KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING
School of Data Science and Analytics

Choice Exploration in Data Wrangling and the Information Corral

Nathaniel Jones

Faculty Advisor: Dr. Jennifer Priestley

INTRODUCTION

The goal of this project is to use customer performance to build a model that maximizes profitability of subprime credit lending. To do this a binary variable, ‘goodbad,’ is created to indicate whether a customer is a good credit risk (0) or a bad credit risk (1). A logistic regression model is then created from the best independent variables after wrangling and preparing the data. The best models are chosen by simplicity, performance, and profitability.

Initial Data Structure:

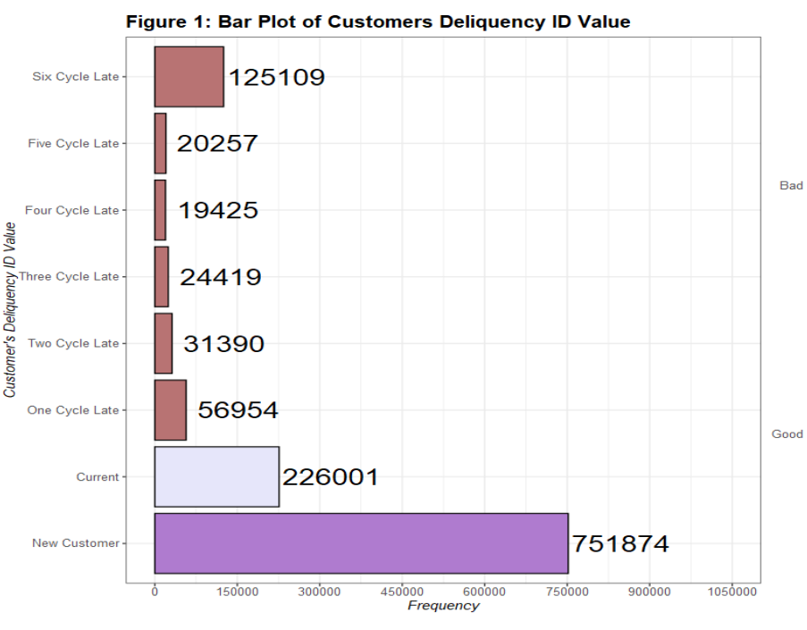
The company has provided two datasets that contain an ID variable, ‘MATCHKEY’, which represents unique customers and will be used to join the datasets together:

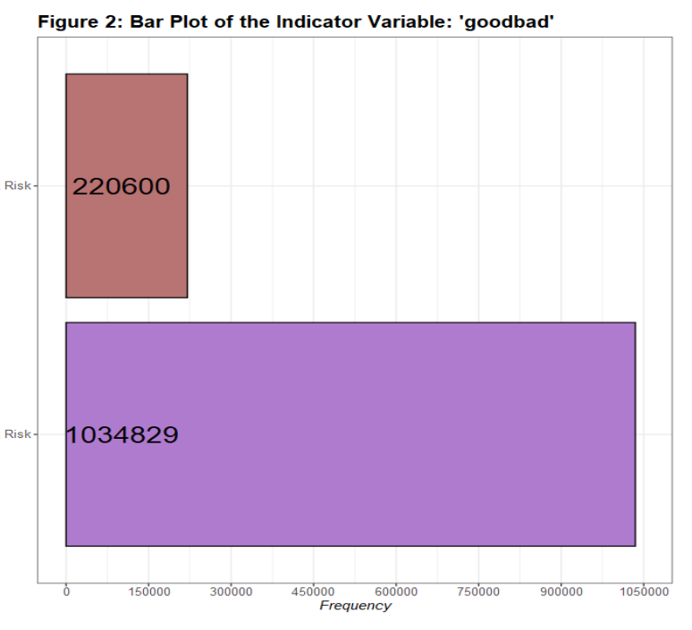
- **CPR dataset:** This dataset observed 1,462,955 customers and collected 338 features from each customer. One variable, ‘MATCHKEY’, identifies unique customers. This file contains potential predictors of credit performance, but the variables have differing levels of completeness.
- **PERF dataset:** This dataset observed 17,244,104 repayment outcomes and collected 18 post hoc features on customer performance. Customers in this dataset are observed monthly, and for each observation the following features were used for this analysis:
  - **MATCHKEY:** Customer’s identifier.
  - **DELQID:** The number of cycles a customer’s payment is late.
  - **CRELIM:** The customer’s credit limit.

The result of merging CPR and PERF together gives a dataset with 337 independent variables, MATCHKEY, DEQLID, and CRELIM for 1,255,429 unique customers.

Creation of the Dependent Variable:

Since the PERF dataset contains monthly records, the observation with the highest value of DELQID will be used to create a goodbad for each customer. Customers with a value of two or less are assigned a value of 0, indicating a good credit risk, while customers with a value greater than two are assigned a value of 1, indicating a bad credit risk.





Data Wrangling

Wrangling the data into a structure good for modelling is required for the merged dataset. The variables within this dataset contain both coded and extreme values that must be removed and replaced with meaningful values. By setting the tolerance of imputed values, the number of variables will be reduced.

- **Coded Values:** Some values contained within the dataset are coded by the company to a value ending with a number between 92-99. There are five types of coding used depending on the number of digits of the variables maximum value (Table 1).
- **Imputation:** Observations with coded values must be replaced with a value that is meaningful but will not introduce new bias to the variable. Imputing to the median will allow the retention of these observations and will not introduce any new variation into the data.
  - **Tolerance of Imputed Values:** Each variable contains a differing percent of coded values (Figure 3). Since variables that require imputing more than 50% of the data will result in a column where most of the rows are the median, these variables contain very little information and need to be dropped from the dataset. 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.
- **Extreme Values:** Apart from the coded values, many of the variables are extremely skewed (Figure 4 Before). These values may truly belong to the domain of the variable but can also be caused by collection error. In either case, extreme values can skew the results and must be handled before modelling. By imputing these values to the median, a portion of the true variance can be lost. Instead, Chebyshev’s theorems can be used to project extreme values to a less extreme part of the domain while keeping the original structure of the variable (Figure 4 After).
- **Chebyshev’s Theorems of Inequalities:** Chebyshev found that for most distributions, at least 75% of the observations are within two standard deviation steps from the mean. In general, the percentage of values within *k* standard deviation steps from the mean is equal to $1-\left(\frac{1}{k^2}\right)$, while the percent of values beyond *k* steps is equal to $\left(\frac{1}{k^2}\right)$.

For this project, a value of 42.7% was chosen to be the tolerance of imputed values through finding an inflection point of the function that interpolates the points in Figure 3. The resulting data frame consisted of 198 independent variables, where the percentage of coded values is 42.7% or less. Furthermore, the interval 4 – 6.5 standard deviation steps was chosen to be the projection range for the extreme values.

Variable Reduction & Preparation

Many of the remaining 198 variables are linearly dependent, which can cause their variance to inflate. Models built with variables that are linearly related have an issue known as multicollinearity and can result in the model having difficulty independently estimating the relationship between the independent variables and the dependent variable. To reduce this issue, a variable clustering process was conducted on the 198 independent variables to select the best set of representatives that retain the most amount of variance.

- **Variable Clustering:** By first grouping variables that are most correlated into clusters, two values can be computed for each variable. One value represents the within cluster R^2 while the other value represents the R^2 for the next closest cluster.
  - **Selecting the Best Representative of a Cluster:** The representative of each cluster should explain most of variance contained within the cluster it represents. Additionally, this representative should also minimally explain the variance contained within the next closely related cluster. Thus, the representative of each cluster should be the variable with the lowest $1-R^2$ ratio, computed by

$$\frac{1-R^2{}_{within}}{1-R^2{}_{Next\ Closest}}.$$

  - **Selecting the Number of Clusters:** Selecting the number of clusters to reduce the data into will result in a loss of variance observed by the full set of variables. Figure 6 displays the variance retained by selecting differing values for the number of clusters. The company has advised that at least 70% of the original variance should be retained after this step, which would result in the data being put into 35 clusters. Requiring the retention of variance to be 90% would result in the data being put into more than double the number of clusters while retaining only 20% more variance. The optimal choice will maximize the proportion of retained variance while minimizing the number of clusters.
- **Discretization Process:** Discretization of a continuous variable is the process of binning the domain of the variable into bins that are equal in interval width, or bins that contain an equal number of observations. In either process, the desired transformation should display a monotonic relation between the ordinal rank of the bins and the probability of default for customers in that bin. After discretizing the variable, two more transformations, the odds ratio and log odds ratio, are created from the probability of default of each bin.

For this project, the data was clustered into 56 cluster representatives which retained 80.11% of the variance from the full set. After selecting the representatives of each cluster, an additional 21 variables were dropped in order to further reduce variance inflation factors.

Model Development and Performance

The logistic regression model will be built from a subset of the 35 independent variables and their transformations, but only one version of each variable should be used at a time. From only the pool of the original 35 independent variables, there are 2^{35}, or 34 billion, different models that can be created. Furthermore, choosing which version of the variable to use can add up to six more model variants per variable. Brute forcing the best model is computationally expensive and cumbersome.

- **Feature Selection:** One method of selection, backward selection, creates a model from the full list of variables and removes variables found to be insignificant. Another method of selection, forward selection, begins with an empty model and individually enters the most important variable. The Wald Chi-Square Statistic, which is the squared ratio between the variable’s estimate and the standard error produced by the model, is used to evaluate the importance of each variable, but maximizing concordance is equally important to selecting the best subset of model features.
- **Classification Cutoff:** The classification cut point determines whether a customer in the dataset is classified as a one or a zero. Setting this point to differing values across the range of probabilities output by the model changes the number of correctly and incorrectly predicted customers, and in turn the model’s profitability.
- **Model Evaluation:** Evaluating the fit and predictability of the model is important to developing a useful model. Model fitness is evaluated by selecting model variants that produce lower values for the AIC statistic. The model’s predictability is assessed by selecting variants that produce higher values for the percent of concordant pairings (C-statistic), the estimated Area Under the ROC Curve, and the Kolmogorov-Smirnov (KS) statistic.
  - **Concordance:** The outcome of a logistic regression model is a value that indicates the probability that the observation is a one (1), or, for this project, a bad credit risk. If the model performs as expected, then customers observed to have a goodbad value of 1 should produce higher probabilities than customers observed with a value of 0. A “concordant” pair is when the probability value produced by a good credit risk customer is lower than the probability value produced by a bad credit risk customer. In the opposite case, the pair is “discordant.”
  - **ROC Curve:** The area under the ROC curve can be used to gauge the performance of a model by assessing the sensitivity and specificity of the model. The vertical axis is the True Positive Rate (Sensitivity) while the horizontal axis is the False Positive Rate (1- Specificity).
  - **Kolmogorov-Smirnov (KS) Test:** The standard measure of model strength and performance in credit scoring is the Kolmogorov-Smirnov statistic. It measures the maximum difference in the shape, center, and spread between the distribution of cumulative good credit risk customers and cumulative bad credit risk customers. Higher KS statistics suggest better model performance and discriminatory power.
- **Profitability:** Profitability of a model is heavily influenced by the classification cutoff point. Choosing the correct cut off point can gain or lose the company money. The company has given information that correctly identifying good credit risk customers as good credit risks (True Negatives) earns the company approximately $250 and incorrectly identifying bad credit risk customers as good credit risks (False Negatives) loses the company approximately half of the credit limit given to that customer. For example, if a bad credit risk customer was predicted to be a good credit risk and given a $1500 credit line, then the company is positioned to lose $750 due to misclassification. Overall, the profitability of a model can be computed by:

$$profit = (\$250\ *True\ Negative) + (\$-750\ *False\ Negative) + (\$0\ *True\ Positive) + (\$0\ *False\ Positive)$$

CONCLUSIONS

- **Project Choices:** Choosing the correct model ultimately depends on the goal of the project. The goal of this project was to build a model that maximizes profitability while remaining interpretable.
  - To maximize profitability conservative choices, such as using the customer’s highest DELQID value to create the dependent variable, were chosen to limit risk.
  - To retain interpretability, certain methods to reduce the number of variables, such as Principal Component Analysis, were not used. In addition, during the feature selection phase greater consideration was given to the original form of the variables over the transformed versions.
- **Model Performance:**
  - **Concordance:** Table 4 displays the percentage of concordant, discordant, and tied pairs for the 12 best models. The best concordance (85%) was found using 12 variables. This table shows that the models using seven or more variables returned minimal additional concordance relative to the model created from six.
  - **ROC Curve and KS Test:** The model that produced the highest Kolmogorov-Smirnov statistic was the model containing 12 variables, with a difference of 53.21% found between the 3rd and 4th decile. In comparison, the model created from six variables produced a difference of 51.1% in the same decile.
  - **Profitability:** Figure 8 displays the overall profit curve across the each of the model outputs. The models created from less than four variables have an overall profit curve less than using one of the models with four or more variables. Table 5 numerically displays the overall profit for each model. From this table, models using more than six variables gain little additional profit relative to the model using six. For this reason, the model using six variables was the chosen as the best model.
- **K-means Clustering:** A K-means clustering algorithm was ran on the twelve-variable model after standardizing the data using the Median Absolute Deviation measure. It was found that four clusters produced the best CCC, Pseudo F, and Pseudo T-Squared statistics (Figure 13).

LinkedIn

GitHub

RESULTS

Figure 3: The Number of Variables by the Tolerance of Coded Values

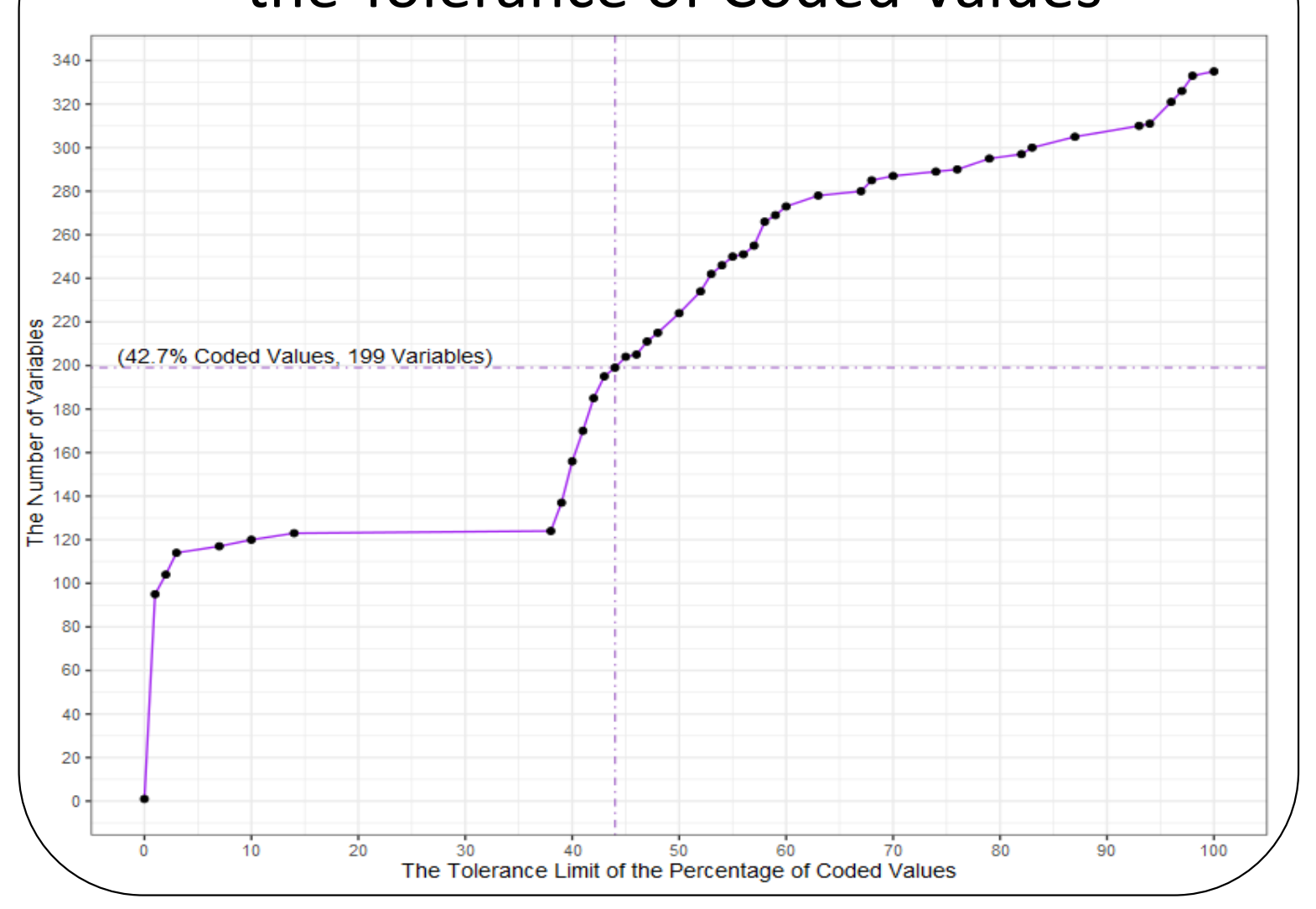


Table 1: Descriptive Statistics of Five Variables Before Decoding

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

Table 2: R-Square 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 3: The R-Square 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

Figure 6: Proportion of Variation Explained by the Number of Clusters

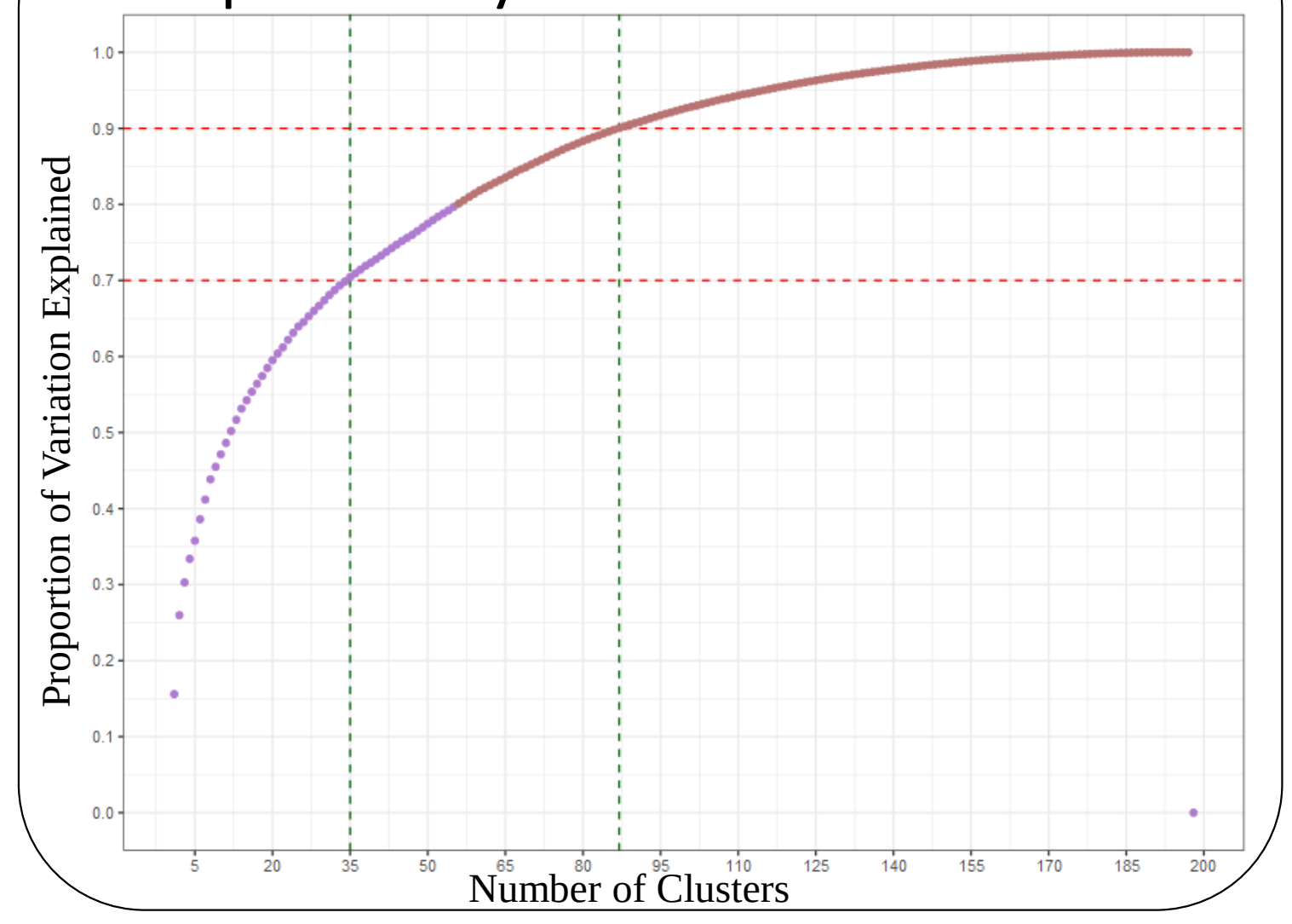


Figure 7: Line Plots for the Probability of Default by different Ordinal Rankings of RBAL (Equal Width)

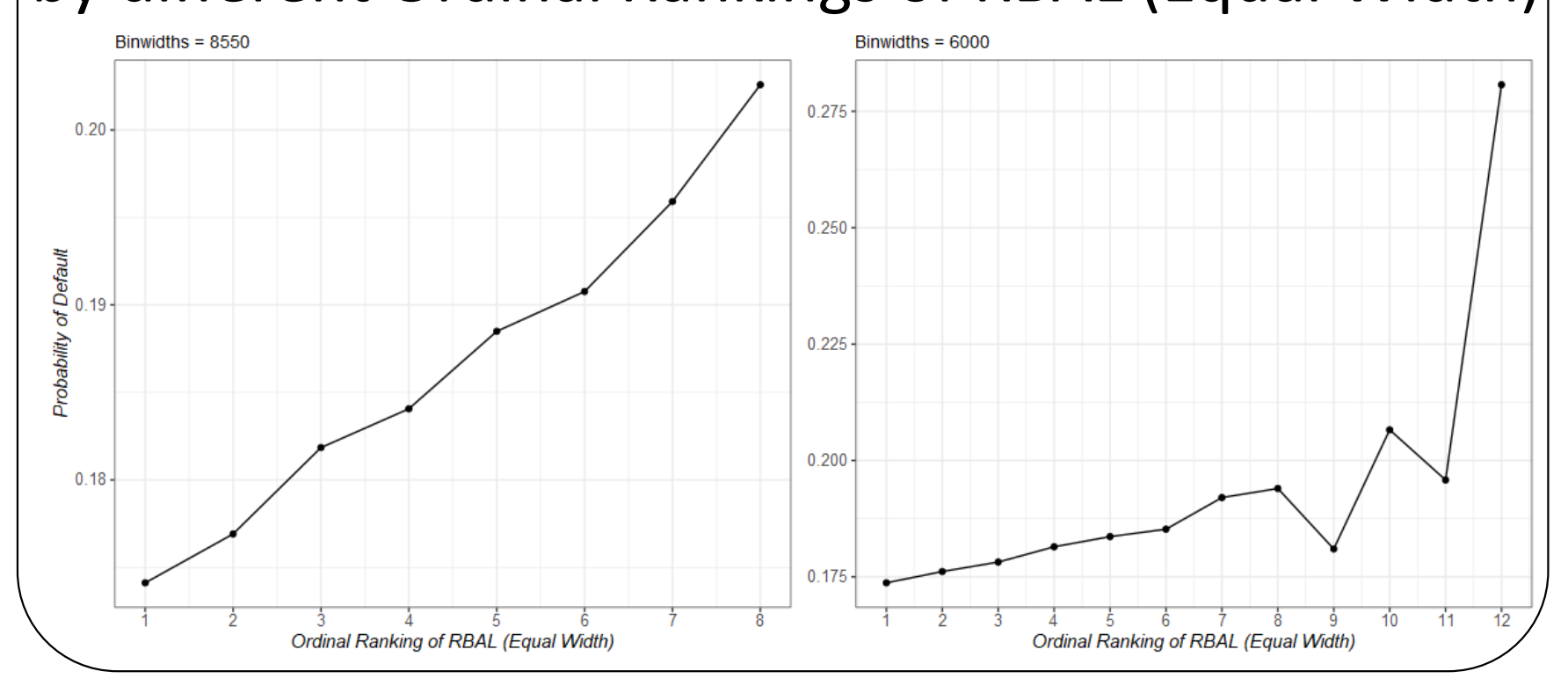


Table 4: Percent of Concordant, Discordant, and Tied Pairs of the 12 Best Models

Number of Variables in the Model	Variable List	% of Concordant Pairs	% of Discordant Pairs	% of Tied Pairs
1	[RADB6]	75.0%	24.8%	0.2%
2	[RADB6, BRR4524_ORD_EQFREQ]	79.5%	20.5%	0.1%
3	[RADB6, BRR4524_ORD_EQFREQ, TRR23]	81.4%	18.6%	0%
4	[RADB6, BRR4524_ORD_EQFREQ, TRR23, …]	82.2%	17.8%	0%
5	[RADB6, BRR4524_ORD_EQFREQ, TRR23, …]	83.2%	16.8%	0%
6	[RADB6, BRR4524_ORD_EQFREQ, TRR23, …]	83.9%	16.1%	0%
7	[RADB6, BRR4524_ORD_EQFREQ, TRR23, …]	84.1%	15.9%	0%
8	[RADB6, BRR4524_ORD_EQFREQ, TRR23, …]	84.3%	15.7%	0%
9	[RADB6, BRR4524_ORD_EQFREQ, TRR23, …]	84.6%	15.4%	0%
10	[RADB6, BRR4524_ORD_EQFREQ, TRR23, …]	84.8%	15.2%	0%
12	[LOCINQS, odds_BRR4524_EQWID, …]	85.0%	15.0%	0%

Figure 4: Histogram and Boxplot of RADB6 Before and After Decoding

(Before) (After)

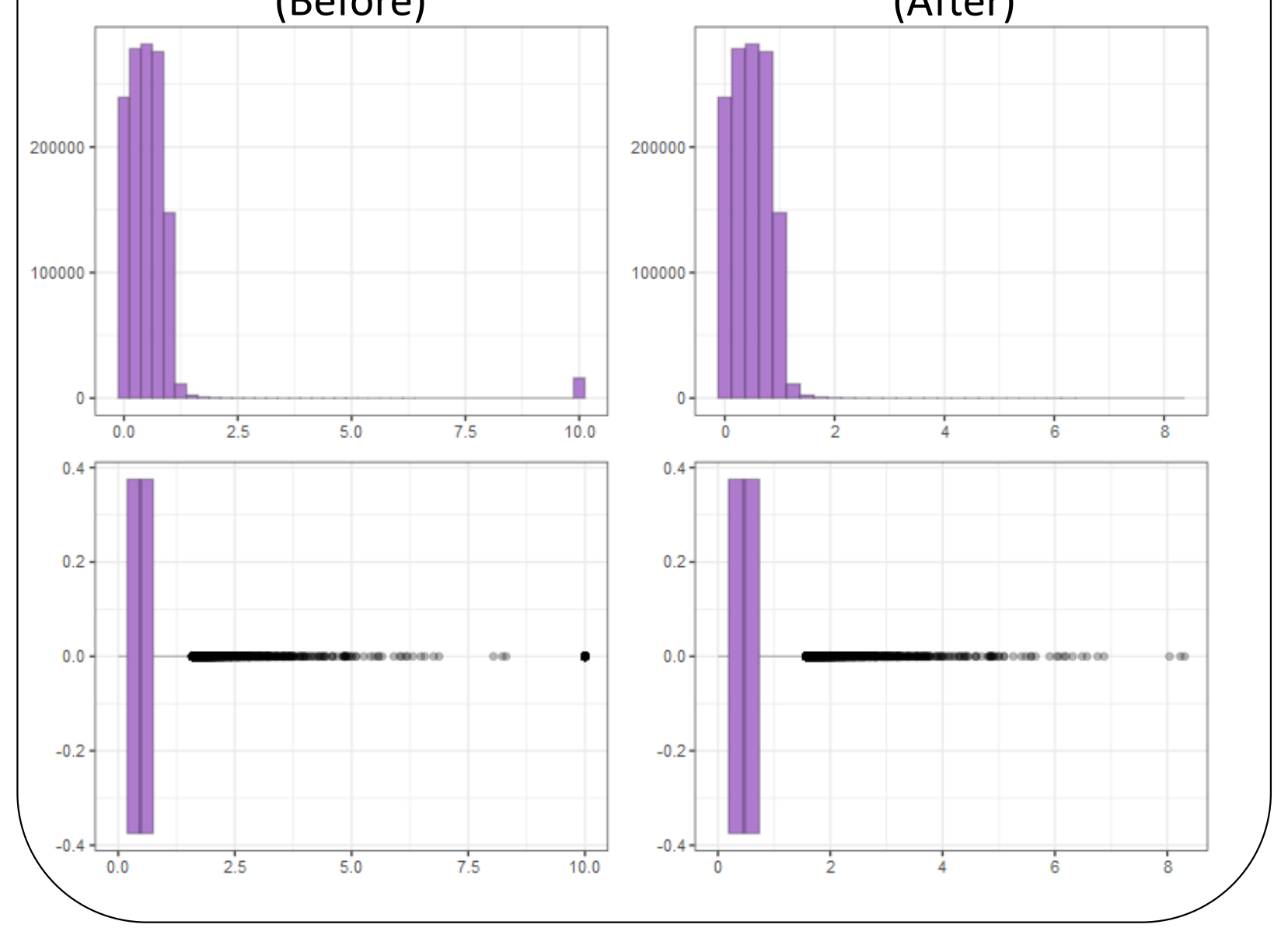


Figure 11: K-Means Clustering of the Model using 12 Variables (k=4)

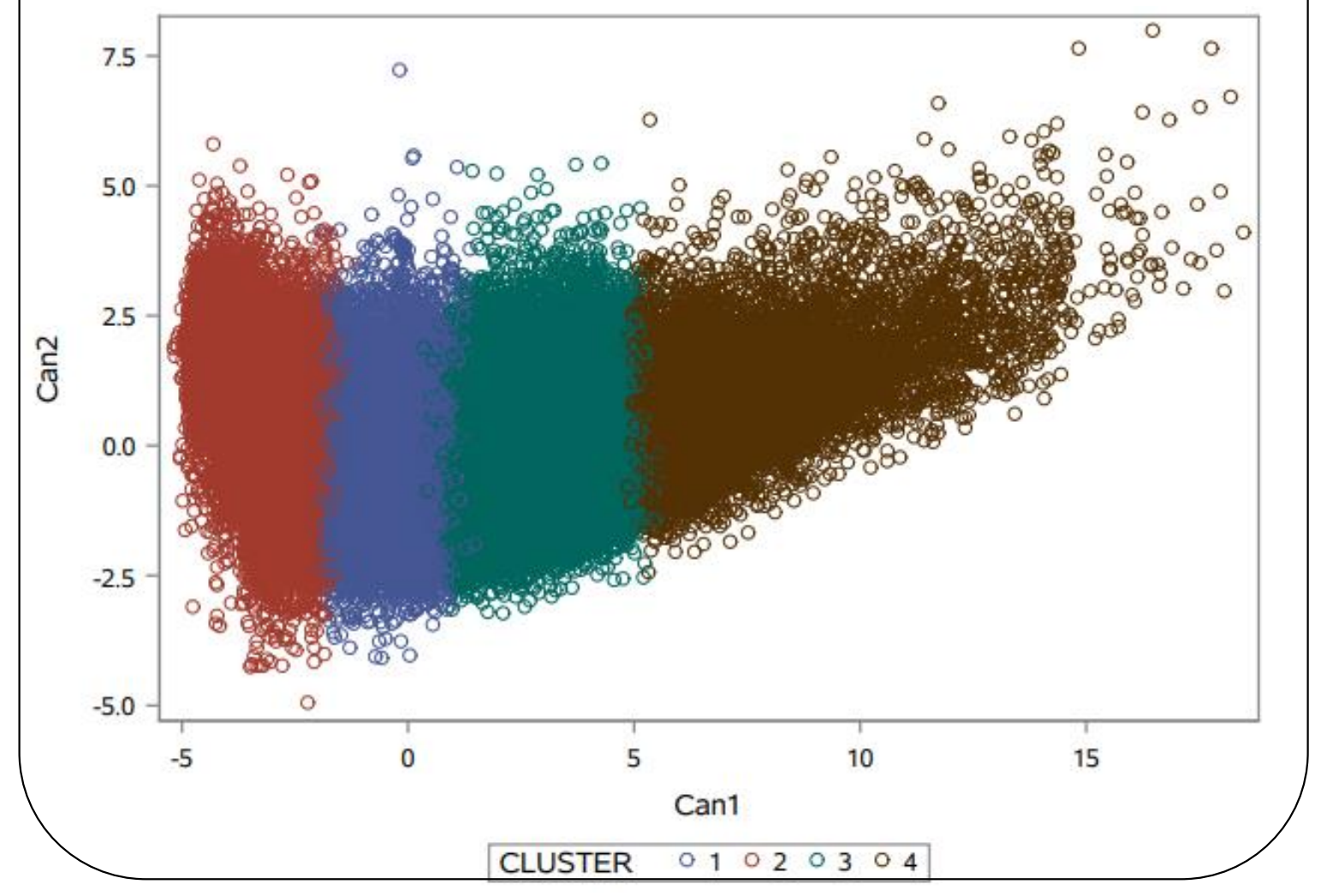


Figure 12: K-Means Clustering of the Model using 12 Variables (k=3)

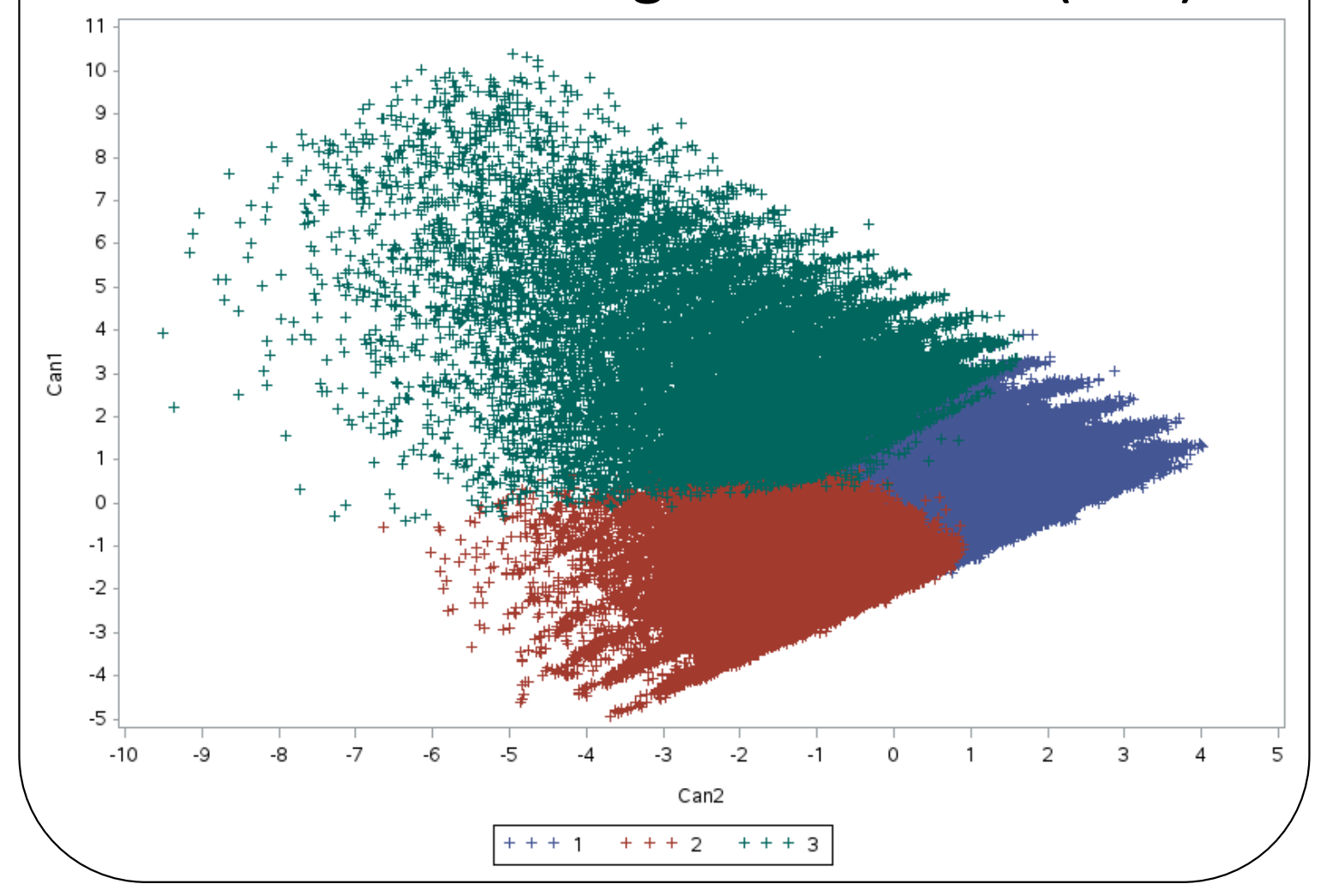


Figure 8: Overall Profit by the probability of Default

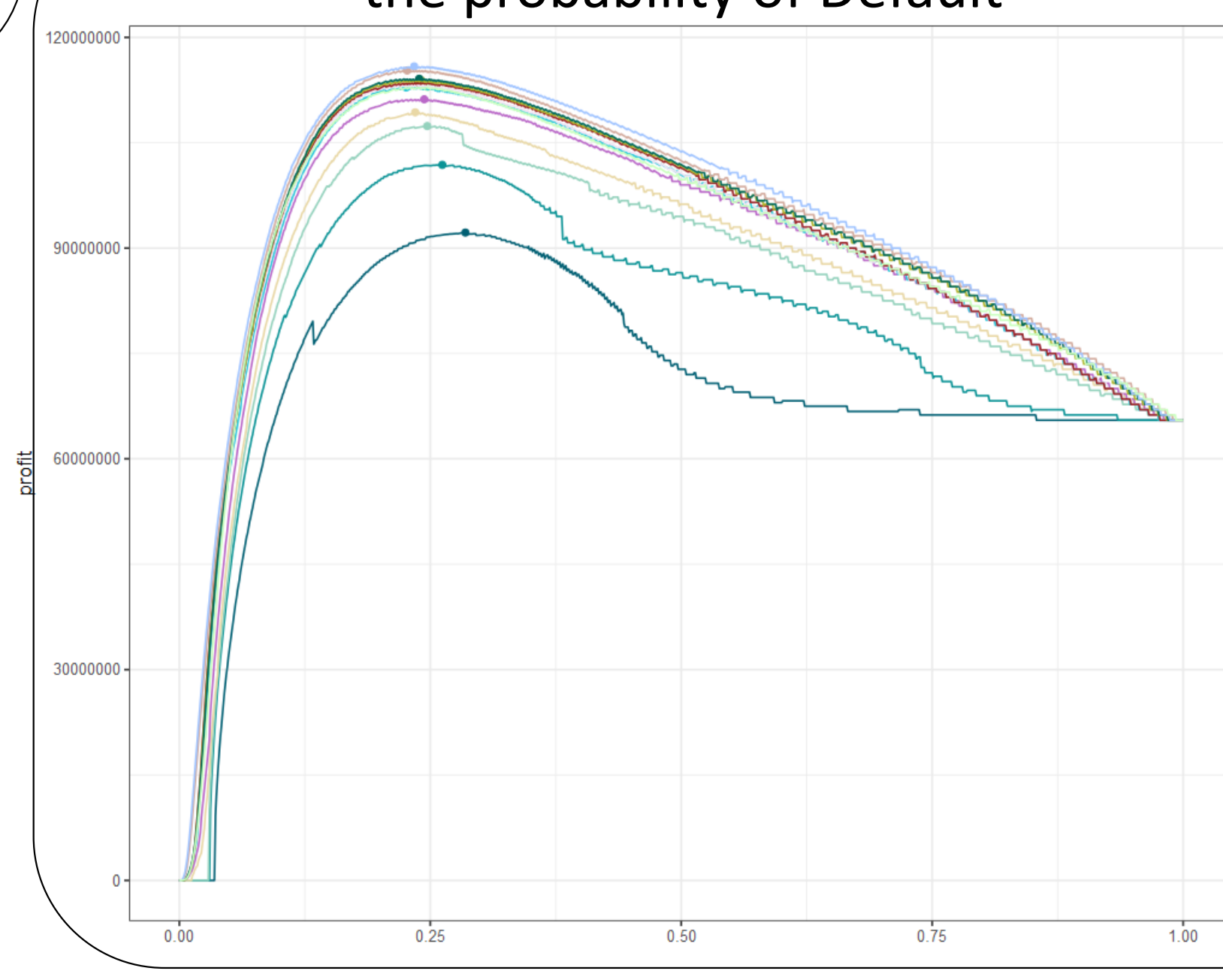


Figure 9: ROC Curves for the Best Models

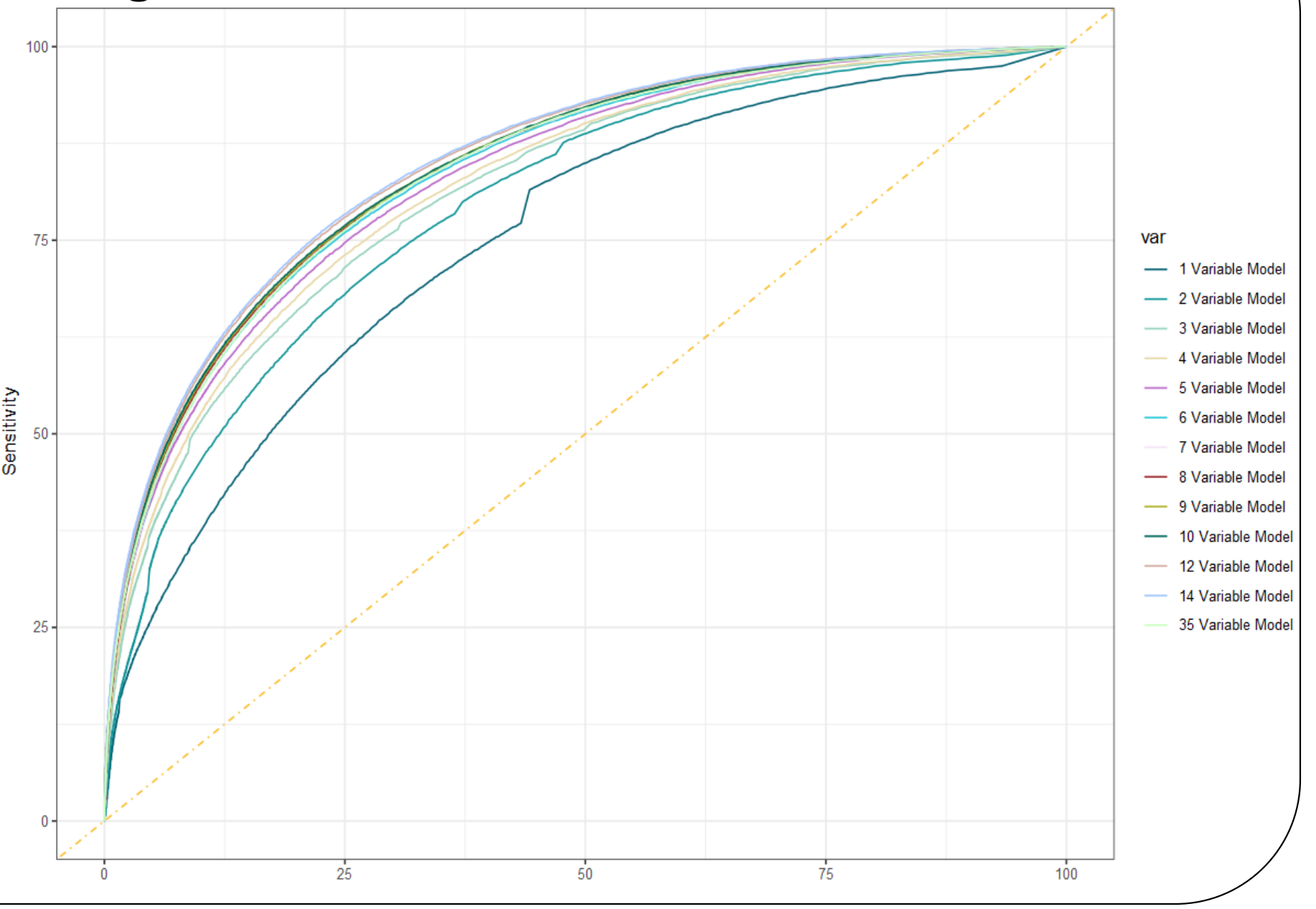


Figure 5: Histogram of RADB6 at differing Standard Deviation Cutoff Points

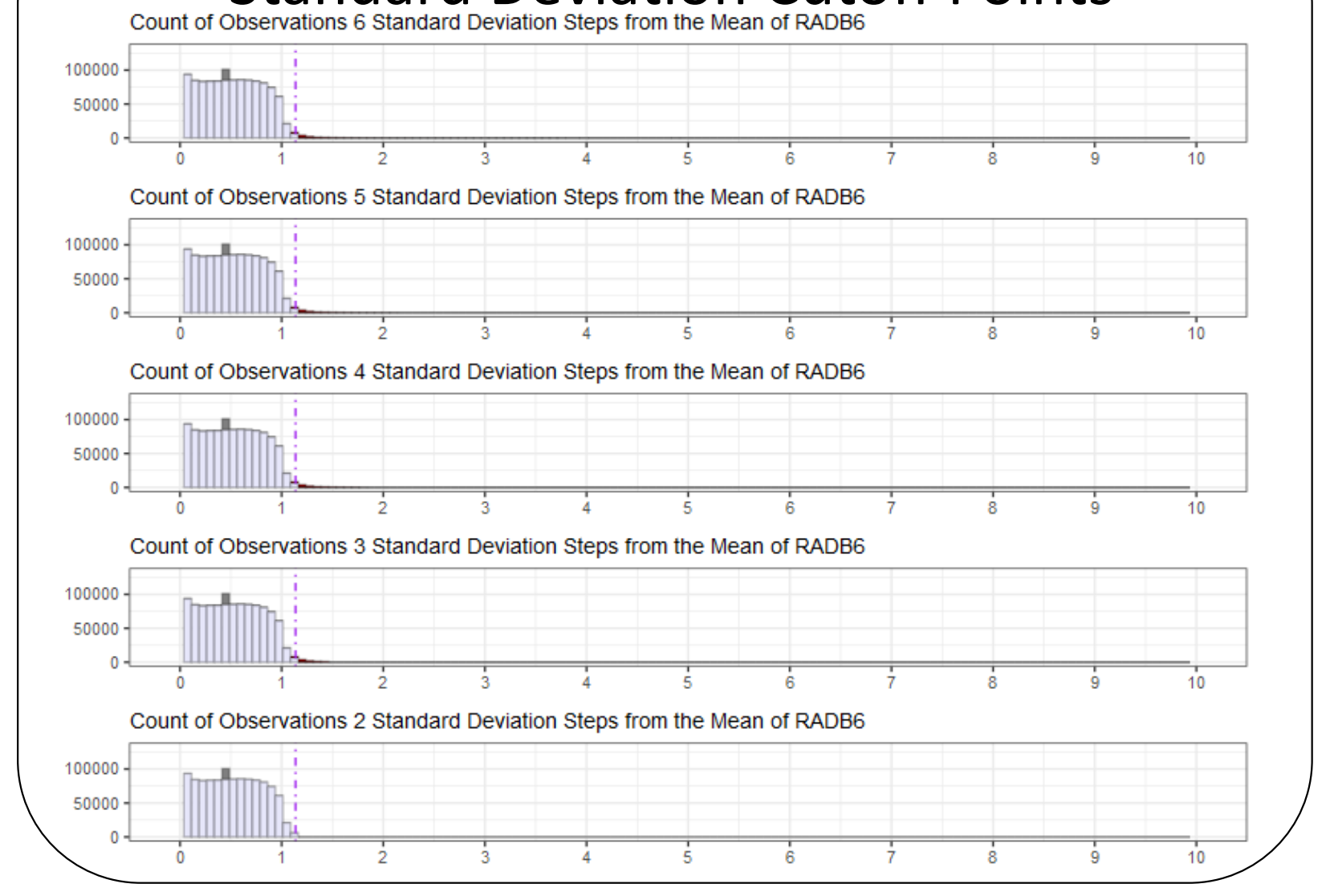


Figure 13: Cubic Clustering Criteria, Pseudo F, and Pseudo T-Squared by the Number of Clusters

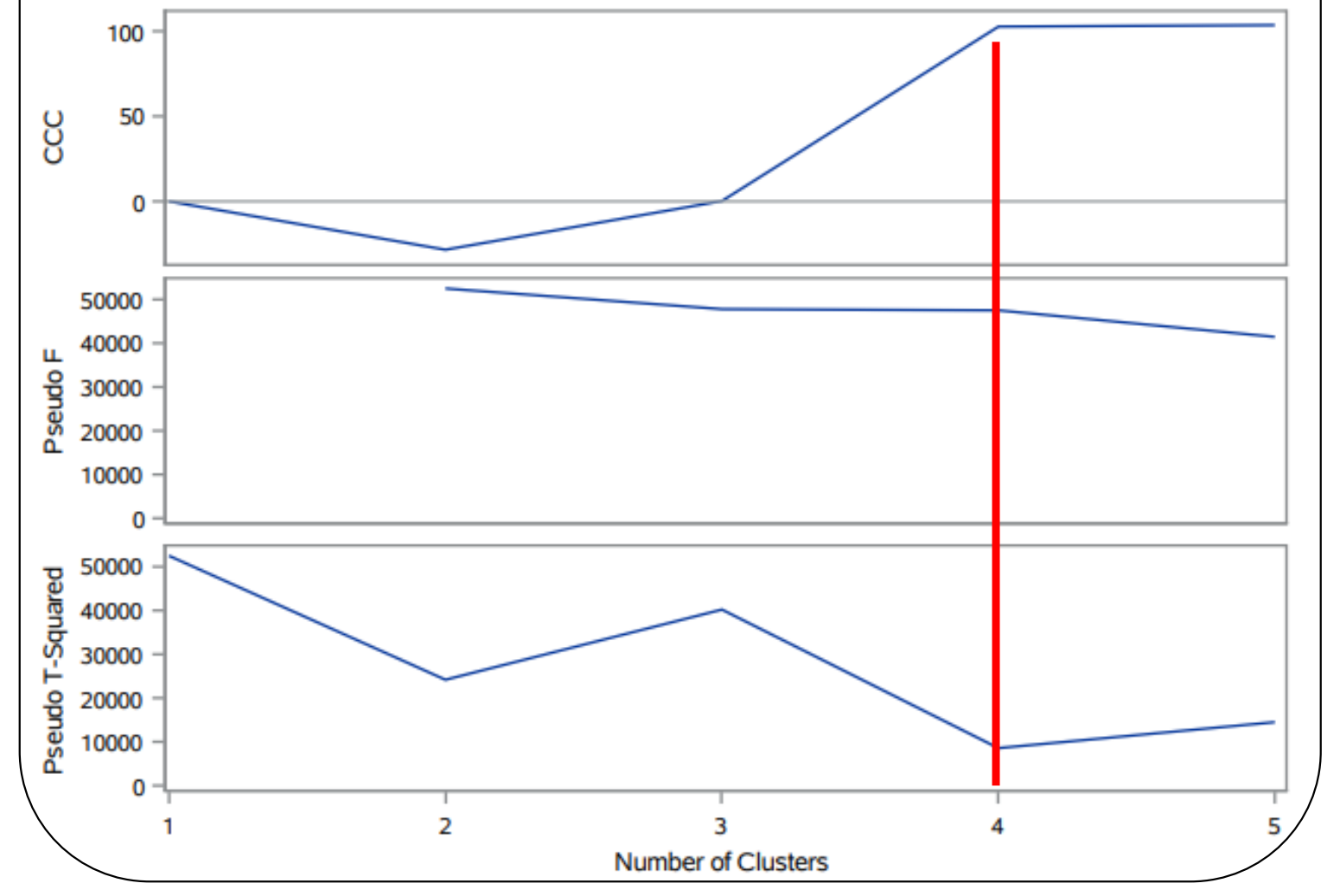


Figure 14: K-Means Clustering of the Model using 12 Variables (k=7)

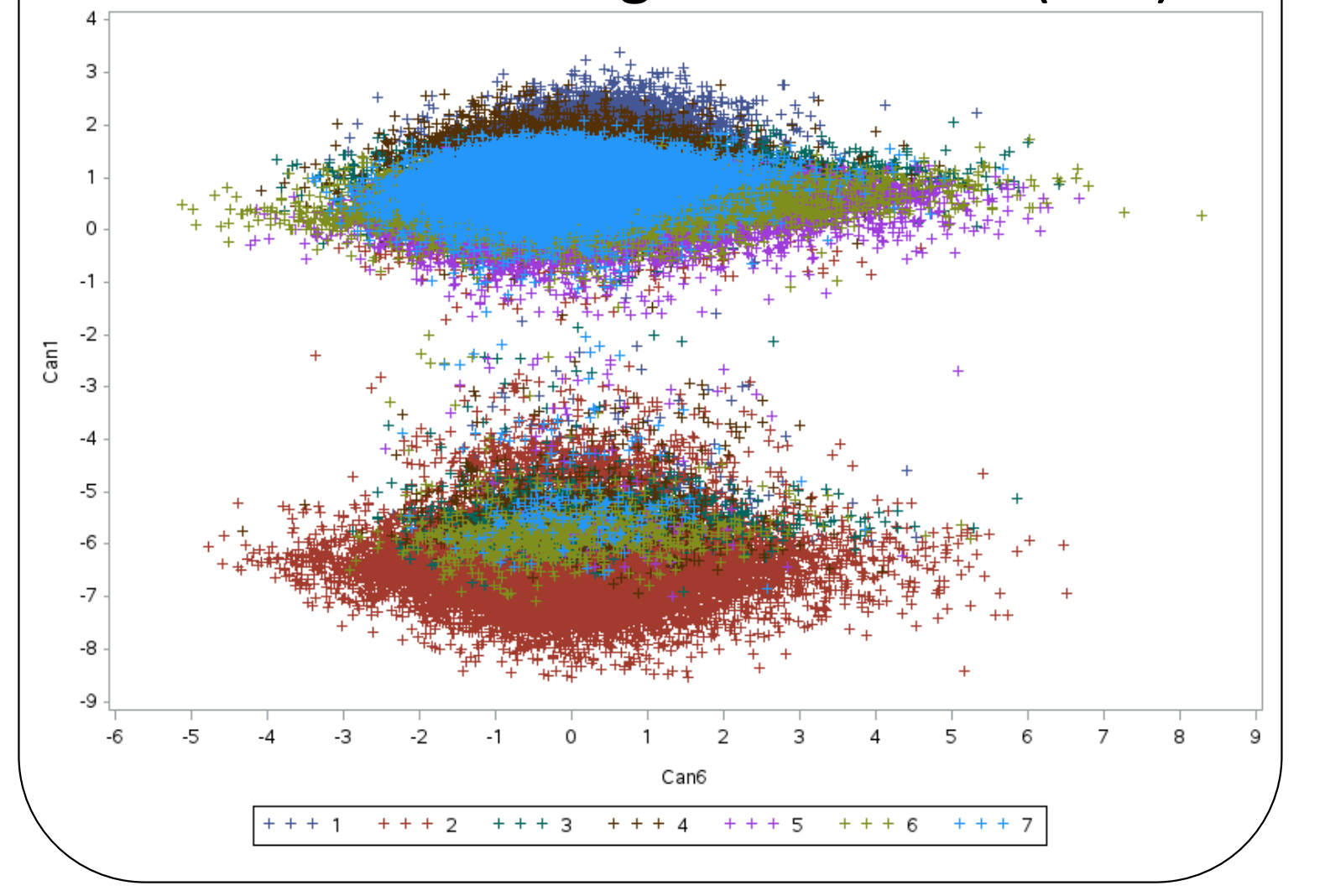


Table 5: Profitability of the 12 Best Models

Number of Variables in the Model	Variable List	Profit Gain	Profit Loss	Overall Profit	Profit Per 1000 Customers
1	[RADB6]	$ 32,631,500	-$15,691,742	$ 16,939,758	$ 89,955
2	[RADB6, BRR4524_ORD_EQFREQ]	$ 33,346,250	-$14,260,781	$ 19,085,469	$ 101,349
3	[RADB6, BRR4524_ORD_EQFREQ, TRR23]	$ 34,004,250	-$13,666,589	$ 20,337,662	$ 107,999
4	[RADB6, BRR4524_ORD_EQFREQ, TRR23, …]	$ 33,444,750	-$13,125,895	$ 20,318,855	$ 107,899
5	[RADB6, BRR4524_ORD_EQFREQ, TRR23, …]	$ 33,947,000	-$13,377,277	$ 20,569,724	$ 109,231
6	[RADB6, BRR4524_ORD_EQFREQ, TRR23, …]	$ 33,252,000	-$12,175,965	$ 21,076,036	$ 111,920
7	[RADB6, BRR4524_ORD_EQFREQ, TRR23, …]	$ 33,100,000	-$11,943,467	$ 21,156,534	$ 112,347
8	[RADB6, BRR4524_ORD_EQFREQ, TRR23, …]	$ 33,759,500	-$12,447,019	$ 21,312,481	$ 112,028
9	[RADB6, BRR4524_ORD_EQFREQ, TRR23, …]	$ 33,759,500	-$12,580,097	$ 21,179,404	$ 112,469
10	[RADB6, BRR4524_ORD_EQFREQ, TRR23, …]	$ 33,706,750	-$12,471,630	$ 21,235,120	$ 112,764
12	[LOCINQS, odds_BRR4524_EQWID, …]	$ 33,243,500	-$11,655,185	$ 21,588,316	$ 114,640
14	[LOCINQS, logodds_TRR23_EQFREQ, …]	$ 33,463,750	-$11,733,008	$ 21,730,743	$ 115,396

Figure 10: Kolmogorov-Smirnov (KS) Curve for the Model Containing 12 Variables

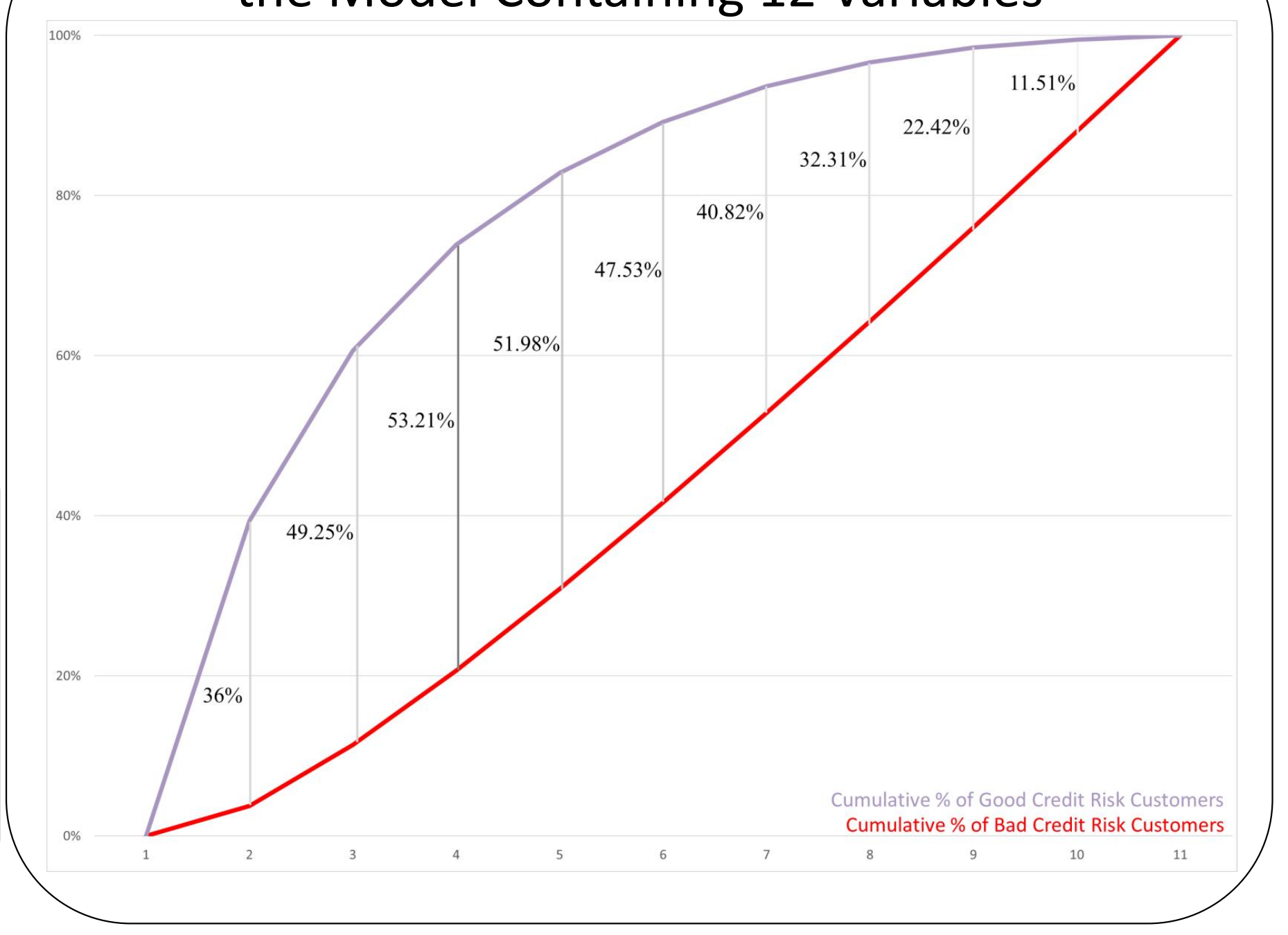


Table 6: Correct/Incorrect Classifications for the 12 Best Models

Number of Variables in the Model	Variable List	Actual 0, Predicted to be a 0 (True Negative)	Actual 0, Predicted to be a 1 (False Positive)	Actual 1, Predicted to be a 0 (False Negative)	Actual 1, Predicted to be a 1 (True Positive)
1	[RADB6]	130,526	24,698	17,219	15,871
2	[RADB6, BRR4524_ORD_EQFREQ]	133,385	21,839	15,443	17,647
3	[RADB6, BRR4524_ORD_EQFREQ, TRR23]	136,017	19,207	14,754	18,336
4	[RADB6, BRR4524_ORD_EQFREQ, TRR23, …]	133,779	21,445	13,436	19,654
5	[RADB6, BRR4524_ORD_EQFREQ, TRR23, …]	135,788	19,436	13,525	19,565
6	[RADB6, BRR4524_ORD_EQFREQ, TRR23, …]	133,008	22,216	12,049	21,041
7	[RADB6, BRR4524_ORD_EQFREQ, TRR23, …]	132,400	22,824	11,754	21,336
8	[RADB6, BRR4524_ORD_EQFREQ, TRR23, …]	134,174	21,050	12,337	20,753
9	[RADB6, BRR4524_ORD_EQFREQ, TRR23, …]	135,038	20,186	12,501	20,589
10	[RADB6, BRR4524_ORD_EQFREQ, TRR23, …]	134,827	20,397	12,387	20,703
12	[LOCINQS, odds_BRR4524_EQWID, …]	132,974	22,250	11,358	21,732
14	[LOCINQS, logodds_TRR23_EQFREQ, …]	133,855	21,369	11,472	21,618