three or four may actually catch up on payments and not default but are still considered a high default risk customer (GOODBAD = 1). Due to the nature of the industry, this is preferable to providing someone credit that will default (Type I error).

Variable Reduction

Missing values are still a concern along with many coded values. The following is provided by the client (slightly modified by the author) and presented here for informational purposes toward coded values:

> A variable that is 2 digits long (such as AGE), each value between 0 and 91 represents its actual value, 92 represents its actual value plus all actual values above it, and 93 through 99 represent coded values for internal purposes. These defaults should be thought of as codes rather than numeric values.
>
> NOTE: 92 is the highest possible numerically significant value for a 2-digit variable.
>
> Likewise, for a variable that is 7 digits long, each value between 0 and 9,999,991 represents its actual value, 9,999,992 (the highest possible numerically significant value) represents its actual value plus all values above it, and 9,999,993 through 9,999,999 represent coded values. The same applies for 3-digit and 4-digit variables.
>
> The 5 digit variables have 4 decimal places. In that case, each value between 0.0000 and 9.9991 represents its actual value, 9.9992 represents its actual value plus all actual values above it, and 9.9993 through 9.9999 represent coded values.
>
> The 1 digit variables, which are for rating of an account, have different meaning:
>
> 9 - no trade on file
> 8 - all items that could be considered fell into an exclusion, such as disputed trades
> 7 - no industry-specific trade, but there is trade of any other type
> 6 - charge-off, repossession, account in credit counseling, included in bankruptcy
> 5 - 120 days late
> 4 - 90 days late
> 3 - 60 days late
> 2 - 30 days late
> 1 - current
> 0 - too new to rate

To account for the missing and coded values, multiple imputations methods were considered including but not limited to:

- Mean Imputation
- Median Imputation
- Regression Imputation
- Stratified Imputation

The increased number of variables precludes more advanced methods like regression imputation to be efficient for this process. Median imputation was implemented due to the nature of many variables be right skewed indicating a mean value greater than the median. This will minimize impact on variable distributions. Figures 2 and 3 below show two examples of variables (RBAL and TRADES) prior to and after imputation. RBAL is the total balance on open revolving trade accounts. TRADES is the number account on file. Notice the severe effect coded values have on RBAL distorting the pre-imputation histogram as somewhat meaningless. After imputation, the distributional variances are compressed for a more interpretable histogram. For efficiency purposes, these examples are generated using only a sample

of the full dataset to represent effect for the reader. The actual results may differ slightly from the implemented macro referenced below. Tables 3 and 4 highlight the effects on variable descriptive statistics. Notice significant adjustments made to RBAL while TRADES is affected to a much lesser extent.

Figure 2: Histogram of RBAL & TRADES (Pre-Imputation)

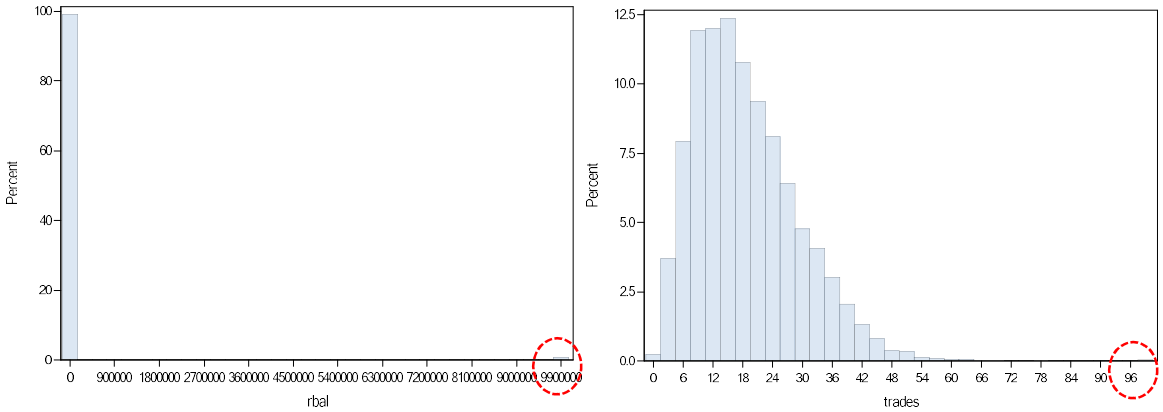


Figure 3: Histogram of RBAL & TRADES (Post-Imputation)

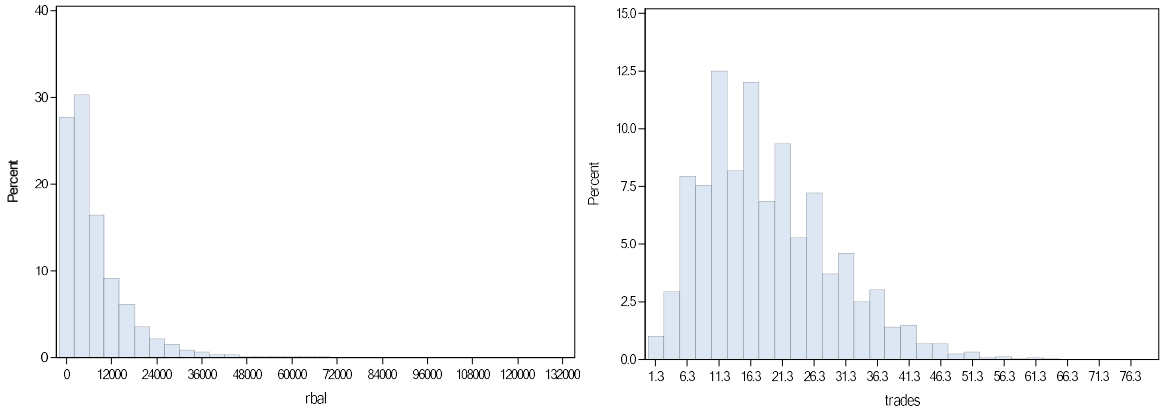


TABLE 3: DESCRIPTIVE STATISTICS (PRE-IMPUTATION)

VARIABLE	Mean	S.D.	Min	Median	Max
RBAL	$101,628	$964,320	$0	$4,778	$9,999,999
TRADES	18.76	10.46	1	17	99

TABLE 4: DESCRIPTIVE STATISTICS POST-IMPUTATION)

VARIABLE	Mean	S.D.	Min	Median	Max
RBAL	$7,672	$9,453	$0	$4,695	$133,891
TRADES	18.72	10.3	1	17	79

For efficiency, a macro is implemented that analyzes all variables in the FINAL dataset (which is a working copy duplicate of the FINALMERGED dataset referenced above). Per variable, the macro addresses:

- Missing values
- Coded values
- Variable dispersion

Missing values and coded values are addressed using an input parameter specifying a cutoff percent. For example, if the parameter is set to 0.3 then any variable with 30% of observations recoded, due to missing values or the coded values discussed above, will be removed from the output dataset FINALOUT. Figure 4 and Table 5 below show how variables remaining is affected by this parameter specification. Specifications of .2 and .3 yield the same number of variables, 130. A specification of .5 and .6 yield 239 and 284 variables, respectively, which is more than preferred. A specification of .4 is utilized in the process leaving 165 variables, approximately a 50% reduction in variables. Specifications .2 or .3 were strongly considered. However due to the early stage in the process, more variables are retained. Further reduction will occur through cluster analysis and variance inflation factor analysis below.

Figure 4: Variable Reduction

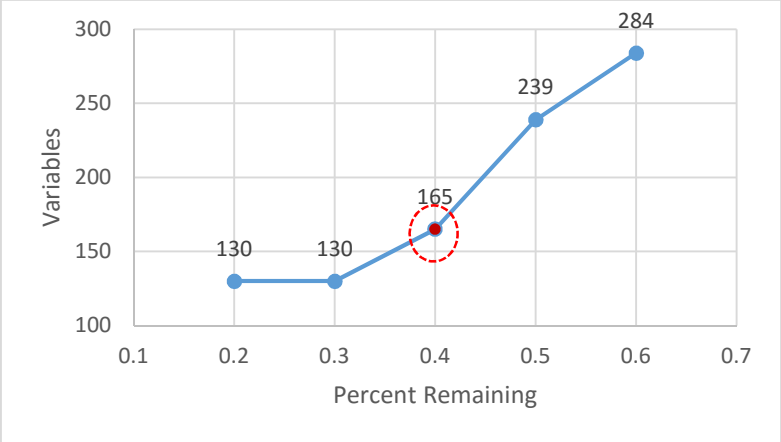


TABLE 5: VARIABLE SELECTION

PCTREM	Variables
0.6	284
0.5	239
0.4	165
0.3	130
0.2	130

Variable dispersion is addressed by specifying a parameter for maximum standard deviation. Any observation beyond this specification is recoded to the median value as missing and coded values are above. For example, if the parameter is set at three, all observations greater than three standard deviations from the mean are recoded to the median value. This parameter helps to prevent over fitting the model to the sample data. The final model is only relevant for data that fall within the recoded variable ranges in the sample data. For example, the model should not be used for predicting default probability rates for values beyond four mean standard deviations. Holding PCTREM constant, this parameter has no effect on variable reduction.

Table 6 below compares descriptive statistics for three and four max standard deviations. Figures 5 – 7 provide better visual representations for specifications of RBAL and TRADES at specifications of three, four and five max standard deviations affecting distributions. As the max standard deviation allowed decreases, more 'outliers' are recoded to the median value compressing the right tail toward the median. Considering many variables have right-skewed distributions, a max standard deviation of four is chosen to allow some distributional spread in the right tails.

TABLE 6: MAXIMUM MEAN STANDARD DEVIATION OPTION EFFECTS

	Mean Standard Deviation = 3				Mean Standard Deviation = 4			
VARIABLE	**Mean**	**StDev**	**Min**	**Max**	**Mean**	**StDev**	**Min**	**Max**
RBAL	$6,595	$6,842	$0	$35,457	$6,931	$7,557	$0	$44,786
TRADES	18.5	9.8	1.0	49.0	18.7	10.1	1.0	59.0

Figure 5: Max Standard Deviation of 5 for RBAL and TRADES

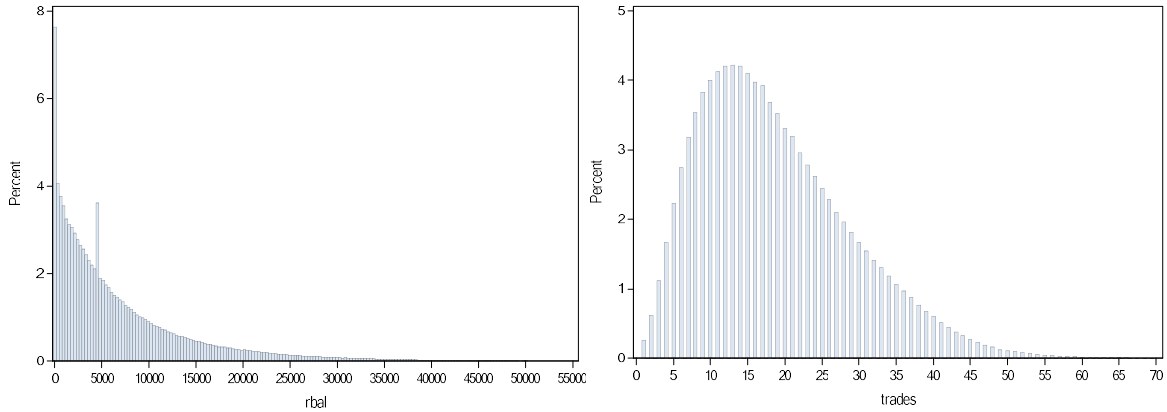


Figure 6: Max Standard Deviation of 4 for RBAL and TRADES

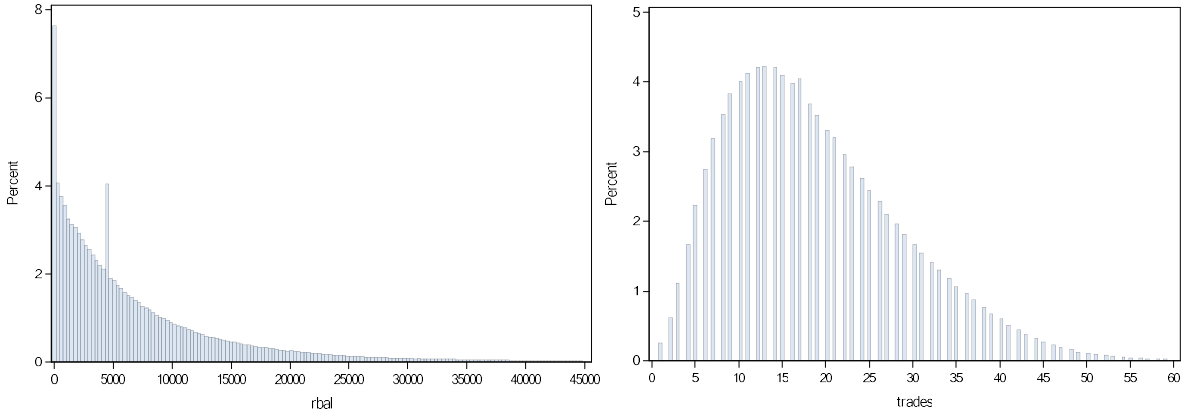
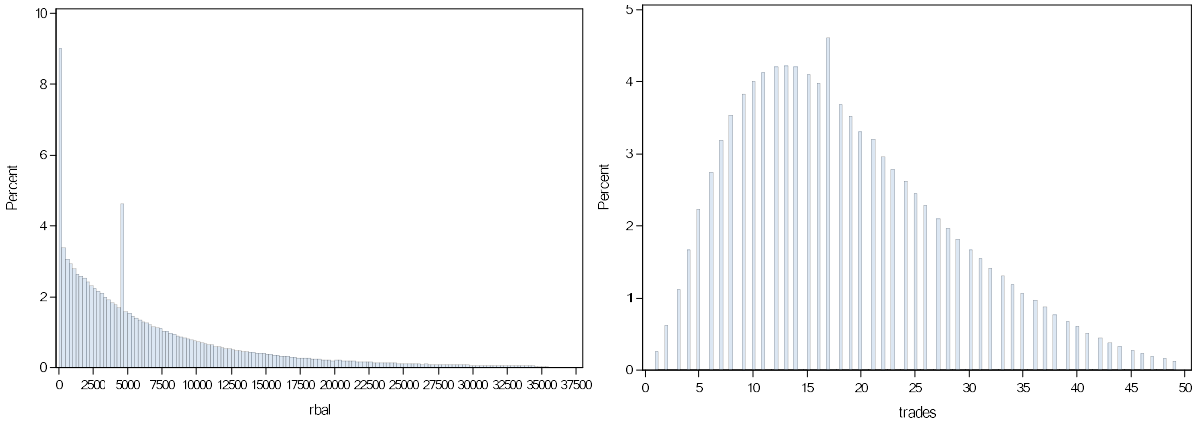


Figure 7: Max Standard Deviation of 3 for RBAL and TRADES



Variable Clustering

Variable cluster analysis reduces multicollinearity by identifying the most representative variables to retain per cluster. The following variables are removed for the cluster analysis since they will not be used as model predictors.

- BEACON (100% missing)
- AGE (prohibited by law)
- DELQID
- CREMLIM
- MATCHKEY
- GOODBAD

Figure 8 below shows a scatter plot for proportion of variance explained at the different number of clusters. As clusters increase the proportion of variation increases. The optimal recommendation is 32 clusters where the maximum second eigenvalue becomes less than 1.00. However, 43 clusters are retained due to explaining slightly more than .80 proportion of the variance. These points are highlighted in red. Table 7 shows the proportion of variance explained at different cluster levels from 30 to 45.

Figure 8: Variation Explain by Clusters

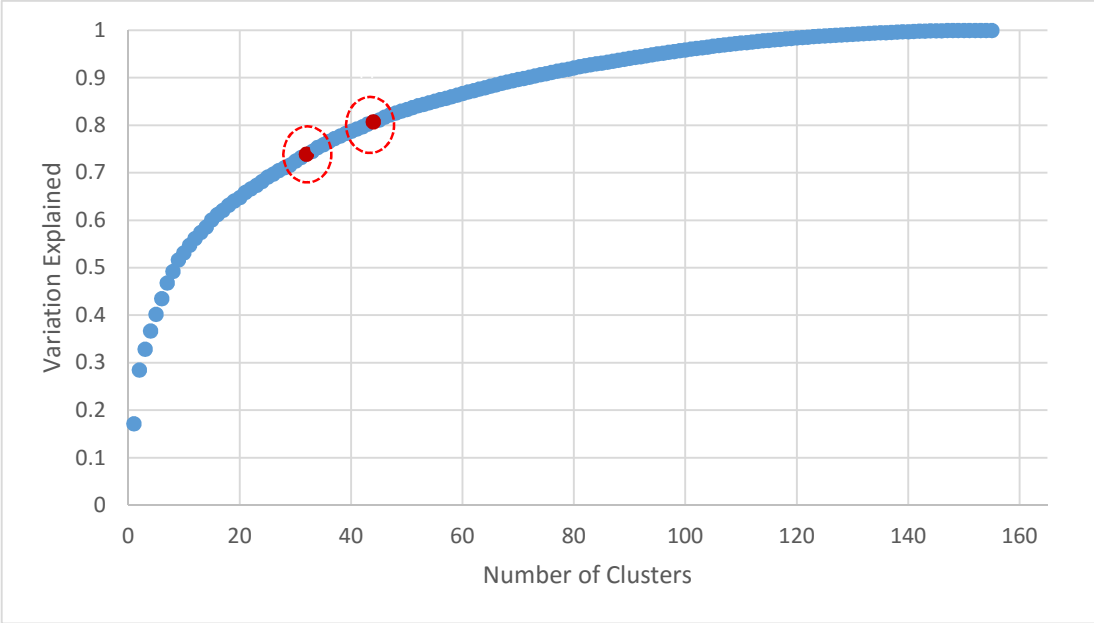


TABLE 7: PROPORTION OF VARIANCE EXPLAINED

NUMBER OF CLUSTERS	Proportion of Variance Explained
30	.725
31	.733
32	.739
33	.745
34	.753
35	.760
36	.766
37	.772
38	.778
39	.783
40	.788
41	.793
42	.797
43	.802
44	.807
45	.811

Table 8 compares correlations between variables in cluster one and cluster two. Correlation are high within clusters and low across clusters. An optimal variable from each cluster is selected to move forward. By selecting a single variable from each cluster, where there are two or more variables, some information will be lost. However, multicollinearity is reduced. Potential predictor variables are reduced approximately 70% to 44.

Table 9 shows variables in clusters one, two and three along with variable within cluster R^2, the R^2 with the next closest cluster and the $1 - R^2$ ratio between these two statistics (within cluster / next closest). A higher R^2 indicates increased collinearity. The most representative variable will have the lowest $1 - R^2$ ratio within each cluster. The optimal variables for clusters one, two, and three are highlighted in blue.

$$1 - R^2Ratio = \frac{(1 - R^2Own\ Cluster)}{(1 - R^2Next\ Cluster)}$$

Table 8: Variable Correlations in Cluster 1 and Cluster 2

	DCR7924	DCN90P24	DCR39P24	BRTRADES	BROPEN	BRCRATE1
DCR7924	1	0.88	0.807	0.031	-0.042	-0.024
DCN90P24	0.88	1	0.916	0.025	-0.066	-0.047
DCR39P24	0.807	0.916	1	0.023	-0.073	-0.051
BRTRADES	0.031	0.025	0.023	1	0.768	0.92
BROPEN	-0.042	-0.066	-0.073	0.768	1	0.805
BRCRATE1	-0.024	-0.047	-0.051	0.92	0.805	1

TABLE 9: CLUSTER VARIABLE SELECTION

CLUSTER	Variable	OwnCluster	NextClosest	(1 - RSquareRatio)
CLUSTER 1	DCCRATE7	.892	.510	.221
	DCRATE79	.892	.511	.220
	DCR7924	.875	.531	.267
	DCN90P24	.844	.620	.412
	DCR39P24	.755	.633	.669
	DCCR49	.881	.565	.272
CLUSTER 2	BRTRADES	.892	.463	.201
	BROPEN	.800	.584	.482
	BRCRATE1	.939	.514	.126
	BRRATE1	.903	.507	.197
	BRR124	.835	.585	.397
	BROPENEX	.892	.463	.201
CLUSTER 3	BRWCRATE	.619	.347	.584
	BRCRATE7	.856	.370	.228
	BRRATE79	.855	.370	.230
	BRR7924	.568	.365	.680
	BRR49	.775	.677	.700
	BRCR39	.810	.626	.507
	BRCR49	.873	.551	.283

For the remaining 44 potential predictor variables, variance inflation factors (VIF) are assessed to further test for multicollinearity. Table 10 shows the top VIFs. A VIF of 10 signals issues with multicollinearity and is computed by running each predictor in the model as a linear combination of all other predictors. The highest VIF observed is 5.69 implying some correlation among a small number of variables but not enough to indicate further multicollinearity concerns.

The variables below previously removed have been added to back the MASTER dataset which now has 48 variables.

- DELQID
- CRELIM
- GOODBAD
- MATCHKEY

TABLE 10: VARIANCE INFLATION FACTORS

VARIABLE	VIF
TOPEN	5.69
TROPENEX	4.10
TCR1BAL	3.57
BRCRATE1	3.55
TRCR49	3.32
BADPR1	3.15
TPCTSAT	2.99
BRR39P24	2.84
RBAL	2.52

Variable Preparation

The remaining 44 potential predictor variables are evaluated for transformations under two separate processes creating ordinal versions of the variables. Discretization 1 (DISC 1) is a supervised process distributing transformed variables into equal width bins. DISC 1 requires user evaluation and consecutive bin combination decisions due to lack of separation in bin default probabilities and bin frequencies. Discretization 2 (DISC2) is an unsupervised process distributing transformed variables into equal frequency bins. DISC 2 utilizes a t-test to evaluate consecutive ranks (similar to bins and discussed further below) for significance and combining lower ranks into higher ranks where no significance is discovered. For each discretization process, the ordinal variable is used to calculate the default rate per bin allowing for odds ratio and log odds ratio transformations to be evaluated. The odds ratio is constructed per bin by dividing the probability of default by one minus the probability of default. The log of the odds ratio is calculated by taking the log of the constructed odds ratio per bin.

$$Odds\ Ratio = \frac{Probability\ of\ Default}{1 - (Probabilty\ of\ Default)}$$

$$Log\ (Odds\ Ratio) = Log\ \left(\frac{Probability\ of\ Default}{1 - (Probabilty\ of\ Default)}\right)$$

For example, a bin with a default .25 default rate yields the following:

$$Odds\ Ratio = \frac{.25}{.75} = .333$$

$$Log(Odds\ Ratio) = -.478$$

An applicant in this bin is three times more likely to not default vs defaulting.

Variables originally ordinal were still evaluated for lack of separation attempting to construct more relevant bins. Some variables are inherently binary and are recoded as such. Originally binary variables do not go through DISC 1 or DISC 2, while inherently binary variables only go through DISC 1. However, the odds and log odds are constructed for each group. Figure 9 shows the original binary distribution of BKP (indicating if an applicant has previously filed for bankruptcy), the original ordinal distribution for LOCINQS (the number of local inquiries in previous six months), and the original inherently binary distribution of LAAGE (age of last activity).

Initially, for DISC 1, variables were discretized across a range of eight to twelve bins depending on the variable. Each variable is manually analyzed for further improvement based on lack of separation in default rates or low bin counts. Where consecutive bins displayed lack of separation or low bin counts, the bins were combined.

For example, RBAL's distribution is heavily skewed to the right as seen in Figures 5 – 7 above. Figure 10 below shows the average default rate per original and final ORDRBAL (DISC 1 ordinal transformation of RBAL) bin, while Tables 11 and 12 show the original and final bin counts. Due to lack of separation, bins three through nine were combined. Three bins remain to evaluate the reversal in default rates at higher revolving balances. RBAL is likely not a good predictor due to the probabilities of

default falling between .15 and .21 where the overall average is .175. Additionally, the trend reversal for higher revolving balances presents issues of a non-linear relationship with default rates. The cutoffs are $2500 for bin 1 and $5000 for bin 2 while everything greater than $4999 is as assigned to bin 3.

Figure 9: Histogram of BKP, LOCINQS, LAAGE

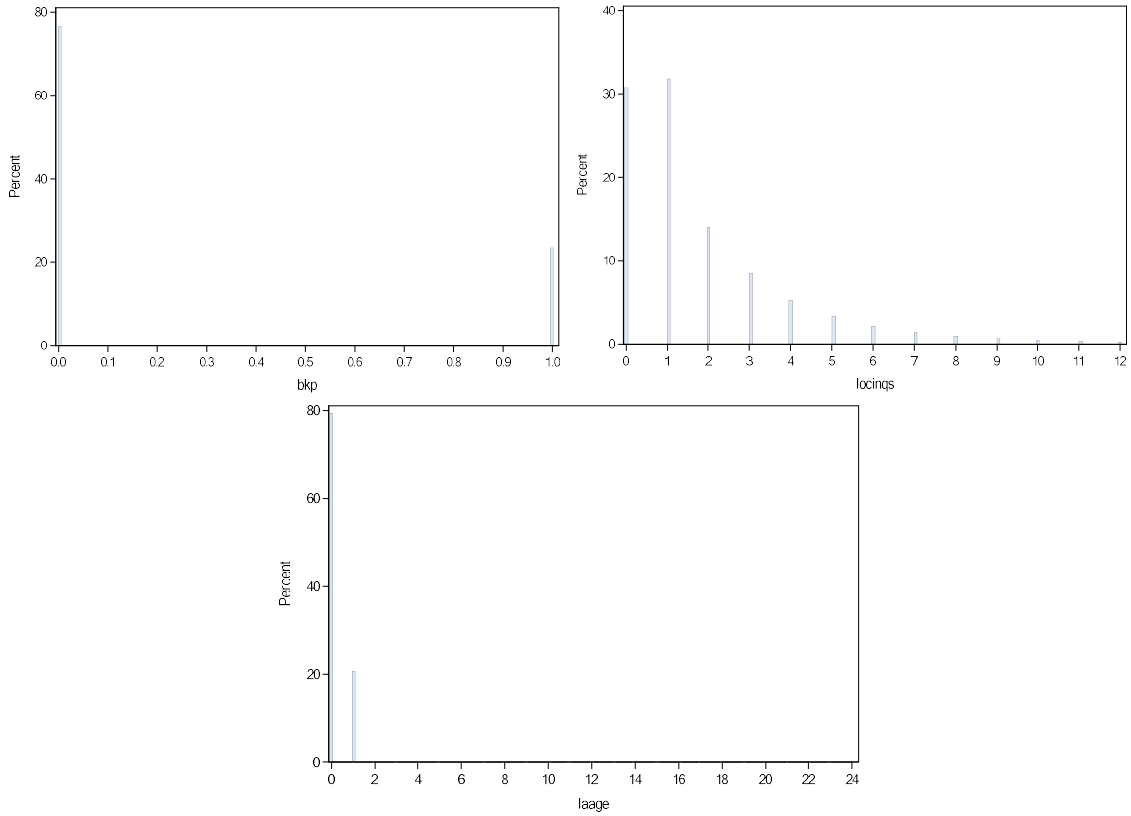


Figure 10: Initial & Final ORDRBAL Bin Default Rates (DISC 1)

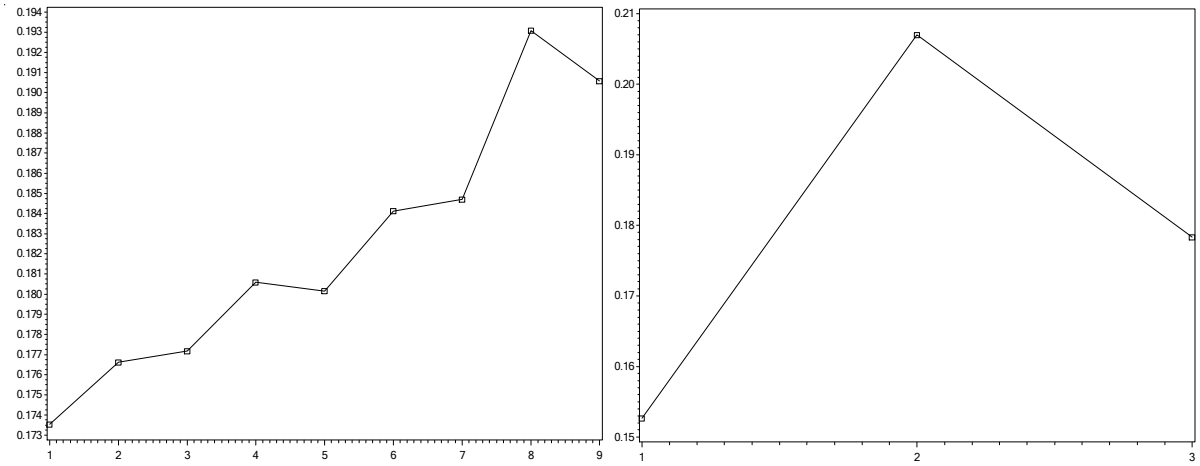


TABLE 11: ORDRBAL INITIAL BINS (DISC 1)

BINS	Count
1	685933
2	273803
3	135099
4	71452
5	39896
6	22652
7	13573
8	8162
9	4859

TABLE 12: ORDRBAL FINAL BINS (DISC 1)

BINS	Count
1	422642
2	263291
3	569496

For another example, BADPR1's distribution is originally ordinal and possibly inherently binary as seen in Figure 11 below. Figure 12 below shows the average default rate per original and final ORDBADPR1 (DISC 1 ordinal transformation of BADPR1) bin, while Tables 13 and 14 show the original and final bin counts. The original bins have a linear association up through bin twelve where the association flattens out to bin fourteen. While, this variable is potentially inherently binary, bins twelve to fourteen are only combined in DISC 1 allowing DISC 2, described in more detail below, to explore further bin reductions. Generally, ORDBADPR1 appears to be a decent predictor. The trend line is consistently linear with default rates between .09 and .40.

Figure 11: Histogram of BADPR1

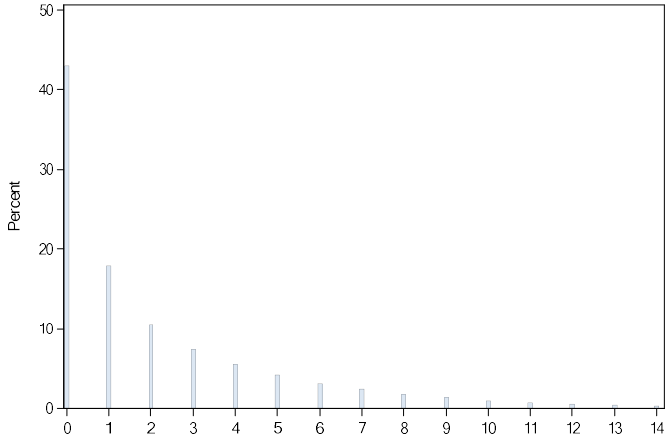


Figure 12: Initial & Final ORDBADPR1 Bin Default Rates (DISC 1)

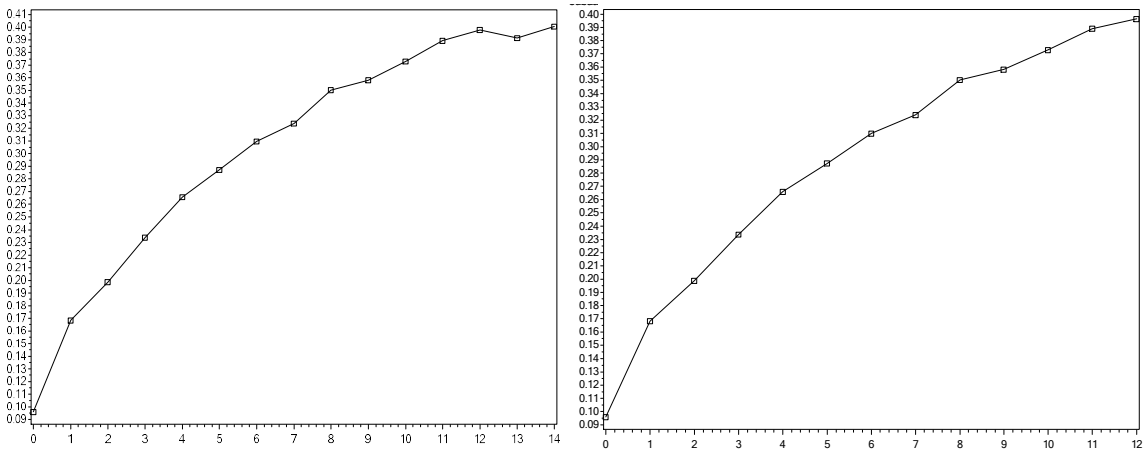


TABLE 13: ORDBADPR1 INITIAL BINS (DISC 1)

BINS	Count
0	539168
1	225179
2	132547
3	93146
4	69525
5	52440
6	39484
7	29877
8	22202
9	16246
10	12044
11	8894
12	6410
13	4749
14	3518

TABLE 14: ORDBADPR1
FINAL BINS (DISC 1)

BINS	Count
0	539168
1	225179
2	132547
3	93146
4	69525
5	52440
6	39484
7	29877
8	22202
9	16246
10	12044
11	8894
12	14677

DISC 2 followed a similar process, initially discretizing across 10 ranks. However, evaluating lack of separation in consecutive ranks is automated utilizing a SAS rank procedure allowing ranks to be combined if not significantly different with ties going to the higher rank. Final ranks are relabeled as bins after each variable evaluation. Note that four variables were fully collapsed into a single rank during DISC 2 and were excluded from completing the process. These variables are:

- BRCRATE4 (# of bank revolving accounts currently 90+ days past due)
- BRR324 (# of bank revolving account 60 days past 24 months late)
- DCRATE79 (# of department store account ever having a bad debt)
- FFCR49 (# of consecutive financial trades currently 90+ days past due)

For example, Figure 13 below shows the average default rate per original and final ORDEQRBAL (DISC 2 ordinal transformation of RBAL) bins. Due to lack of separation after the initial rank procedure, rank one is manually combined with rank five while rank eight is manually combined with rank nine. Where DISC 1 reduced to three bins with less than desired separation, DISC 2 maintains four bins with a wider default probability range of approximately .09 to .22. ORDEQRBAL still maintains the trend reversal at higher revolving balances, however a very strong linear relationship exists from rank zero up to rank five. Tables 15, 16 and 17 respectively show the original rank counts, combined rank counts, and final bin counts. The cutoffs are $321 for bin one, $4504 for bin five and $5925 for bin six while everything greater than $5925 is as assigned to bin ten.

Figure 13: Initial & Final ORDEQRBAL Rank Default Rates (DISC 2)

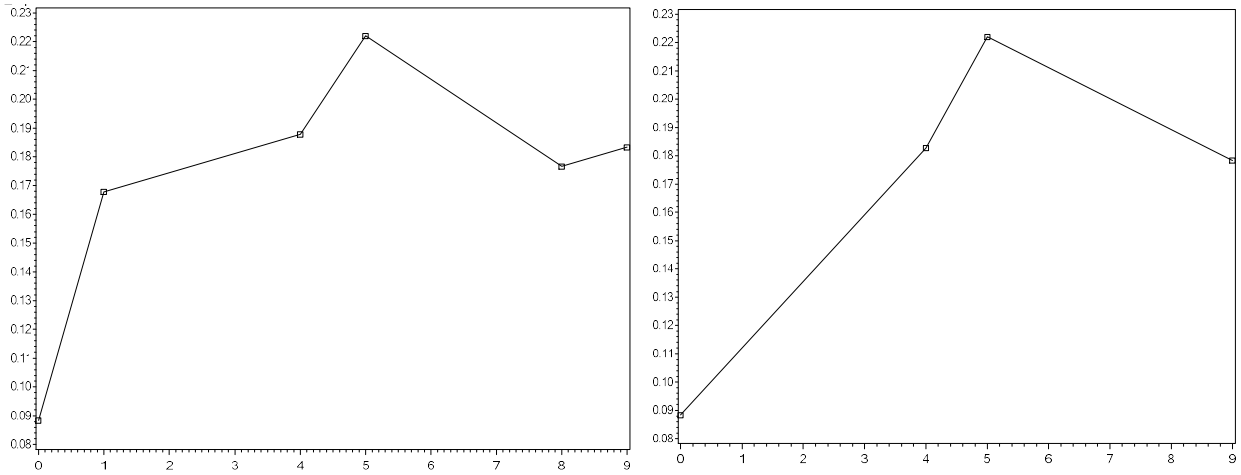


TABLE 15: ORDEQRBAL INITIAL RANKS

RANK	Count
0	125498
1	125541
2	125557
3	125511
4	118359
5	132732
6	125597
7	125524
8	125562
9	125548

TABLE 16: ORDEQRBAL COMBINED RANKS

RANK	Count
0	125498
1	125541
4	369427
5	132732
8	376683
9	125548

TABLE 17: ORDEQRBAL FINAL BINS

BINS	Count
1	125498
2	494968
3	132732
4	502231

For another example, Figure 13 below shows the average default rate per original ORDEQBADPR1 (DISC 2 ordinal transformation of BADPR1) rank. These ranks have a linear association and desirable separation. Tables 18 and 19 show the original rank counts and final bin counts. While the variable appeared inherently binary in Figure 11 above, DISC2 maintained five bins. ORDEQBADPR1 is likely to be a good predictor with a strong linear trend line and nice separation between ranks. However, some information is lost due to less bins and a narrower range of bin default rates compared with ORDBADPR1 in DISC 1.

Figure 13: BADPR1 Rank Default Rates DISC 2)

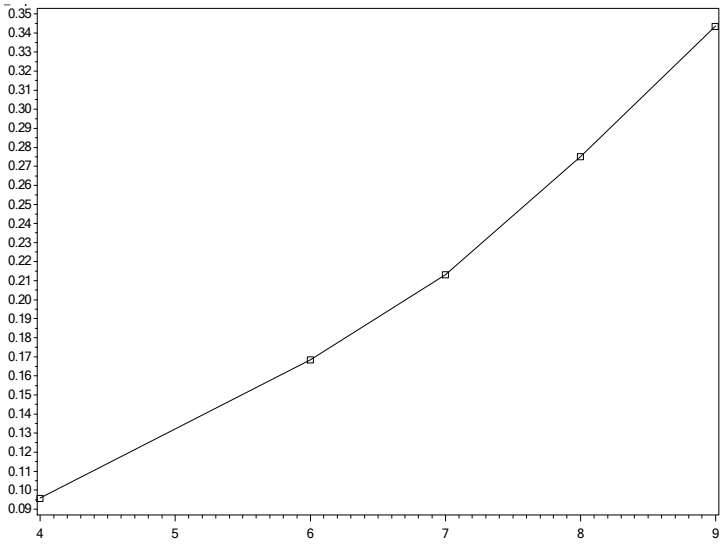


TABLE 18: BADPR1 INITIAL RANKS

RANK	Count
4	539168
6	225179
7	225693
8	121965
9	143424

TABLE 19: BADPR1 FINAL BINS

BINS	Count
1	539168
2	225179
3	225693
4	121965
5	143424

The MASTER dataset has approximately 250 variables. Next steps involve logistically modeling GOODBAD with remaining potential predictors and model validation.

Modeling

The data set is split into a training dataset (70% of MASTER dataset) and a validation dataset pool (remaining 30% of MASTER dataset) of observations to construct five different validation datasets. The remaining potential predictors are evaluated to logistically model GOODBAD via a backward selection process on the training dataset. Backward selection is an iterative process beginning with all potential predictors in the model and removing the least significant predictor per iteration. This process repeats until all remaining predictors are significant at the .05 significance level. (i.e. all remaining predictors' p-values are less than .05.) A forward selection process is not used due to potential predictors not being evaluated in the model. A stepwise selection process is not used due to the number of potential predictors, computational resources, and time required. Table 20 displays the top twelve predictors from the backward selection logistic regression model. While there is no set cutoff point for predictors, approximately eight to ten were expected to be kept for operational purposes. The chi-squared values are used as the gauge for significance (with the high value being the most significant). Twelve are kept for the final model due to the noticeable percent drop in the chi square value from the twelfth to the thirteenth variable where the incremental percent drops for the eight to twelve were marginal.

Coefficients are interpretable similar to a multiple regression model with the primary difference being the logit function (i.e. the log of the odds ratio) involved with logistic regression. For example, the coefficient of CRATE2 is 0.4273. For a single unit increase in CRATE2, the log of the odds is 0.4273. The odds ratio is 1.53 (e^0.4273). Therefore, for each additional unit increase in CRATE2, the odds of default increase 1.53 times.

TABLE 20: ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATES

PARAMETER	DF	Estimate	Std. Error	Chi-Square	Pr > ChiSq
ORDRADB6	1	0.4292	0.00349	15147.1948	<.0001
LODSEQTOPENB75	1	0.5552	0.00692	6440.2938	<.0001
CRATE2	1	0.4273	0.00632	4570.2359	<.0001
ORDBRR4524	1	0.8795	0.0104	7218.3086	<.0001
LODSEQBRCRATE1	1	0.7009	0.00851	6774.7374	<.0001
BRCRATE3	1	1.0478	0.0143	5366.704	<.0001
LODSEQBRMINB	1	0.3288	0.00752	1913.1348	<.0001
ODSEQBRAGE	1	3.8637	0.0618	3910.3534	<.0001
LOCINQS	1	0.0862	0.00145	3521.8592	<.0001
BRCRATE4	1	0.4266	0.00998	1829.0249	<.0001
TPCTSAT	1	-1.0071	0.0156	4182.2141	<.0001
ODDSBRRATE2	1	2.0349	0.0484	1764.9455	<.0001

Tables 21 displays results from the full model initially evaluated from the approx. 250 predictors. The C-statistic is used to assess model adequacy. The C-statistic represents the percentage of concordant pairs (calculated by adding Percent Concordant to ½ of Percent Tied shown below). Concordance is evaluated during the logistic regression procedure by splitting the data into two portions, one where GOODBAD = 1 and another where GOODBAD = 0. Predicted probabilities are predicated for each observation in both portions of the dataset. Each observation for one portion is then compared to all observations in the other portion. A concordant pair implies the probability of a compared observation's default probability for GOODBAD = 0 is less the default probability for an observation's default probability for GOODBAD = 1. In other words, lower default probabilities are expected for customers who do not default (GOODBAD = 0) vs. customers who default (GOODBAD = 1). This occurs 84.1% of the time with the full model as displayed in Table 21.

Table 22 displays results from the final model. The C-statistic is 0.826 which is a marginal drop considering the decrease in model predictors included. Twelve predictors is much easier to operationalize than the 250 in the full model.

TABLE 21: PERCENT CONCRODANT (FULL MODEL)

PERCENT CONCORDANT	84.1
PERCENT DISCORDANT	15.9
PERCENT TIED	0.0
PAIRS	111745868316
C	0.841

TABLE 22: PERCENT CONCRODANT (FINAL MODEL)

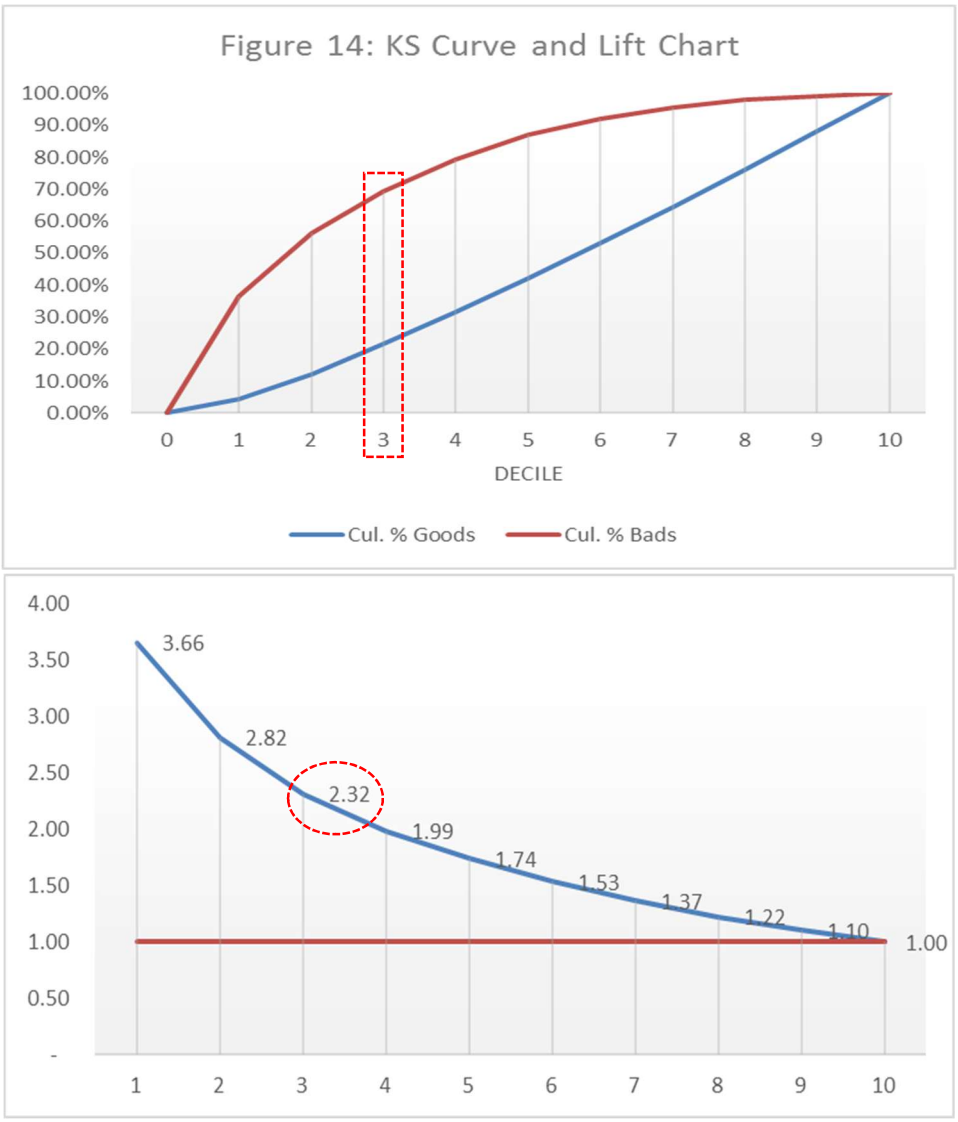
PERCENT CONCORDANT	82.4
PERCENT DISCORDANT	17.3
PERCENT TIED	0.3
PAIRS	111745868316
C	0.826

Model adequacy is also measured by a KS curve and a lift chart. Table 23 displays the calculations for the KS curve seen along with the lift chart in Figure 14. The KS compares, across deciles, the cumulative percent good versus the cumulative percent bad. The optimal KS stat,47.9 is the found in the third decile maximizing separation on the KS curve for the two lines plotted (% Tot. Good vs % Tot. Bad). Optimal KS stats occurring in the initial few deciles indicate the model can appropriately distinguish GOODBADs = 1 from GOODBADs = 0. In the top three deciles, the model identifies 69.5% of GOODBADs = 1 and 21.6% GOODBADS = 0. Notice for a marginal drop in KS at the fourth decile, (KS of 47.8), the models identifies 79.4% of GOODBADs = 1 and 31.6% of GOODBADs = 0.

The lift curve represents how much more likely our model will predict a default in any given decile vs not using model. For example, in the third decile, the model is 2.32 time more likely to predict a default when compared with not using a model.

TABLE 23: KS STATISTIC CALCULATIONS

DECILE	Min Score	Max Score	# of ACCTS	# Goods	% Goods	Cul. % Goods	# Bads	% Bads	Cul. % Bads	KS
1	445	995	31,385	11,150	4.31%	4.31%	20,235	36.57%	36.57%	32.3
2	271	445	31,386	20,472	7.92%	12.23%	10,914	19.73%	56.30%	44.1
3	192	271	31,386	24,091	9.32%	21.55%	7,295	13.19%	69.49%	47.9
4	144	192	31,386	25,897	10.02%	31.57%	5,489	9.92%	79.41%	47.8
5	106	144	31,385	27,225	10.53%	42.10%	4,161	7.52%	86.93%	44.8
6	77	106	31,386	28,526	11.03%	53.13%	2,859	5.17%	92.10%	39.0
7	54	77	31,386	29,392	11.37%	64.50%	1,994	3.60%	95.70%	31.2
8	36	54	31,386	30,143	11.66%	76.16%	1,243	2.25%	97.95%	21.8
9	22	36	31,386	30,664	11.86%	88.02%	722	1.31%	99.25%	11.2
10	6	22	31,385	30,972	11.98%	100.00%	413	0.75%	100.00%	0.0



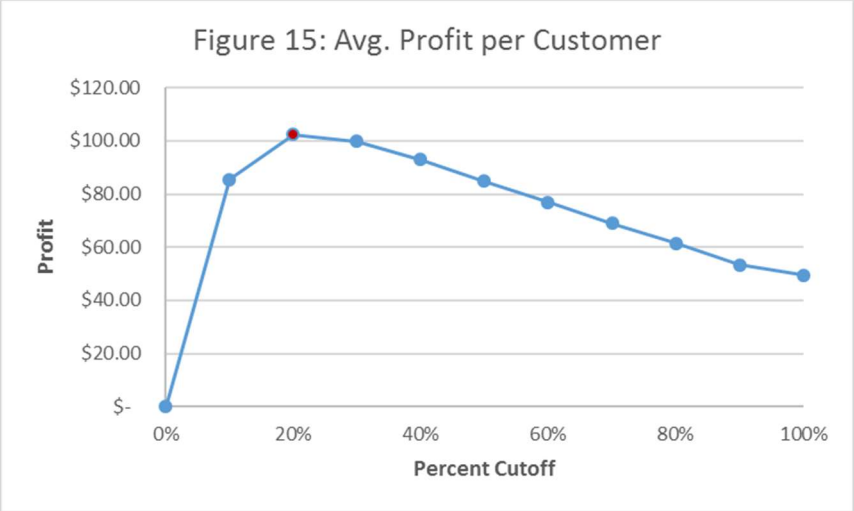
The five validation datasets (each extracted from the validation dataset pool and containing 25% of the MASTER dataset) are scored against the final training model using twelve predictors. Profitability is maximized from the training model to determine a cutoff probability for default, and the results are compared to the five training datasets. Figure 15 shows a high level evaluation of estimated profit across the different default probabilities. The optimal cutoff appears around 20%. Figure 16 drills down to calibrate this optimal cutoff to 21% for an estimated profit of $102.65 per customer. This cutoff is the point that determines if a customer is issued credit or not based on their default probability score. For example, customers determined to have less than or equal a 21% probability of default will be issued credit, while customers determined to have greater than 21% probability of default will not be issued credit. An estimated profit function, given a customer was issued credit, was provided by the client to calculate Figures 15 and 16 below and is approximately:

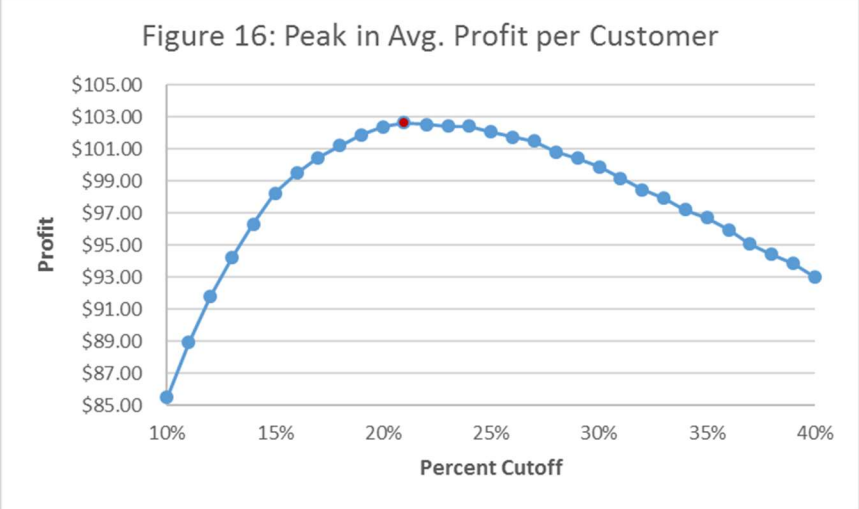
$$Pr \qquad [250 * (GOODBAD_{0|0})] - [800 * (GOODBAD_{1|0})]$$

Four potential outcomes are considered to determine profitability.

1) A customer not issued credit but would not have defaulted (type II error).
    a. Opportunity cost of $250 profit on average but no actual loss realized.
2) A customer not issued credit that would have defaulted (% Valid 1).
    a. Correct decision resulting in not losing $800 on average.
3) A customer issued credit and does not default (% Valid 2).
    a. Resulting in $250 profit on average.
4) A customer issued credit but defaults (type I error).
    a. Resulting in $800 loss on average.

If the cutoff default probability is set too low, very few customers are issued credit. While this minimizes the risk of losing money, it also diminishes profitability. If the cutoff default probability is set too high, too many customers are issued credit. While this allows for potentially high profits, it includes very high risk of customers defaulting, again diminishing profitability due to the increased average loss vs the average profit (average loss is 3.2 times higher than average profit, i.e. 800 / 250 = 3.2). Figures 15 and 16 show optimal cutoff default probability recommendations considered the above.





Tables 24 display results comparing five validation datasets along with the training dataset. Generally, the results appear consistent across the table. Type I error rates range from 5.90% to 5.95% while type II error rates range from 15.47% to 15.61%. Correctly predicted outcomes (% Valid 1 and % Valid 2) also contain narrow ranges. Approximately 27% are predicted to default and are not issued credit compared with 17.5% average default rate stated above. Given the conservative decision to use the highest customer DELIQ to determine GOODBAD, this increased type II error is expected. Remember the objective is to reduce type I errors due to the average loss per default being 3.2 times higher than the average profit per non-default.

TABLE 24: MODEL VALIDATION COMPARISONS

DATASET	% Valid 1	% Valid 2	% Type I Error	% Type II Error	Est. Profit per 1000
TRAINING	11.60%	67.00%	5.93%	15.47%	$102,650
VALID 1	11.68%	66.85%	5.95%	15.53%	$110,432
VALID 2	11.69%	66.83%	5.91%	15.61%	$110,895
VALID 3	11.68%	66.84%	5.90%	15.58%	$110,757
VALID 4	11.67%	66.82%	5.95%	15.57%	$110,473
VALID 5	11.72%	66.79%	5.92%	15.58%	$110,792

The Cost of Simplicity

Table 25 displays results from a simplified version of the final model. In the simplified final model, the original versions of the final model twelve predictors are used vs any transformations (e.g. BRCRATE1 is used vs LODSEQBRCRATE1). The C-statistic is exactly the same, 0.826 where the percent concordant actually increases (82.6% vs 82.4%). The average profit per customer increases to $103.22 (vs. 102.65).

TABLE 25: PERCENT CONCORDANT (SIMPLIFIED FINAL MODEL)

PERCENT CONCORDANT	82.6
PERCENT DISCORDANT	17.4
PERCENT TIED	0
PAIRS	111745868316
C	0.826

Transaction Analysis

Transactions are tracked for customer who received credit. Many SIC codes were analyzed with 9223 and 501 discussed below. Table 26 and 27 display transactional descriptive statistics for bail bondsman payments (9223) and motor vehicle parts / supplies (501). The average profit from customer bail bonds payments is more than five times higher than the average profit. However, the average loss from customer motor vehicle supply / parts purchases is approximately $14.72 indicating these customers may be unprofitable.

TABLE 26: DESCRIPTIVE STATISTICS FOR SIC CODE 9223

MEAN	Std Dev	Minimum	Maximum
$544.85	$709.48	-$2,265.00	$5,000.00

TABLE 27: DESCRIPTIVE STATISTICS FOR SIC CODE 501

MEAN	Std Dev	Minimum	Maximum
-$14.72	$12.56	-$187.44	$30.59

Conclusion

The initial full model results in a C-statistic of .841. The final model reduces the number predictors to twelve resulting in a C-statistic of .826. The C-statistic reduction is marginal compared to the 200+ variables removed from the model. An average profit per 1000 customers scored for the final model is $102,650. A simplified version of the final model, using original versions of the final twelve predictors, results in the same C-statistic (.826) for an increased average profit per 1000 customer scored of $103,220. Operationalizing the simplified model is ideal since it provides increased profit, and is much easier to communicate to customers / clients.

The primary consideration in developing the model is evaluating different default probability cutoff points to maximize expected profit. Minimizing defaults is accomplished with lower cutoffs, but this strategy yields suboptimal expected profit. Maximizing potential profit is accomplished with higher cutoffs, but this strategy also maximizes potential loss from defaults. Having an accurate profit function is critical in evaluating all possible cutoffs. Profit is approximated across customers receiving credit (+$250 per correctly predicted non-defaults and -$800 per incorrectly predicted non-defaults).

Evaluating the opportunity cost of incorrectly predicting customer defaults should also be considered. Additionally, customer segmentation would be beneficial in deploying segment specific profit functions to target high profit and low profit segments or high risk or low risk segments.