Binary Classification Modeling Final Deliverable

Using Logistic Regression to Build Credit Scores

Nathaniel Thomas Jones

Supervised by Jennifer Lewis Priestley, Ph.D.

Kennesaw State University

Submitted 5/6/2022 to fulfill a requirement for STAT 8330

Executive Summary

The purpose of this report is to maximize the profitability of extending credit to subprime customers. The dataset that the company has provided contained 1,255,429 customers with 340 independent variables. Both SAS and R was used to generate the analysis and the graphics used in this project. A binary variable was created from the customers payment delinquency. After wrangling the data into a structure that is useful to prediction, the number of variables were reduced to find a subset of predictors that achieve the highest concordance and profit.

Table 23: The Percent of Concordant, Discordant, and Tied Pairs for the Best Model

Number of Variables in the Model	Percent of Concordant Pairs	Percent of Discordant Pairs	Percent of Tied Pairs	C	Profit per 1000 Customers
12	85%	15%	0%	0.850	$ 114,540.22

Table 28: The Percent of Concordant, Discordant, and Tied Pairs for the Chosen Best Least Model

Number of Variables in the Model	Percent of Concordant Pairs	Percent of Discordant Pairs	Percent of Tied Pairs	C	Profit per 1000 Customers
6	83.9%	16.1%	0.0%	0.839	$ 112,082.18

The best model overall was achieved a concordant percentage of 85% and acquired a $114,540 profit per 1,000 customers (Table 23). If the company desires a more parsimonious model, then the model using six variables would be the best choice. The six variable model achieved $112,082 profit per 1,000 customers and a concordant percentage of 83.9%. Using less variables will significantly impact the profitability of the model (Figure 18a and 18b).

Figure 18a:
Profitability by the Predicted Probabilities for the Least Variable Models

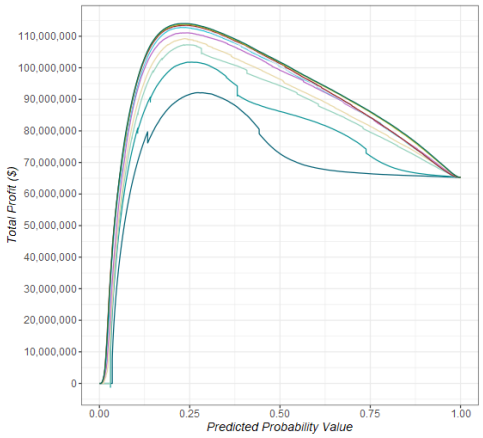


Figure 18b:
Profitability by the Predicted Probabilities for the Least Variable Models

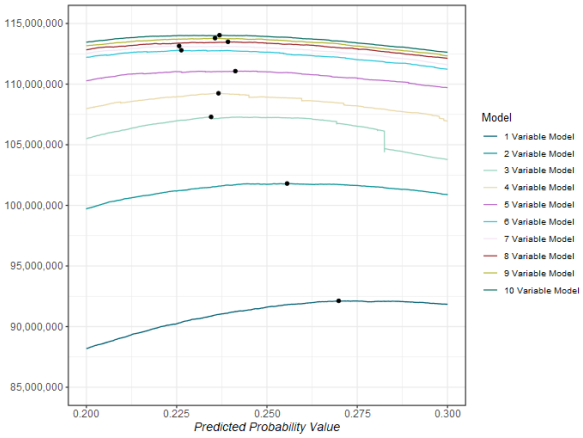


Table 22: Best Model Parameter Estimates, Standard Error, and Wald Chi-Square test

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	P-value
Intercept	1.1766	0.0253	2164.83	< 0.0001
LOCINQS	0.0557	0.0014	1650.75	< 0.0001
odds_BRR4524_EQWID	-0.1089	0.0021	2808.57	< 0.0001
OT12PTOT	0.9225	0.0140	4369.52	< 0.0001
CRATE45	0.3488	0.0049	5142.18	< 0.0001
BRAVGMOS	-0.0110	0.0002	5190.31	< 0.0001
logodds_BRADBM_EQFREQ	-0.3871	0.0052	5604.74	< 0.0001
OBRPTAT	-1.4799	0.0185	6405.97	< 0.0001
odds_TRATE1_EQFREQ	-0.2160	0.0026	7056.56	< 0.0001
BRCR1BAL	0.1413	0.0017	7152.07	< 0.0001
RADB6	1.4283	0.0145	9707.36	< 0.0001
BRPCTSAT	-1.5901	0.0158	10086.98	< 0.0001
BRR23	0.7033	0.0063	12395.01	< 0.0001

Table 28: Six Variable Model Parameter Estimates, Standard Error, and Wald Chi-Square test

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	P-value
Intercept	1.7187	0.0371	2144.41	< 0.0001
RADB6	1.6211	0.0138	13897.80	< 0.0001
BRR4524_ORD_EQFREQ	0.8497	0.0096	7792.59	< 0.0001
TRR23	0.4202	0.0038	12519.91	< 0.0001
logodds_BRAVGMOS_EQW	-2.3249	0.0162	20490.48	< 0.0001
BRPCTSAT	-1.7240	0.0139	15278.61	< 0.0001
logodds_BRADBM_EQWID	-0.4425	0.0049	8151.51	< 0.0001

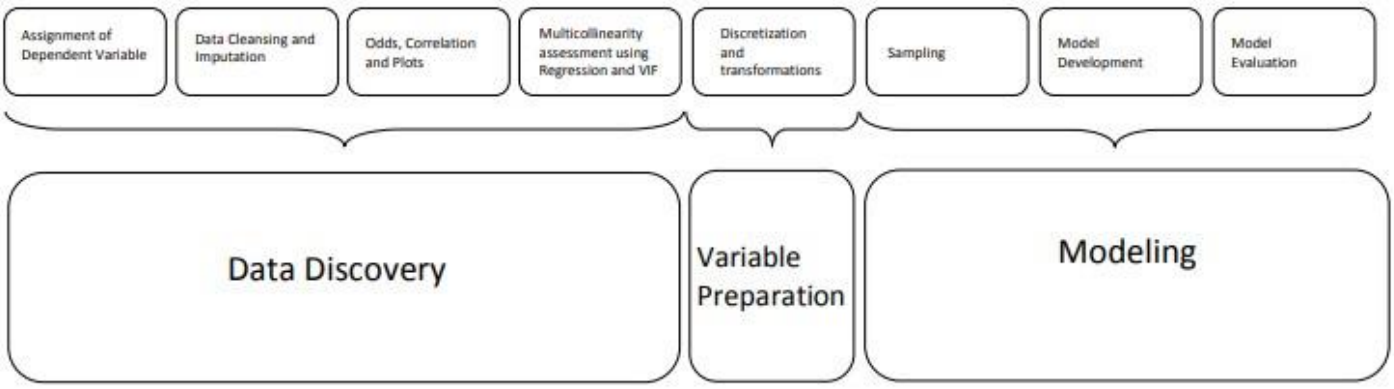
Introduction

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 (1,462,955 observations and 338 variables)
  - Each observation represents a unique customer. This file contains the potential predictors of credit performance.
- 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 PERF dataset contains the dependent variable "DELQID" that will be transformed into a binary indicator, "goodbad". The CPR file contains the independent variables that will be used to build a logistic regression model. The process for the project includes many steps before and after building a model. Each of these processes will be discussed in turn.



Data Discovery

The first step of this project is to create a target variable that will be used classify whether a customer is a ‘good’ or ‘bad’ credit risk. The PERF dataset contains monthly outcomes for customer credit lines. Since the data is recorded monthly, each customer may be observed multiple times. Each monthly outcome includes a MATCHKEY and CYCLEDT, distinguishing the associated customer and date of the observation. Additionally, the PERF dataset contains a variable, DELQID, representing the number of cycles a customer is late on payment. Then this variable is used to assign a binary variable, “goodbad”, indicating a customer’s credit risk. A value of 1 represents “bad” credit risks while 0 represents “good” credit risks.

The values observed for DELQID range from 0 up to 7 (Figure 1a), where a 0 indicates the customer is too new to rate, 1 indicates the customer is in good standing, 2 indicates the customer is one cycle late, 3 indicates the customer is two cycles late, and so on. The company has advised that customers observed 2 cycles late or later are credit risks, and so the variable goodbad will be created by assigning a 1 if the representative of DELQID is 3 or greater.

Figure 1a:
Bar Plot of Customers Deliquency ID Value
Conservative Approach

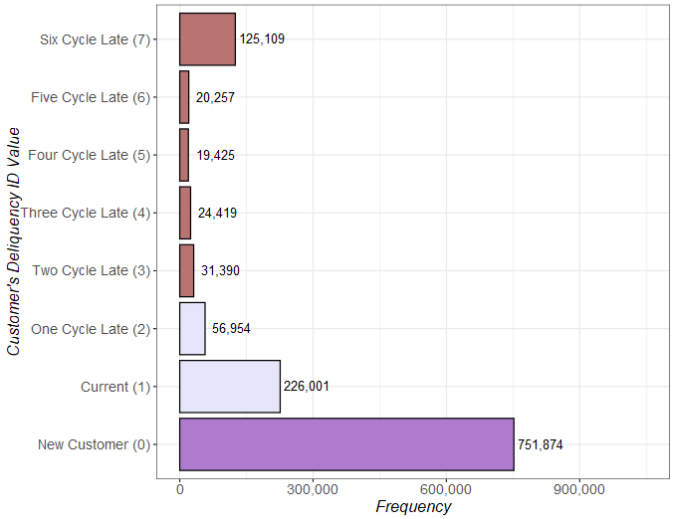
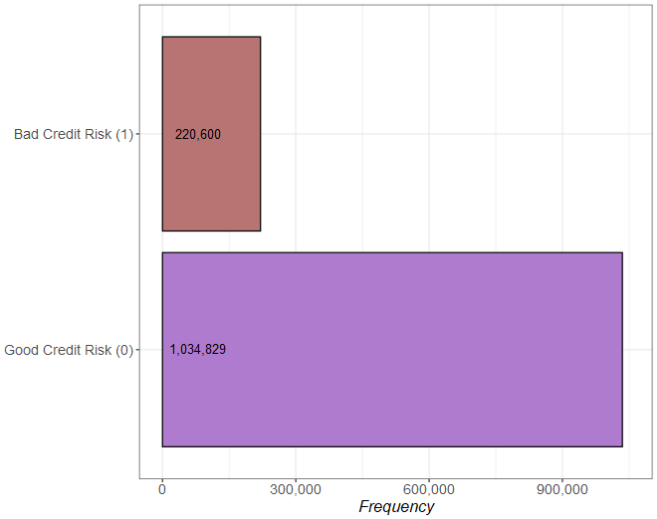


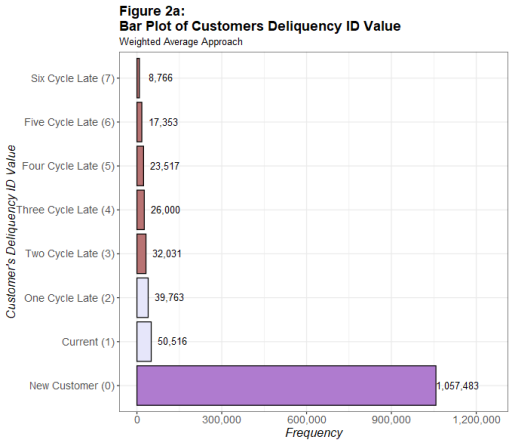
Figure 1b:
Bar Plot of the Indicator Variable: 'goodbad'
Conservative Approach

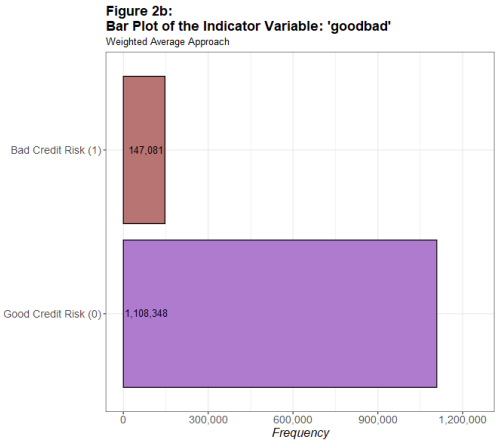


A conservative approach to creating the representative of DELQID is to assign the maximum value of DELQID to be the representative of each customer. For example, suppose a customer is recorded twelve times in this dataset and was observed with the following DELQID’s: [1,1,4,5,6,7,1,1,1,1,1,1]. By this approach, the maximum value 7 is assigned to represent the customer, and goodbad is then assigned a value of 1 (indicating a “bad” credit risk).

The result of assigning the highest DELQID to be the representative observation of each customer can be seen in Figure 1a. The highest DELQID for customers observed in any of the bottom three categories is less than 3 and will be assigned goodbad values of 0. These categories add up to be the 1,034,829 customers in the “good” credit risk bar in Figure 1b. In comparison, there are 220,600 customers that were assigned values of 1. These customers came from the top five categories in Figure 1a and were observed with DELQID values 3 or greater.

Another approach to creating goodbad is to compute a temporally weighted average for each customer, and assign goodbad a value of 1 to customers with weighted averages greater than 2. This may include customers that previously displayed poor payment history in goodbad's "good" credit risk category. Doing so can negatively impact a model's prediction accuracy. Figure 2a displays the frequency of customers' DELQID values rounded down to the nearest whole number. Like Figure 1a and 1b, the bottom three categories consist of customers with values less than 3 while the top five have values 3 or greater. Interestingly, this approach increased the number of New Customers (1,057,483 vs 751,874). Figure 2b displays 1,108,348 customers were assigned a value of 0 while 147,081 customers a value of 1.





Using a temporally weighted average as a customer's DELQID value increased (+73,519) the number of customers assigned a goodbad value of 0 ("good" credit risk). In the conservative approach, these customers are assigned values of 1 ("bad" credit risk). Assigning these customers values of 0 may introduce customers who are truly "bad" credit risks into the "good" credit risk category. Training a logistic regression model may then result in more false negative predictions where "bad" credit risk customers are given credit. This means the weighted average approach could potentially have more Type I error than the conservative approach, which can be costly to the company and result in a loss in profit.

On the other hand, the conservative approach could potentially have more Type II error. Some of the 73,519 "bad" customers may truly be "good" credit risks. Training a model with these customers in the "bad" credit risk category may result in more false positive predictions where "good" credit risk customers are not given credit. As indicated by the company, giving "bad" credit risk customers credit probabilistically results in losing half of the credit line given to the customer. In addition, not giving credit to a "good" credit risk customers may result in the loss of opportunity to earn more profit but would not result in any profit loss. Considering the nature of the risk the company is taking and the financial implications that can result, the conservative approach seems to be intuitively the better option in reducing the amount of risk.

The next step of this phase will merge the PERF dataset to the CPR. Like the PERF dataset, the CPR dataset contains multiple observations per customer. After sorting the CPR dataset by MATCHKEY, the last unique customer's MATCHKEY was used to join the two datasets together. The resulting dataset included a total of 340 independent variables and 1,743,505 observations. Some observations in the CPR and PERF dataset were not found in the other which resulted in either a missing value for goodbad or for all the CPR variables. Dropping these rows resulted in a dataset with 340 variables and 1,255,429 observations.

Data Wrangling

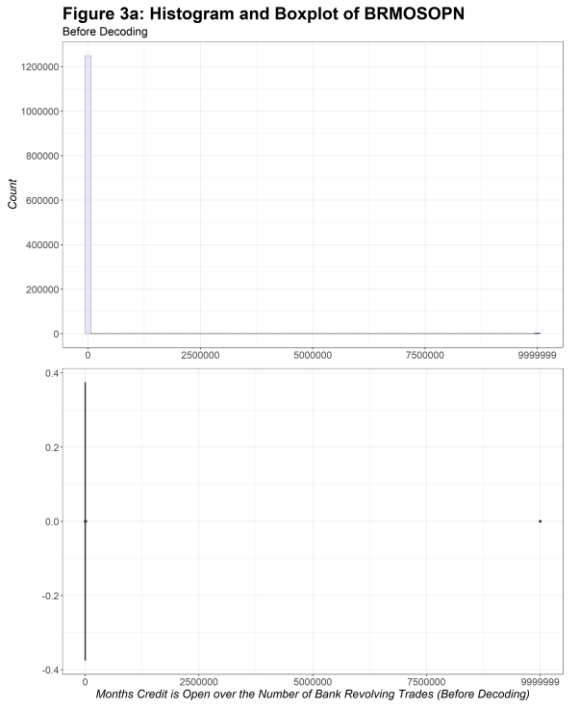
This phase of the project will wrangle the independent variables into a structure good for modelling. Due to the illegality associated with using a customer's age to predict their credit, the age variable will be dropped from the dataset. Additionally, the variable BEACON contains only missing values and will be dropped. The next step in wrangling the data is to handle the coded and extreme values. Within the dataset the last two digits ending in 93 or above are coded in one of several ways: invalid past due, invalid high credit, etc.. The following list details the five unique coded cases:

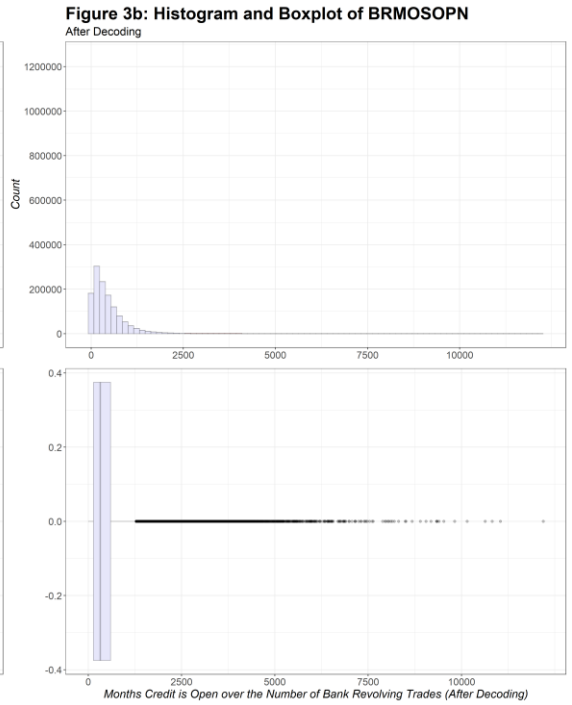
- Variables that are less than ten and continuous have coded values for 9.9993 – 9.9999.
- Variables with a two-digit max value have coded values for 93 – 99.
- Variables with a three-digit max value have coded values for 993 – 999.
- Variables with a four-digit max value have coded values for 9993 – 9999.
- Variables with a seven-digit max value have coded values for 9999993 – 9999999.

Table 1: Descriptive Statistics of Five Variables Before Imputation

Variable	Min	Q1	Median	Mean	Q3	Max	n
TADB	0	0.32	0.58	0.56	0.78	9.9999	1,255,429
PRDEROG	1	1	7	49.03	99	99	1,255,429
AVGMOS	0	35	56	59.24	77	999	1,255,429
PRAGE	0	19	113	4,855.10	9,999	9,999	1,255,429
TSHIC	0	11,500	24,787	53,650.02	44,360	9,999,999	1,255,429

For example, the variable BRMOSOPN has a max value that is seven digits long. This means the interval between 9999993 – 9999999 are coded. The histogram in Figure 3a below displays the visual effects of the coded values. Figure 3b displays the histogram of the variable BRMOSOPN after removing the coded values. These values must be removed from the dataset and replaced with a meaningful value.



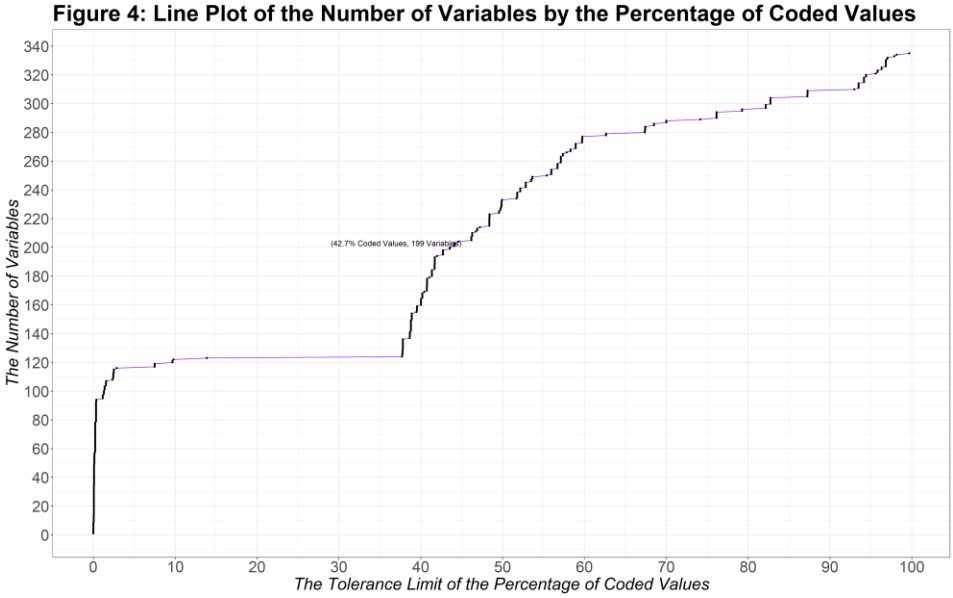


The median value is the center of the distribution and is not affected by the skew of the distribution. Distributions that follow a bell-curve have approximately equal mean and median values. So, in either case the median value is meaningful to the center of distribution. For these reasons, the median was chosen over the mean to replace the coded values. Other imputation methods, such as KNN and regression, were considered but not used due to the number of variables containing a moderate proportion of coded values.

Table 2: Percent of Coded Values Sorted in Descending Order

Variable	True Values	Coded Values	Percent True	Percent Coded	Median	Variance
ATPDBAL	3,826	1,251,603	0.3%	99.7%	1	0.07
BIPDBAL	23,101	1,232,328	1.8%	98.2%	1	0.33
AFPDBAL	26,757	1,228,672	2.1%	97.9%	1	0.14
CUPDBAL	36,865	1,218,564	2.9%	97.1%	1	0.41
ORPDBAL	37,971	1,217,458	3.0%	97.0%	1	0.45

Figure 4: Line Plot of the Number of Variables by the Percentage of Coded Values



From Table 2, imputing the variable ATPDBAL would result in 1,251,603 of the 1,255,429 observations equaling the median value of 1, but this variable would contain virtually no variation and may introduce bias in modeling. For this reason, this variable will be dropped from the dataset along with any variables containing 50% or more coded values. The line graph below in Figure 4 displays the number of variables remaining, on the vertical axis, and the percentage of coded values, on the horizontal axis. This plot displays that as the tolerance in the percentage of coded values decreases, the number of variables decreases. Setting the tolerance of coded values to 30% would result in 124 variables, while setting the tolerance to a value of 50% would result in 235 variables. Choosing the correct tolerance level is imperative in wrangling the data into a structure that is good for modeling and true to the original distribution.

Since this project will go through several steps that will further eliminate variables, a value of 42.7% was chosen as the tolerance of coded values, due to it being a local max of the function that interpolates each of the points displayed in Figure 4. This level of tolerance will result in 198 independent variables, where a maximum percentage of coded values is 42.7% for each variable. Variables with more than 42.7% coded values will be dropped from the dataset. The remaining variables will be further reduced during the variable clustering step of this project. Another value to consider in this step is how many standard deviations an observation is allowed to be from the mean before imputing the value.

Table 3a: Percentage of the Retained Variance for TSHIC Before/After Handling Extreme Values

Variable	Percent Variance	Total Variance	Count of True Values	Count of Imputed Values	Median	Mean	Standard Deviation	Min	Max	Before/ After
TSHIC	100%	868,259,107	1,252,745	2,684	24,725	32,340.02	29,466.24	0	2,317,866	Before
TSHIC	90.99%	790,008,988	1,248,413	7,016	24,686	32,038.57	28,107.10	0	209,137	After

A method in handling extreme values is to impute observations that are determined extreme to the median. For example, the variable TSHIC before handling extreme values was observed to have a maximum of 2,317,866 (Table 3a, before) which is nearly 94 times the median. Keeping a value of this magnitude may result in poor model performance and can have costly outcomes for the company. Suppose it was decided to cull values for this variable that are greater than six standard deviation steps from the mean. Then a new maximum will be observed less than or equal to 209,137 (Table 3a, after). By performing this culling, the count of imputed values will increase by 4,332 (from 2,684 to 7,016). This means that the median of this variable (24,725) will be used to impute 0.6% of the values, resulting in a new distribution that will retain approximately 91% of the original variance.

Figure 5: Cumulative Distribution of TSHIC

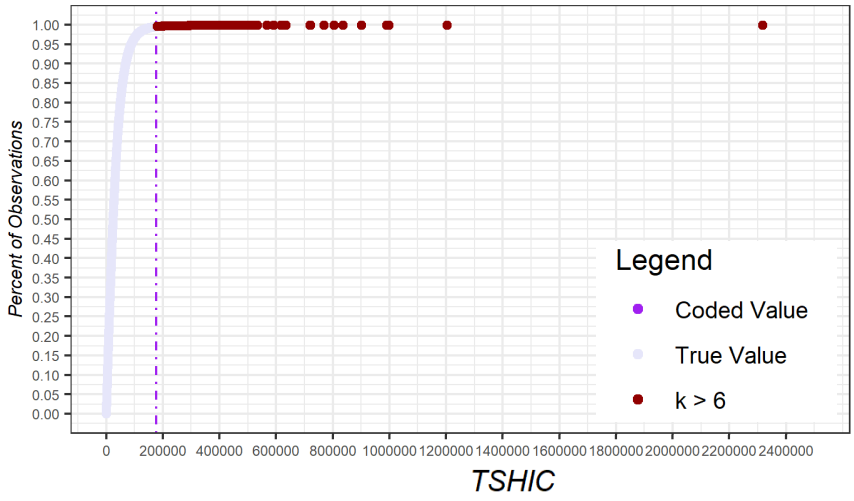


Figure 5 displays the cumulative distribution for the variable TSHIC before imputing the outliers. A line was placed six standard deviation steps from the mean and observations to the right of this line were considered extreme values. Setting the line six standard deviation steps from the mean is a choice that should be carefully considered. These values will be altered during this step and a portion of them may be true to the distribution of the variable.

Figure 6: Histograms of TSHIC at Differing Cutoff Points

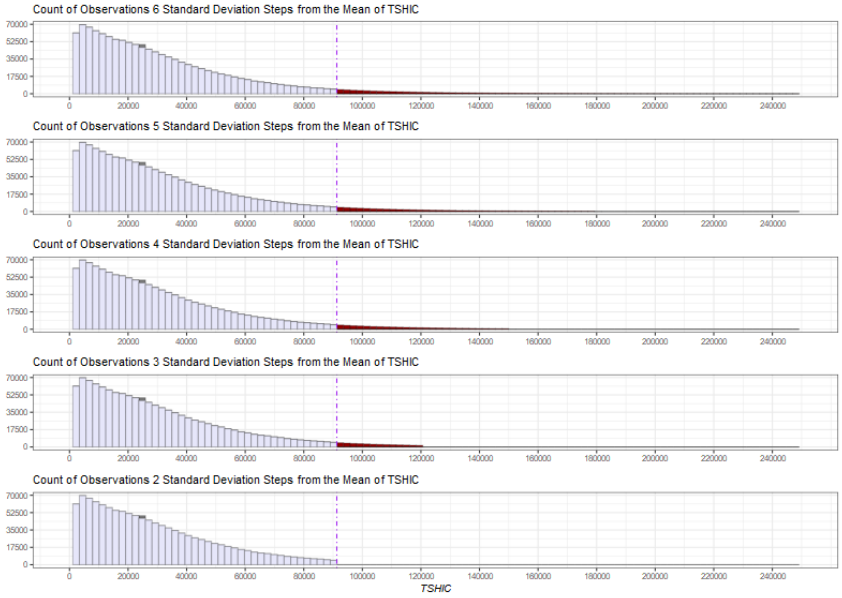


Figure 6 displays histograms used in determining the maximum number of standard deviation steps allowed. A line was placed two standard deviation steps from the mean and observations to the right of this line were considered retained by setting the cutoff point greater than two. The top panel displaying the results of setting the cutoff point to six. Each panel down reduces the cutoff point by one which reduces the maximum number of standard deviations an observation is allowed to be from the mean.

From these two graphs, it becomes clear that values greater than 100,000 increasingly become more isolated and rare. Chebyshev's theorems would suggest that at least 75% of the data lies within two standard deviation steps from the mean but using two standard deviation steps as the cutoff point may result in the loss of important true values (Figure 6). This loss can be numerically observed in the percentage of variance retained (Table 3).

Table 3: Percentage of the Retained Variance at Varying Max Standard Deviations

Variable	k_step	Percentage of Retained Variance	Retained Variance	New Min	Mew Max	Count of Imputed Outliers
TSHIC	1	28%	242,605,254	2,874	61,806	225,871
	2	51%	445,098,627	0	91,272	58,232
	3	68%	591,046,173	0	120,739	23,182
	4	79%	688,527,669	0	150,205	10,961
	5	86%	750,387,444	0	179,671	6,305
	6	91%	790,008,988	0	209,137	4,332
	7	94%	813,531,135	0	238,604	3,501
	8	95%	827,419,979	0	268,070	3,133

Table 3 displays the percentage of variance retained at varying max standard deviations steps from the mean. At six standard deviation steps, TSHIC retains 91% of its original variance and will result in a new maximum value of 209,137. If instead of six steps, three steps from the mean are used, then an additional 23,182 observations will be assigned to the median value, along with 2,684 coded values. In doing so, 68% of the original variance will be retained in the dataset and the new maximum will be less than or equal to 120,739. Since a later step in this project will bin the variables to create ordinal versions of the data, a value of six standard deviation steps was chosen to be the allowed maximum standard deviation. By choosing this value, the data will retain 91% of its original variance.

Table 4: Percent Within/Beyond k Standard Deviation Steps

k_steps	% within k	% beyond k
2	75%	25%
3	89%	11%
4	94%	6%
5	96%	4%
6	97%	3%
7	98%	2%

For a real number k, greater than one, Chebyshev's inequality suggests that at least $1 - \left(\frac{1}{k^2}\right)$ of data from a sample must fall within k standard deviations from the mean. Reciprocally, this theorem gives that less than $\left(\frac{1}{k^2}\right)$ of data from a sample will be beyond k steps. Using this theorem, less than 3% of observations are beyond six steps from the mean and less than 6% of observations are beyond 4 steps (Table 4). Table 4 displays the percentage within or beyond k steps from the mean. Instead of imputing extreme observations to the median value, Chebyshev's theorem can be used to construct a less extreme value that preserves the overall structure of the data by projecting observations beyond the chosen cutoff point (k = 6) to a value in the range between 4 and 6.5 steps. To do this for an extreme observation x_i:

1. First, compute the number of k_i steps x_i is from the mean by $k_i = \frac{(x_i - \mu)}{\sigma}$.
2. Then use $1 - \left(\frac{1}{k_i^2}\right)$ to get the floor percentage of values p_i, within k_i steps.
3. Let p_4 be the floor percentage of values within 4 steps.
4. Find the difference between p_i and p_4: $p_i - p_4 = p_d$.
5. Then, use $\left(\frac{1}{\sqrt{p_d}}\right)$ to revert p_d back to standard deviation steps, k_{p_d}.
6. For extreme values far from the cutoff point, the value of k_{p_d} will be closer to the lower bound of the desired range (4) than extreme values near the cutoff point (Table 5: step 4, pg. 10). To correct this issue, find the difference between the upper bound of the desired range (6.5) and k_{p_d} and then add the lower bound (4): $(6.5 - k_{p_d}) + 4 = k_{newi}$.
7. Finally, revert k_{newi} back to the domain of the variable by $(k_{newi} * \sigma) + \mu$.

Table 5: Result of Projecting an Extreme Observations Far/Near the Cutoff Point

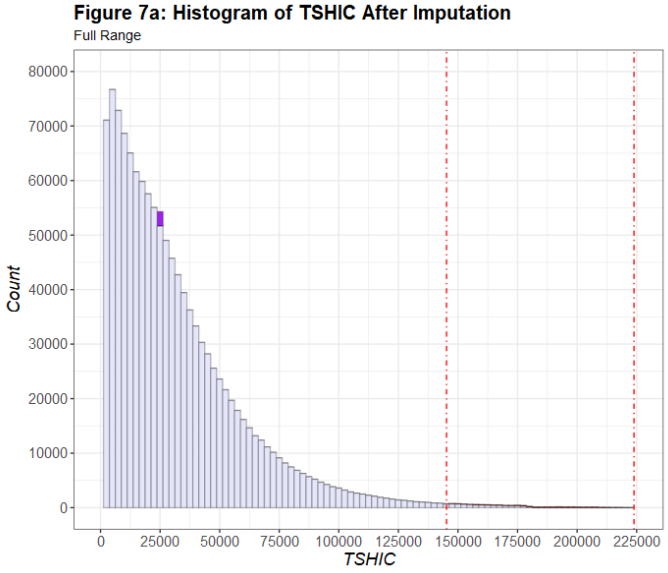
Variable: TSHIC			*Cutoff value:* 176,797			
Start Value	**Step One**	**Step Two**	**Step Three**	**Step Four**	**Step Five**	**End Value**
2,317,866	77.56	0.999	0.062	4.01	6.49	223,713.5
250,000	7.39	0.982	0.044	4.76	5.74	201,536.0

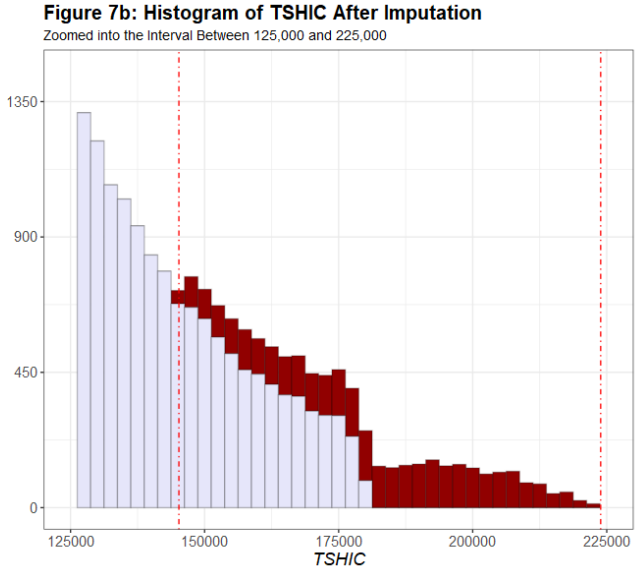
Table 5 above displays the results of this projection for an extreme observation both far and near the cutoff point for the variable TSHIC. Table 6 below displays the before and after for the variables with the largest original maximums. The original max of the variable BRMAXH is 4 steps from the original mean. For each variable other than BRMAXH, the max after imputing is within the upper bound (6.5) of the desired range.

Table 6: Before and After Imputing Values Beyond Five Standard Deviation Steps

	Before					**After**				
Variable	**max**	**median**	**mean**	**Std Dev**	**Number of Std Dev Steps from the Original Mean**	**max**	**median**	**mean**	**Std Dev**	**Number of Std Dev Steps from the Original Mean**
TSHIC	2317866	24725	32340	29466	77.6	223714	24725	32200	28493	6.5
TSBAL	899254	12006	17710	18974	46.5	140759	12006	17614	18356	6.5
RHIC6	635317	12150	18118	19746	31.3	145812	12150	17889	18502	6.5
BRHIC	625117	8700	14111	17192	35.5	125418	8700	13863	15773	6.5
RBAL6	279747	4499	7464	9325	29.2	67724	4499	7352	8701	6.5
RBAL	279747	4505	7470	9329	29.2	67754	4505	7362	8710	6.5
BRBAL	278232	3433	6058	8182	33.3	59003	3433	5941	7509	6.5
BRMAXH	25000	4000	5452	4843	4.0	25000	4000	5417	4789	4.0
BRMINH	25000	500	931	1080	22.3	7882	500	912	980	6.4
BRMINB	25000	314	686	1219	19.9	8512	314	659	1041	6.4

Figure 7 below displays the histogram for TSHIC after imputing coded values to the median and projecting values beyond six standard deviation steps from the mean to a value between 4 and 6.5 standard deviations. On the left in Figure 7a, two lines were placed at the upper and lower bound of the desired range. These lines correspond to the lines in the plot on the left (Figure 7b). Figure 7b zooms into this interval to display the results of projecting the extreme values. In either plot, the lighter bars are the true values observed for TSHIC while the darker bars are the projected and imputed values. The results of projecting the extreme values worked as expected and visually kept the structure of the original distribution. At the end of this phase of the project, the dataset contained 198 independent variables, MATCHKEY, CRELIM, DELQID, and goodbad for 1,255,429 customers.





Variable Clustering

The next step in the process is to further reduce the number of variables that will be used in the variable preparation process. Since the final model of this project must be interpretable, methods such as Principal Component Analysis and Factor Analysis were not used to reduce the number of variables. Instead, the variables were clustered into k groups and a representative variable was chosen from each cluster. Many of the 198 variables that entered this part of the project explain similar phenomena which can result in multicollinearity between one or more variables. Clustering each of the variables and selecting the best representative from each cluster should reduce both the amount of multicollinearity and the number of variables. The maximum number of clusters we can assign is 198, or one cluster per variable. Although this would retain 100% of the original variation in the data, it would not reduce the amount of multicollinearity. Figure 8a below displays the percentage of variation retained on the vertical axis and the number of clusters on the horizontal axis.

Figure 8a:
Proportion of Variation Explained by the Number of Clusters
Full Range

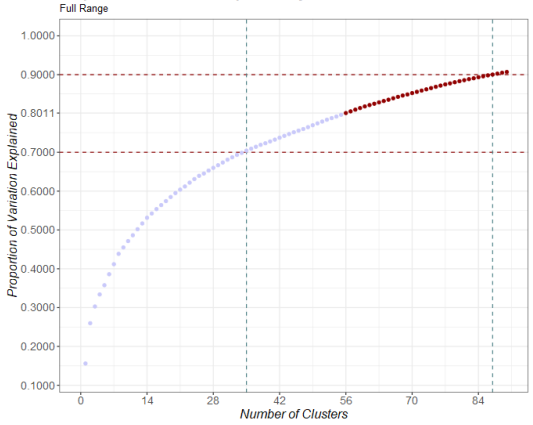
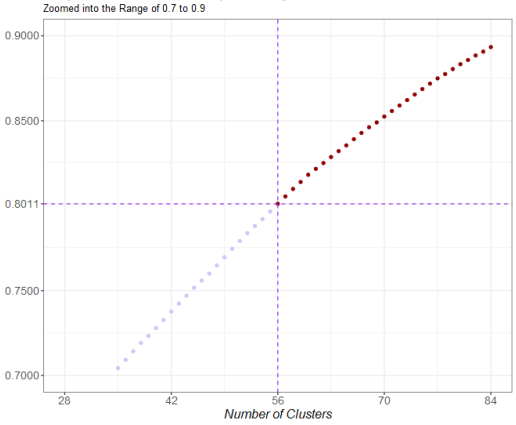


Figure 8b:
Proportion of Variation Explained by the Number of Clusters
Zoomed into the Range of 0.7 to 0.9



It was advised that at least 70% of the variation of the original data should be retained in this process. If it was desired to keep at least 90% of the original variation, then the minimal number of clusters needed would be 87. From this point, adding additional clusters would result in minimal additional retention of explained variation. Figure 8b zooms into the region between 70% and 90% explained variation. The value of 80% was chosen to be the desired percentage of explained variation by finding a local maximum of the function that interpolates the points. The corresponding number of clusters that would produce the desired percentage was found to be 56. This means that the 198 variables that entered this step of the process will be put into 1 of 56 clusters where a representative variable will be chosen from each of the clusters.

SAS produces these clusters by first running all of the variables in one cluster. The second eiganvalue for the cluster is computed. Larger second eiganvalues indicate that at least two components account for a large amount of variation among the variables and that cluster should be split. The variables are then obliquely rotated and, for each variable in the clusters, an R^2 is computed for the variables within the cluster and the variables in the next closest cluster. Since the representative of each cluster should be collinear with the variables within the cluster, the R^2 within the cluster should be higher than the R^2 with the next closest cluster. Thus the representative for each of the clusters will be the variable with the lowest $(1- R^2)$ ratio.

Table 7: Rsquare Ratio for each of the First Five Cluster Representatives

Number of Clusters	Cluster	Variable	Rsquare Ratio
56	1	TRR49	0.08
	2	ROPEN	0.20
	3	BRRATE79	0.16
	4	TOPEN6	0.28
	5	TROPENEX	0.07

Table 8: The Rsquare Ratio for the Variables in the First Cluster

Cluster	Variable	Own Cluster	Next Closest	Rsquare Ratio
1	TRR49	0.961	0.538	0.0844
	TRCR39	0.952	0.523	0.101
	TRCR49	0.951	0.519	0.102
	TRR39	0.935	0.527	0.138
	TRATE79	0.905	0.486	0.184
	CRATE79	0.904	0.484	0.186
	TR7924	0.839	0.657	0.47
	TRR29	0.782	0.549	0.483
	TN90P24	0.847	0.683	0.484
	TR39P24	0.802	0.652	0.569
	WCRATE	0.465	0.316	0.782

Increased Rsquare Ratio

The clustering process ends when the max number clusters (56) allowed is met. Table 7 displays the representatives of the five clusters when the max number of clusters is set to 56. For Cluster 1, the variable TRR49 was selected to be the representative with an R^2 ratio of 0.08. The representative of Cluster 2, ROPEN, displayed an R^2 ratio of 0.20. Table 8, above on the right, displays all of the variables within Cluster 1. The representative of this cluster had an R^2 of 0.961 within the cluster, 0.538 with the next closest cluster, and a (1- R^2) ratio of 0.0844.

Figure 9:
Correlation Matrix of the Variables Within the First Cluster and the Representative of the Second Cluster

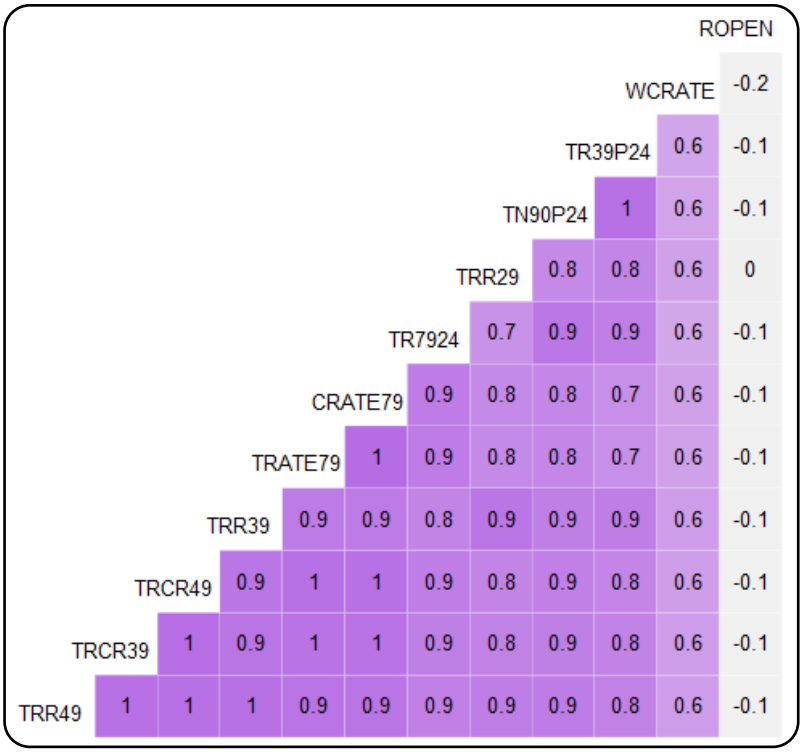


Figure 9 below displays the correlation between the variables within Cluster 1 with the representative of Cluster 2. In this plot, darker boxes correspond to higher correlation and lighter boxes correspond to weaker correlations. Variables that are highly correlated have values greater than 0.7 in magnitude while variables that are weakly correlated have values less than 0.4 in magnitude. This plot displays variables within Cluster 1 are highly correlated with each other while being weakly correlated with the representative of Cluster 2, ROPEN. The variable WCRATE had the highest correlation in magnitude with ROPEN (-0.2), but these variables still exhibit a weak negative correlation. The variable with the weakest correlation with ROPEN was TRR29 (0).

Table 9: Top Five Variables with the Highest VIF (Before)

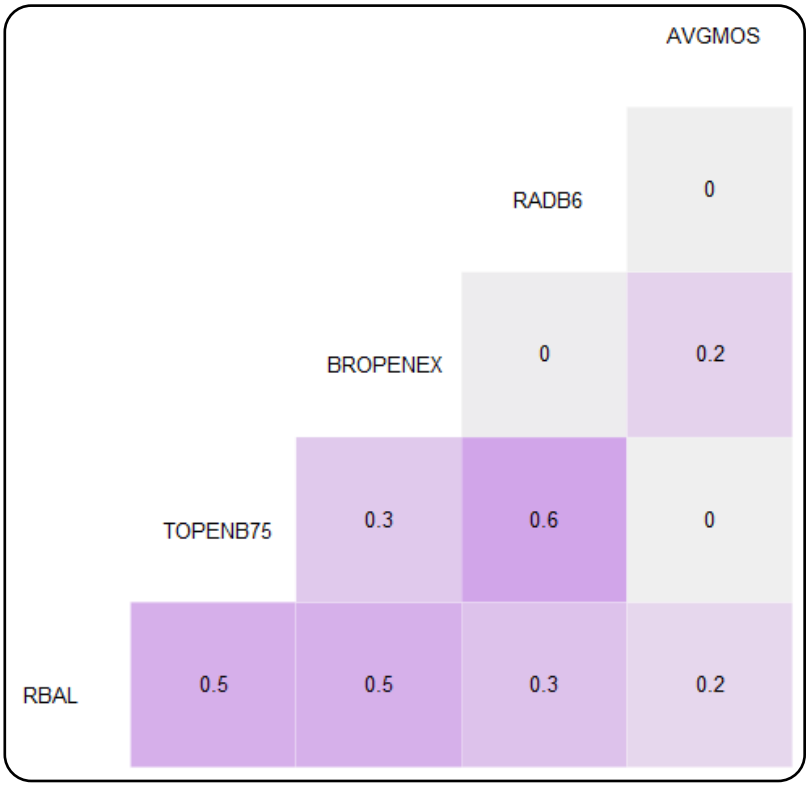
Variable	VIF
ROPEN	6.37
BRR39P24	6.06
TRR49	5.72
BROPENEX	5.42
TROPENEX	5.31

Table 10: Top Five Variables with the Highest VIF (After)

Variable	VIF
RBAL	2.12
TOPENB75	2.10
BROPENEX	2.03
RADB6	2.02
AVGMOS	1.99

After clustering process concludes and the representative variable of each cluster is selected, the Variance Inflation Factor, VIF, is calculated for each of the 56 representatives by running each of the variables through a linear regression. The R^2 is then used to compute the VIF of each variable. Table 9 displays the top five highest VIF's of the 56 cluster representatives. The variable ROPEN displayed a VIF of 6.37 which suggests that multicollinearity may still be an issue amongst the selected representatives. Higher VIF values indicate that the variance of a particular coefficient is larger than what would be expected if there was no correlation with the other predictors. In general, VIF values higher than ten suggest high correlation and are cause for concern. The company has suggested that a cutoff value of two should be used as the deciding factor to ensure the weak to moderate collinearity between the predictors.

Figure 10: Correlation Matrix for the Variables with the Highest VIF's



To reduce multicollinearity, the variable with the highest VIF will be dropped from the dataset and the remaining variables will have their VIF's recalculated. This process continues until the remaining variables have a VIF approximately less than 2. Table 10 displays the top five variables with the highest VIF's after dropping the variables one by one. Figure 10 displays the correlation between the five variables seen in Table 10. The highest correlation value of 0.6 was found between the variables TOPENB75 and RADB6. Their corresponding VIF's were 2.10 and 2.02, respectively (Table 10). The variable clustering process resulted in a dataset that includes 37 independent variables, MATCHKEY, DELQID, CRELIM, and goodbad for 1,255,428 observations.

Variable Preparation

The next phase of this project will transform the variables through one of two methods to create an ordinal ranking for the values of each continuous variable. The output of the variable clustering phase of this project was 37 independent variables. Ten variables were either binary or already discrete (Table 11). These variables will only include odds and log-odds ratio transformations and, if needed, variables with extremely infrequent categories will be truncated to include at most three new variables: the truncated variant, the odds ratio, and the log odds ratio. The other 27 variables are continuous and will go through two discretization methods.

First Discretization Method

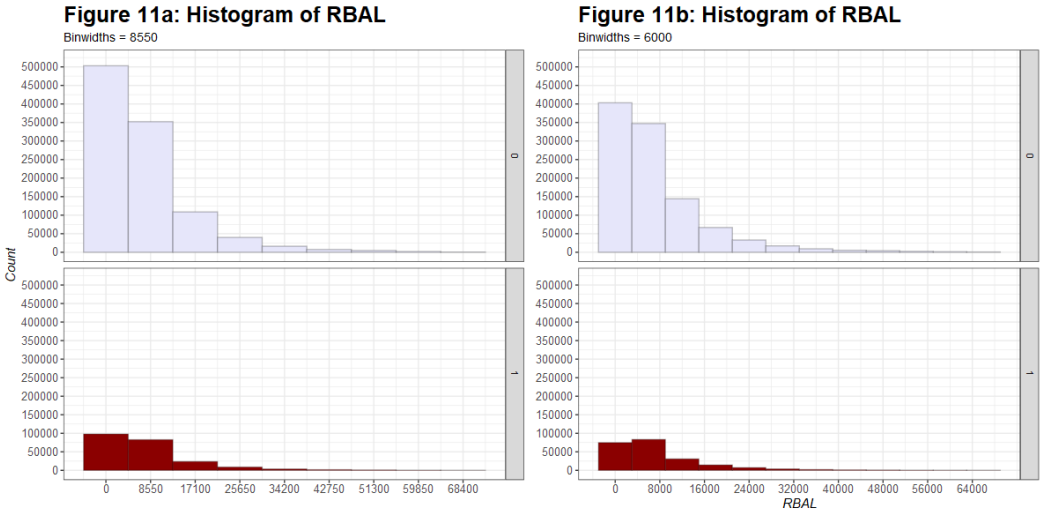
In one method, the variables will be binned into equal intervals across the domain. The width of the interval will be user defined in a logical manner and is labelled with a discrete integer [1 to n]. The second method will utilize a function to construct bins that contain approximately an equal amount of the observations. Each bin is labelled with a discrete integer [1 to n]. These variables are then individually used to construct odds and log-odds ratios from the average of goodbad for each bin. The result of both methods will be six new variables: 3 from the equal interval width process and 3 from the equal observational frequency process.

In the first discretization process, each variable will be discretized into approximately equal interval width bins defined by the user. In addition, the target variable, goodbad, will be used to create the probability of default by finding the average of goodbad for each bin. The desired transformed variable should produce a monotonic probability of default over the created bins. The best transformations will visually display a straight line across the bins and should equally cover as much of the domain as possible for as many bins as necessary.

Table 11: Thirty-Seven Variables from the Variable Clustering Phase

Variable	Count	Min	Median	Mean	Max	Standard Deviation	Number of Distinct Values
BKP	1,255,429	0	0	0.23	1	0.42	2
FFR324	1,255,429	0	0	0.02	4	0.14	5
DCCRATE2	1,255,429	0	0	0.01	5	0.13	6
DCR4524	1,255,429	0	0	0.03	7	0.19	8
FFRATE2	1,255,429	0	0	0.08	7	0.30	8
FFR39P24	1,255,429	0	0	0.09	9	0.35	9
DCRATE2	1,255,429	0	0	0.11	8	0.37	9
DCWCRATE	1,255,429	0	1	3.74	9	2.89	10
BRCRATE3	1,255,429	0	0	0.06	18	0.34	15
PFRATE2	1,255,429	0	0	0.08	16	0.32	15
DCCR39	1,255,429	0	0	0.16	4.74	0.52	17
BRRATE3	1,255,429	0	0	0.23	4.25	0.59	17
BRRATE2	1,255,429	0	0	0.53	6.83	0.95	19
CRATE2	1,255,429	0	0	0.23	4.51	0.60	20
PFRATE45	1,255,429	0	0	0.05	3	0.26	23
LAAGE	1,255,429	0	0	0.21	3.58	0.42	25
BRRATE79	1,255,429	0	0	0.46	7.57	1.03	27
DCOPENEX	1,255,429	1	2	2.23	13.62	1.36	27
TOPEN6	1,255,429	0	1	1.06	10.21	1.40	28
TR4524	1,255,429	0	0	0.33	6.95	0.90	32
FFCRATE1	1,255,429	0	1	1.45	10.12	1.03	35
BNKINQ2	1,255,429	0	4	4.46	27.21	3.64	36
TOPENB75	1,255,429	0	2	3.18	22.87	3.00	47
BADPR2	1,255,429	0	1	1.77	20.06	2.70	68
BROPENEX	1,255,429	1	7	7.61	36.96	4.55	69
PFRATE1	1,255,429	0	2	2.41	25.87	2.40	85
INQ12	1,255,429	0	3	3.73	31.06	3.90	89
PRMINQS	1,255,429	0	17	19.03	80.74	12.11	92
FFLAAGE	1,255,429	0	1	3.05	88.54	10.41	120
DCPCTSAT	1,255,429	0	1	0.85	1	0.30	135
BRNEW	1,255,429	0	11	15.95	127.41	16.88	251
AVGMOS	1,255,429	0	56	58.66	256.53	30.80	371
FFAVGMOS	1,255,429	0	47	58.87	462.24	50.93	554
BRMINB	1,255,429	0	314	658.67	8511.62	1040.99	11635
RADB6	1,255,429	0	0.4657	0.48	2.60	0.33	16614
RBAL	1,255,429	0	4505	7361.70	67753.51	8709.92	50937

Figure 11 below displays the histogram for RBAL stratified by good or bad credit risk customers. In Figure 11b, the variable was binned into intervals of 6,000. The top plot displays customers that are a good credit risk (goodbad = 0); the bottom plot displays customers that are bad credit risks (goodbad = 1). This plot displays the count of good or bad credit risk customers. The count of bad credit risk customers in the bin starting at 0, in Figure 11b, is less than the count of the next bin starting at 6,000. By shifting the bin width to 8,550, the histogram of RBAL displays a decreasing frequency of bad credit risk customers as the variable, RBAL, increases.



Using a bin width of 8,550 to create an ordinal ranking of RBAL produced a starting point to finding a monotonic relation between the probability of default and the ordinal ranking of RBAL (Figure 12a). Figure 12a below displays the results of using a bin width of 8,550 while Figure 12b displays the results of using a bin width of 6,000. The plot in Figure 12a displays the probability of default increasing as the ordinal rank of RBAL increases. The ordinal bins given the labelled 9 and 11 in Figure 12b do not follow a monotonic relation. At these bins the probability of default dips below the value of the bin before it and would need to be collapsed to achieve monotonicity over the ordinal bins.

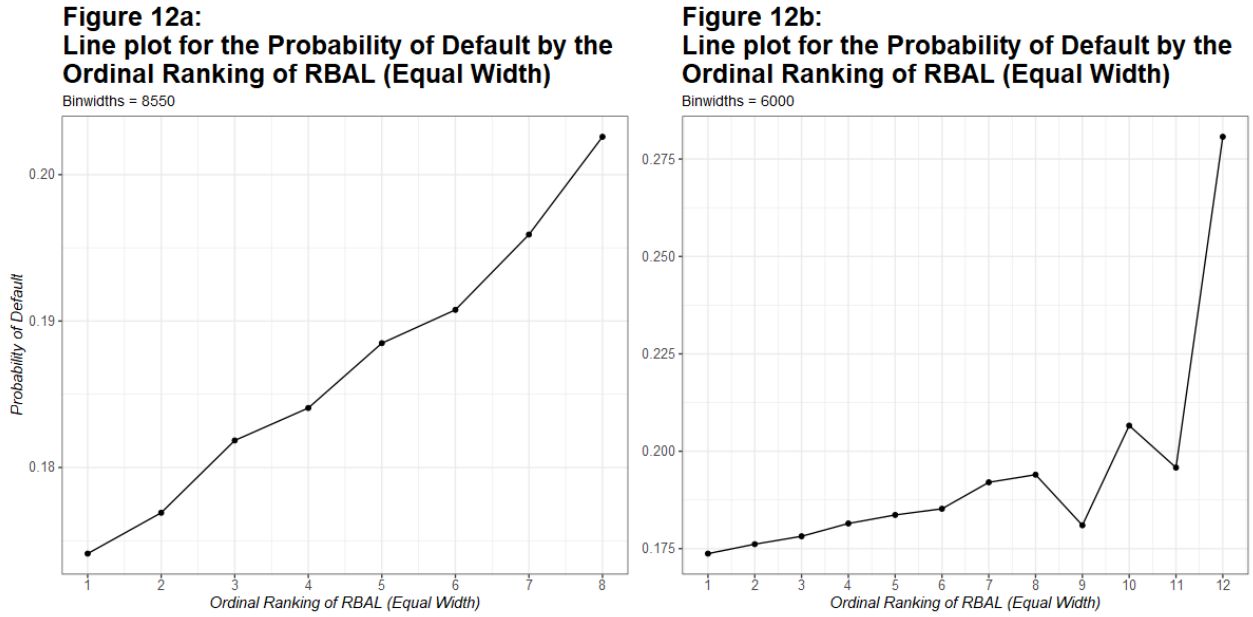


Table 12a: Transformation of RBAL from the First Discretization Process

Bin Width set to 8,550								
Ordinal Ranking of RBAL	Probability of Default	Bin Count	Min RBAL	Max RBAL	Bin Range	Weight of Evidence	Information Value	Total Information Value
8	0.2026	1935	59852.66	67753.51	7900.849	-0.175	0.00005	0.000534
7	0.1959	3568	51302	59844.04	8542.038	-0.134	0.00005	0.000534
6	0.1908	7538	42753	51299.26	8546.261	-0.101	0.00006	0.000534
5	0.1885	12892	34200	42749	8549	-0.086	0.00008	0.000534
4	0.1841	30897	25650	34198	8548	-0.057	0.00008	0.000534
3	0.1818	79309	17100	25649	8549	-0.042	0.00011	0.000534
2	0.1769	230728	8550	17099	8549	-0.008	0.00001	0.000534
1	0.1741	888562	0	8549	8549	0.011	0.00009	0.000534

Table 12a above displays the results of using a bin width of 8,550 to transform the variable RBAL during the first discretization process. Since the first process looks to create an ordinal ranking with approximately equal interval widths, the count of observations in each bin does not necessarily need to be equal. Rank 8 covers less of the domain of the variable than the other seven ranks but is not an issue since this bin collects the upper end of the variable. Another metric that aided in deciding bin width is the information value of each bin, which is defined as the product between the bin's weight of evidence, and the difference between the percent of good credit risk customers and bad credit risk customers. The weight of evidence for each bin can be calculated by taking the natural log of the ratio between the percent of good credit risk customers and bad credit risk customers. By summing together each bin's information value we get the total information value of the ordinal ranking variable. Using a bin width of 8,550 produces a total information value of 0.000534 (Table 12a), while using an information value of 6,000 would produce a total information value of 0.000635 (Table 12b). This suggests that using a bin width of 6,000 is more valuable than a bin width of 8,550, but, as seen in Figure 12b, the relation between the ordinal ranks and the probability of default is not monotonic and collapsing bins 9 through 12 would be necessary to achieve monotonicity. Additionally, information values less than 0.1 have weak predictive power which suggests this variable may not be a useful predictor.

Table 12b: Transformation of RBAL from the First Discretization Process

Bin Width set to 6,000								
Ordinal Ranking of RBAL	Probability of Default	Bin Count	Min RBAL	Max RBAL	Bin Range	Weight of Evidence	Information Value	Total Information Value
12	0.2807	171	66012.7	67753.5	1740.81	-0.605	0.00006	0.000635
11	0.1958	1716	60010.7	65993.6	5982.89	-0.133	0.00003	0.000635
10	0.2066	1762	54000	60000	5999.98	-0.200	0.00006	0.000635
9	0.1810	4431	48001	53994	5993	-0.036	0.00000	0.000635
8	0.1940	5722	42000	47998	5998	-0.121	0.00007	0.000635
7	0.1920	8509	36000	41999	5999	-0.109	0.00008	0.000635
6	0.1852	15526	30001	35999	5998	-0.064	0.00005	0.000635
5	0.1837	28935	24000	29999	5999	-0.054	0.00007	0.000635
4	0.1815	56920	18000	23999	5999	-0.039	0.00007	0.000635
3	0.1782	118238	12000	17999	5999	-0.017	0.00003	0.000635
2	0.1762	266651	6000	11999	5999	-0.003	0.00000	0.000635
1	0.1737	746848	0	5999	5999	0.014	0.00011	0.000635

Since a bin width of 8,550 produces bins that cover the full domain of RBAL and each bin is approximately equal in interval width, this value was the chosen bin width to create the ordinal ranking of RBAL in the first discretization process. After the creation of the ordinal variable, the odds and log odds ratio were computed for each bin by first dividing the probability of default by 1 minus the probability of default to produce the odds ratio:

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

This will produce a value for each bin that can be interpreted as the odds of customer being a bad credit risk. If a value is less than 1, the reciprocal can be used in its place, but the interpretation will change to be the odds of a customer being a good credit risk. For example, customers assigned to bin 6 during the equal width discretization of RBAL had a probability of default value of 0.1895 (Table 13). The odds ratio for this bin is $\frac{0.1895}{(1-0.1895)} = 0.23$. Since this value is less than 1, the reciprocal will be used in its place: $\frac{1}{0.23} = 4.28$.

Thus, customers assigned to bin 6 are 4.28 times more likely to be a good credit risk than a bad credit risk. Table 13 on the right displays that as the probability of default decreases, the odds ratio increases. Next, the log odds ratio can be computed for each bin by taking the natural log of the odds ratio (Table 13). For the variable RBAL, the first discretization process resulted in three transformations: a discrete ordinal ranking, an odds ratio, and a log odds ratio. A similar process was conducted for the other 26 variables that came from the variable clustering phase which resulted in 81 variable transformations added to the dataset.

Table 13: Odds and Log Odds Ratio of the Created Bins

Ordinal Ranking of RBAL	Probability of Default	Odds Ratio	Log Odds Ratio
8	0.2026	3.94	1.37
7	0.1959	4.10	1.41
6	0.1908	4.24	1.45
5	0.1885	4.31	1.46
4	0.1841	4.43	1.49
3	0.1818	4.50	1.50
2	0.1769	4.65	1.54
1	0.1741	4.74	1.56

Second Discretization Method

In the second discretization method, each variable will be discretized into bins of approximately equal frequency. Like before, a desirable transformation during this process would be an ordinal ranking that exhibits a monotonic relation between the ranks and the probability of default. A function cut the variables into bins with an approximately equal number of observations. The probability of default was then computed for each bin. Adjacent bins that are not significantly different were collapsed into a single bin. Table 14a displays the initial results of the second discretization process for the variable RBAL. Notably, the total information value observed for RBAL was 0.064 which indicates that the predictive power is 'too good to be true.' For comparison, the ordinal ranking of RBAL created during the first discretization process had a total information value of 0.000534 (Table 12a, pg. 16) which suggested weak predictive power.

Table 14a: Transformation of RBAL from the Second Discretization Process

Before Bin Collapse								
Ordinal Ranking of RBAL	Probability of Default	Min RBAL	Max RBAL	Bin Range	Bin Count	Weight of Evidence	Information Value	Total Information Value
11	0.1849	18,757	67,754	48,996.5	114,141	-0.062	0.000356	0.063778
10	0.1786	12,499	18,756	6,257	114,131	-0.020	0.000036	0.063778
9	0.1763	9,109	12,498	3,389	114,148	-0.004	0.000002	0.063778
8	0.1750	6,832	9,108	2,276	114,161	0.005	0.000002	0.063778
7	0.1779	5,130	6,831	1,701	114,084	-0.015	0.000021	0.063778
6	0.2316	3,938	5,129	1,191	114,196	-0.346	0.012156	0.063778
5	0.1869	2,802	3,937	1,135	114,089	-0.075	0.000525	0.063778
4	0.1903	1,834	2,801	967	114,174	-0.098	0.000897	0.063778
3	0.1892	983	1,833	850	114,176	-0.091	0.000767	0.063778
2	0.1572	256	982	726	114,004	0.134	0.001550	0.063778
1	0.0849	0	255	255	114,125	0.832	0.047465	0.063778

This process split the domain of RBAL into eleven ranks, each with approximately 114,000 observations (Table 14a). The probability of default for ranks 3 and 4 were not significantly different (0.1892 vs. 0.1903) and were collapsed into a single bin. Additionally, ranks 7 through 11 were significantly less than rank 6 which is displayed by the mountain like spike in Figure 13a. After collapsing ranks to only three ranks, a monotonic relationship was found between the probability of default and the ordinal ranks created in this process (Figure 13b), but the total information value reduced from 0.0638 (Table 14a) to 0.0546 (Table 14b). Once this process was completed, the probability of default was used to compute the odds and log odds ratios for each rank. A similar process was conducted for the other 26 variables that came from the variable clustering phase of this project, which resulted in 81 variable transformations added to the dataset.

Figure 13a:
Line plot for the Probability of Default by the Ordinal Ranking of RBAL (Equal Frequency)

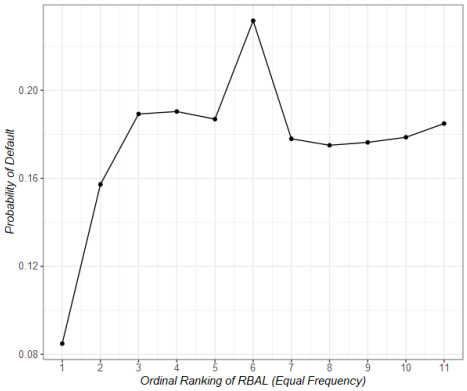


Figure 13b:
Line plot for the Probability of Default by the Ordinal Ranking of RBAL (Equal Frequency)

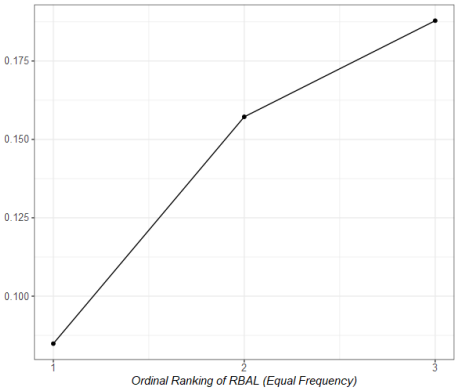


Table 14b: Transformation of RBAL from the Second Discretization Process

After Bin Collapse								
Ordinal Ranking of RBAL	**Probability of Default**	**Min RBAL**	**Max RBAL**	**Bin Range**	**Bin Count**	**Weight of Evidence**	**Information Value**	**Total Information Value**
3	0.1879	983	67753.5	66,770.5	1,027,300	-0.082	0.00561	0.054621
2	0.1572	256	982	726	114,004	0.134	0.00155	0.054621
1	0.0849	0	255	255	114,125	0.832	0.04746	0.054621

Due to their discrete nature, the ten variables at the top of Table 11 did not go through either of the processes described above. Instead, only the odds ratio and log odds ratio transformations were computed for the natural categories of the variable. Table 15 below displays the transformation of the variable BKP. A probability of default of 0.1873 was observed for customers with a value of 1 for the variable BKP while customers that had a value of 0 had a probability of default of 0.1722 (Table 15). The corresponding odds and log odds ratio were computed resulting in an additional two variables added to the full dataset.

Table 15: Transformation of BKP into the Odds Ratio and Log Odds Ratio

BKP	**Probability of Default**	**Bin Count**	**Weight of Evidence**	**Information Value**	**Total Information Value**	**Odds Ratio**	**Log Odds Ratio**
1	0.1873	293,586	-0.08	0.00145037	0.00191	4.34	1.47
0	0.1722	961,843	0.02	0.00045755	0.00191	4.81	1.57

Some variables, such as, DCCRATE2, contained discrete categories with very few observations (Table 16). Only two customers were observed to have a value of 5 for the variable DCCRATE2. Since both customers also had a value of 1 for goodbad, the computation of the weight of evidence measure was unable to produce a value for this category (Table 16a). After reassigning these customers to a value of 4, the weight of evidence and information value was computed for each category. In doing so, the total information value of this variable was observed to be 0.03 (Table 17) which indicates this variable has medium predictive power.

Table 16a: Frequency of Observed Values in the Variable DCCRATE2

DCCRATE2	Probability of Default	Bin Count	Weight of Evidence	Information Value	Total Information Value
5	1.0000	2	-Inf	Inf	Inf
4	0.8333	12	-3.16	0.00014	Inf
3	0.7540	126	-2.67	0.00107	Inf
2	0.6370	956	-2.11	0.00511	Inf
1	0.4324	14790	-1.27	0.02659	Inf
0	0.1722	1,239,543	0.02	0.00058	Inf

Table 17: Frequency of Observed Values in the Variable DCCRATE2

DCCRATE2	Probability of Default	Bin Count	Weight of Evidence	Information Value	Total Information Value
5	0.8571	14	-3.34	0.00018	0.0335
4	0.7540	126	-2.67	0.00107	0.0335
3	0.6370	956	-2.11	0.00511	0.0335
2	0.4324	14790	-1.27	0.02659	0.0335
1	0.1722	1,239,543	0.02	0.00058	0.0335

There were 1,096 customers that were observed with DCCRATE2 values two or greater. These customers could be reassigned to be included with the customers observed with a value of 1. In doing so, the total information value remained at 0.03 (Table 16b) suggesting the predictive power is not hindered by this truncation.

Table 16b: Frequency of Observed Values in the Variable DCCRATE2

DCCRATE2	Probability of Default	Bin Count	Weight of Evidence	Information Value	Total Information Value
2	0.4480	15,886	-1.34	0.0317	0.0323
1	0.1720	1,239,543	0.02	0.000576	0.0323

The corresponding odds and log odds ratio were computed for the truncated variable. The truncation of the variable DCCRATE2 was added to the dataset with the suffix "_ORD" added to the variable's name. The original variable was kept in the dataset in addition to the truncated version, the odds ratio, and the log odds ratio. A similar process was conducted for the other 8 variables which resulted in a total of 27 variable transformations added to the dataset.

The variable preparation phase of this project resulted in a dataset that includes the variables MATCHKEY, CRELIM, DELQID, goodbad, and 226 predictor variables for the 1,255,429 customers. The next phase will develop a model from the pool of 226 predictors.

Logistic Regression

Each of the 37 original variables could have up to six transformations, but due to multicollinearity, only one version of each variable should be allowed in the model. This forms seven families of the variables (Table 18) that were used to gain initial model results. Each variable belongs to one of three cases. Twenty-seven variables were continuous and have seven variants. Three of the ten discrete or binary variables had only three transformations while the remaining seven discrete variables have four variants. Table 18 below displays the three different variable cases and their associated transformations. The question arises of how to choose a subset of variables that will produce the best model. Brute forcing the solution is near-impossible due to the enumerative process associated with the Axiom of Choice.

Table 18: The Seven Variable Families Created During the Variable Preparation Phase

ORIGINAL	ORDEQWID	ORDEQFREQ	LOGODDSEQWID	LOGODDSEQFREQ	ODDSEQWID	ODDSEQFREQ
BRRATE79	BRRATE79_ORD_EQWID	BRRATE79_ORD_EQFREQ	logodds_BRRATE79_EQWID	logodds_BRRATE79_EQFREQ	odds_BRRATE79_EQWID	odds_BRRATE79_EQFREQ
DCCRATE2	DCCRATE2_ORD	DCCRATE2_ORD	logodds_DCCRATE2	logodds_DCCRATE2	odds_DCCRATE2	odds_DCCRATE2
FFR324	FFR324	FFR324	logodds_FFR324	logodds_FFR324	odds_FFR324	odds_FFR324

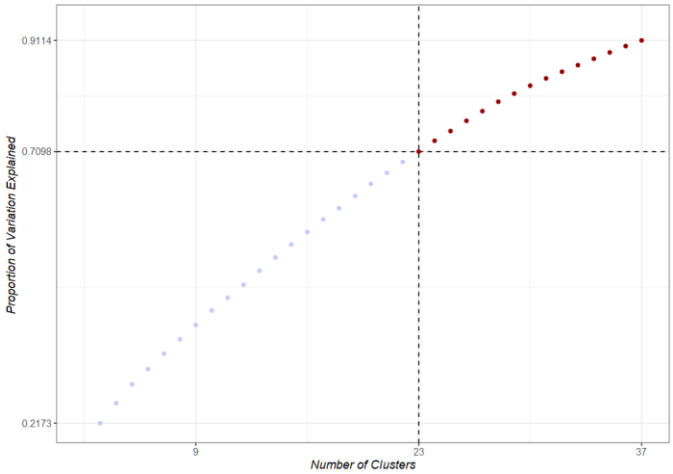
An initial logistic regression model was created for the seven sets of the variables. The model created by the original forms of the variables initially produced the highest percentage of concordant pairs (84.1%, Table 19) while the lowest was produced by the model created by the equal frequency odds ratio forms (82.9%, Table 19). The next step in this phase is to reduce the number of variables in the model to be between 10 and 15. To do this, thirty-seven predictors must be selected from the 226 different variants with no familywise duplications. Once the 37 best predictors are chosen, individually select the least significant or impactful variable and remove it from the dataset. After reevaluating the logistic regression procedure repeat the selection and removal process until ten predictors remain. An additional stopping criterion was used for when the percentage of concordant pairs reduced significantly but only when the number of predictors were 15 or less.

Table 19: The Percent of Concordant and Discordant Pairs and the C-Statistics for the Model Created by each Family

Variable Family	Percent Concordant	Percent Discordant	C
Original	84.1	15.9	0.84
Ordinal (Equal Width)	83.8	16.2	0.84
Ordinal (Equal Frequency)	83.3	16.7	0.83
Odds (Equal Width)	83.5	16.5	0.84
Odds (Equal Frequency)	83.5	16.5	0.84
Log Odds (Equal Width)	82.7	17.3	0.83
Log Odds (Equal Frequency)	82.9	17.1	0.83

As stated above, only one version of each variable will be allowed in the model at a time. Allowing more than one version would result in a model built with collinear variables which could inflate the variance of the predictors. It can also increase the standard error of the estimate in the model and/or flip the sign associated with each parameters estimate value. At an earlier step of this project, multicollinearity was reduced using a variable clustering process where minimally collinear representative variables were selected from a cluster of collinear variables. This process greatly reduced the number of collinear variables during that phase of the project. A second variable clustering process will be conducted here to aid in selecting the best variant of each variable.

Second Variable Clustering Process



A second variable clustering process was conducted on the 226 predictor variables in this phase. Like before, the number of clusters that this procedure will result in will retain at least 70% of the variation. In Figure 14, a vertical line was placed at 23 and a horizontal line was placed at the proportion of variation explained by 23 clusters (0.7098). Since the 226 predictors are associated with at least 1 of the 37 original variables, the max number of clusters allowed will be 37. Choosing this value would explain 91.1% of the variation in the data (Figure 14). For the second variable clustering process, the chosen number of clusters was 37. Choosing this value results in 37 representative variables that retain 91.1% of the explained variation. After selecting the representative variables, the VIF's were rechecked and the top five variables with the highest VIF values are displayed in Table 20. Since the VIF values were all approximately close to 2, multicollinearity was determined to not be an issue.

Table 20: Top Five Variables with the Highest VIF

Variable	VIF
odds_BRPCTSAT_EQWID	2.28
logodds_TROPENEX_EQFREQ	2.06
logodds_BADPR2_EQWID	1.72
logodds_BRRATE79_EQWID	1.65
logodds_AVGMOS_EQWID	1.65

A model was built from the 37 representative variables using the logistic regression procedure in SAS with goodbad as the target variable. The percent of concordant pairs produced by this model was 85.4% which is higher than the models created from the transformation families in Table 19. A backward selection procedure was used to ensure that each of the 37 representative variables begin in the model. Using a significance level of 0.05 as the cutoff point, variables are removed if the p-value computed for the Wald Chi-Square test is greater than the significance level. Since there are over a million observations in the dataset, the significance of the variables will be easily achieved even if the variable is not useful to predicting credit risk.

The next measure to consider is the Wald Chi-Square statistic which is computed by the squared ratio of the model's estimate and standard error for each parameter. Variables in the model that are observed with a near-zero standard error and a large estimate will produce higher Wald score. Since model predictors that are accurate and impactful are desired, the variables with larger Wald score are more desirable than variables with smaller Wald score.

The last measures to consider are the concordance and discordance of the model, which measure the amount of agreement between the prediction and goodbad. For any 0 - 1 pairing of goodbad, if the predictive probability of the 1 is higher than the predicted probability of the 0 then the 0 - 1 pair is concordant, else it is discordant. Models with higher concordance exhibit more agreement between the prediction and goodbad. Therefore, a significant reduction in concordance while removing variables should be carefully analyzed and evaluated.

Table 21a: The Logistic Regression Procedure's Bottom Five Variables According to the Wald Ch-Square Value

Before Dropping LOCINQS				
Variable	Parameter Estimate	Standard Error	Wald Chi-Square	P-value
LOCINQS	0.0258	0.0026	102.37	< 0.0001
DCWCRATE	0.0159	0.0013	142.42	< 0.0001
odds_TRATE1_EQFREQ	-0.0556	0.0044	159.98	< 0.0001
INQ12	0.0203	0.0015	175.31	< 0.0001
TOPEN6	0.0502	0.0034	224.92	< 0.0001

The 37 representative variables will be removed one by one by first removing the most insignificant variable or by removing the variable with the lowest Wald score. In Table 21a, the five variables with the smallest Wald Chi-Square statistics are displayed with the Wald score, their p-value, the parameter estimate of the model, and standard error associated with each of the variables. This table displays the variables all have p-values less than 0.0001 suggesting that all five of the Wald score were found to be significant. The first variable that will be removed is LOCINQS, it had the lowest Wald score (102.37). The logistic regression procedure will be reevaluated with the remaining 36 variables. Table 21b below displays the resulting five predictors with the smallest Wald score after dropping LOCINQS. Two variables, DCWCRATE and odds_TRATE1_EQFREQ, were observed in both Table 21a and 21b. In Table 21b, the variable DCWCRATE displays an increase in both the estimate (0.0161 vs. 0.0159) the Wald score (145.34 vs. 142.42), but the standard error remained the same.

Table 21b: The Logistic Regression Procedure's Bottom Five Variables According to the Wald Ch-Square Value

After Dropping LOCINQS				
Variable	Parameter Estimate	Standard Error	Wald Chi-Square	P-value
DCWCRATE	0.0161	0.0013	145.34	< 0.0001
odds_TRATE1_EQFREQ	-0.0550	0.0044	156.75	< 0.0001
BRRATE3	-0.0951	0.0062	236.24	< 0.0001
INQ12	0.0330	0.0009	1387.39	< 0.0001
TOPEN6	0.0619	0.0031	388.09	< 0.0001

Since DCWCRATE has the smallest Wald score, this variable will be removed from the dataset. The logistic regression procedure will be reevaluated with the remaining 35 predictors. This iterative process will continue until the number of predictors is between 10 and 15 but stop if concordance falls below 85%. After removing another 24 variables from the model, eleven variables remained. The model created from these eleven variables produced a concordance value of 84.8%. Before optimizing the model, a final iterative process was conducted where the 26 removed variables were individually put back into the model and a twelve variable model was reevaluated by the logistic regression procedure. If the addition of that variable increased the percent of concordant pairs, then it was left in the model, else it was removed from the dataset once more. This resulted in an additional variable to be placed back into the model due to the percent of concordant pairs increasing to 85.0%. Out of all the models created during this phase, the twelve variable model produced the highest percent of concordant pairs with a minimal number of predictor variables. A small number of predictors is highly desirable to the company since the company will have to both collect all the data needed to do the prediction and interpret the model's decision for customers seeking credit.

The Best Model

The best model that was found during this project consists of 12 variables where nine are in the original form of the variable, two are odds ratios, and one is a log odds ratio (Table 22). The model parameter observed with the highest standard error (0.0158) was BRPCTSAT. This variable also produced the second highest Wald score (10,086.98). Similar to a general linear regression, the model's parameter estimates are interpretable. Since this is a logistics regression model, the interpretation of the parameters estimate is expressed through the logit function (Log Odds of Default). Using Euler's number, the parameter estimates can be exponentiated ($e^{-1.5901} = 0.20$) which would result in an odds ratio of default. This value can then be interpreted as follows: For every one unit increase in the variable BRPCTSAT, a customer is 0.20 times more likely to default on their credit line.

Table 22: Best Model Parameter Estimates, Standard Error, and Wald Chi-Square test

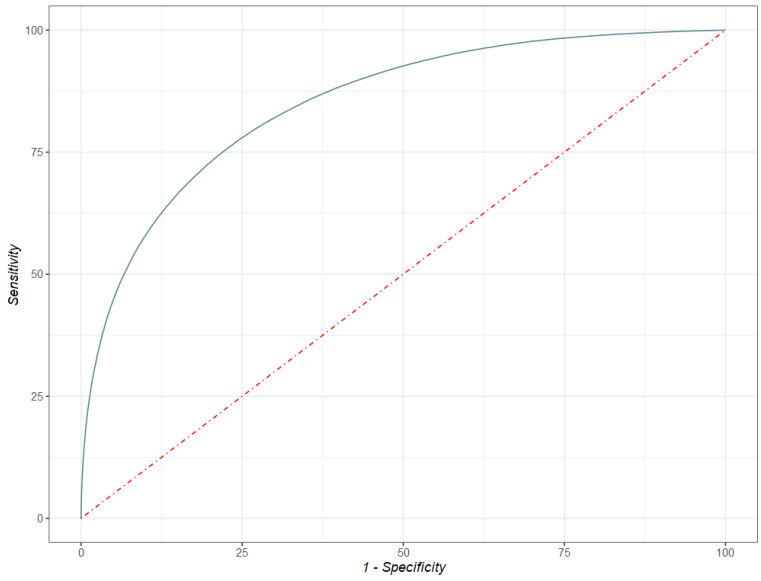
Variable	Parameter Estimate	Standard Error	Wald Chi-Square	P-value
Intercept	1.1766	0.0253	2164.83	< 0.0001
LOCINQS	0.0557	0.0014	1650.75	< 0.0001
odds_BRR4524_EQWID	-0.1089	0.0021	2808.57	< 0.0001
OT12PTOT	0.9225	0.0140	4369.52	< 0.0001
CRATE45	0.3488	0.0049	5142.18	< 0.0001
BRAVGMOS	-0.0110	0.0002	5190.31	< 0.0001
logodds_BRADBM_EQFREQ	-0.3871	0.0052	5604.74	< 0.0001
OBRPTAT	-1.4799	0.0185	6405.97	< 0.0001
odds_TRATE1_EQFREQ	-0.2160	0.0026	7056.56	< 0.0001
BRCR1BAL	0.1413	0.0017	7152.07	< 0.0001
RADB6	1.4283	0.0145	9707.36	< 0.0001
BRPCTSAT	-1.5901	0.0158	10086.98	< 0.0001
BRR23	0.7033	0.0063	12395.01	< 0.0001

Table 23: The Percent of Concordant, Discordant, and Tied Pairs for the Best Model

Number of Variables in the Model	Percent of Concordant Pairs	Percent of Discordant Pairs	Percent of Tied Pairs	C	Profit per 1000 Customers
12	85%	15%	0%	0.850	$ 114,540.22

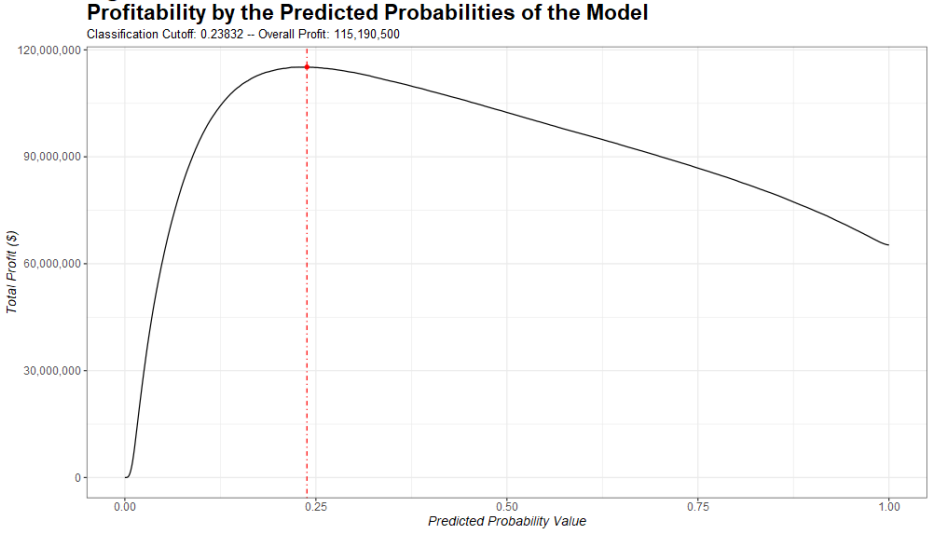
The best model agreed on 85% of the 0 - 1 pairings of goodbad and disagreed on only 15% of pairs. In addition, this model achieves $114,540 of profit for every thousand customers it predicts (Table 23). The C statistic relates to the percent of concordance and indicates whether a model is better or worse at correctly classifying outcomes. It can be visualized through the Receiver Operating Characteristic (ROC) curve, displayed in Figure 16. In Figure 16, the sensitivity of the model (True Positive Rate) is plotted on the vertical axis while for the horizontal axis the model's 1-specificity (False Positive Rate) is plotted. The line starting at the origin and diagonally traveling to the upper right side of the plot represents a random classifier. The performance of the model can be gauged by how close the curve is to the diagonal line and better classifiers capture more of the space above the diagonal line. This model achieved a C statistic of 0.85.

Figure 16: Receiver Operating Characteristic Curve



Another tolerance cutoff choice arises at this step of the project. This choice requires the user to choose the predicted probability value that will be used to round to either a one or a zero. The profitability of the model is calculated by using the correctly predicted nonevents ("good" credit risk customers predicted to be a "good" credit risk) and the incorrectly predicted nonevents ("bad" credit risk customers predicted to be a "good credit risk). As the company has advised, this can be done by multiplying the number of correctly predicted "good" credit risk customers by $250 and the incorrectly predicted "bad" credit risk customers by -$750. Setting the rounding point of the predicted probabilities that come out of the logistic regression model greatly changes the number of the correct and incorrect predicted customers and in turn the profitability of the model. Figure 15 displays the profitability of the model at different probability cutoff points. A vertical line is placed at the predicted probability that produces the highest total profit (0.23832) which achieves a total profit of $115,190,500. Setting the predicted probability value to any other value will reduce the total profit of the model.

Figure 15:
Profitability by the Predicted Probabilities of the Model



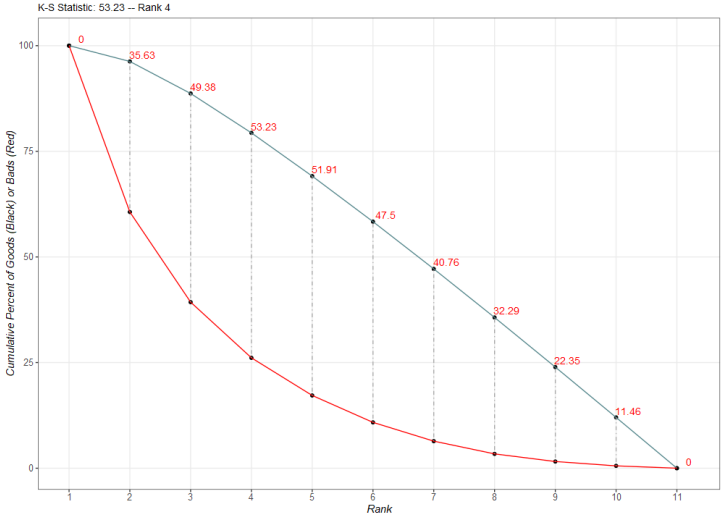
The last measure that will be considered is the Kolmogorov-Smirnov goodness-of-fit statistic and the associated KS-curve. The Kolmogorov-Smirnov goodness-of-fit test measures the discriminatory power of a model and further gives us an idea of the model's ability to distinguish between "good" and "bad" credit risks. Higher values for the KS statistic are better and indicates a larger difference between the cumulative "good" credit risk and the cumulative "bad" credit risk. Table 24 above displays the table used to build the KS curve in Figure 17. From Table 24, the KS statistic was observed in the decile rank 4 and had a value of 53.2. This value corresponds to the distance between the two lines in Figure 17.

Table 24: Kolmogorov-Smirnov Test for the Best Model

Number of Variables Inmodel: 12

Rank	Quantile	Min Score	Max Score	Number of ACCTS	Number of Goods	Percent of Goods	Cumulative Percent of Goods	Number of Bads	Percent of Bads	Cumulative Percent of Bads	Interval Badrate	G/B Ratio	KS
10	91-100%	1	20	37,662	37,296	12.01%	12.01%	366	0.55%	0.55%	0.98	21.72	11.5
9	81-90%	20	33	37,663	36,984	11.91%	23.93%	679	1.03%	1.58%	1.84	11.61	22.3
8	71-80%	33	49	37,663	36,467	11.75%	35.67%	1,196	1.81%	3.39%	3.28	6.50	32.3
7	61-70%	49	69	37,663	35,666	11.49%	47.16%	1,997	3.02%	6.40%	5.60	3.81	40.8
6	51-60%	69	94	37,663	34,725	11.19%	58.35%	2,938	4.44%	10.84%	8.46	2.52	47.5
5	41-50%	94	128	37,663	33,446	10.77%	69.12%	4,217	6.37%	17.22%	12.61	1.69	51.9
4	31-40%	128	177	37,663	31,769	10.23%	79.35%	5,894	8.91%	26.12%	18.55	1.15	53.2
3	21-30%	177	262	37,663	28,941	9.32%	88.68%	8,722	13.18%	39.30%	30.14	0.71	49.4
2	11-20%	262	469	37,663	23,546	7.58%	96.26%	14,117	21.33%	60.63%	59.95	0.36	35.6
1	0-10%	469	1,000	37,662	11,608	3.74%	100.00%	26,054	39.37%	100.00%	224.45	0.09	0

Figure 17: Kolmogorov-Smirnov Curve for the Best Model



The Cost of Simplicity

Table 25: All of the Best Least Models Found in this Project

Number of Variables in the Model	Percent of Concordant Pairs	Percent of Discordant Pairs	Percent of Tied Pairs	AIC	Profit per 1000 Customers	KS Statistic
1	75.0%	24.8%	0.2%	709,478	$ 91,200.70	37.25
2	79.5%	20.5%	0.1%	664,616	$ 102,450.70	42.90
3	81.4%	18.6%	0.0%	636,504	$ 109,269.98	46.21
4	82.2%	17.8%	0.0%	624,574	$ 108,437.96	48.14
5	83.2%	16.8%	0.0%	608,692	$ 109,691.95	49.60
6	83.9%	16.1%	0.0%	600,557	$ 112,082.18	51.09
7	84.1%	15.9%	0.0%	598,266	$ 112,423.91	51.50
8	84.3%	15.7%	0.0%	595,914	$ 112,516.78	51.66
9	84.6%	15.4%	0.0%	588,347	$ 112,817.36	51.88
10	84.8%	15.2%	0.0%	586,277	$ 113,172.85	52.02
12	85.0%	15.0%	0.0%	584,012	$ 114,540.22	53.23

The model discussed in the previous section consists of twelve variables in total. It is desired to have both a model that performs well and a model that is parsimonious. The chosen best model achieves this, but what are the costs associated with a simpler model? Using the same process described at the beginning of this phase, ten additional models were found. Table 25 displays the percentage of concordant, discordant, and tied pairs, the profit per 1,000 customers, and the KS statistic for each model. Table 26, at the top of the next page, displays the list of variables that create each of these models. From Table 25, using the model with only one predictor results in a significant reduction in the performance. This model would achieve $91,200 of profit per 1,000 customers. Adding the second predictor increases the percentage of concordant pairs and achieves an additional $11,250 (102,450 – 91,200) in profit per 1,000 customers. Adding a third, fourth, fifth and sixth predictor significantly increases the amount of profit acquired by the model and achieves $112,082 in profit per 1,000 customers. After adding the sixth predictor to the model, the profit return greatly decreases with additional predictors. The model generated using seven predictors achieves $112,424 in profit per 1,000 customers. This means the addition of the seventh predictor would only generate an extra $342 in profit for every 1,000 customers. Since using less than six predictors would result in a major reduction in profit (compared to the twelve-variable model), the six variable model is considered the best of the least variable models found. Since the next significant gain in profit is achieved by the twelve-variable model, the models generated with either seven, eight, nine, or ten variables will not be considered better than either the six or twelve variable models. Figure 18a displays the profitability of the 10 different models at different predicted probability cutoff points. Figure 18b zooms into the top of the range of the curve and aids in visualizing the profitable differences between the 10 least models. Figure 19 displays the ROC curvers for each of the 10 least variable models. Better classifiers will produce a curve further away from the diagonal line that runs from the origin of the plot to the upper right corner. In each of these plots, the models generated from more than six variables performed similarly to each other. A distinguishable difference can observed between the six and twelve variable models.

Figure 18a:
Profitability by the Predicted Probabilities for the Least Variable Models

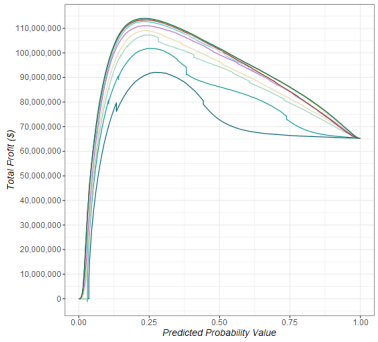


Figure 18b:
Profitability by the Predicted Probabilities for the Least Variable Models

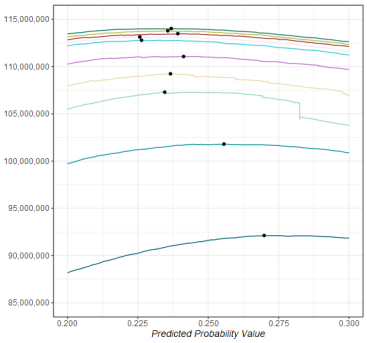


Figure 19:
Receiver Operating Characteristic Curve for the Least Variable Models

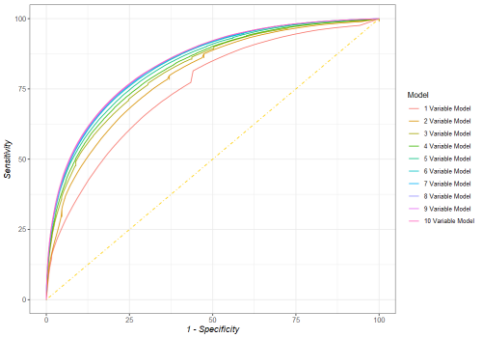


Table 26: All of the Best Least Models Found in this Project

Number of Variables in the Model	Variable List
1	[RADB6]
2	[RADB6, BRR4524_ORD_EQFREQ]
3	[RADB6, BRR4524_ORD_EQFREQ, TRR23]
4	[RADB6, BRR4524_ORD_EQFREQ, TRR23, logodds_BRAVGMOS_EQWID]
5	[RADB6, BRR4524_ORD_EQFREQ, TRR23, logodds_BRAVGMOS_EQWID, BRPCTSAT]
6	[RADB6, BRR4524_ORD_EQFREQ, TRR23, logodds_BRAVGMOS_EQWID, BRPCTSAT, logodds_BRADBM_EQWID]
7	[RADB6, BRR4524_ORD_EQFREQ, TRR23, logodds_BRAVGMOS_EQWID, BRPCTSAT, logodds_BRADBM_EQWID, INQ12_ORD_EQWID]
8	[RADB6, BRR4524_ORD_EQFREQ, TRR23, logodds_BRAVGMOS_EQWID, BRPCTSAT, logodds_BRADBM_EQWID, INQ12_ORD_EQWID, BROPEN_ORD_EQWID]
9	[RADB6, BRR4524_ORD_EQFREQ, TRR23, logodds_BRAVGMOS_EQWID, BRPCTSAT, logodds_BRADBM_EQWID, INQ12_ORD_EQWID, BROPEN_ORD_EQWID, CRATE45]
10	[RADB6, BRR4524_ORD_EQFREQ, TRR23, logodds_BRAVGMOS_EQWID, BRPCTSAT, logodds_BRADBM_EQWID, INQ12_ORD_EQWID, BROPEN_ORD_EQWID, CRATE45, MOSOPEN]
12	[LOCINQS, odds_BRR4524_EQWID, OT12PTOT, CRATE45, BRAVGMOS, logodds_BRADBM_EQFREQ, OBRPTAT, odds_TRATE1_EQFREQ, BRCR1BAL, RADB6, BRPCTSAT, BRR23]

The Least Variable Model

The best of the least variable models was chosen to be the model created using the six variables in Table 28. Three of the variables are in their original form while the other three are transformed variants. Exactly as before with the twelve-variable model, the interpretation of the parameter estimates can be expressed through the logit function. In this model, the estimate for the variable BRPCTSAT is -1.724. Using Euler's number, the interpretation of this estimate is: for every one unit increase in the variable BRPCTSAT, a customer is ($e^{-1.724} = 0.18$) 0.18 times more likely to default on their credit line. This model agreed on 83.9% of 0 - 1 pairs and achieves $112,082 of profit per 1,000 customers (Table 28). This model uses half the number of variables as the best model with only a $2,458 reduction in the profit per 1,000 customers.

Table 28: Six Variable Model Parameter Estimates, Standard Error, and Wald Chi-Square test

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	P-value
Intercept	1.7187	0.0371	2144.41	< 0.0001
RADB6	1.6211	0.0138	13897.80	< 0.0001
BRR4524_ORD_EQFREQ	0.8497	0.0096	7792.59	< 0.0001
TRR23	0.4202	0.0038	12519.91	< 0.0001
logodds_BRAVGMOS_EQW	-2.3249	0.0162	20490.48	< 0.0001
BRPCTSAT	-1.7240	0.0139	15278.61	< 0.0001
logodds_BRADBM_EQWID	-0.4425	0.0049	8151.51	< 0.0001

Table 28: The Percent of Concordant, Discordant, and Tied Pairs for the Chosen Best Least Model

Number of Variables in the Model	Percent of Concordant Pairs	Percent of Discordant Pairs	Percent of Tied Pairs	C	Profit per 1000 Customers
6	83.9%	16.1%	0.0%	0.839	$ 112,082.18

Figure 20 displays the total profit of the six-variable model at differing cutoff points. A vertical line is placed at the predicted probability value that produces the highest total profit (0.22629) which achieves a total profit of $112,796,000 (a $2,394,500 reduction in profit when compared to the twelve-variable model).

Figure 20: Profitability by the Predicted Probabilities of the Six Variable Model

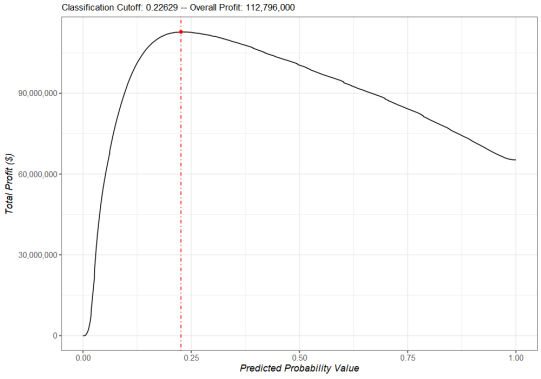


Table 29: Kolmogorov-Smirnov Test for the Chosen Best Least Model

Number of Variables Inmodel: 6

Rank	Quantile	Min Score	Max Score	Number of ACCTS	Number of Goods	Percent of Goods	Cumulative Percent of Goods	Number of Bads	Percent of Bads	Cumulative Percent of Bads	Interval Badrate	G/B Ratio	KS
10	91-100%	5	26	37,600	37,112	11.95%	11.95%	488	1.30%	0.74%	1.31	9.21	11.2
9	81-90%	26	36	37,725	36,989	11.91%	23.87%	736	1.95%	1.85%	1.99	6.11	22.0
8	71-80%	36	52	37,663	36,358	11.71%	35.58%	1,305	3.46%	3.82%	3.59	3.38	31.8
7	61-70%	52	73	37,663	35,484	11.43%	47.01%	2,179	5.79%	7.11%	6.14	1.98	39.9
6	51-60%	73	99	37,663	34,499	11.11%	58.12%	3,164	8.40%	11.89%	9.17	1.32	46.2
5	41-50%	99	130	37,663	33,147	10.68%	68.80%	4,516	11.99%	18.72%	13.62	0.89	50.1
* 4	31-40%	130	176	37,663	31,593	10.18%	78.98%	6,070	16.12%	27.89%	19.21	0.63	51.1 *
3	21-30%	176	257	37,663	29,165	9.39%	88.37%	8,498	22.56%	40.73%	29.14	0.42	47.6
2	11-20%	257	465	37,663	23,871	7.69%	96.06%	13,792	36.62%	61.57%	57.78	0.21	34.5
1	0-10%	465	997	37,662	12,230	3.94%	100.00%	25,432	67.53%	100%	207.95	0.06	0

Figure 21: Kolmogorov-Smirnov Curve for the Six Variable Model

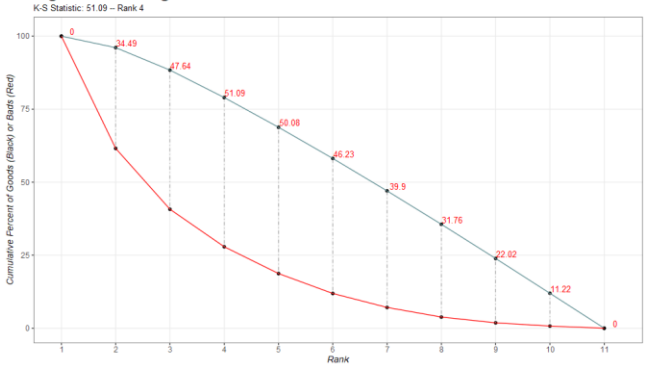


Table 29 displays the results from the KS test conducted on the six-variable model. The KS statistic value of 51.1 was observed in the decile rank 4 and is only 2.1 less than the KS statistic produced by the twelve-variable model. In Figure 21, this value is the distance between the points on the two lines at rank 4.

Conclusion

The best model overall was achieved using twelve variables. This model had a concordant percentage of 85% and acquired a total profit of $115,190,500. If the company desires a more parsimonious model, then the model using six variables would be the best choice. The six variable model achieved a total profit of $112,796,000 and a concordant percentage of 83.9%. Using less variables will significantly impact the profitability of the model (Table 25). Most of the predictors that generated both the twelve and six variable models were in their original forms. This is beneficial due to the extra step in interpretation that comes with including a transformed variant.

Table 25: All of the Best Least Models Found in this Project

Number of Variables in the Model	Percent of Concordant Pairs	Percent of Discordant Pairs	Percent of Tied Pairs	AIC	Profit per 1000 Customers	KS Statistic
1	75.0%	24.8%	0.2%	709,478	$ 91,200.70	37.25
2	79.5%	20.5%	0.1%	664,616	$ 102,450.70	42.90
3	81.4%	18.6%	0.0%	636,504	$ 109,269.98	46.21
4	82.2%	17.8%	0.0%	624,574	$ 108,437.96	48.14
5	83.2%	16.8%	0.0%	608,692	$ 109,691.95	49.60
6	83.9%	16.1%	0.0%	600,557	$ 112,082.18	51.09
7	84.1%	15.9%	0.0%	598,266	$ 112,423.91	51.50
8	84.3%	15.7%	0.0%	595,914	$ 112,516.78	51.66
9	84.6%	15.4%	0.0%	588,347	$ 112,817.36	51.88
10	84.8%	15.2%	0.0%	586,277	$ 113,172.85	52.02
12	85.0%	15.0%	0.0%	584,012	$ 114,540.22	53.23

Many choices were made while working on this project. These choices impact the model's ability to predict the dependent variable and ultimately the profitability of lending credit. One choice, the choice in the classification cutoff, directly resulted in either more or less people receiving credit. Setting the cutoff point to low would result in more Type II error and fewer people would receiving credit. This is only a loss in potential profit gain and does not result in losing any money. On the other hand, setting the cutoff point to high would result in more Type I error and more people would receive credit. This would result in losing money if the people that received credit defaulted on their credit line. The company provided a profitability equation that accounted for true negative predictions returning $250 per customer and false positive predictions losing approximately $750 per customers (or half of the customers credit line). True positive and false negative predictions do not result in the customer receiving credit and do not result in the loss or return of profit.

Figure 20:
Profitability by the Predicted Probabilities of the Six Variable Model