Binary Classification Modeling Deliverable Three
Using Logistic Regression to Build Credit Scores

Nathaniel Jones
Supervised by Michael Frankel

Spring 2021 – May 7th, 2021

Executive Summary:

The purpose of this report is to create a model that predict a customer's credit score to maximize the profitability of granting credit. The original data contained over 1.2 million observations with over 300 parameters for each. SAS and Python was used during this project. A binary predictor was used to indicate whether a customer was considered a credit risk. This predictor was created from the customers' delinquency status on payments. Initially, the data needed to be recoded and cleaned of missing and coded values. Parameters with a low enough proportion of missing or coded data had the median calculated and put in place of any missing or coded value. Parameters with a proportion too high were removed to avoid imputing a significant portion of the parameter. The remaining parameters underwent the process of variable clustering to further reduce the number of parameters that will be used in the model.

The next phase transformed the parameters to create additional forms. The forms were included to see if they yielded significant results in the final model. Some of these transformations included creating discrete indicators for variables which were originally continuous and calculating the odds of the variables. Two methods were used to in creating these transformations: an automatic process using a procedure in SAS and a user process.

The last phase of the project, a logistic regression model was run on the parameters and the different transformed versions in the dataset. The main metric used to assess the performance of the models created were the percent of concordant observations, the Kolmogorov-Smirnov statistic, the Chi-Squared statistic, and most importantly, profit. From this logistic regression, two simpler models were created to reduce model complexity. To assess the consistency of the model performance it was split into a training and a validation subset.

An analysis of the profitability was the conducted by looking at the number of correct customers predicted to not default, which earns the company a mean revenue of $250 per customer. Of those that do default but were predicted to be low rick customers, it was determined that every 1000 customers who are extended credit yield an average profit of $115,303.14. In addition to the model, a reduced model that included twenty parameters yield an average profit of $112,777 per 1000 customers. This less than $3000 off the initial models profit per 1000 customers and this model is much less complex meaning it is easier to interpret and explain to potential customers.

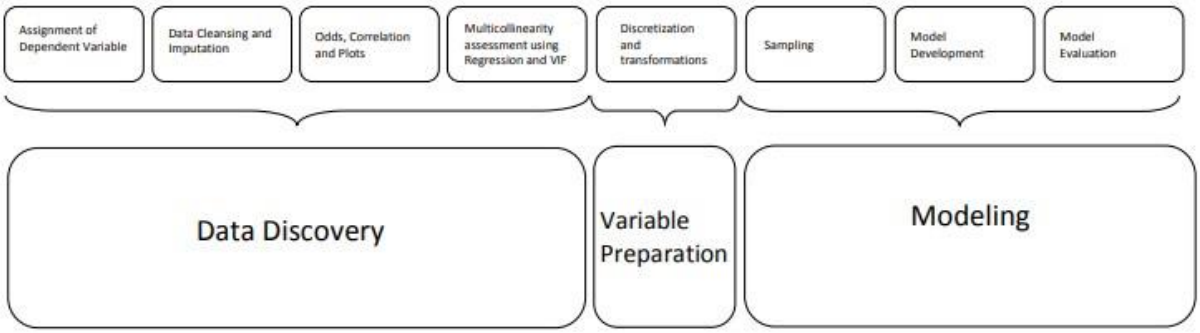
Introduction:

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

The data for this project came from a Sub-Prime lender and was provided in three Data Frames:

- CPR. 1,462,955 observations and 338 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 an ID column “MATCHKEY” that will be used to merge the Data Frames into one Dataset. A unique customer’s “MATCHKEY” is consistent across the three frames.



Data Discovery:

This report will begin by cleaning the data of missing and coded values by imputing those values to be the median of the column feature. Then this report will reduce feature redundancy by analyzing the collinearity of the column features and select the column that best represents the variable cluster. In a later phase of the project, a model will be made. The CPR and PERF datasets are merged together by MATCHKEY to yield a dataset with 1,255,429 observations and 356 features. All observations that are missing either AGE or DELQID were deleted during the merge procedure. The PERF dataset contains a variable labeled DELQID (the number of cycles that a customer is late on payment), which is used to assign a binary variable "goodbad" indicating whether a customer is considered a credit risk (0 = 'good' or 1 = 'bad'). Since the PERF dataset contains monthly data for customers during the observation period, an individual customer may be assigned to multiple observations. Each unique observation can have differing DELID values. The maximum value of DELQID is chosen to represent each customer in the creation of the variable "goodbad." A customer is assigned a value ranging from 0 to 7, where 0 means the customer is too new to rate, 1 means the customer paid within the current pay cycle, 2 means the customer is one cycle late, 3 means the customer is two cycles late, 4 means the customer is three cycles late, 5 means the customer is four cycles late, 6 means the customer is five cycles late, and 7 means the customer is six cycles late (1 cycle = 30 days). Suppose a customer is tracked for twelve months in the performance period and at some point, during that twelve months that customer is late on a payment by four cycles. This customer would be assigned a DELQID of 5 regardless of the payment activity in the remaining months. The variable "goodbad" will then be assigned a value of 1, indicating 'bad' for this customer. In general, if a customer has a DELQID of 3 or more, indicating that this customer at one point during the performance period paid two or more cycles late, then that customer will be assigned a "goodbad" value of 1. If a customer has a DELQID of 2 or less, indicating that this customer is at most one cycle late or too new to rate, then that customer will be assigned a "goodbad" value of 0. Table 1 below shoes the distribution of "goodbad."

Table 1: *Distribution of the variable "goodbad".*

goodbad	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1034829	82.43	1034829	82.43
1	220600	17.57	1255429	100.00

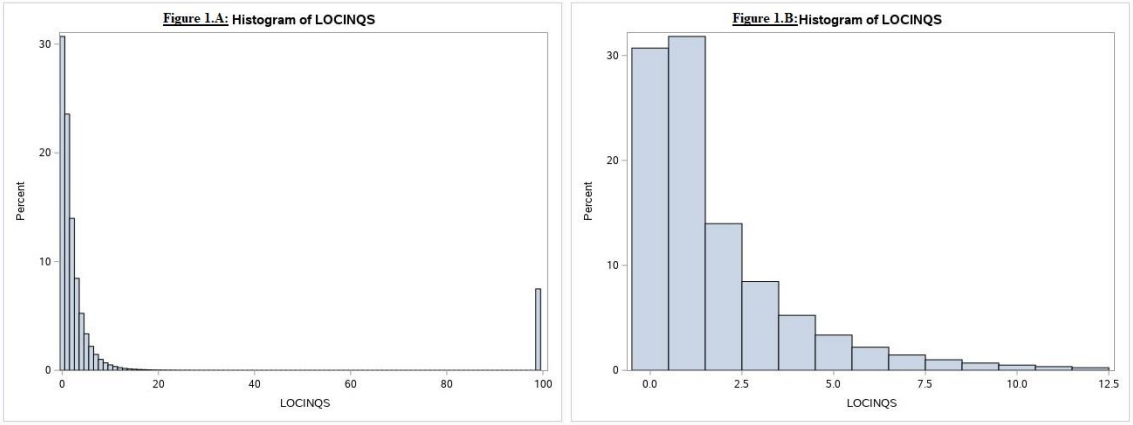
From table 1, the percentage of customers that are labeled ‘bad’ is 17.57%. This label corresponds to a customer being two or more cycles late. This is chosen as the cutoff point between a "good" or "bad" customer because credit card companies can make large profits on customers who are continually only 1 cycle late on payment. Once a customer is 2 or more cycles late, they generally tend to fall further and further behind on payments. One notable implication of assigning the DELQID variable to be the observed maximum value for a customer during their performance period is that the Type II error may increase. Since customers who may have gone three cycles late but caught up on their payments and finished the performance period paying on time are still assigned a "1" and considered a bad customer, there will be customers who are predicted to default and denied a credit card, even though they will pay off their accounts. The alternative would be extending credit cards to too many customers and losing money on the ones who default, which would result in a Type I error. The safer route will be taken, to try to reduce the probability of giving a credit card to a customer that will actually default.

During the merge of the CPR and PERF datasets, some of the observations between the two datasets did not match. There were 17,230 observations without DELQID. These are the number of customers who are in the CPR file, but not the PERF file. Since these customers will not have the binary indicator "goodbad", these observations are deleted from the dataset. There were also 470,846 customers without an age. These are the number of customers who are in the PERF file, but not the CPR file. These customers have none of the predictors that will be used to assess whether the customer is good or bad, so these observations are also deleted. All the performance variables except for DELQID and CRELIM (credit limit) are dropped for the cleaning process since these variables should not be changed. The dataset now contains 1,255,429 observations and 340 variables.

Variable Reduction:

The remaining 340 variables contains many missing and coded values. Variables with values ending 92 and above are coded. Values ending in 93 or more represent one of seven labels: invalid past due, invalid high credit, invalid balance, etc. Since these values are coded as such, the actual value associated with the numbers ending in 93 or more will be assigned the ending 92. To handle these values for this report, an imputation procedure was performed on the data. This procedure replaced the coded values with a 'meaningful' value. The median was chosen over the mean because the coded values greatly skew the data. Figure 1 below shows the variable LOCINQS, the number of local inquires in past six months, before (1.A) and after (1.B) the imputation procedure.

Figure 1: *Histogram of LOCINQS BEFORE (1.A) and AFTER (1.B) the imputation procedure.*



Many Histograms can be made from variables in this dataset that will show the same skewness as Figure 1.A. As a result of the coded values, the mean of each variable is very different from that variables median. The median is a better representative of the actual data values for customers, because the coded values greatly affect the mean value. Table 2 on the next page shows the descriptive statistics for five variables before (2.A) and after (2.B) the imputation procedure. Before the imputation, the variable LOCINQS had a mean value of 9.1836 (Table 2.A). After the imputation, the mean reduced to 1.7230 (Table 2.B). One can see that the mean is greatly reduced after imputation with the median. Note that the imputation is done with the median of the non-coded data.

Table 2.A: *Descriptive statistics of five variables before imputation.*

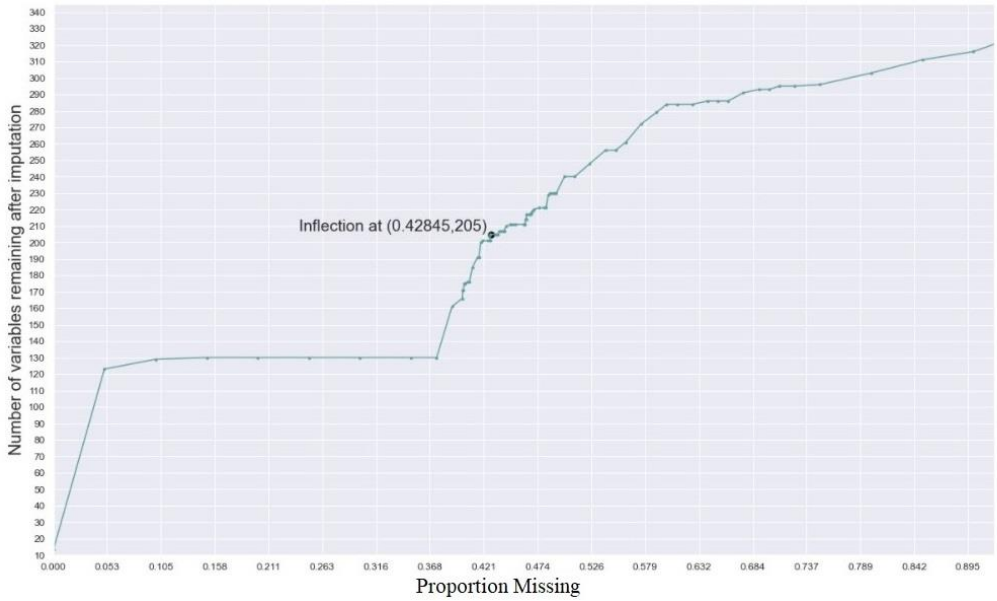
Variable	Mean	Std Dev	Maximum	N
BKPOP	0.0032371	0.0568037	1.0000000	1255429
LOCINQS	9.1836376	25.6771677	99.0000000	1255429
FFAVGMOS	440.8662091	458.7901233	999.0000000	1255429
CRDPTH	154.0329704	256.7546303	9999.00	1255429
TSBAL	38891.94	459737.00	9999999.00	1255429

Table 2.B: *Descriptive statistics of five variables after imputation.*

Variable	Mean	Std Dev	Maximum	N
BKPOP	0.0032371	0.0568037	1.0000000	1255429
LOCINQS	1.7230062	2.0802877	12.0000000	1255429
FFAVGMOS	56.9368184	44.7358637	331.0000000	1255429
CRDPTH	147.2582814	95.7045086	540.0000000	1255429
TSBAL	16978.59	16807.17	93603.00	1255429

A macro is used to go through each of the variables to decide which will remain in the dataset. One of the parameters in this macro sets the limit for the percentage of coded values within each variable. Suppose this parameter was set to (0.428451). The macro would not include variables with 42.8451% or more coded values in the output. Figure 2 displays the number of variables remaining when the limit is varied between 0 and 1 on a scatterplot.

Figure 2: ***Scatterplot of the Number of variables remaining after imputation by proportion of missing values.***



During the creation of Figure 2, eighty-six values were observed for the percentage of coded values. A limit of .42845 is chosen by interpolating the data via the method of polynomial least-squares approximation and finding an inflection point from the chosen polynomial. An adjusted-R^2 value was found for each polynomial and the highest was chosen. Table 3 on the next page depicts the adjusted-R^2 values for the polynomials of degree less than or equal to eleven that interpolates the data. On the right side of Table 3, the adjusted-R^2 values were sorted by highest to lowest. One can easily see that the polynomial of degree 10 has the highest score and was therefore chosen to be the polynomial used for finding an inflection point.

Table 3: ***Display of the adjusted R-Squ values for the interpolating polynomial of degree less than or equal to 11.***

Unsorted Adj_R-Squ	Degree	*Sorted* Adj_R-Squ	Degree
0.891072	1	0.984782	10
0.901099	2	0.984371	8
0.924023	3	0.984327	7
0.923089	4	0.984218	9
0.962673	5	0.984036	6
0.984036	6	0.962673	5
0.984327	7	0.924023	3
0.984371	8	0.923089	4
0.984218	9	0.901099	2
0.984782	10	0.891072	1
0.523222	11	0.523222	11

To find the inflection point from the interpolated polynomial, the bisection method was performed with starting values chosen at .4 and .5. These points were chosen due to leveling off of the graph for values less than or greater than this interval. Figure 3, on the next page, displays the polynomial of degree 10 that interpolates the data, as well as the second derivative of that polynomial. Figure 4 magnifies the x-axis to give a better visual of the range used in the bisection method. After successfully reaching a tolerance of 10^{-9} we received a value of (.428451). Next, a second and first derivative test was performed to check that this point is in fact an inflection point. Two relatively close values were chosen for the test, .4284 and .4285, and evaluated. Table 4 shows that the second derivative of the polynomial produced values of opposite signs when evaluated at the chosen points. Furthermore, the midpoint (.428451) produces the largest value when evaluated by the polynomials first derivative. Based on this evidence and the relative closeness to zero, a conclusion was made that the value produced by the bisection method is sufficiently close to the true inflection point and will be used as such in the imputation procedure.

Table 4: ***Displays the results of the Second-Derivative test (LEFT) and the First-Derivative test (RIGHT).***

	Second-Derivative test				*First-Derivative test*		
	xk	f``(x)	f``(xk)-f``(xk-1)		xk	f`(xk)	f`(xk)-f`(xk-1)
0	0.428400	1.396984e-09	1.862645e-09	0	0.428400	673.680723	0.000079
1	0.428451	3.259629e-09	0.000000e+00	1	0.428451	673.680802	0.000000
2	0.428500	-2.328306e-10	-3.492460e-09	2	0.428500	673.680727	-0.000075

Figure 3: *Scatterplot of the Number of variables remaining after imputation by proportion of missing values with a plot of the interpolating polynomial of degree 10.*

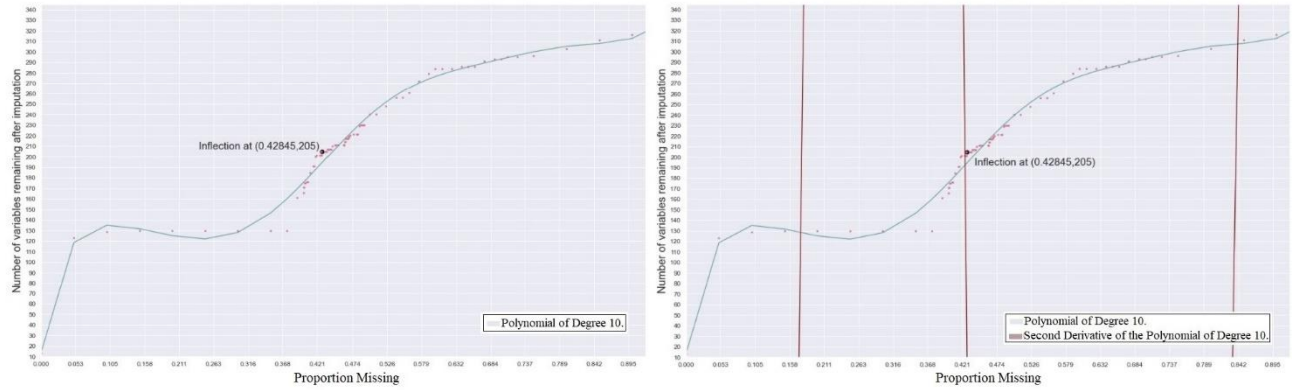
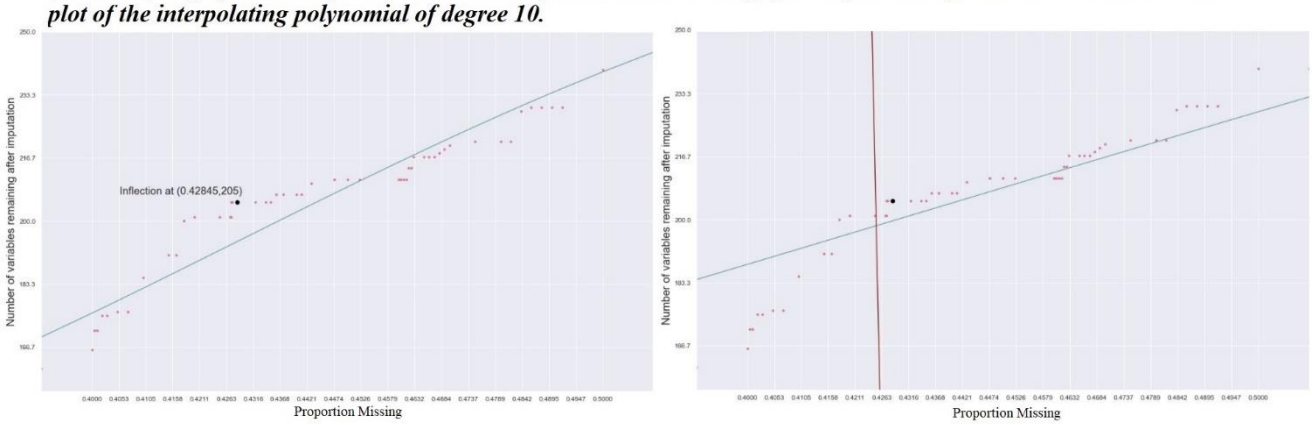


Figure 4: *Magnified scatterplot of the Number of variables remaining after imputation by the proportion of missing values with plot of the interpolating polynomial of degree 10.*



After imputing the variables at a limit of .42845, the number of remaining variables was 204. Before moving on to the next phase in the process, the remaining variables were checked with a means procedure for any coded values or errors. Table 6 displays twelve variables that have a minimum, maximum, mean, and standard deviation of zero. Since these variables were just columns of zeros, they were dropped from the list of remaining variables. Note that some variables, such as beacon, contained only missing values. Since these variables cannot provide any information, they were deleted. Furthermore, the variable age will be removed on the basis that this variable cannot be used to assess credit.

Table 6: *Variables that are columns of zeros.*

Variable	N	Mean	Std Dev	Minimum	Maximum
DCCRATE2	1255429	0	0	0	0
DCCRATE3	1255429	0	0	0	0
DCCRATE4	1255429	0	0	0	0
DCR324	1255429	0	0	0	0
FFCRATE2	1255429	0	0	0	0
FFCRATE3	1255429	0	0	0	0
FFCRATE4	1255429	0	0	0	0
FFRATE45	1255429	0	0	0	0
FFR324	1255429	0	0	0	0
FFR4524	1255429	0	0	0	0
DCR23	1255429	0	0	0	0
FFR23	1255429	0	0	0	0

Variable Clustering:

The next step in the process is clustering the variables into groups to see which variables have the most similarities. Once in clusters, the variable that is selected to be the cluster's representative will be chosen by selecting the variable that explains the largest proportion of variation within that variable's cluster. The determination of how many clusters will be chosen is based on percent variation explained for different number of clusters. Currently, there are 192 variables remaining in the dataset, four of which are the target variables: DELQID, CRELIM, goodbad, and MATCHKEY. These variables will be left out of the variable clustering portion. There will be 188 variables that will go through the variable clustering phase. One can see from Figure 5 that a higher number of variable clusters corresponds to a higher percent of variation explained. This is to be expected because when each variable is assigned their own clusters, the percent of variation explained will be 100%. Figure 5 displays the percent variation explained at different Number of Clusters. Around 61 variable clusters the percent of variation explained exceeds 80%. After 61 variable clusters the graph swiftly levels off. For this reason, 78 was chosen as the number of clusters with 85.9369% variation explained. This choice further reduces the number of starting variables (188) to under half.

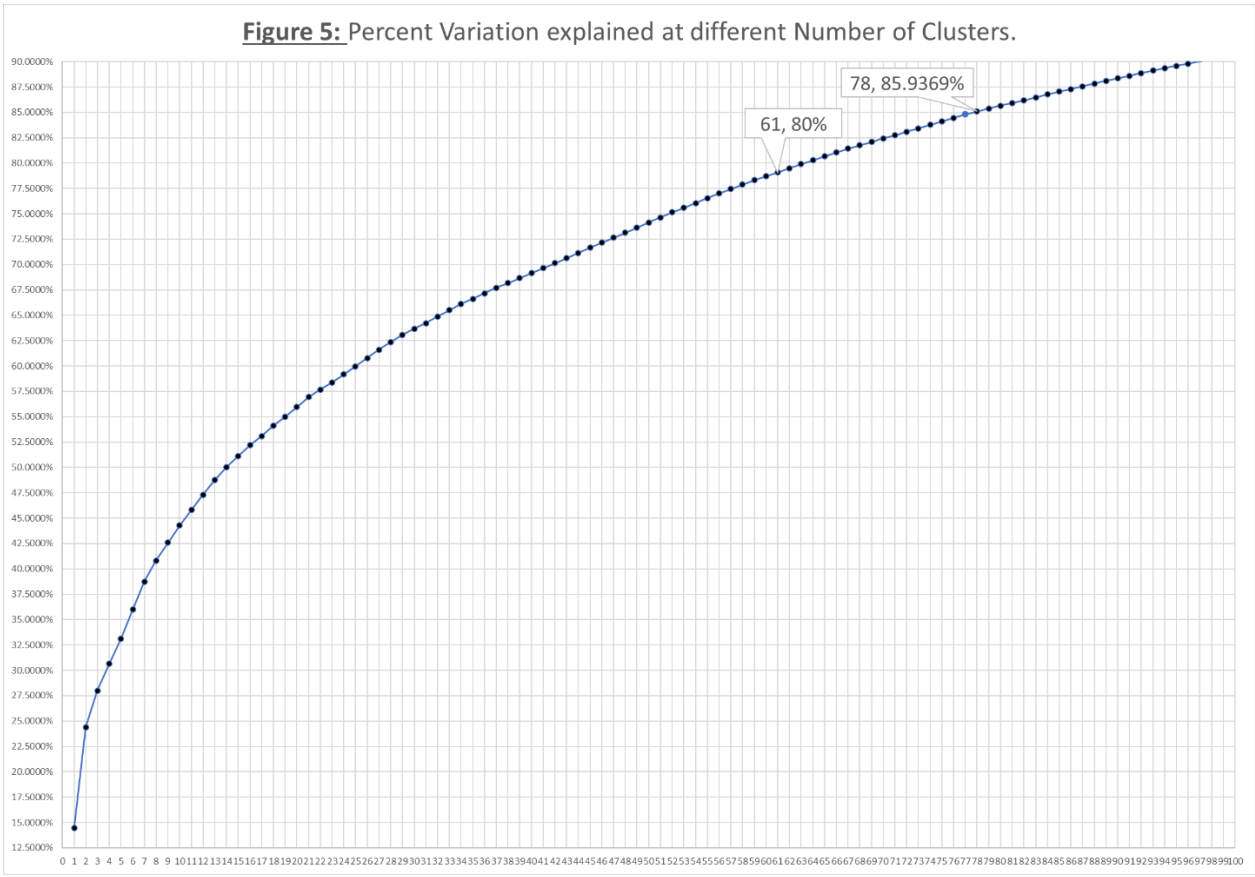


Figure 5: Percent Variation explained at different Number of Clusters.

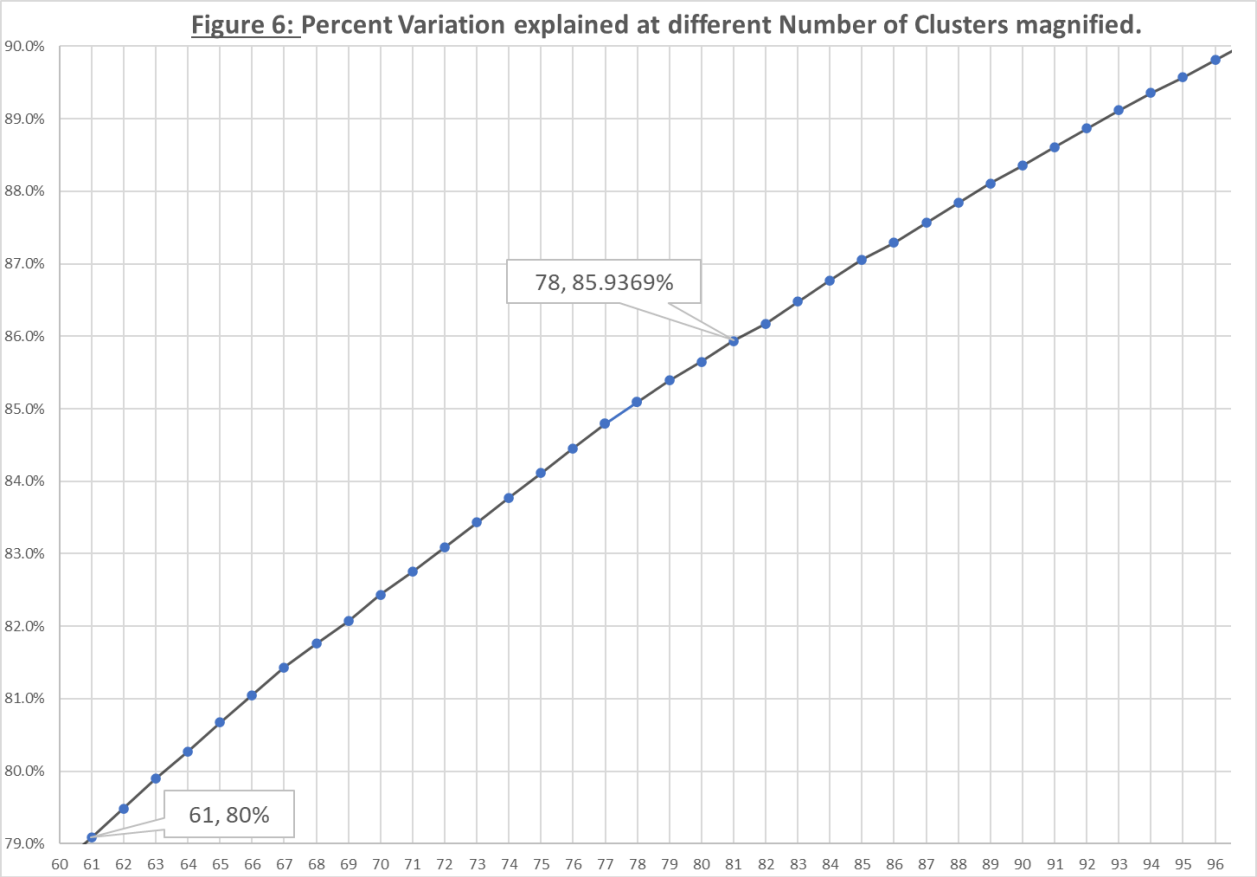


Figure 6 displays the same information as Figure 5 only magnified. In either figure, one can see a steady levelling off after 78 clusters. Once the clusters of variables are created and the number of variable clusters has been decided, a representative is selected from the cluster of variables by a ratio constructed from the variable's proportion of variation explained within its own cluster and the proportion of variation explained with the next cluster. Since the ratio is constructed to find the variable with the highest proportion of variation explained within its own cluster and the lowest proportion of variation explained with the next cluster, the variable with the lowest R^2 ratio will be chosen at the representative of that variables cluster. Table 7 shows the first three clusters with the representative variable highlighted in red. As one can see, the representative has the lowest value in the RSquareRatio column.

Table 7: *Displays all of the variables within a cluster.*

NumberOfClusters	Cluster	Variable	OwnCluster	NextClosest	RSquareRatio
78	Cluster 1	TRCR49	0.9454	0.6062	0.1386
78	Cluster 1	TRCR39	0.9393	0.6156	0.1579
78	Cluster 1	TRR49	0.9386	0.6355	0.1686
78	Cluster 1	TRATE79	0.8939	0.5403	0.2309
78	Cluster 1	CRATE79	0.8932	0.5381	0.2313
78	Cluster 1	TRR39	0.8868	0.6417	0.3159
78	Cluster 1	TRR29	0.7197	0.5567	0.6322
78	Cluster 1	WCRATE	0.5263	0.3771	0.7604
78	Cluster 2	BRCRATE1	0.9390	0.5184	0.1267
78	Cluster 2	BRRATE1	0.9027	0.5079	0.1977
78	Cluster 2	BROPENEX	0.8923	0.4631	0.2006
78	Cluster 2	BRTRADES	0.8920	0.4629	0.2012
78	Cluster 2	BRR124	0.8354	0.5949	0.4063
78	Cluster 2	BROPEN	0.7998	0.6040	0.5055
78	Cluster 3	BRCR39	0.8942	0.6262	0.2831
78	Cluster 3	BRCR49	0.8951	0.6543	0.3036
78	Cluster 3	BRR49	0.8438	0.6774	0.4843
78	Cluster 3	BRWCRATE	0.6876	0.4109	0.5304

Table 8: *Displays the variable clusters and its representative.*

NumberOfClusters	Cluster	Variable	OwnCluster	NextClosest	RSquareRatio	NumberOfClusters	ControlVar	Cluster	Variable	OwnCluster	NextClosest
78	Cluster 1	TRCR49	0.9454	0.6062	0.1386	78	Cluster 40	DCRATE2	1.0000	0.3772	0.0000
78	Cluster 2	BRCRATE1	0.9390	0.5184	0.1267	78	Cluster 41	BRCR1BAL	0.8734	0.4643	0.2364
78	Cluster 3	BRCR39	0.8942	0.6262	0.2831	78	Cluster 42	FFPCTSAT	0.7612	0.2383	0.3134
78	Cluster 4	TOPEN12	0.8156	0.4463	0.3331	78	Cluster 43	BKPOP	1.0000	0.0106	0.0000
78	Cluster 5	TROPENEX	0.9599	0.5359	0.0865	78	Cluster 44	PFRATE45	1.0000	0.1323	0.0000
78	Cluster 6	DCRATE79	0.9739	0.6223	0.0692	78	Cluster 45	PFRATE3	1.0000	0.1001	0.0000
78	Cluster 7	FFCR49	0.9745	0.5933	0.0627	78	Cluster 46	FFRATE2	1.0000	0.4773	0.0000
78	Cluster 8	BRR324	0.7449	0.1788	0.3107	78	Cluster 47	PFRATE2	1.0000	0.1797	0.0000
78	Cluster 9	TOPENB75	0.8628	0.3992	0.2284	78	Cluster 48	BNKINQ2	1.0000	0.3807	0.0000
78	Cluster 10	FFCRATE1	0.9146	0.2179	0.1092	78	Cluster 49	FFCR1BAL	1.0000	0.1141	0.0000
78	Cluster 11	DCOPENEX	0.9267	0.4027	0.1227	78	Cluster 50	BRPOPEN	0.8633	0.4352	0.2420
78	Cluster 12	PFR49	0.9191	0.4566	0.1489	78	Cluster 51	FFLAAGE	1.0000	0.0677	0.0000
78	Cluster 13	RADB6	0.9004	0.3704	0.1582	78	Cluster 52	PRMINQ2	1.0000	0.3807	0.0000
78	Cluster 14	BRR4524	0.7875	0.3307	0.3175	78	Cluster 53	DCRATE3	1.0000	0.0944	0.0000
78	Cluster 15	BRRATE2	0.7400	0.0862	0.2846	78	Cluster 54	FFR224	1.0000	0.4773	0.0000
78	Cluster 16	AVGMOS	0.7743	0.5570	0.5094	78	Cluster 55	TOPEN	0.9443	0.5581	0.1261
78	Cluster 17	RBAL	0.9163	0.3831	0.1357	78	Cluster 56	OBRPTAT	1.0000	0.0894	0.0000
78	Cluster 18	LOCINQS	0.8136	0.1512	0.2197	78	Cluster 57	BINQ12	1.0000	0.1541	0.0000
78	Cluster 19	BRMOSOPN	0.8370	0.4284	0.2852	78	Cluster 58	ORRATE3	1.0000	0.1051	0.0000
78	Cluster 20	PFRATE1	0.9320	0.1009	0.0756	78	Cluster 59	DCCR1BAL	1.0000	0.2267	0.0000
78	Cluster 21	BADPR1	0.9143	0.5423	0.1874	78	Cluster 60	BKP	1.0000	0.2040	0.0000
78	Cluster 22	FFN90P24	0.9296	0.3258	0.1044	78	Cluster 61	FFAGE	1.0000	0.6887	0.0000
78	Cluster 23	BRHIC	0.8781	0.3751	0.1951	78	Cluster 62	BRCRATE4	0.8338	0.3038	0.2388
78	Cluster 24	DCR4524	0.8513	0.1132	0.1677	78	Cluster 63	FININQS	1.0000	0.1871	0.0000
78	Cluster 25	OT6PTOT	0.8105	0.5299	0.4030	78	Cluster 64	FFMOSOPN	1.0000	0.6596	0.0000
78	Cluster 26	PRMINQS	0.8451	0.3597	0.2419	78	Cluster 65	FFWCRATE	1.0000	0.1710	0.0000
78	Cluster 27	BRR23	0.8031	0.4180	0.3384	78	Cluster 66	CRDPTH	0.8022	0.5501	0.4397
78	Cluster 28	BRR39P24	0.8560	0.5361	0.3104	78	Cluster 67	TOPEN6	0.8518	0.5527	0.3314
78	Cluster 29	DCR29P24	0.8390	0.5774	0.3809	78	Cluster 68	CPAF29	0.8454	0.2926	0.2185
78	Cluster 30	LAAGE	0.7222	0.0506	0.2926	78	Cluster 69	TSBAL	1.0000	0.2536	0.0000
78	Cluster 31	BRMINB	1.0000	0.1167	0.0000	78	Cluster 70	DCR49	0.8865	0.5920	0.2781
78	Cluster 32	BRCRATE2	0.7579	0.3680	0.3830	78	Cluster 71	FFAVGMOS	1.0000	0.6887	0.0000
78	Cluster 33	OT24PTOT	0.8207	0.3169	0.2625	78	Cluster 72	BRCRATE3	0.7974	0.2479	0.2694
78	Cluster 34	DCR224	1.0000	0.3772	0.0000	78	Cluster 73	BRNEW	0.7058	0.3077	0.4249
78	Cluster 35	DCWCRATE	1.0000	0.1946	0.0000	78	Cluster 74	BRMINH	1.0000	0.1167	0.0000
78	Cluster 36	COLLS	1.0000	0.0864	0.0000	78	Cluster 75	TADB25	1.0000	0.4374	0.0000
78	Cluster 37	DCLAAGE	1.0000	0.0317	0.0000	78	Cluster 76	BRRATE79	0.9449	0.6492	0.1571
78	Cluster 38	TPOPEN	0.8769	0.4098	0.2086	78	Cluster 77	DCMOSOPN	1.0000	0.4548	0.0000
78	Cluster 39	FFRATE3	1.0000	0.1643	0.0000	78	Cluster 78	FFR29	0.8759	0.4950	0.2456

Table 8 displays the variables that will represent the other variables their cluster. Once these representatives are selected from the cluster, the dataset will have 78 predictor variables and four target variables bringing the total to 82 variables. One final check is performed on the remaining predictor variables, a linear regression is run to inspect the variation inflation factor, or VIF. In general, a VIF of ten or more indicates multicollinearity among variables. For this reason, any variables with a VIF of ten or more will be dropped. Table 9 shows the variables that have the highest VIF. None of the 78 predictor variables had a VIF greater than 10. The variables TOPEN, TROPENEX, and BRCRATE1 had the highest VIF's. This suggests that these variables have a correlated relationship with one or more of the other variable cluster representatives. Table 10 displays the correlation table between these variables. One can see that the variables that are the most correlated are TOPEN and TROPENEX. However, since the cutoff point was set at VIF = 10, these three variables will be included going into the next phase. The final dataset to take into the discretization process contains 82 variables (78 predictor and 4 target variables).

Table 9: Linear Regression of the remaining variables.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	VIF
TOPEN	1	0.0022	0.0002	13	<.0001	9.2205
TROPENEX	1	-0.0014	0.0001	-16.94	<.0001	7.1043
BRCRATE1	1	-0.0107	0.0002	-58.91	<.0001	6.7714
AVGMOS	1	-0.0005	0.0000	-19.96	<.0001	5.6831
FFAGE	1	0.0000	0.0000	-2.22	0.0261	5.2082

Table 10: Correlations between variables with the largest VIF's.

Pearson Correlation Coefficients, N = 1255429 Prob > |r| under H0: Rho=0					
	TOPEN	TROPENEX	BRCRATE1	AVGMOS	FFAGE
TOPEN	1.00000	0.72561 <.0001	0.67187 <.0001	0.01355 <.0001	0.12076 <.0001
TROPENEX	0.72561 <.0001	1.00000	0.62792 <.0001	0.30909 <.0001	0.26177 <.0001
BRCRATE1	0.67187 <.0001	0.62792 <.0001	1.00000	0.17562 <.0001	0.15288 <.0001
AVGMOS	0.01355 <.0001	0.30909 <.0001	0.17562 <.0001	1.00000	0.48093 <.0001
FFAGE	0.12076 <.0001	0.26177 <.0001	0.15288 <.0001	0.48093 <.0001	1.00000

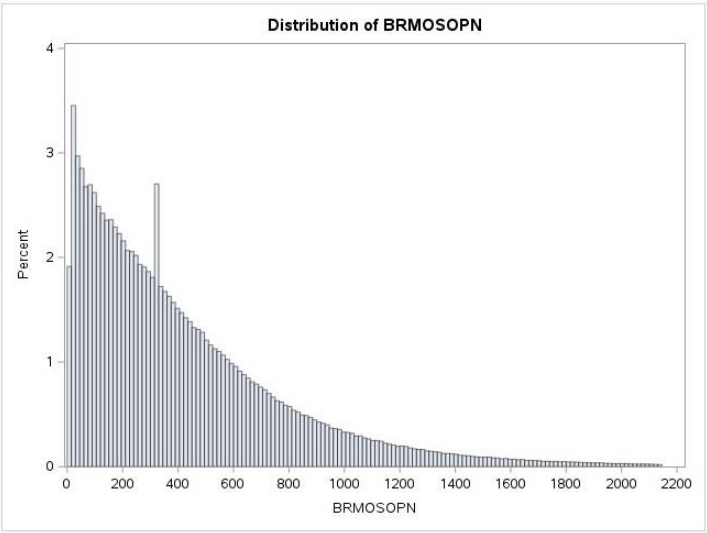
Variable Preparation:

In the next phase of this report, two methods will be used to create an ordinal ranking for the values of each variable. The two methods, labelled discretization process 1 and discretization process 2, will attempt to bin each variable into bins of equal size. The discretization process 1 was performed before the discretization process 2 and is done by manually setting the bins to be equal interval widths. Discretization process 2 utilizes the rank procedure in SAS to bin the variable into a maximum of ten bins. For each bin, an aggregation of the mean proportion of 'bads' (goodbad = 1) is calculated. This number represents the probability of default. An odds ratio is then created from the mean proportion of 'bads' in each bin. This value is calculated by dividing the mean of each bin by the quantity (1 - mean) and represents the odds of default for the values contained within the corresponding interval. Suppose a variable has a bin with a default value of .2. The odds ratio for the variable is $(\frac{(0.2)}{(1-(0.2))} = .25)$. Since this value is less than one, the reciprocal of this value will be used in its place: $(\frac{1}{0.25} = 4)$. Thus, a customer in this bin is four times more likely to not default than to default. Finally, a log transformation is created from the odds ratio for use in the logistic function during the next phase. At the end of this phase 36 of the 78 variables will undergo at least one discretization process that will result in up to six additional transformations of the variable. Table 11 depicts the names of the 36 variables that will go through the process of discretization. The last four variables, FFLAAGE, DCLAAGE, LAAGE, and TRCR49, were found to not produce a useful discrete variable after the first discretization process. The ordinal variables created for these variables was deleted but the original variables remained during the next phase. The remaining 46 variables were already discrete and will also be left in during the modelling phase.

Table 12: Discretization process 1 variables

Variable	Max	N Miss	Median	Variable	Max	N Miss	Median
TSBAL	93603	0	12006	TADB25	20	0	3
BRHIC	82876	0	8700	BNKINQ2	19	0	4
RBAL	44786	0	4505	PFRATE1	17	0	2
BRMINB	5564	0	314	TOPENB75	15	0	2
BRMINH	5252	0	500	BADPR1	14	0	1
BRMOSOPN	2143	0	321	FININQS	13	0	2
DCMOSOPN	1190	0	129	LOCINQS	12	0	1
FFMOSOPN	658	0	81	TOPEN12	11	0	2
CRDPTH	540	0	130	OT24PTOT	2.2222	0	0.5
FFAGE	422	0	60	RADB6	1.7967	0	0.4657
FFAVGMOS	331	0	47	TPOPEN	1	0	0.56
AVGMOS	182	0	56	OBRPTAT	1	0	0.5
BRNEW	85	0	11	BRPOPEN	1	0	0.7143
PRMINQS	67	0	17	OT6PTOT	0.8	0	0.0625
TROPENEX	55	0	16	FFLAAGE*	61	0	1
PRMINQ2	41	0	10	DCLAAGE*	56	0	1
TOPEN	32	0	9	LAAGE*	24	0	0
BRCRATE1	24	0	6	TRCR49*	16	0	1

Figure 7: Distribution of the variable BRMOSOPN



In the first discretization process, the researcher logically defines the bins by inferential cut points given by the histogram in Figure 7. They then collapse the first or last intervals at the end that have similar probabilities of default. Figure 7 displays the distribution of the variable BRMOSOPN (Months open for per bank revenue trades). As one can see, the distribution of this variable is heavily skewed to the right. It was found that in 5 bins this variable displayed a negative relationship between the variable and the probability of default suggesting that the probability of default decreases as the value of this parameter increases. From table 11, one can see that from bin 1 (values of BRMOSOPN ranging from 0 to 199) to bin 2 (values of BRMOSOPN ranging from 200 to 399) the probability of default decreases from (0.2035189) to (0.1754772).

Table 11: Transformation of the variable BRMOSOPN from the first discretization process.

Rank	Frequency	Percent	Cumulative Percent	Probability of Default	Min of BRMOSOPN	Max of BRMOSOPN
1	427631	34.06%	34.063%	0.2035189	0	199
2	318976	25.41%	59.470%	0.1754772	200	399
3	205596	16.38%	75.847%	0.1680237	400	599
4	126577	10.08%	85.929%	0.1543724	600	799
5	176649	14.07%	100.00%	0.1330944	800	2143

Bin 5 contains values greater than or equal to 800. The researcher initially chose to partition this variable into ten bins. It was noticed that values greater than 800 had a default rate close to 13%. Due to this, a cutoff point was chosen at 800. This bin had the lowest probability of default of the remaining bins. Figure 8 displays the probability of default by each bin. One can see a clear negative relationship between the probability of default and the variable.

Figure 8: Probability of Default by the bins of BRMOSOPN

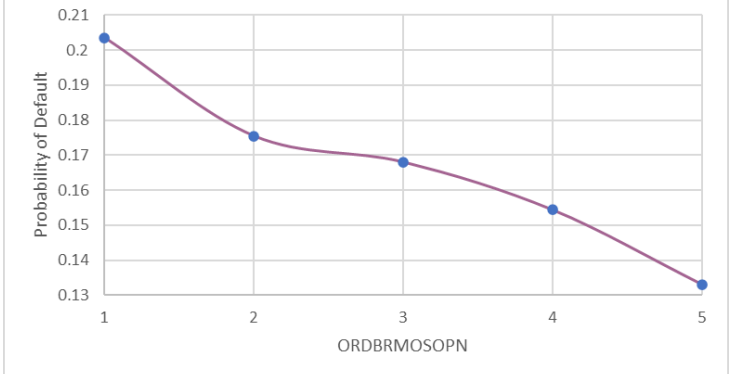


Figure 9: Histogram and line plot of the variable FFLAAGE and the ordinal variable created in the first discretization process.

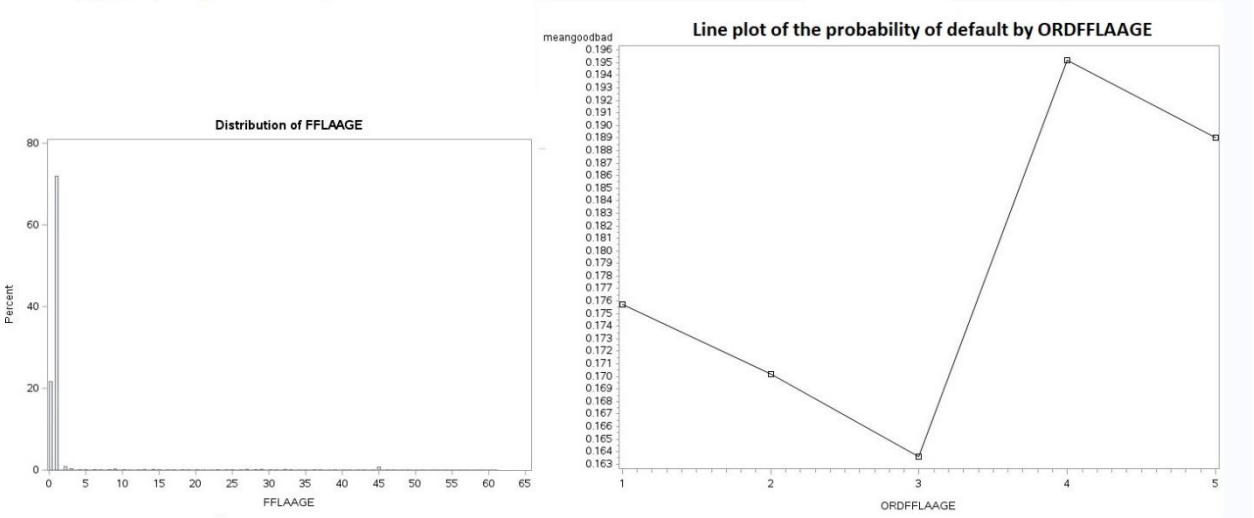


Table 12 on page 13 of this report depicts four variables that could not be discretized into uniformly distributed bins. In figure 9, a histogram and line plot for the variable FFLAAGE and the corresponding ordinal variable, ORDFFLAAGE is shown. As one can see from the histogram, a majority of the observations have a value less than five. If one looks closely around 45 there is an additional spike. From figure 9, one can see that bin 3 has the lowest probability of default of the five bins (from table 13 the probability of default = 0.1636327). This bin corresponds to the interval of values greater than or equal to 24 and less than 40 where the density of observations is disseminated evenly. Neither collapsing bins into bin 3 nor changing the length of the interval resulted in a desirable distribution of this variable. For this reason, this variable will only be represented by the original variable, FFLAAGE.

Table 13: Transformation of the variable FFLAAGE from the first discretization process.

Rank	n	percent	cumalitive percent	Probability of Default	Min of FFLAAGE	Max of FFLAAGE
1	1195923	95.26%	95.26%	0.1757153	0	7
2	20452	1.63%	96.89%	0.1702034	8	23
3	20106	1.60%	98.49%	0.1636327	24	39
4	17075	1.36%	99.85%	0.1951977	40	55
5	1873	0.15%	100.0%	0.1890016	56	61

The next discretization process will use the remaining 32 variables for a second time to create an additional three variables. Each of the original variable should be split into bins of equal frequency but due to the discrete nature of many of the variables, this may not be the case for every variable. Figure 10 on the next page displays such a variable. The variable BRMINB displays a negative linear relationship in six bins. To the right of this line plot is the results of the second discretization process. This plot shows that bin 4, 5, and 6 have a higher probability of default than bins 1 to 3 and 7 to 9.

Figure 10: Line plots of the ordinal variables created for the variable BRMINB during the discretization processes.

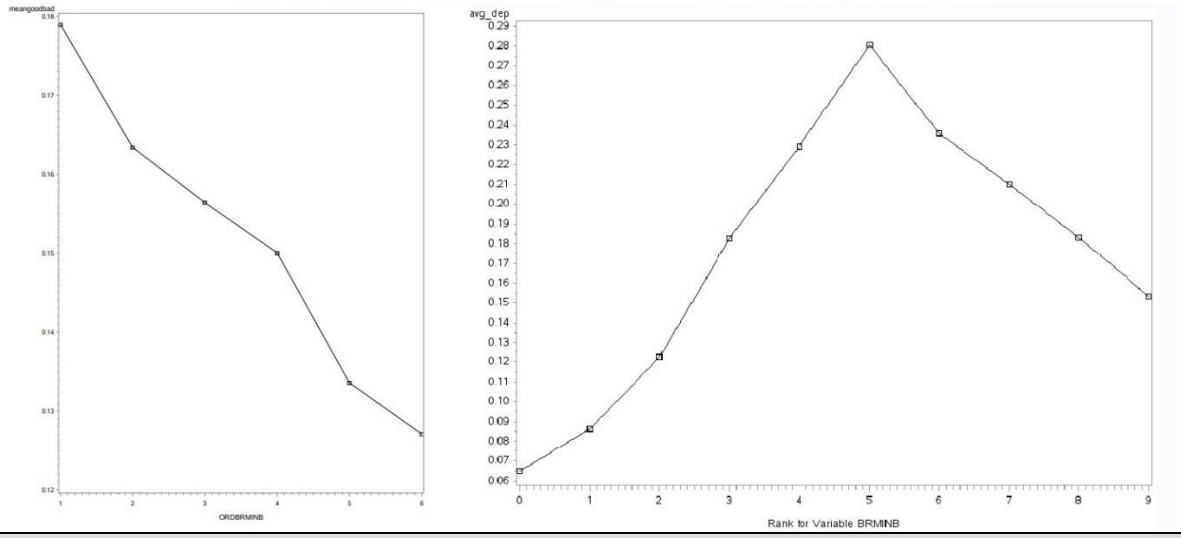


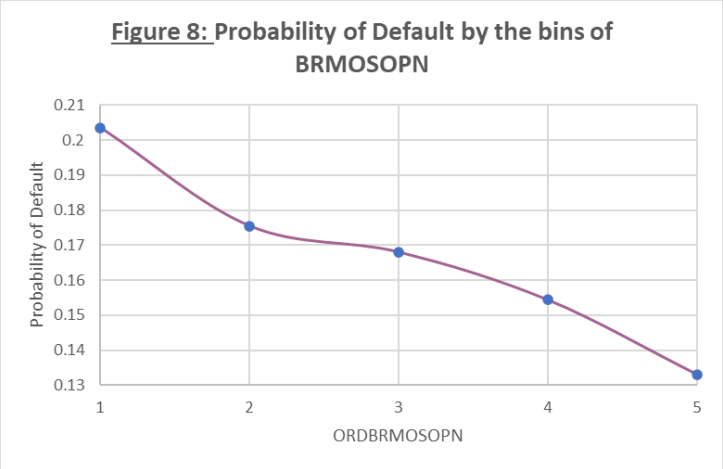
Table 14: Transformation of the variable BRMINB from the first discretization process.

Rank	Frequency	Percent	Cumulative Percent	Probability of Default	Min of BRMINB	Max of BRMINB
1	1066899	84.98%	84.983%	0.178986	0	1099
2	105701	8.42%	93.402%	0.1634232	1100	1999
3	43480	3.46%	96.866%	0.1564397	2000	2899
4	21348	1.70%	98.566%	0.1500375	2900	3799
5	11289	0.90%	99.465%	0.1335814	3800	4699
6	6712	0.53%	100.00%	0.1270858	4700	5564

Table 14 shows the numeric results of the discretization process 1. From this table, the percentage of data that is within the bounds of the interval assigned to bin 1 is 84.98% of the total column sample. This interval corresponds to values of BRMINB less than 1100 and a Probability of Default of 0.178986. This corresponds to Rank 1 through 9 for the results of the second discretization process on table 15. The Probability of Default starts below 10% in Rank 1 and 2 and rapidly increases to 0.28072 in Rank 6. After Rank 6 the Probability of Default begins to slowly decrease down to 0.15308. Since the first discretization process yields an ordinal variable with a linear trend and the second did not, only the original variable and the ordinal variable, odds ratio, and log transformation of the odds ration from the first discretization process will be included in the modelling phase of this report.

Table 15: Transformation of the variable BRMINB from the second discretization process.

Rank	Frequency	Percent	Cumulative Frequency	Cumulative Percent	Probability of Default	Min of BRMINB	Max of BRMINB
1	120136	9.57%	120136	9.57%	0.06498	0	5
2	128682	10.25%	248818	19.82%	0.0862	6	58
3	127718	10.17%	376536	29.99%	0.12243	59	148
4	124461	9.91%	500997	39.91%	0.18256	149	237
5	111011	8.84%	612008	48.75%	0.22941	238	313
6	141126	11.24%	753134	59.99%	0.28072	314	408
7	125212	9.97%	878346	69.96%	0.23612	409	546
8	125768	10.02%	1004114	79.98%	0.21022	547	868
9	125737	10.02%	1129851	90.00%	0.18318	869	1559
10	125578	10.00%	1255429	100.0%	0.15308	1560	5564



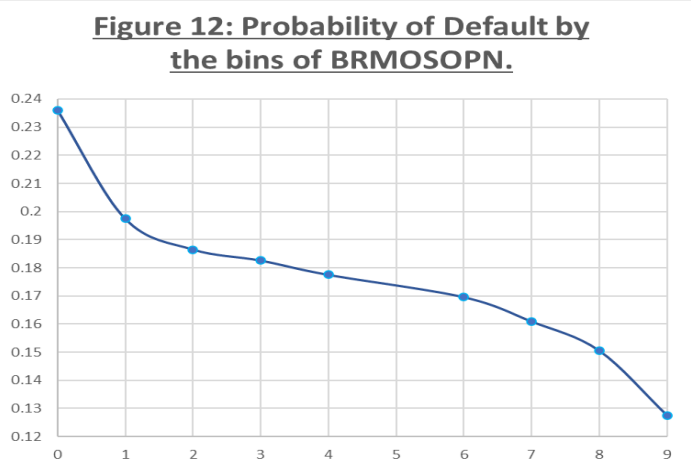


Figure 8 describing the ordinal variable of BRMOSOPN created during the first discretization procress was copied and placed next to the results of the second discretization process (Figure 12) for comparison. Both figures display a similar shape and trend. Bin 3 in figure 8 has a probability of default of 0.1680237 and ranges between 400 and 599. This bin corresponds to the Ranks 7 and 8 which range from 321 to 516 and 517 to 669, respectively. Of these two ranks, Rank 7 has the highest probablity of default at 0.1609.

Table 11: Transformation of the variable BRMOSOPN from the first discretization process.

Rank	Frequency	Percent	Cumulative Percent	Probability of Default	Min of BRMOSOPN	Max of BRMOSOPN
1	427631	34.06%	34.063%	0.2035189	0	199
2	318976	25.41%	59.470%	0.1754772	200	399
3	205596	16.38%	75.847%	0.1680237	400	599
4	126577	10.08%	85.929%	0.1543724	600	799
5	176649	14.07%	100.00%	0.1330944	800	2143

Table 18: Transformation of the variable BRMOSOPN from the second discretization process (AFTER rank collaspe).

Rank	Frequency	Percent	Cumulative Frequency	Cumulative Percent	Probability of Default	Min of BRMOSOPN	Max of BRMOSOPN
1	123743	9.86%	123743	9.86%	0.23609	0	52
2	125602	10.00%	249345	19.86%	0.19748	53	108
3	125831	10.02%	375176	29.88%	0.18649	109	171
4	125544	10.00%	500720	39.88%	0.18256	172	241
6	125308	9.98%	626028	49.87%	0.17749	242	320
7	251904	20.07%	877932	69.93%	0.16954	321	516
8	126316	10.06%	1004248	79.99%	0.1609	517	669
9	125305	9.98%	1129553	89.97%	0.15054	670	924
10	125876	10.03%	1255429	100.0%	0.12759	925	2143

Logistic Regression:

In this phase, a logistic regression model will be built from all 270 predictor variables with the goodbad variable as the target. A stepwise selection approach to this regression was performed based off a cutoff point at a significance level of .05. Initially, the regressor with the largest correlation with the target variable will selected. Next, a single regressor will be added to the model and the regressors in the new model will be evaluated for a change in significance with the target variable.

Suppose the evaluation of the new model indicates either no change or an increase in significance for each of the regressors in the new model. Then the procedure will continue by selecting another variable from the list of unused regressors and reevaluate the newest model. Suppose the evaluation of the new model indicates one or more regressors had a reduction in significance. Then the procedure will remove the insignificant variables and reevaluate the newest model. This procedure will finish when no variables left meet the entry criterion.

Table 19: Top 15 variables from Stepwise Selection Logistic Regression

Parameter	Estimate	Standard Error	Wald Chi-Sq	Pr > Chi-Sq
BRMINH	-0.00041	0.000010	1716.7726	0.0001
BRR4524	0.3252	0.008140	1596.5739	0.0001
BRRATE2	0.1647	0.004130	1588.6752	0.0001
BRCR1BAL	0.1078	0.002730	1558.2272	0.0001
BRR23	0.4425	0.012000	1349.9849	0.0001
BRPOPEN	-2.7806	0.091500	923.6404	0.0001
BRCRATE4	0.3268	0.011400	826.1276	0.0001
BRNEW	-0.0132	0.000573	526.6001	0.0001
BRR324	0.2148	0.009460	515.7815	0.0001
RADB6	1.2372	0.056500	479.0043	0.0001
BRRATE79	-0.1188	0.005450	475.3289	0.0001
ORDBRNEW	0.1973	0.009200	459.9005	0.0001
ordeqBRHIC	0.0998	0.004720	447.2199	0.0001
ordeqBRPOPE	0.1634	0.007740	445.7173	0.0001
BRCRATE2	0.3373	0.016100	439.5258	0.0001

The parameter estimates in the model cannot be interpreted like a regular regression model. Instead, the log odds of the estimates must be computed before interpretation. Since this is a logistic regression, this interpretation can be expressed as the log odds of default. For example, the estimate for the variable BRR4524 (the number of bank revenue accounts 90+ days older than 25 months) is 0.3252. This means that for each additional payment cycle a customer is delinquent on a bank revenue account 90 or more days older than 24 months, a customer is 1.3843 ($e^{0.3252}$) times more likely to default.

The results of the stepwise selection procedure resulted in a model containing 130 regressors. One of the statistics used to assess model performance is the c statistic. This statistic represents the percent of concordant pairs in the model. This percentage is found by grouping the data by goodbad and splitting the groups into two partitions: one contains the customers that are not in default, and the other contains the customers that are in default. The model is then used to compute the predicted probabilities of default for each observation in both group partitions. Each observation in one group partition is then paired with an observation in the other group partition. A pair of observations is concordant when the predicted probability of default for an observation with a goodbad value of 0 is less than the predicted probability of default for an observation with a goodbad value of 1. This means that the model should predict lower probabilities for customers that did not default and higher probabilities for customers that did default. The c statistic is the metric used to assess how well the model predicts whether a customer did or did not default. Table 20 displays the percent of concordant pairs for the full model of 130 regressors. The value of 84.7% indicates that this model predicts 4 out of 5 observations correctly.

Table 20: Percent of Concordant Pairs in the Full Model.

Percent Concordant	84.7		Somers' D	0.694
Percent Discordant	15.3		Gamma	0.694
Percent Tied	0		Tau-a	0.201
Pairs	99440320848		c	0.847

The next step in this phase is to reduce the model by removing the parameter with the lowest chi-square statistic one at a time until three variables remain. Table 21 displays the three parameters left after reducing the model. This model produced a c statistic of 78.1% which means that this model predicts 3 out of 4 observations correctly. After reevaluation of each of the remaining parameters completed, the parameter labels, the chi-square statistics, and their respective p-values were collected to aggregate a mean chi-square statistic for each parameter. The parameters' frequencies were also collected over the 127 different model evaluations. Once completed, different permutations were created from the top 20 variables with the highest aggregated chi-square. Table 22 at the top of the next page displays the top 20 variables with the highest aggregated Chi-Square.

Table 21: Results of Reducing the Full Model to Three Variables.

Parameter	df	Estimate	Std. Err.	Chi-Sq	Pr > Chi-Sq
RADB6	1	2.7585	0.0113	59699.7913	<.0001
BRR23	1	1.0393	0.0074	19517.2605	<.0001
BRR4524	1	0.8398	0.0057	21936.3714	<.0001
c statistic for this model				80%	

Table 22: Aggregated mean Chi-Square after reducing the full model to three parameters.

var	n	Average of wald
lodseqRADB6	100	6473.69
lodseqBRCRATE1	99	5433.06
BRR23	100	3470.93
BRR4524	100	3016.95
BRPOPEN	97	2254.94
BRMINH	95	2027.36
BRCR1BAL	91	1895.27
BRRATE2	92	1772.11
AVGMOS	96	1362.52
TOPENB75	98	1272.55
LOCINQS	94	1020.62
BRCRATE4	90	1014.29
BRNEW	88	744.84
loddsBRMINH	81	716.31
BRR324	93	700.69
BRCRATE1	89	578.15
ordeqBRPOPEN	84	575.72
ORDBRNEW	87	571.02
RADB6	74	507.01
BRRATE79	82	497.39

From these 20 variables seen in Table 22, different permutations were created and modeled. These permutations ranged from three to twenty variables, and the variables with the highest c statistic were chosen to be analyzed in the next phase of this report. Interestingly to note, only two of the three variables in the fully reduced model appeared in the top three positions in Table 22. After evaluating each permutation of parameters, two models, one containing three variables and the other containing twenty, were selected to proceed to the next phase of this report. Table 23 displays the percent concordance of these two models.

Table 23: Percent of Concordant Pairs

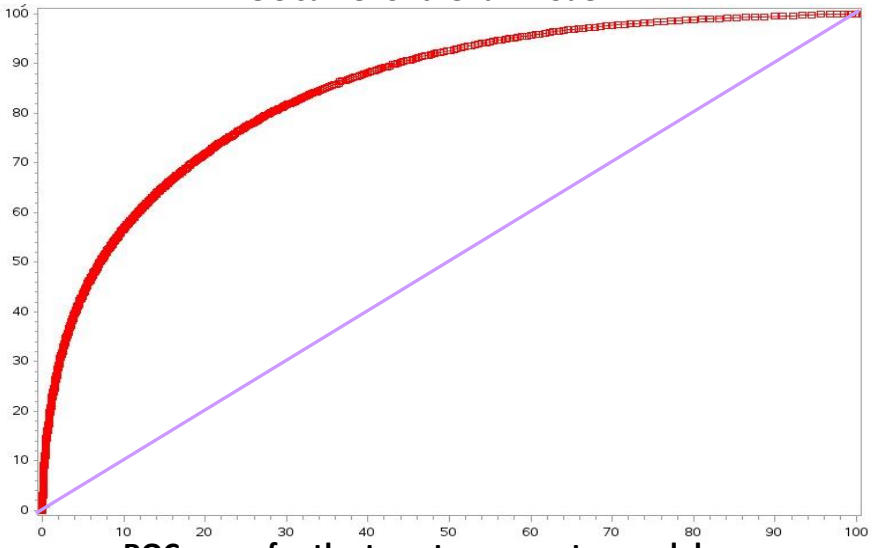
Percent of Concordant Pairs in the Twenty-Parameter Model.				Percent of Concordant Pairs in the Three-Parameter Model.			
Percent Concordant	84	Somers' D	0.679	Percent Concordant	80.3	Somers' D	0.606
Percent Discordant	16	Gamma	0.679	Percent Discordant	19.6	Gamma	0.607
Percent Tied	0	Tau-a	0.197	Percent Tied	0.1	Tau-a	0.176
Pairs	99440320848	c	0.84	Pairs	99440320848	c	0.803

A ROC curve is a visual representation for the c-statistic of a given model. For each of the models discussed above, a ROC curve was created from the probabilities, sensitivity, and specificity listed on the classification table output of the logistic regression procedure. Before specificity can be used, the difference between 100 and specificity must be computed. This is because specificity represents the percent of nonevents correctly classified as nonevents. Subtracting this into a whole, or 100%, will make this the logical negative. The area between the purple linear line and the red curved line is approximately the value of the c statistic for the model.

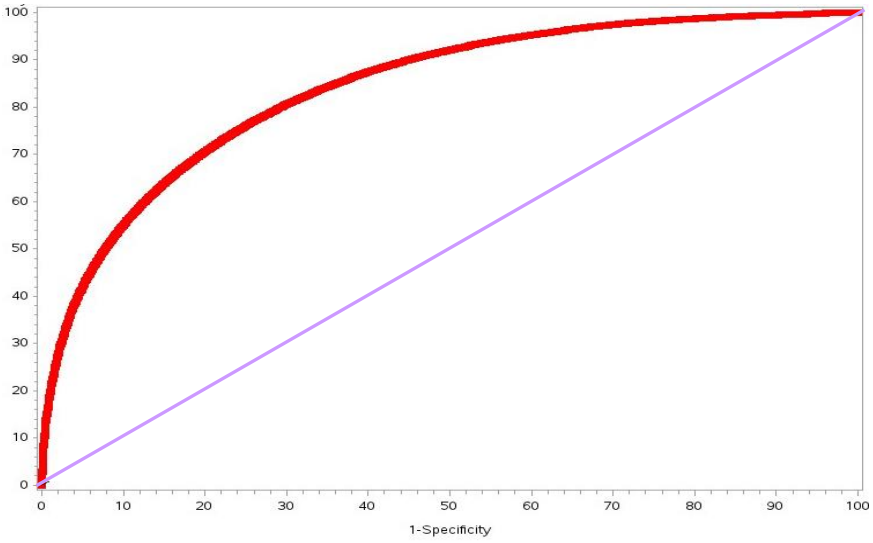
Notably, the twenty-parameter model and the full model have nearly identical plots. This is to be expected because the full model produced a c statistic of 84.7% while the twenty-parameter model produced a c statistic of 84%. The difference between these two models' c statistics is sufficiently small enough to not cause a visual difference between the two models' ROC curves.

From the ROC curve for the three-parameter model, one can see that this curve is slightly skinner than the other two curves. This is to be expected since the c statistic related to this model is at least 4% less than the c statistics for the other two models.

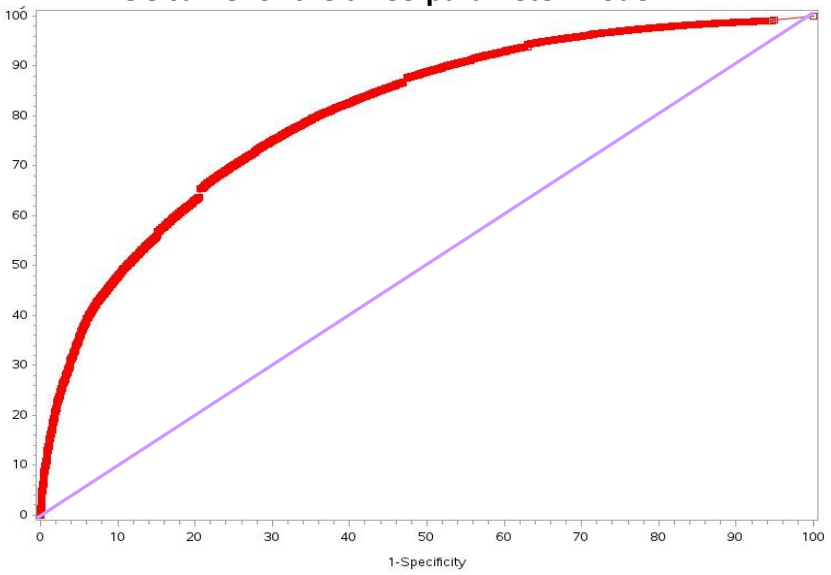
ROC curve for the full model



ROC curve for the twenty-parameter model



ROC curve for the three-parameter model



The next metric used to assess the performance of the models is the Kolmogorov–Smirnov (KS) statistic. To compute this value, the dataset is split into deciles, or ten bins of equal frequencies. The cutoff points for each of the ten bins is determined by the score, which is the probability of default multiplied by one thousand. Within each bin, the percent of customers that are labelled good and bad are calculated alongside the cumulative percentage of good and bad customers across the decile. The KS statistic is the maximum difference between the cumulative percent of good and bad customers. The larger this statistic, the greater the difference between the cumulative good and the cumulative bad customers, which indicates a better model. Table 24 below displays the KS calculations for the full model. One can see that the maximum difference between deciles is located in the 3rd decile with a value of 52.1%. This means that the full model identifies 20.84% of all customers that defaulted while identifying 72.95% of the customers that did not default. The lift, which is the cumulative percentage of customers in default divided by the cumulative percent of customers in the decile, is also calculated. This value indicates how much more likely the model is to predict a default. Figure 13 visually displays the KS statistic. One can see that the maximum difference between the dotted and solid lines occurs between the 3rd and 4th deciles.

Table 24: Kolmogorov-Smirnov calculations for the full model.

Decile	Min	Max	n	number goods	goods %	cumulative goods %	number bad	bad %	cumulative bad %	ks stat	lift
1	469.24	998.93	42684	13423	3.82%	3.82%	29261	39.01%	39.01%	35.20%	2.62
2	275.90	469.23	42685	27152	7.72%	11.53%	15533	20.71%	59.72%	48.19%	1.73
3	188.40	275.90	42684	32764	9.31%	20.84%	9920	13.23%	72.95%	52.10%	1.44
4	134.69	188.40	42685	35704	10.15%	30.99%	6981	9.31%	82.26%	51.26%	1.29
5	96.08	134.69	42685	37705	10.72%	41.71%	4980	6.64%	88.90%	47.19%	1.20
6	66.50	96.08	42684	39179	11.14%	52.84%	3505	4.67%	93.57%	40.72%	1.14
7	44.55	66.50	42685	40433	11.49%	64.34%	2252	3.00%	96.57%	32.23%	1.09
8	28.31	44.55	42684	41254	11.73%	76.06%	1430	1.91%	98.48%	22.42%	1.05
9	16.57	28.31	42685	41915	11.91%	87.97%	770	1.03%	99.50%	11.53%	1.02
10	1.31	16.57	42684	42312	12.03%	100.00%	372	0.50%	100.00%	0.00%	1

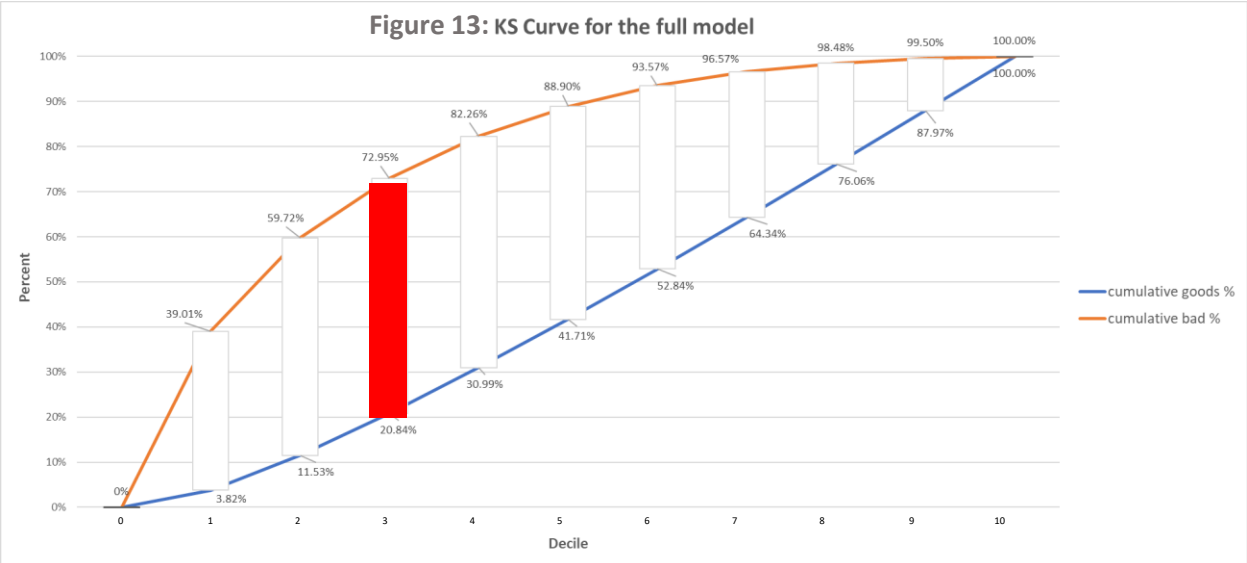


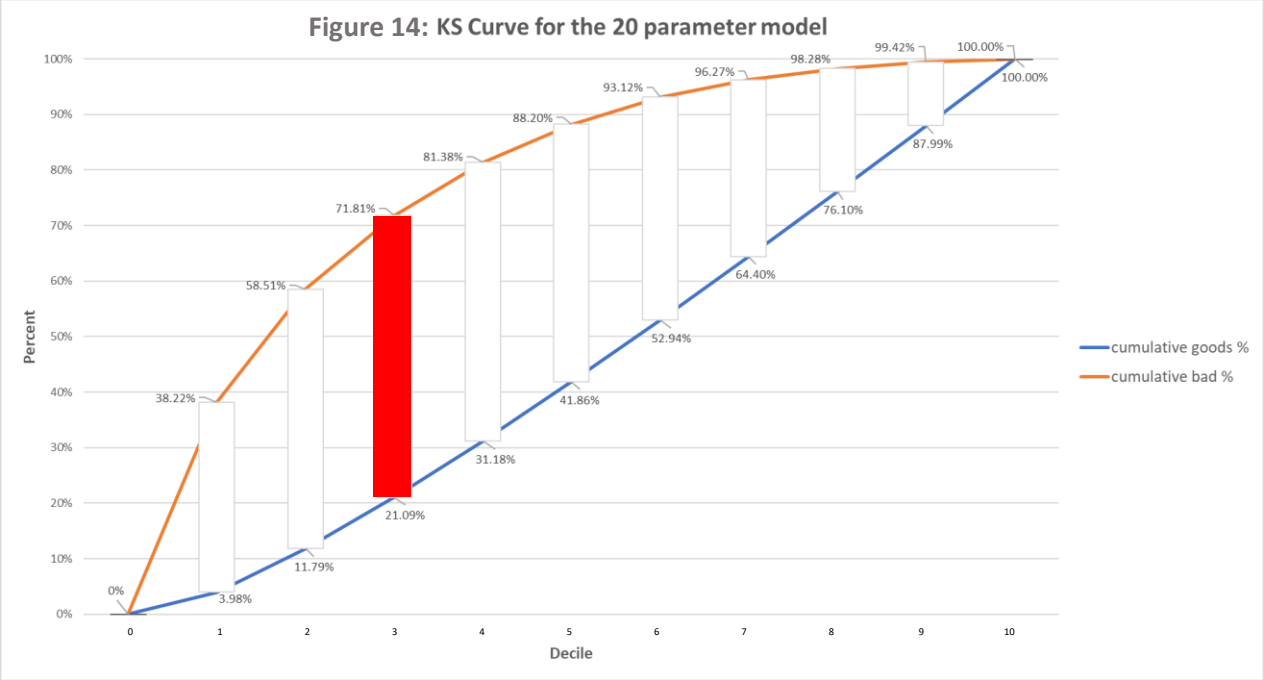
Table 25 displays the KS calculations for the model containing the top 20 variables with the highest aggregated mean chi-square statistic, while Table 26 displays the KS calculations for the model containing three variables. Figures 14 and 15 on the next page display the twenty-parameter and three-parameter models' KS curves, respectively. Regardless of the chosen model, the 3rd decile was found to have the largest difference between cumulative good and cumulative bad. Notably, the full model produced the largest KS statistic (52.1%) of the three models, but the model containing 20 variables produced a KS statistic (50.72%) that was not significantly lower than 52.1%. The model containing 3 variables had the lowest KS statistic of the three models with a value of 44.92%. Thus far, the twenty-parameter model gives the best metrics for the least variables of the three selected models. This was decided on the basis that the number of variables is significantly less than the full model (20 vs 130) while maintaining a c statistic of 84% and a KS statistic of 50.72%.

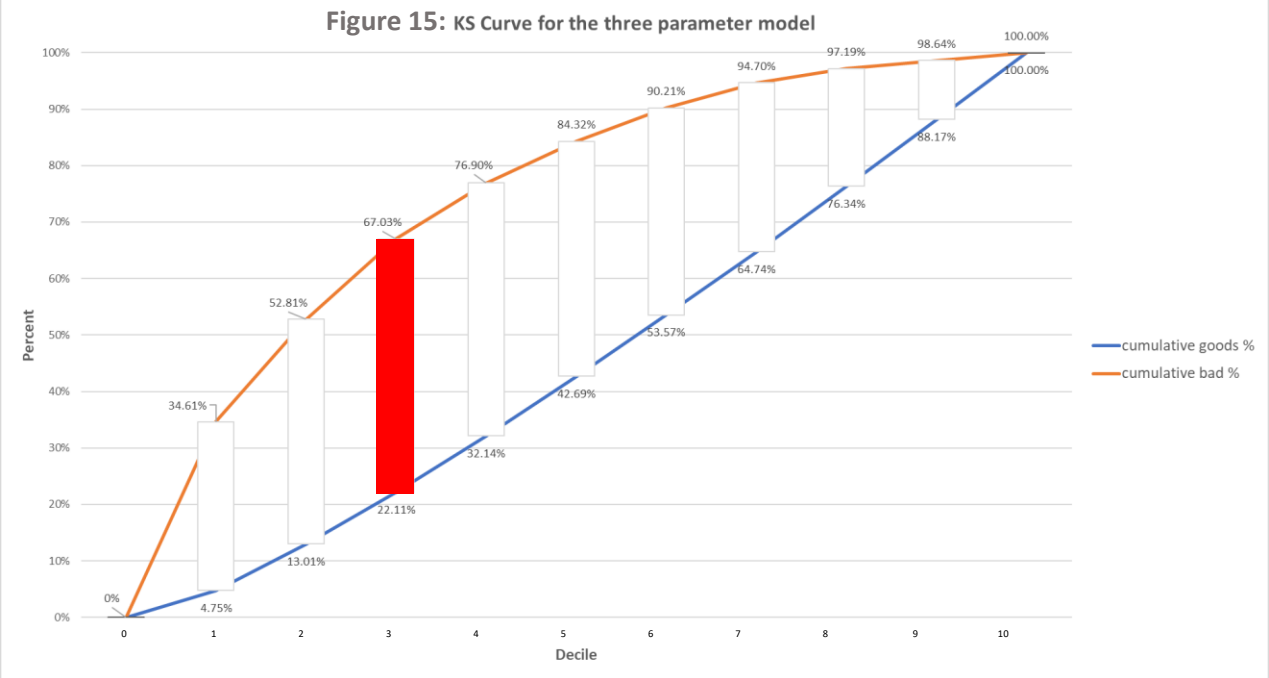
Table 25: Kolmogorov-Smirnov calculations for the twenty parameter model.

Decile	Min	Max	n	number goods	goods %	cumulative goods %	number bad	bad %	cumulative bad %	ks stat	lift
1	461.21	997.20	42684	14016	3.98%	3.98%	28668	38.22%	38.22%	34.24%	2.51
2	274.19	461.19	42685	27466	7.81%	11.79%	15219	20.29%	58.51%	46.72%	1.70
3	187.21	274.19	42684	32710	9.30%	21.09%	9974	13.30%	71.81%	50.72%	1.42
4	135.11	187.21	42685	35504	10.09%	31.18%	7181	9.57%	81.38%	50.21%	1.28
5	98.13	135.11	42685	37572	10.68%	41.86%	5113	6.82%	88.20%	46.35%	1.19
6	69.82	98.13	42684	38996	11.08%	52.94%	3688	4.92%	93.12%	40.18%	1.13
7	48.41	69.82	42685	40325	11.46%	64.40%	2360	3.15%	96.27%	31.86%	1.09
8	32.60	48.41	42684	41176	11.70%	76.10%	1508	2.01%	98.28%	22.17%	1.05
9	20.53	32.60	42685	41829	11.89%	87.99%	856	1.14%	99.42%	22.17%	1.02
10	1.33	20.53	42684	42247	12.01%	100.00%	437	0.58%	100.00%	22.17%	1

Table 26: Kolmogorov-Smirnov calculations for the three parameter model.

Decile	Min	Max	n	number goods	goods %	cumulative goods %	number bad	bad %	cumulative bad %	ks stat	lift
1	374.88	997.52	42685	16726	4.75%	4.75%	25959	34.61%	34.61%	29.86%	2.10
2	260.13	374.86	42699	29046	8.26%	13.01%	13653	18.20%	52.81%	39.80%	1.54
3	201.02	260.11	42681	32018	9.10%	22.11%	10663	14.22%	67.03%	44.92%	1.36
4	152.69	201.01	42680	35276	10.03%	32.14%	7404	9.87%	76.90%	44.77%	1.24
5	114.98	152.67	42699	37137	10.56%	42.69%	5562	7.42%	84.32%	41.63%	1.17
6	87.04	114.97	42684	38265	10.88%	53.57%	4419	5.89%	90.21%	36.64%	1.12
7	65.18	87.02	42696	39327	11.18%	64.74%	3369	4.49%	94.70%	29.96%	1.08
8	46.84	65.16	42670	40805	11.60%	76.34%	1865	2.49%	97.19%	20.85%	1.05
9	34.43	46.83	42706	41613	11.83%	88.17%	1093	1.46%	98.64%	10.48%	1.02
10	30.59	34.42	42645	41628	11.83%	100.00%	1017	1.36%	100.00%	0.00%	1



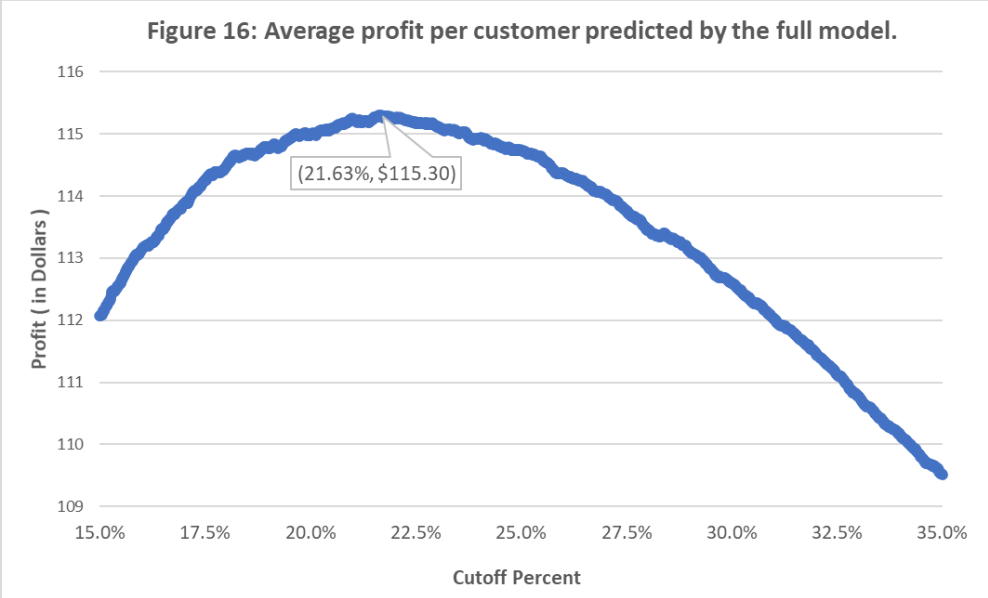


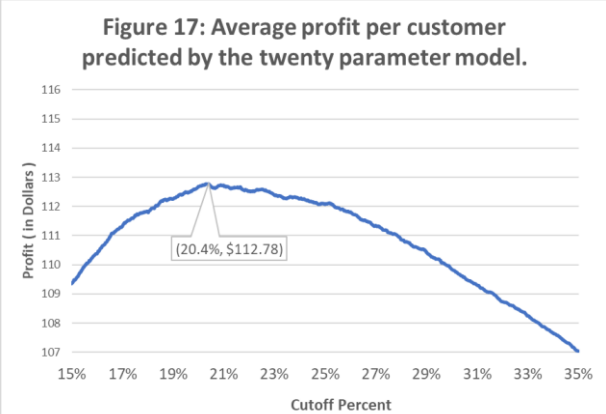
Once the model evaluation is complete, it will be scored against one validation dataset to ensure model performance. The full, the twenty-parameter, and the three-parameter models were all built on the training dataset. A simple random sample split the dataset into two partitions: one partition comprised of 66% of the data and the other partition of the remaining 34%. This sample was taken by generating a random number from the uniform distribution on the range from 0 to 1 and taking numbers that were less than 0.66.

To maximize profitability, a cutoff point for the predicted probability of default will be chosen such that if the predicted probability of default for a customer is greater than the chosen value, then this customer will not receive credit. This will result in four possible outcomes which contribute to the profits of the company. A customer that did not default but has a predicted probability of default that is less than the chosen cutoff point will be predicted to default and would not receive credit. This type of error is labeled as Type II Error. A customer that does default and is predicted by the model to default would not receive credit either. These two outcomes correspond to $0 in profit or loss that can be made from these customers.

The next outcome is labeled as Type I Error and happens as a result of a customer that was predicted to not default by the model and did default. This is a major error to account for since these customers result in a mean loss of half of their credit limit. The last outcome to account for comes from customers that do not default on their loans and are predicted to not default by the model. This outcome is considered a correct decision and the mean profit the company can make from these customers is $250.

These four outcomes outline why the cutoff point must be chosen with caution. Maximizing profitability without minimizing Type I Error will either lead to a loss in opportunity for more profit, or worse: a loss in actual profit. If the cutoff point chosen is too small, then only customers with a predicted probability of default less than the cutoff point will receive credit; this strategy is not optimal and may lead to a loss of opportunity. If the cutoff point chosen is too large, then some of the customers will receive credit when they should not have, resulting in a decrease in profits for the company. Figure 16 displays the effect on the profit per customer when choosing this cutoff point for the full model. Visually, one can see a peak between the values of 20% and 22.5%. After analyzing this range further, the value of 21.63% was chosen to be the cutoff point for the full model. This means any customer with a predicted probability of default greater than 21.63% will not be issued credit. From the results of the full model with the chosen cutoff point, the mean profit per customer is $115.30 while the profit per 1,000 customers is expected to be $115,303.14.





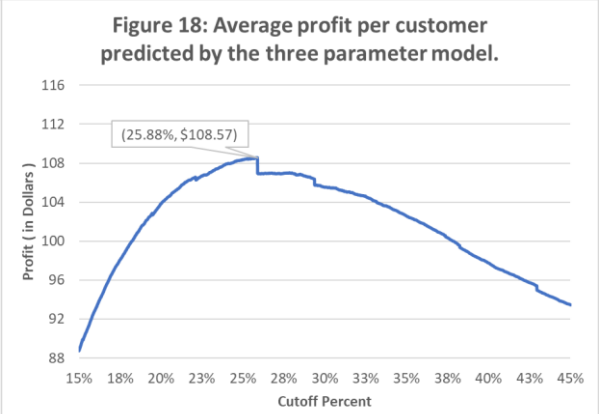


Figure 17 and 18 displays the change in the profit per customer at different cutoff percentages for the twenty-parameter model (on the left), and the three-parameter model on the right. For the twenty-parameter model, choosing a cutoff point of 20.4% results in an expected profit per customer of $112.78, and for 1000 customers the expected profit is $112,777.90. For the three-parameter model, the value of 25.88% was found to return the highest expected profit of $108.57 per customer, and for 1000 customers the expected profit is $108,573.01.

Before comparing the expected profit per customer, each of the models were scored by the validation dataset and compared to see how the models perform in validation. Using the chosen cutoff point for each model, a 'hit' table is created to see the percentages of correct and incorrect predictions made in scoring the validation dataset. Table 27 displays the hit chart for the full model over ten cutoff points chosen by the expected profit per customer. Valid I and Valid II are the two correct decisions discussed earlier in this report. Valid I represents the customers that did default and were correctly predicted to default, while Valid II represents the customers that did not default and were correctly predicted to not default. Since the Valid I customers are correctly predicted to default, these customers will not be given credit and thus the company will receive a net loss of $0. The Valid II customers were correctly predicted to not default, resulting in a mean profit of $250 per Valid II customer. The percent of Type I Error customers for the full model is 5.55% while the percent of Valid I customers is 12.02%. This suggests that the full model is 2.17 times more likely to predict a customer that did default correctly than incorrectly.

Table 27: Hit/Miss table for the full models' top 10 most profitable cutoff points.

Predicted probability	(Valid II) Actual 0 predicted as 0	(Type I Error) Actual 1 predicted as 0	(Type II Error) Actual 0 predicted as 1	(Valid I) Actual 1 predicted as 1	Profit per customer	Profit per 1000 customers
21.63%	68.31%	5.55%	14.12%	12.02%	$ 115.303	$ 115,303.14
21.69%	68.37%	5.57%	14.06%	12.00%	$ 115.299	$ 115,299.03
21.68%	68.36%	5.56%	14.07%	12.01%	$ 115.297	$ 115,297.27
21.62%	68.30%	5.55%	14.13%	12.02%	$ 115.296	$ 115,296.10
21.70%	68.38%	5.57%	14.05%	12.00%	$ 115.293	$ 115,292.50
21.65%	68.33%	5.56%	14.10%	12.02%	$ 115.289	$ 115,289.48
21.64%	68.32%	5.55%	14.11%	12.02%	$ 115.289	$ 115,288.72
21.61%	68.29%	5.54%	14.14%	12.03%	$ 115.288	$ 115,288.49
21.75%	68.42%	5.58%	14.01%	11.99%	$ 115.286	$ 115,285.65
21.80%	68.47%	5.60%	13.96%	11.97%	$ 115.285	$ 115,285.02

The customers that are labelled Type I and Type II Error are the customers that are incorrectly predicted by the model. Customers that receive a label of Type I Error were predicted to not default by the model but actually did default. These customers will result in a mean profit loss of half of the customers' credit line. For this reason, it is important to minimize this value while maximizing the value corresponding to the label Valid II. The last value is labelled Type II Error. These customers did not default but were predicted to default. Since the Type II Error customers are predicted to default, these customers will not be given credit. Although the company will potentially lose an opportunity to make more profit from these customers, denying them credit will ensure minimal profit loss.

As stated before, minimizing the number of customers labelled Type I Error while maximizing the number labelled Valid II will maintain minimal losses. By adding the values of Valid I and Type II Error from the first row, one can see that close to 26.12% of customers will not be given credit since these customers are predicted to default. This value is higher than the true default rate, found in Table 1, of 17.57%. This is to be expected since the procedure performed in this report took a conservative approach when assigning the maximum value of DELQID to each customer. It is foreseeable that the Type II Error would increase after making this decision. On the other hand, 73.68% (Valid II + Type I Error: 68.31% + 5.55%) of customers were predicted to not default by the full model and would be issued credit. These customers will fall into one of two groups: either the customer will default, or they will not default. To maintain the mean profit per 1000 customer of $115,303.14, the full model must correctly predict 12,308 customers that do not default for every 1000 customer that the model inaccurately predicted to not default. This is because customers that default result in a profit loss that is on average half of their credit limit, while customers who do not default will result in profiting on average $250.

Similarly, in Table 28 a hit table is displayed for the model containing the twenty parameters. One can see that the percentage of customers that will be denied credit by this model is 27.55% while the percentage of customers that will be granted credit is 72.45%. This indicates that, to maintain the expected mean profit per 1000 customers of $112,777.90, the twenty-parameter model must correctly predict 12,318 customers to not default for every 1000 customers that are incorrectly predicted to not default. A customer that did default is 2.23 times more likely to be predicted to default by this model than not default. This model produced the highest odds in correctly predicting customers that default of the three selected models.

Table 28: Hit/Miss table for the twenty-parameter models' top 10 most profitable cutoff points.

Predicted probability	(Valid II) Actual 0 predicted as 0	(Type I Error) Actual 1 predicted as 0	(Type II Error) Actual 0 predicted as 1	(Valid I) Actual 1 predicted as 1	Profit per customer	Profit per 1000 customers
20.40%	67.01%	5.44%	15.42%	12.13%	$ 112.778	$ 112,777.90
20.32%	66.93%	5.42%	15.50%	12.15%	$ 112.775	$ 112,775.29
20.29%	66.89%	5.41%	15.53%	12.16%	$ 112.775	$ 112,775.10
20.41%	67.02%	5.45%	15.41%	12.13%	$ 112.771	$ 112,771.35
20.30%	66.91%	5.42%	15.52%	12.16%	$ 112.771	$ 112,770.77
20.31%	66.92%	5.42%	15.51%	12.15%	$ 112.768	$ 112,767.90
20.33%	66.94%	5.42%	15.49%	12.15%	$ 112.768	$ 112,767.60
20.34%	66.95%	5.43%	15.48%	12.14%	$ 112.767	$ 112,767.13
20.37%	66.98%	5.44%	15.45%	12.14%	$ 112.767	$ 112,766.92
20.28%	66.88%	5.41%	15.55%	12.16%	$ 112.766	$ 112,765.73

In contrast to the full model, the twenty-parameter model must correctly predict fewer customers to balance the incorrect prediction. Furthermore, the full model had a higher percentage of customers predicted to not default. Although only 67.01% of these predictions were correctly made by the twenty-parameter model, the full model had a higher percentage of Type I Error (5.55% vs 5.44%). In a later phase of this report, we will analyze the difference in the expected profit of these two models and compare them to the expected profit of the three-parameter model.

Table 29: Hit/Miss table for the three-parameter models' top 10 most profitable cutoff points.

Predicted probability	(Valid II) Actual 0 predicted as 0	(Type I Error) Actual 1 predicted as 0	(Type II Error) Actual 0 predicted as 1	(Valid I) Actual 1 predicted as 1	Profit per customer	Profit per 1000 customers
25.88%	65.31%	6.08%	17.12%	11.49%	$ 108.573	$ 108,573.01
25.89%	65.32%	6.08%	17.11%	11.49%	$ 108.567	$ 108,567.03
25.86%	65.27%	6.07%	17.15%	11.50%	$ 108.567	$ 108,566.93
25.84%	65.25%	6.07%	17.18%	11.51%	$ 108.563	$ 108,563.30
25.80%	65.19%	6.05%	17.24%	11.52%	$ 108.561	$ 108,561.14
25.83%	65.24%	6.06%	17.19%	11.51%	$ 108.559	$ 108,558.99
25.85%	65.26%	6.07%	17.17%	11.50%	$ 108.558	$ 108,558.43
25.87%	65.29%	6.08%	17.14%	11.50%	$ 108.548	$ 108,548.32
25.81%	65.21%	6.06%	17.22%	11.51%	$ 108.547	$ 108,547.25
25.79%	65.18%	6.05%	17.25%	11.52%	$ 108.547	$ 108,547.25

The three-parameter model had the highest Type I Error, the highest cutoff point for the predicted probability, and the lowest profit per customer. Table 29 above displays the hit table for the ten cutoff points with the highest profit per customer for the model with three-parameters. A customer with a predictive probability of default greater than 25.88% will be denied credit by this model. This model will deny 28.61% of customers credit while granting credit to 71.39% of customers and is expected to profit $108.58 per customer.

For this model to receive a mean profit of $108,573.01 per 1000 customers, it would need to correctly predict 10,742 customers to not default for every 1000 customers that are incorrectly predicted to not default. This model produced the lowest percentage of Valid I customers with a value of 11.49%.

Notably, this model incorrectly predicted 6.08% of customers that went into default. In comparison, the full model incorrectly predicted 5.55% of customers that went into default, and the twenty-parameter incorrectly predicted 5.44%. The Type I Error produced by these two models had a difference of 0.11%. The difference between the Type I Error of the three-parameter model and the full model was 0.53%. This difference is nearly five times larger than the difference between the other two models. In fact, a customer that did default is 1.89 times more likely to be predicted to default than to not default.

Table 30: Profit gained or loss at the top 10 most profitable cutoff points for the full models'.

Predicted probability	Total profit loss due to Type I Error	Profit loss per 1000 customers due to Type I Error	Total profit gain due to Valid II	Profit gain per 1000 customers due to Valid II	Profit per customer	Profit per 1000 customers	Total Profit	n
21.63%	-23677181	-999501	$ 72,893,750	$ 250,000	$ 115.303	$ 115,303.14	$ 49,216,569	426845
21.69%	-23742685	-999061	$ 72,957,500	$ 250,000	$ 115.299	$ 115,299.03	$ 49,214,815	426845
21.68%	-23732187	-999250	$ 72,946,250	$ 250,000	$ 115.297	$ 115,297.27	$ 49,214,063	426845
21.62%	-23668937	-999617	$ 72,882,500	$ 250,000	$ 115.296	$ 115,296.10	$ 49,213,564	426845
21.70%	-23751974	-998905	$ 72,964,000	$ 250,000	$ 115.293	$ 115,292.50	$ 49,212,027	426845
21.65%	-23702012	-999410	$ 72,912,750	$ 250,000	$ 115.289	$ 115,289.48	$ 49,210,738	426845
21.64%	-23692087	-999582	$ 72,902,500	$ 250,000	$ 115.289	$ 115,288.72	$ 49,210,413	426845
21.61%	-23661187	-999712	$ 72,871,500	$ 250,000	$ 115.288	$ 115,288.49	$ 49,210,314	426845
21.75%	-23805145	-998580	$ 73,014,250	$ 250,000	$ 115.286	$ 115,285.65	$ 49,209,105	426845
21.80%	-23857668	-998312	$ 73,066,500	$ 250,000	$ 115.285	$ 115,285.02	$ 49,208,833	426845

Next, this report will analyze the profit loss and gain that can be expected by each of the models. First, in Table 30, the ten cutoff points that produce the highest profit per customer are displayed for the full model. The chosen cutoff point for this model was 21.63%. The total profit for this model is expected to be $49,216,569 with a mean profit per customer of $115.30. This model expects that customers that are incorrectly predicted to not default will lose this company $23,677,181 of profit when they default. However, the expected total profit gain of correctly predicting customers that do not default will be $72,893,750. Of these three models, the full model produced the most profitable model, but explaining to a customer why the company may be denying them credit may prove difficult due to the sheer number of parameters (130), if this model was chosen as the best model.

Table 31: Profit gained or loss at the top 10 most profitable cutoff points for the twenty-parameter models'.

Predicted probability	Total profit loss due to Type I Error	Profit loss per 1000 customers due to Type I Error	Total profit gain due to Valid II	Profit gain per 1000 customers due to Valid II	Profit per customer	Profit per 1000 customers	Total Profit	n
20.40%	-23367067	-1005771	$ 71,505,750	$ 250,000	$ 112.778	$ 112,777.90	$ 48,138,684	426845
20.32%	-23281682	-1005992	$ 71,419,250	$ 250,000	$ 112.775	$ 112,775.29	$ 48,137,569	426845
20.29%	-23246762	-1006179	$ 71,384,250	$ 250,000	$ 112.775	$ 112,775.10	$ 48,137,489	426845
20.41%	-23379865	-1005715	$ 71,515,750	$ 250,000	$ 112.771	$ 112,771.35	$ 48,135,885	426845
20.30%	-23260112	-1006191	$ 71,395,750	$ 250,000	$ 112.771	$ 112,770.77	$ 48,135,639	426845
20.31%	-23272587	-1005991	$ 71,407,000	$ 250,000	$ 112.768	$ 112,767.90	$ 48,134,414	426845
20.33%	-23294215	-1006056	$ 71,428,500	$ 250,000	$ 112.768	$ 112,767.60	$ 48,134,286	426845
20.34%	-23304165	-1006008	$ 71,438,250	$ 250,000	$ 112.767	$ 112,767.13	$ 48,134,086	426845
20.37%	-23337253	-1005743	$ 71,471,250	$ 250,000	$ 112.767	$ 112,766.92	$ 48,133,998	426845
20.28%	-23235512	-1006128	$ 71,369,000	$ 250,000	$ 112.766	$ 112,765.73	$ 48,133,489	426845

After reducing the model to twenty variables, as seen in Table 31, and choosing the cutoff point (20.40%) that produces the most profit, the expected total profit reduced to $48,138,684 with a mean profit per customer of $112.78. Customers that default and are incorrectly predicted to not default will cause the company to lose $23,367,067 while the customers that are correctly predicted to not default will earn the company $71,505,750. The comparison of the twenty-parameter model to the full model will be detailed in the next section of this report.

The model including only three-parameters at the chosen cutoff point of 25.88% is expected to produce the company a total profit of $46,343,846 with a mean profit per customer of $108.57 (Table 32 below). A customer that is incorrectly predicted to not default by this model will cost the company a total loss of $23,345,404 with a mean profit loss per customer of $899.46. Customers that are correctly predicted to not default by this model will earn the company a total of $69,689,250. Although this model is expected to earn the company the least amount of money, in comparison to the other two models in this report, this model may produce the simplest reason as to why a customer was denied credit.

Table 32: Profit gained or loss at the top 10 most profitable cutoff points for the three-parameter models'.

Predicted probability	Total profit loss due to Type I Error	Profit loss per 1000 customers due to Type I Error	Total profit gain due to Valid II	Profit gain per 1000 customers due to Valid II	Profit per customer	Profit per 1000 customers	Total Profit	n
25.88%	-23345404	-899457	$ 69,689,250	$ 250,000	$ 108.573	$ 108,573.01	$ 46,343,846	426845
25.89%	-23359704	-899592	$ 69,701,000	$ 250,000	$ 108.567	$ 108,567.03	$ 46,341,296	426845
25.86%	-23313750	-899450	$ 69,655,000	$ 250,000	$ 108.567	$ 108,566.93	$ 46,341,250	426845
25.84%	-23288550	-899345	$ 69,628,250	$ 250,000	$ 108.563	$ 108,563.30	$ 46,339,700	426845
25.80%	-23229722	-899018	$ 69,568,500	$ 250,000	$ 108.561	$ 108,561.14	$ 46,338,778	426845
25.83%	-23278636	-899275	$ 69,616,500	$ 250,000	$ 108.559	$ 108,558.99	$ 46,337,864	426845
25.85%	-23302125	-899453	$ 69,639,750	$ 250,000	$ 108.558	$ 108,558.43	$ 46,337,625	426845
25.87%	-23334194	-899649	$ 69,667,500	$ 250,000	$ 108.548	$ 108,548.32	$ 46,333,307	426845
25.81%	-23248897	-899168	$ 69,581,750	$ 250,000	$ 108.547	$ 108,547.25	$ 46,332,853	426845
25.79%	-23222397	-898978	$ 69,555,250	$ 250,000	$ 108.547	$ 108,547.25	$ 46,332,853	426845

Cost of Simplicity:

This phase of the report started with 270 variables. The initial stepwise logistic regression procedure resulted in a model containing 130 significant parameters. The c statistic for this model was observed at 84.7% concordance. According to this model, the company can expect to receive a total profit of $49,216,569. After performing a series of stepwise logistic regression procedures that reduced the initial model to three parameters, three models were chosen: the full model (containing the results of the initial stepwise logistic model), the three-parameter model (that was obtained from reducing the initial model by the parameter with the lowest chi-square statistic until three variables remained), and the twenty-parameter model (which was created from the mean chi-square statistic aggregated from the output data of the chi-square test on the different models created when reducing the initial model to three parameters).

The twenty-parameter model contains significantly fewer variables with only a small difference in the percent of concordant pairs (FULL: 84.7% vs TWENTY: 84.0%), a small difference in the KS statistic (52.1% vs 50.72%), and a small difference in the profit per customer ($115.30 vs $112.78). In addition, this model had the highest odds of correctly predicting customers that default. Using this model will result in a small decrease (-$1,077,885.5) in the profits this company can make compared to the profits produced by using the full model. The three-parameter model is the simplest of the three models. It was found that this model has the lowest odds (1.89x) of correctly predicting customers that default, a significant reduction in the percent of concordant pairs (TWENTY: 84.0% vs THREE: 80.3%), and the highest percent of customers that default when the model predicted they would not default. Using this model will decrease the amount of profit the company will receive by $2,872,723 as opposed to if the full model was chosen instead. With this information, one can see the cost of simplicity is a million-dollar venture.

Table 33: Percent Concordant for the three models.			
	Three-parameter model	Twenty-parameter model	Full model
Percent Concordant	80.3%	84.0%	84.70%
Percent Discordant	19.6%	16.0%	15.30%
Percent Tied	0.1%	0.0%	0.00%
Pairs	99440320848.0	99440320848	99440320848

Table 34: Hit Chart for the three models found in this report.								
MODEL	Predicted probability	Total Profit	(Type I Error) Actual 1 predicted as 0	(Type II Error) Actual 0 predicted as 1	(Valid I) Actual 1 predicted as 1	(Valid II) Actual 0 predicted as 0	Profit loss per 1000 customers due to Type I Error	Profit per 1000 customers
FULL	21.63%	49216569.0	5.55%	14.12%	12.02%	68.31%	-999501.08	115303.14
BEST	20.40%	48138683.5	5.44%	15.42%	12.13%	67.01%	-1005770.52	112777.90
LEAST	25.88%	46343846.0	6.08%	17.12%	11.49%	65.31%	-899456.91	108573.01

Table 35: Expected total profit gained or lossed by the three models					
MODEL	Predicted probability	Total Profit	Total profit loss due to Type I Error	Total profit gain due to Valid II	KS
FULL	21.63%	49216569.0	-23677181.0	72893750.00	52.10%
BEST	20.40%	48138683.5	-23367066.5	71505750.00	50.72%
LEAST	25.88%	46343846.0	-23345404.00	69689250.00	44.92%

Transaction Analysis:

Table 36: Descriptive statistics for SIC Code 505: Metals and Minerals, except Petroleum						
SIC	Name of field	number	Mean	Std Dev	Minimum	Maximum
505	Metals and Minerals, except Petroleum	610494	-145.0445488	362.9260045	-13000	4800

While looking at some of the transaction file's SIC codes, the 505 transaction type was found to be interesting. In Table 36, we can see that a company had an average loss of -$145.04 with customers in Metals and Minerals, except Petroleum, with a maximum loss of -$13000 and a maximum profit of only $4800. This does not appear to be a profitable field for creditors, but even so the number of observations (610,494) indicates a large number of credit lines awarded to these types of businesses. These customers clearly had large credit limits, as the maximum amount lost for the company presents, and the amounts seem to vary wildly according to the standard deviation of $362.93. This type of transaction appears to be particularly risky and yet also sought-after.

Table 37: Descriptive Statistics for SIC Code 8111: Legal Services						
SIC	Name of field	number	Mean	Std Dev	Minimum	Maximum
8111	Legal Services	2502	445.6401519	600.5874593	-2457.69	5500

Additionally, SIC code 8111, corresponding to Legal Services, also displayed an interesting pattern. Far fewer observations involved SIC 8111, with only 2502 in the data. The average profit for the company on these customers was $445.64, with a maximum profit of $5500. The maximum loss for Legal Services was -$2457.69, much lower than that for SIC 505, indicating that these customers are far less risky. In fact, these customers produce a far higher average and maximum profit, and only deviate by $600.59. This average profit per customer is far above the average profit found for each of the three models in this report, indicating that these customers, while fewer than other transaction types, are worthy investments.

Cluster Analysis:

After analyzing the three models found during this report, cluster analysis was performed on the output of the full model. Due to the size of the dataset, the decision was made to utilize the fastclus procedure in SAS with the max cluster value ranging from 3 to 9. It was found that a max cluster of 4 produced the highest total profit. Table 38 displays the output of the fastclus procedure with the max cluster equal to four. At a cutoff point of 21.63% this model will deny 1.55% of customers credit while granting credit to 99.11% of customers and is expected to profit $206.33 per customer. For this model to receive a mean profit of $206,327.90 per 1000 customers, it would need to correctly predict 28 customers to not default for every 1 customer that is incorrectly predicted to not default.

Table 38: Profit statistics for the output of the fastclus procedure with max cluster = 4 for the full model							
MODEL	Predicted probability	Total Profit	Total profit loss due to Type I Error	Total profit gain due to Valid II	Profit loss per 1000 customers due to Type I Error	Profit per 1000 customers	Profit per customer
Cluster = 4	21.63%	$ 23,856,663	($3,815,587)	$27,672,250	($976,103.10)	$206,327.90	206.33

Table 39: Hit table for the output of the fast clus procedure with max cluster = 4 for the full model					
MODEL	Predicted probability	(Type I Error) Actual 1 predicted as 0	(Type II Error) Actual 0 predicted as 1	(Valid I) Actual 1 predicted as 1	(Valid II) Actual 0 predicted as 0
Cluster = 4	21.63%	3.38%	0.66%	0.89%	95.73%

Conclusion:

The first model that was created contained only significant parameters from a stepwise selection process. The percent of concordant pairs this model produced was 84.7%. An iterative method was performed to reduce the initial model's parameters until three remained. After reducing the model by the parameter that had the least chi-square statistic, the remaining parameters were reevaluated by the logistic regression procedure and the output of the chi-square test was collected.

Once the model was reduced and the chi-square test output was collected, an aggregation of the chi-square statistic was made for each unique parameter. This data was then sorted by the mean chi-square statistic, and the top 20 variables were selected to produce a model. This model had a concordant percentage of 84%, which is less than a 1% reduction from the full model. Reducing the variables in such a manner is important for facilitating the explanation of why a customer was denied credit. The mean profit gain per customer in this model was $112.78 while the mean profit loss per customer due to Type I Error was $1005.77. It was found that this model predicted customers that did default 2.23 times more often than those that did not default. The twenty-variable model not only minimizes Type I Error and maximizes profit per customer, but it also avoids using a low cutoff point, which would deny credit to customers who would not default. This model helps to make the understanding of credit approval or disapproval easier, and can also help the company increase its profits while avoiding too much opportunity loss, thanks to the consideration of all four possible outcomes when predicting whether a customer will default or not.