ANALYTICS EXPERIENCE 2016

#analyticsx

Using Logistic Regression to Predict Credit Default

Steven Leopard and Jun Song

Dr. Jennifer Priestley and Professor Michael Frankel

Kennesaw State University
College of Science and Mathematics
Department of Statistics and Analytical Sciences

Executive Summary

The proceeding documentation was created over the course of developing a functional model to predict the risk of default in customers seeking a credit loan using data provided by Equifax Credit Union. In doing so, maximum profitability was achieved by determining the necessary risk of defaulted loans over the potential for profit of successful credit extensions in the sub-prime market.

The original inner merge of the pre and post credit extension data contained 1.4 million observations and 336 predictors. Of these predictors, the binary response was created from the delinquent cycles of the observations. The remaining variables were cleansed of all coding and then transformed into 3 or 6 additional versions depending on the variable's naturally binary status. These versions included both SAS and user defined discretization along with the odds ratio of default and the log function of the ratio. All variance inflation factors 4 or greater were removed to prevent multicollinearity.

After transforming and cleansing, the data was split into two separate sets in which to both build and validate the model. A C-statistic of .812 was found after trimming the model down to a more manageable and cost effective 10 variables. The variables were single versions of each original raw data source and were scored to produce a profitability of $107 per person at a 24% risk of default.

Two more additional analyses were preformed to further optimize the model. The KS test reported that the 31-40% decile of observations yielded the largest difference of good and bad credit risks and the cluster analysis found four groups within the dataset. Of these four groups, cluster 2 produced a profit of $140,000. Once finished, the model provided a profitable way to predict credit default, optimize the sample size needed, and distinguish the ideal group in which to target credit extensions.

Methods

The Methods for this project included:

1. **Data Discovery:** Cleansing, merging, imputing, and deleting.
2. **Multicollinearity:** Removing variance inflation factors.
3. **Variable Preparation:** User and SAS defined discretization.
4. **Modeling and Logistic Regression:** Training and validation files created then modeled.
5. **KS testing and Cluster Analysis:** Optimization of profit and group discovery.

Introduction

This research 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:

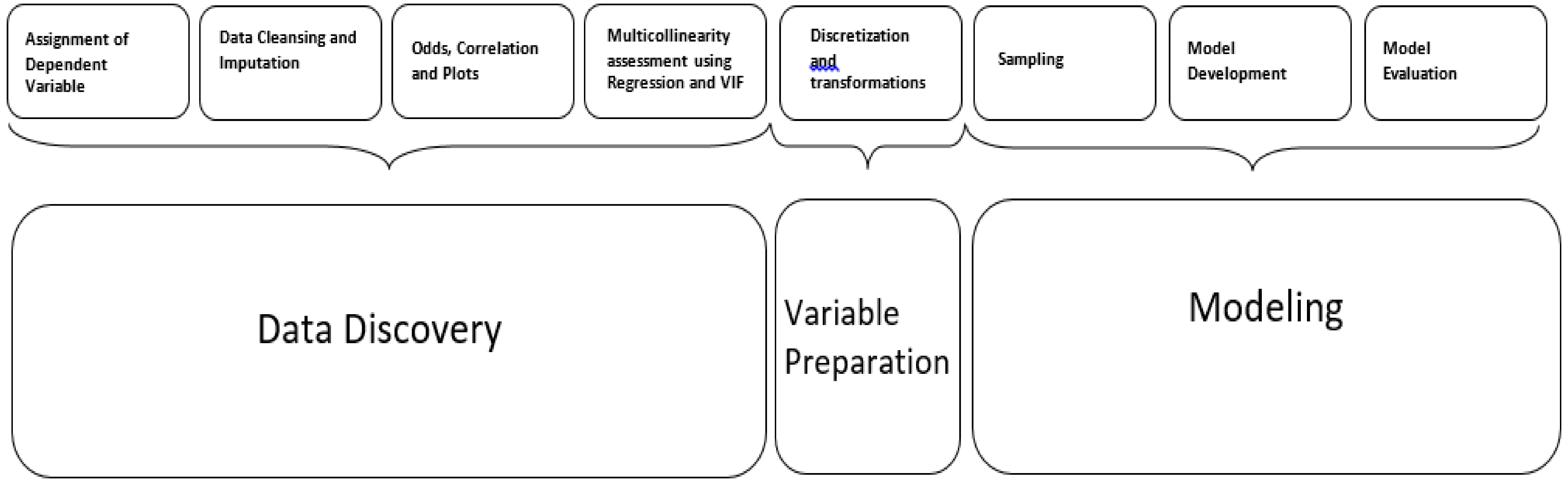
CPR. 1,462,955 observations and 338 variables. Each observation represents a unique customer. This file contains all of the potential predictors of credit performance. The variables have differing levels of completeness.

PERF. 17,244,104 observations and 18 variables. This file contains the post hoc performance data for each customer, including the response variable for modeling – DELQID.

TRAN. 8,536,608 observations and 5 variables. This file contains information on the transaction patterns of each customer.

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

The process of the project included:


Assignment of Dependent Variable
Data Cleansing and Imputation
Odds, Correlation and Plots
Multicollinearity assessment using Regression and VIF
Discretization and transformations
Sampling
Model Development
Model Evaluation
Data Discovery
Variable Preparation
Modeling

ANALYTICS EXPERIENCE 2016

#analyticsx

Kennesaw State University
College of Science and Mathematics
Department of Statistics and Analytical Sciences

Using Logistic Regression to Predict Credit Default

Steven Leopard and Jun Song

Dr. Jennifer Priestley and Professor Michael Frankel

Data Cleansing and Merging

The merge of the raw data was made possible by the ordinal variable MATCHKEY in which customers with the same value for this variable from both datasets were included in an inner merge, or the intersection of the two datasets by the variable MATCHKEY. For the case in which duplicate MATCHKEYs exist, we pick the highest value for DELQID to minimize risk.

The variables in the CPR dataset provide information on clientele before credit is extended and will be used as a basis for prediction. The PERF dataset provides information post credit approval and cannot be used for prediction alone but rather justification of any predictions. More specifically, the PERF variable DELQID will be used as the response or the value we are trying to predict. DELQID is a quantitative variable numbered 0-7 in which each number specifies a costumer's current payment status i.e.; a person given a 0 is too new to rate, 1 signifies the person is in the current payment cycle, 2 signifies that the person is one cycle late, 3 is two cycles late, and so on. After using SAS to merge the two datasets we are left with 1,743,505 observations and 356 variables.

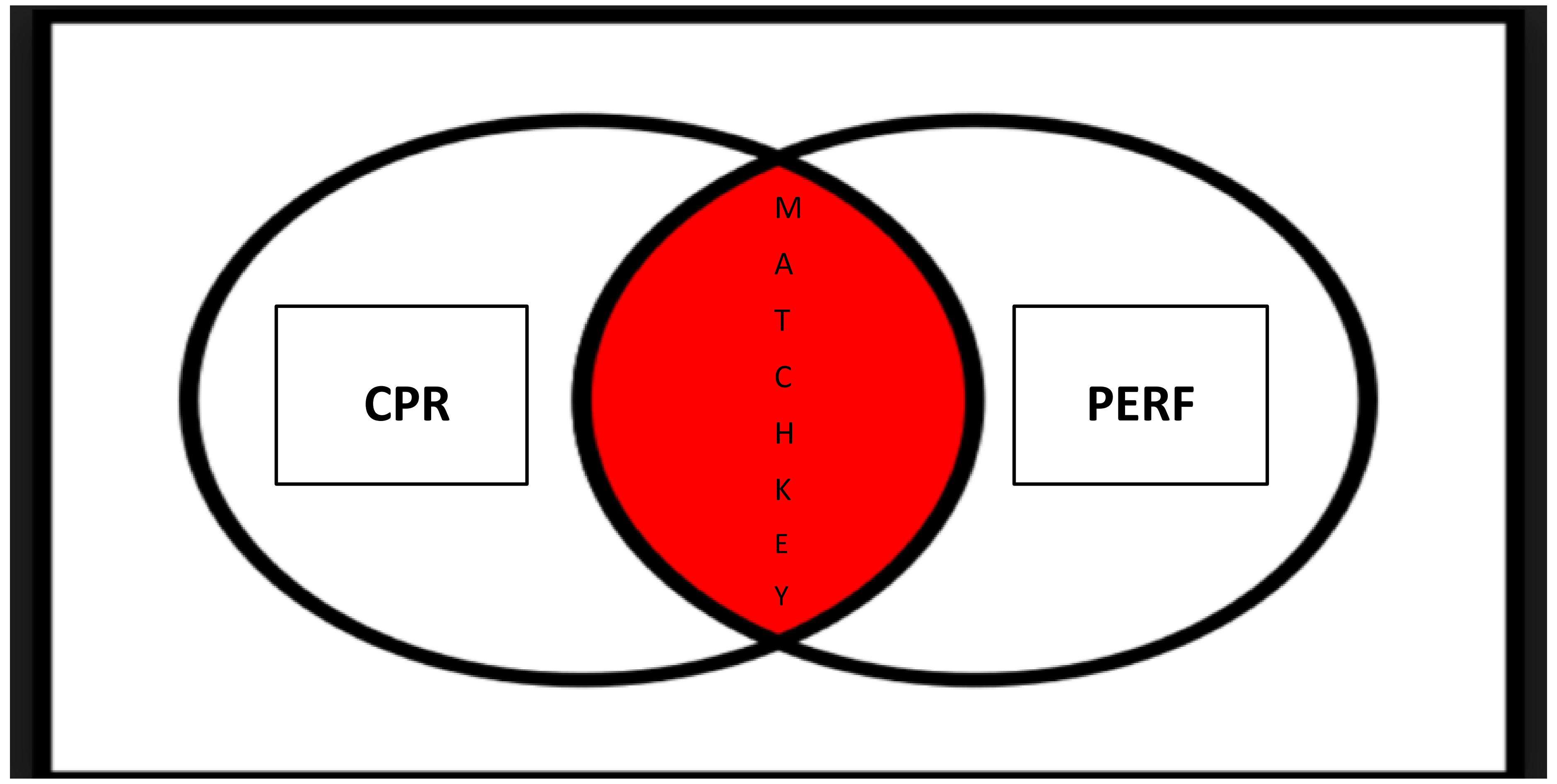
Imputing Coded Variables

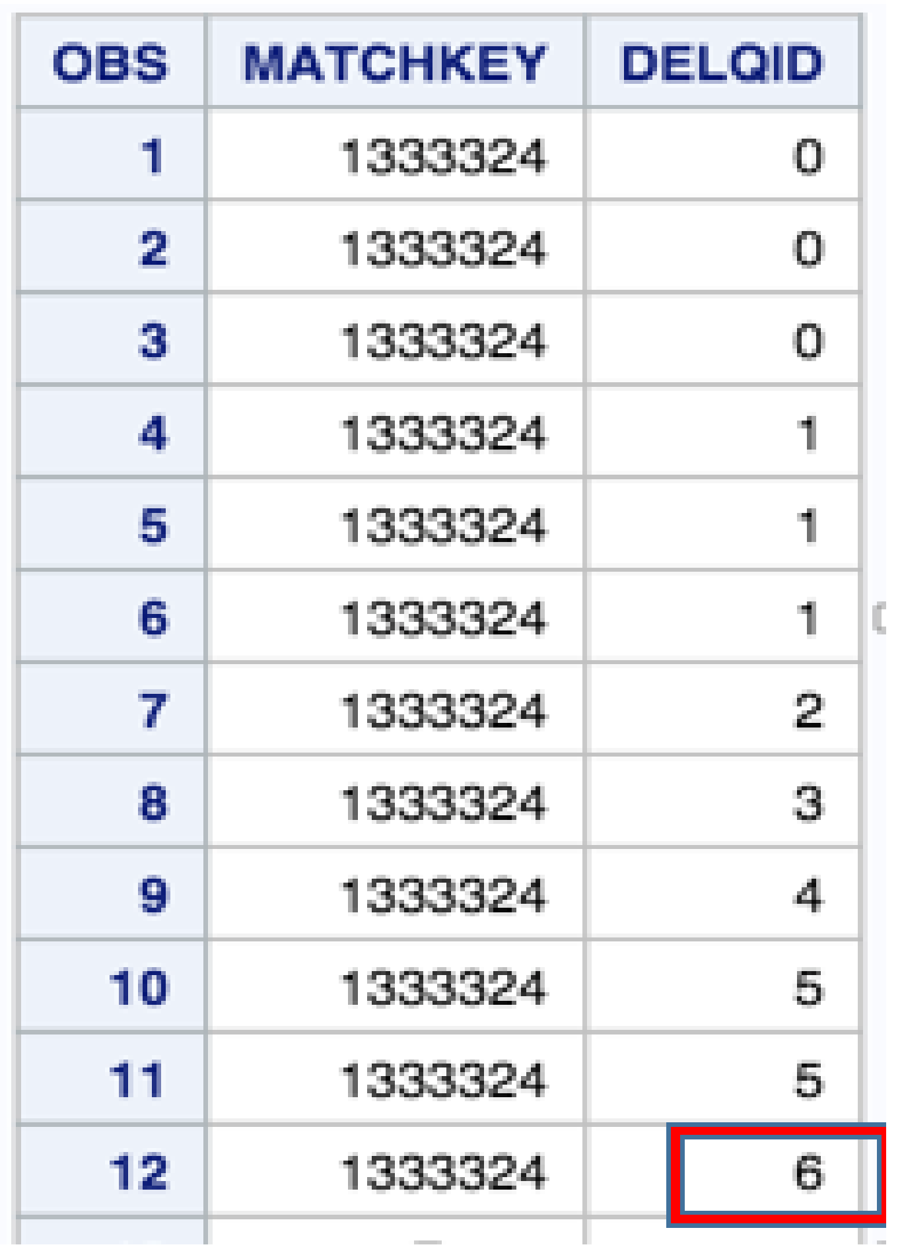
For this part of Data Discovery, we deal with coded values and create our binary response variable. Some imputing was done by hand before we applied the macro and the figures below are an example of AGE before and after. Ages equal to 99 were in fact coded values and reset to the median value of age (47); we can see the spike leave from 99 and go to 47. We use the median in our macro instead of mean since most of the data is skewed and for normally distributed data the mean and median are nearly equal.

For the macro, we choose any coded values above 4 standard deviations to be imputed. We choose 4 so that the bulk of the data and any possible non coded outliers were still preserved for both normally distributed and skewed variables. F

The second step for the macro was to delete any variables that had more than 25% coded. We choose 25% simply by observing a pattern in convergence. Originally, we started at more than 80% delete, than 60%, 40%, 25%, 20%, 5%, and eventually choose 25%.

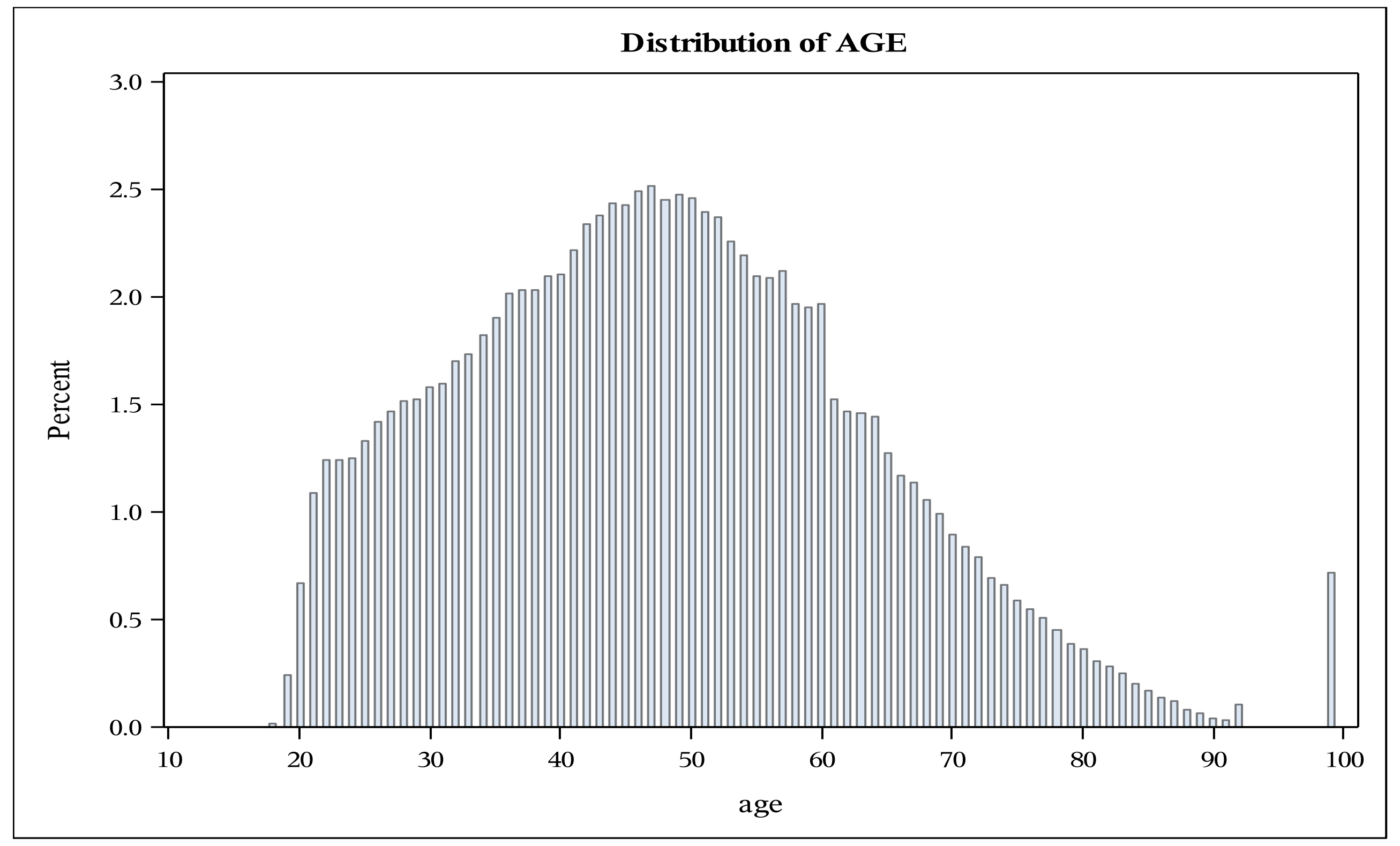
Merging Visualization and Multiple Matchkeys

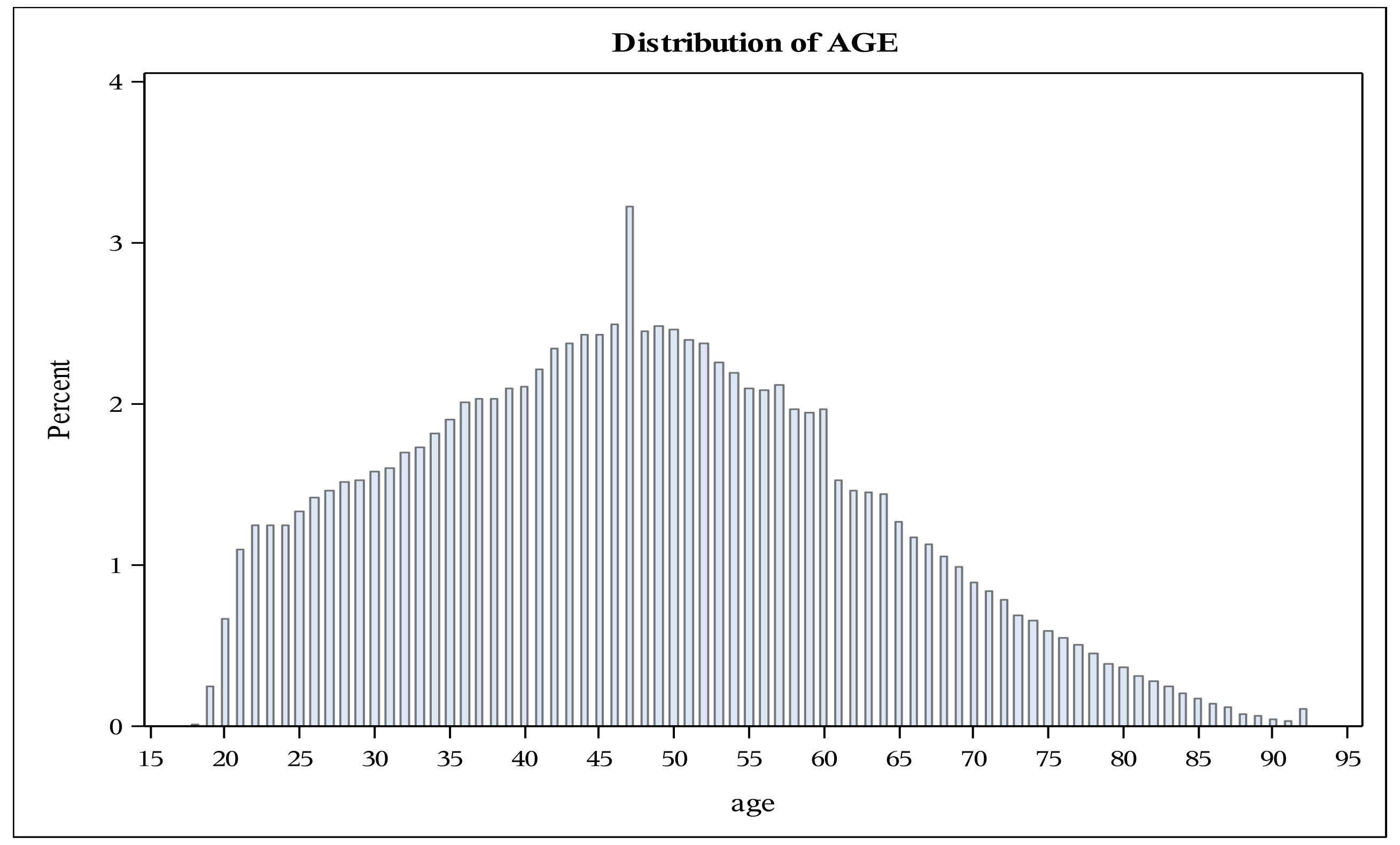




OBS	MATCHKEY	DELQID
1	1333324	0
2	1333324	0
3	1333324	0
4	1333324	1
5	1333324	1
6	1333324	1
7	1333324	2
8	1333324	3
9	1333324	4
10	1333324	5
11	1333324	5
12	1333324	6

Histograms of Age Pre/Post Imputation





ANALYTICS EXPERIENCE 2016

#analyticsx

Kennesaw State University
College of Science and Mathematics
Department of Statistics and Analytical Sciences

Using Logistic Regression to Predict Credit Default

Steven Leopard and Jun Song

Dr. Jennifer Priestley and Professor Michael Frankel

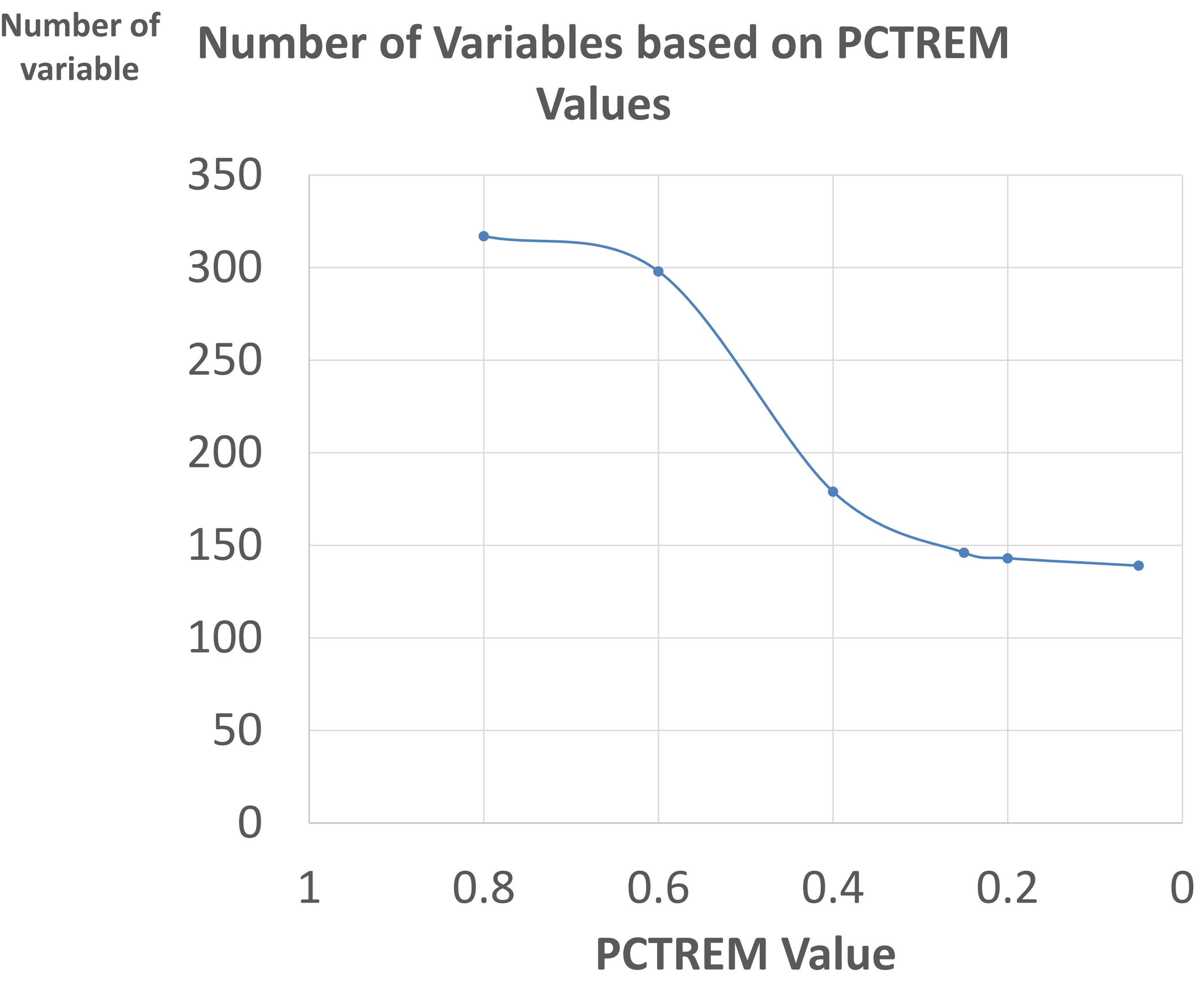
Imputing and/or Deleting Variables

For the macro, we choose any coded values above 4 standard deviations (MSTD) to be imputed. We choose 4 so that the bulk of the data and any possible non coded outliers were still preserved for both normally distributed and skewed variables.

The second step for the macro was to delete any variables that had more than 25% coded. We choose 25% simply by observing a pattern in convergence after running the macro several times. Originally, we started at more than 80% delete, than 60%, 40%, 25%, 20%, 5%, and eventually choose 25%.

Tables Illustrating Cutoff Points for Macro

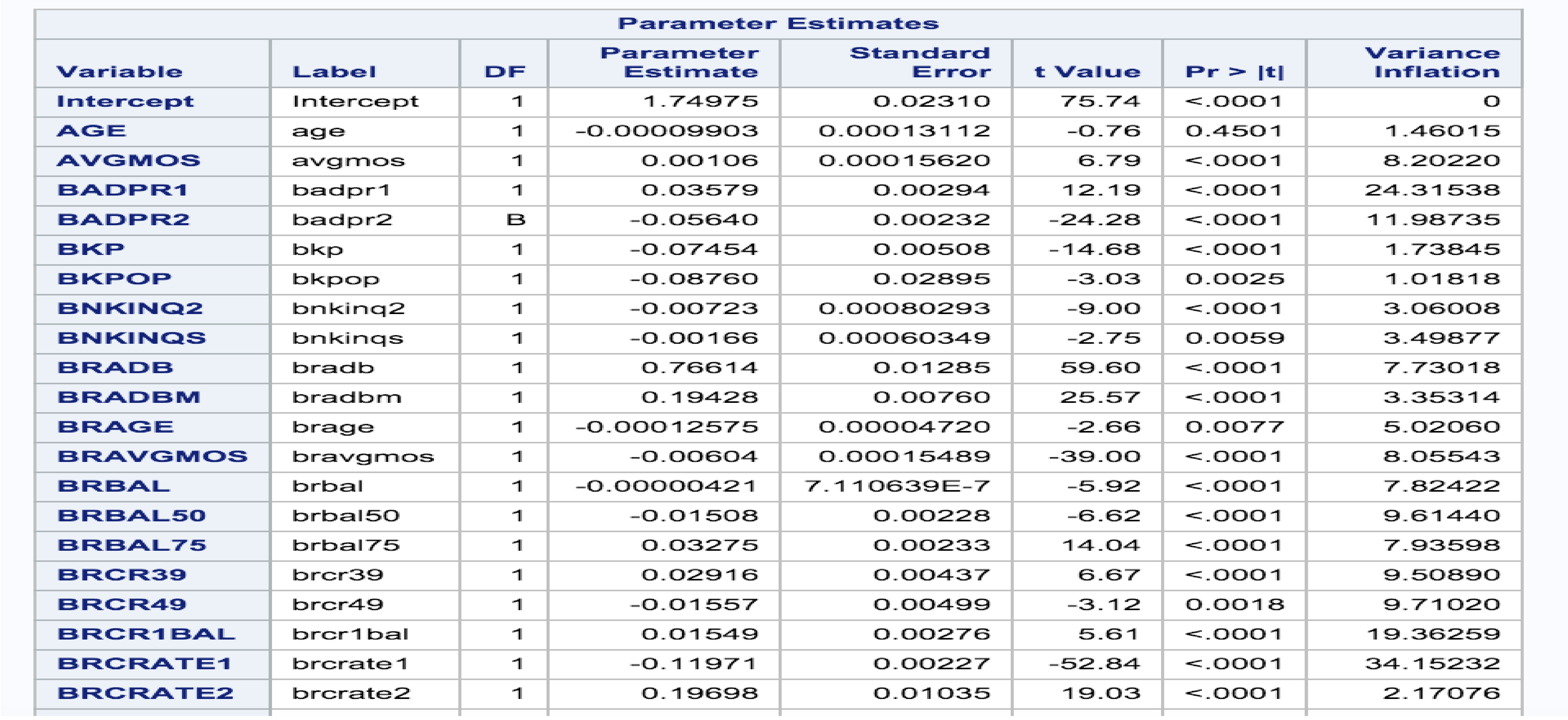
PCTREM	MSTD	Obs.	Variables
0.8	4	1255429	317
0.6	4	1255429	299
0.4	4	1255429	182
0.25	4	1255429	146
0.2	4	1255429	143
0.05	4	1255429	139



Multicollinearity

After the macro has run, we can remove the remaining variables from the PERF dataset. PERF alone had only 18 variables, of which, DELQID is the only relevant one for now. MATCHKEY we will keep too since it's linked to CPR. The last step in the data cleansing is eliminating variance inflation or multicollinearity. We can use SAS to check for VIFs or variance inflation by running a linear regression using all the variables. Some texts say a variable with a VIF of 10 or higher should be removed but others say 5 (e.g., Rogerson, 2001) or even 4 (e.g., Pan & Jackson, 2008). VIFs are calculated as the reciprocal of tolerance or $\frac{1}{1-r^2}$ such that r^2 is the percentage of deviation that can be explained and $1-r^2$ is the percentage of deviation that cannot. When this number gets lower, the reciprocal or the VIF, gets higher. Variance inflation can cause signs to change such that when a beta coefficient should either increase or decrease the predicted value of the response variable it does the opposite. After removing all VIFs and the blank variable BEACON, we are left with 56 variables.

Variance Inflation Factors (VIFs)



Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	1.74975	0.02310	75.74	<.0001	0
AGE	age	1	-0.00009903	0.00013112	-0.76	0.4501	1.46015
AVGMOS	avgmos	1	0.00106	0.00015620	6.79	<.0001	8.20220
BADPR1	badpr1	1	0.03579	0.00294	12.19	<.0001	24.31538
BADPR2	badpr2	B	-0.05640	0.00232	-24.28	<.0001	11.98735
BKP	bkp	1	-0.07454	0.00508	-14.68	<.0001	1.73845
BKPOP	bkpop	1	-0.08760	0.02895	-3.03	0.0025	1.01818
BNKINQ2	bnkinq2	1	-0.00723	0.00080293	-9.00	<.0001	3.06008
BNKINQS	bnkinqs	1	-0.00166	0.00060349	-2.75	0.0059	3.49877
BRADB	bradb	1	0.76614	0.01285	59.60	<.0001	7.73018
BRADBM	bradbm	1	0.19428	0.00760	25.57	<.0001	3.35314
BRAGE	brage	1	-0.00012575	0.00004720	-2.66	0.0077	5.02060
BRAVGMOS	bravgmos	1	-0.00604	0.00015489	-39.00	<.0001	8.05543
BRBAL	brbal	1	-0.00000421	7.110639E-7	-5.92	<.0001	7.82422
BRBAL50	brbal50	1	-0.01508	0.00228	-6.62	<.0001	9.61440
BRBAL75	brbal75	1	0.03275	0.00233	14.04	<.0001	7.93598
BRCR39	brcr39	1	0.02916	0.00437	6.67	<.0001	9.50890
BRCR49	brcr49	1	-0.01557	0.00499	-3.12	0.0018	9.71020
BRCR1BAL	brcr1bal	1	0.01549	0.00276	5.61	<.0001	19.36259
BRCRATE1	brcrate1	1	-0.11971	0.00227	-52.84	<.0001	34.15232
BRCRATE2	brcrate2	1	0.19698	0.01035	19.03	<.0001	2.17076

ANALYTICS EXPERIENCE 2016

#analyticsx

Using Logistic Regression to Predict Credit Default

Steven Leopard and Jun Song

Dr. Jennifer Priestley and Professor Michael Frankel

Kennesaw State University
College of Science and Mathematics
Department of Statistics and Analytical Sciences

Creating the Response

Finally, the creation of the variable GOODBAD was done so we could give a simple yes or no, 0 or 1, answer to the question concerning credit. This variable is made from DELQID such that any values less than 3 are given a 0 and considered good, while the rest are given 1's and considered bad. This is the primary component in binary classifications and Bernoulli's probability distribution. **Table 4** is a frequency chart of GOODBAD, note that over 80% of the observations are good or 0.

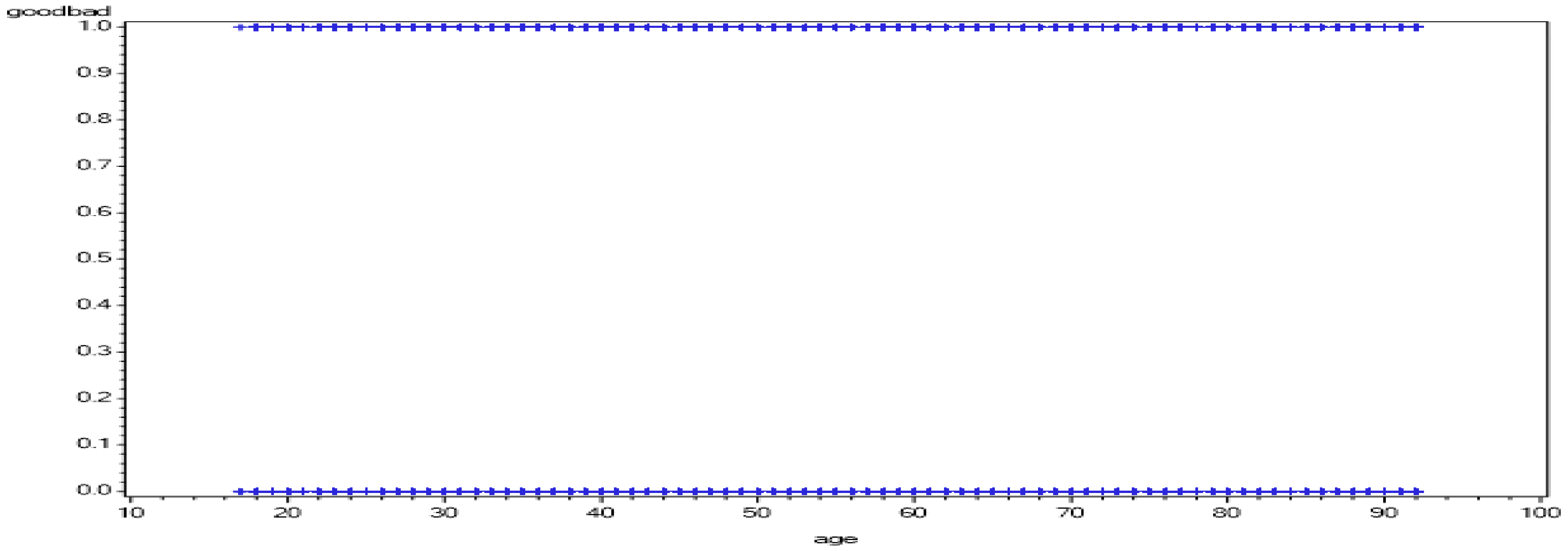
Frequencies of the Response Variable GOODBAD

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

Discretization

This section deals with the transformation of the variables remaining after cleansing the data. Here we will be creating 3 to 6 discrete versions of the variables in a way that better fits the dependent variable GOODBAD. We are trying to map the optimal monotonic transformation of the continuous variable back to the form of the function that is used in the model. Certain variables in their raw form are not useful in binary classifications. This creates a problem when a variable like AGE is graphed with GOODBAD; we see horizontal lines at 0 and 1 as is such in the following figure

Horizontal Plot of Age Before Discretization



User and SAS Defined Discretization

To solve the problem previously described, we need to transform a continuous variable into a bounded, discrete variable otherwise known as discretization. There are two ways we will discretize the data, user defined (Disc1) and SAS defined (Disc2). For the discretization of both, there are 3 transformations of the variables; ordinal, odds, and log of odds; seven total including the original. This holds true unless we come across an 'inherently' binary variable. If so, that variable will have only 4 versions: itself and Disc1 transformations. The following figures show the difference between user and SAS defined discretization of the variable PRMINQS (Number of promotions, account money / revolving inquiries).

ANALYTICS EXPERIENCE 2016

#analyticsx

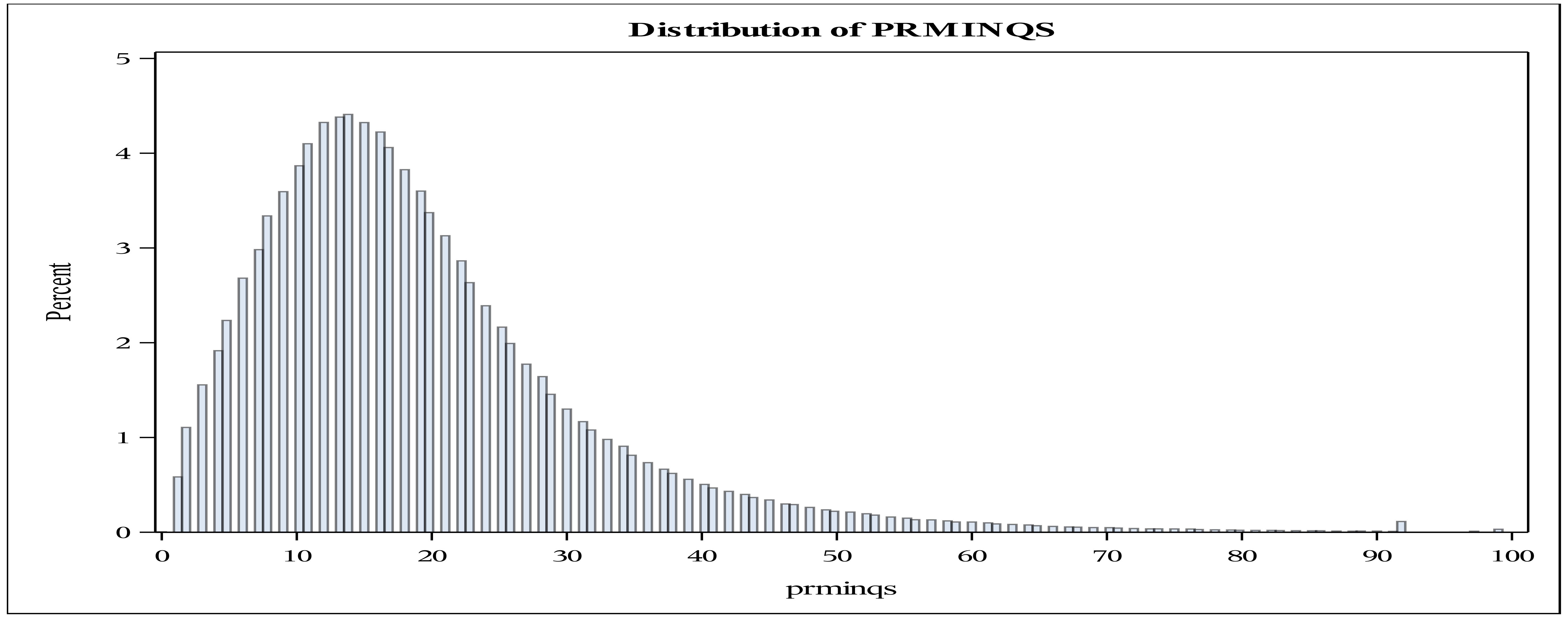
Using Logistic Regression to Predict Credit Default

Steven Leopard and Jun Song

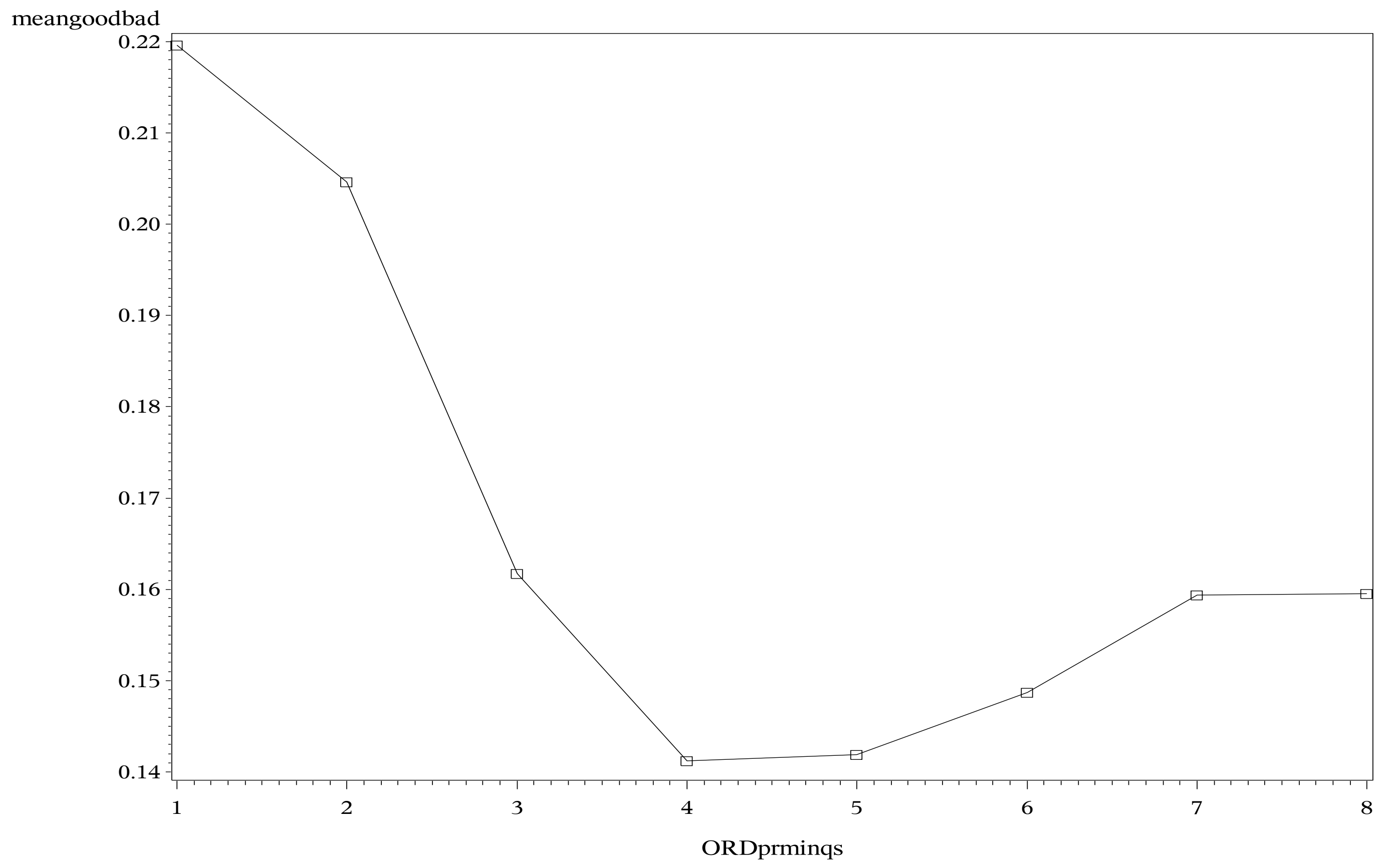
Dr. Jennifer Priestley and Professor Michael Frankel

Kennesaw State University
College of Science and Mathematics
Department of Statistics and Analytical Sciences

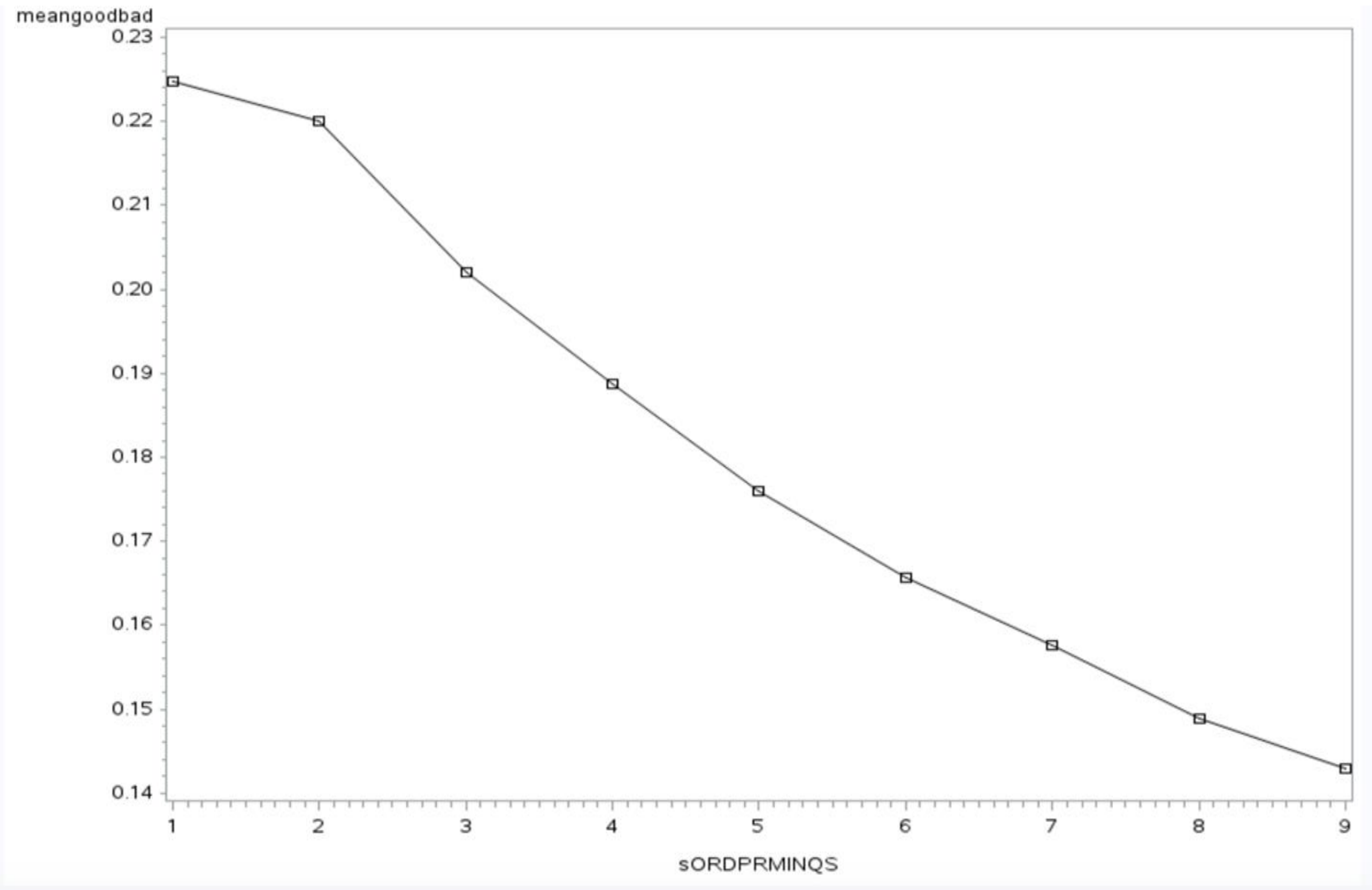
Raw Form of PRMINQS



User Created



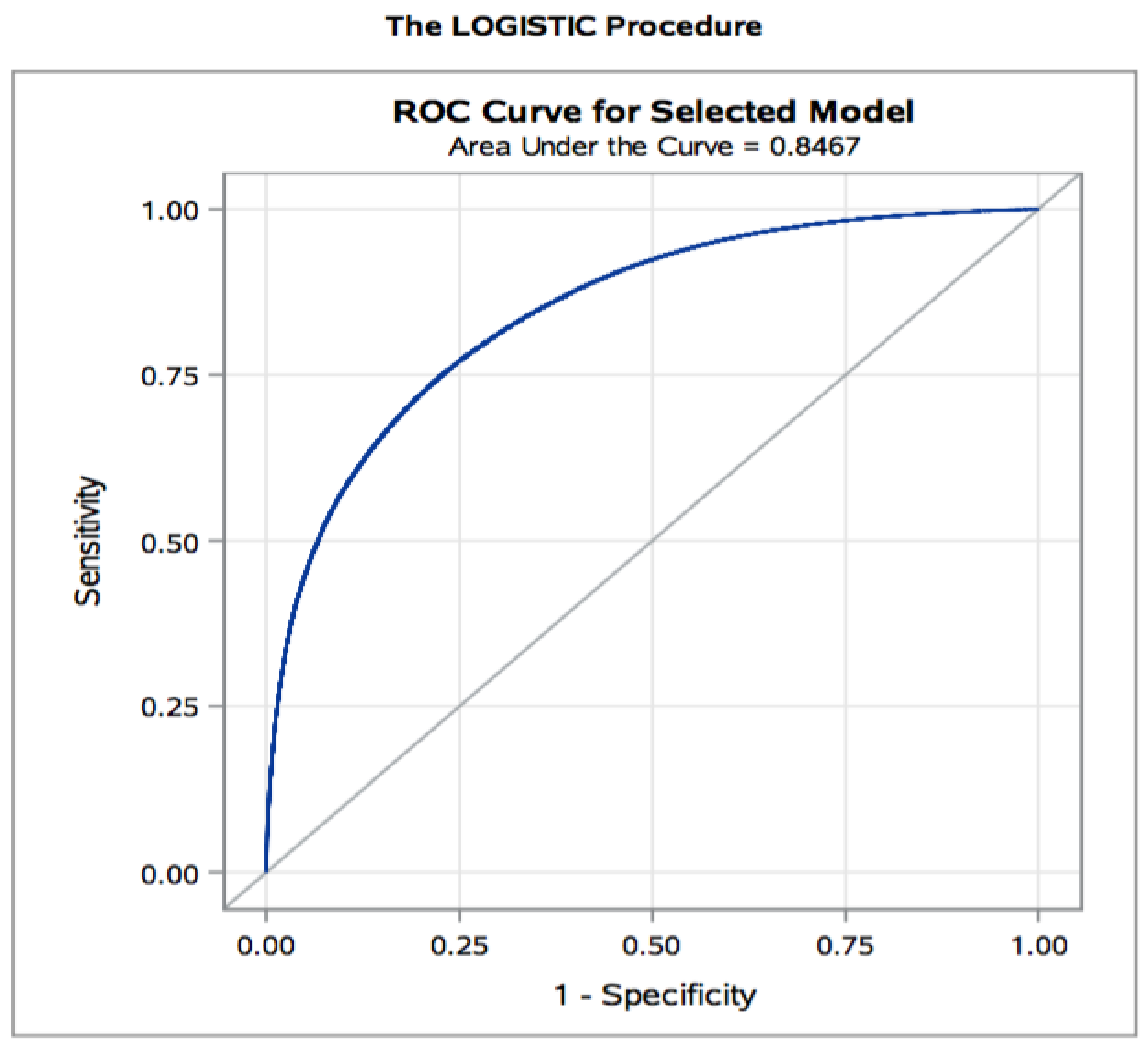
SAS Created



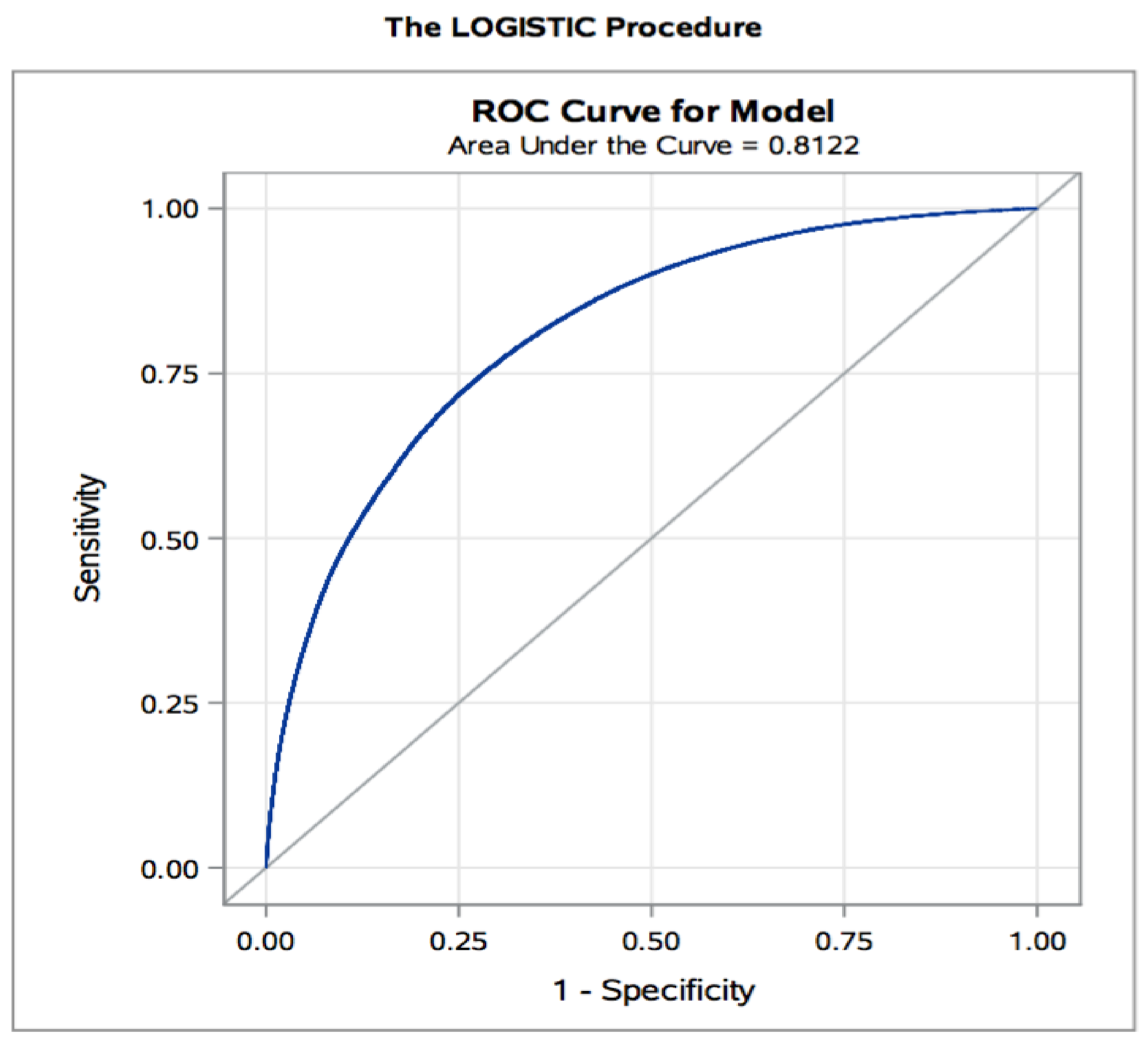
Building the Model

After all the variables have been discretized and variance inflation factors have been removed one additional step remains before modeling. The training and validation datasets must be created to both build and legitimize the model. To do this, a new variable RANDOM will be used in order to split the data 70/30 such that the training dataset will receive the majority of observations. With the training portion, the model can be built then scored by the random, mutually exclusive validation file. This is done so that the model can be tested on a subset of the data different than that used in the creation of the model which theoretically, should work on any subset in which the model is applied. Typically in certain work environments an altogether completely different dataset would be used to validate. After the master file has been split, a proc logistic is ran on the training file. The model was created using backward selection in which 85 redundant or insignificant variables were removed. By using backward selection the model analyzes everything at once then begins removing variables that ultimately enhance the capability of prediction for the model. The ROC curves below show the area or C-stat for the model with all remaining variable transformations and again with only those with the highest chi-square value.

ROC - All Variables 0.85



ROC - 10 Variables 0.81



ANALYTICS EXPERIENCE 2016

#analyticsx

Kennesaw State University
College of Science and Mathematics
Department of Statistics and Analytical Sciences

Using Logistic Regression to Predict Credit Default

Steven Leopard and Jun Song

Dr. Jennifer Priestley and Professor Michael Frankel

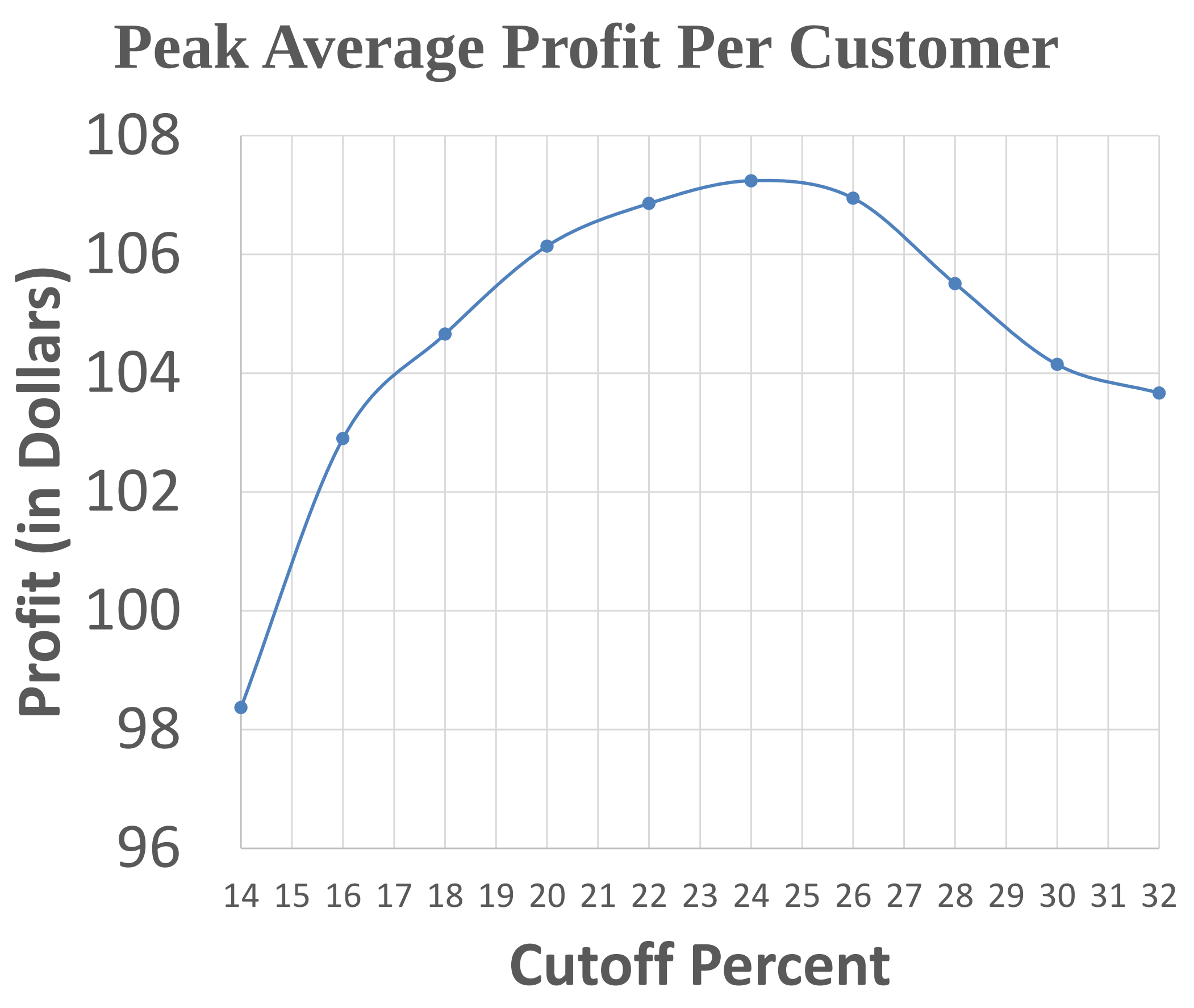
Scoring on the Validation Set

Once the model has been built the validation file can now be used to score the data. When scoring the data, percentiles are created to determine the cutoff point for the probability of default in this way we are able to maximize the profitability of the model. By using the classification table we can create a graph to see where the derivative of the function = 0.

Classification Table and Profit Function

Classification Table

Prob Level	Correct Event	Correct Non-Event	Incorrect Event	Incorrect Non-Event	Percentages Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.000	154E3	0	725E3	0	17.6	100.0	0.0	82.4	.
0.100	137E3	387E3	338E3	17636	59.6	88.6	53.4	71.2	4.4
0.200	105E3	566E3	159E3	49494	76.3	67.9	78.1	60.2	8.0
0.300	79817	639E3	85796	74477	81.8	51.7	88.2	51.8	10.4
0.400	59516	679E3	46001	94778	84.0	38.6	93.7	43.6	12.3
0.500	41719	7E5	25255	113E3	84.3	27.0	96.5	37.7	13.9
0.600	26380	713E3	12121	128E3	84.1	17.1	98.3	31.5	15.2
0.700	13379	72E4	4669	141E3	83.4	8.7	99.4	25.9	16.4
0.800	4225	724E3	1073	15E4	82.8	2.7	99.9	20.3	17.2
0.900	217	725E3	30	154E3	82.5	0.1	100.0	12.1	17.5
1.000	0	725E3	0	154E3	82.4	0.0	100.0	.	17.6

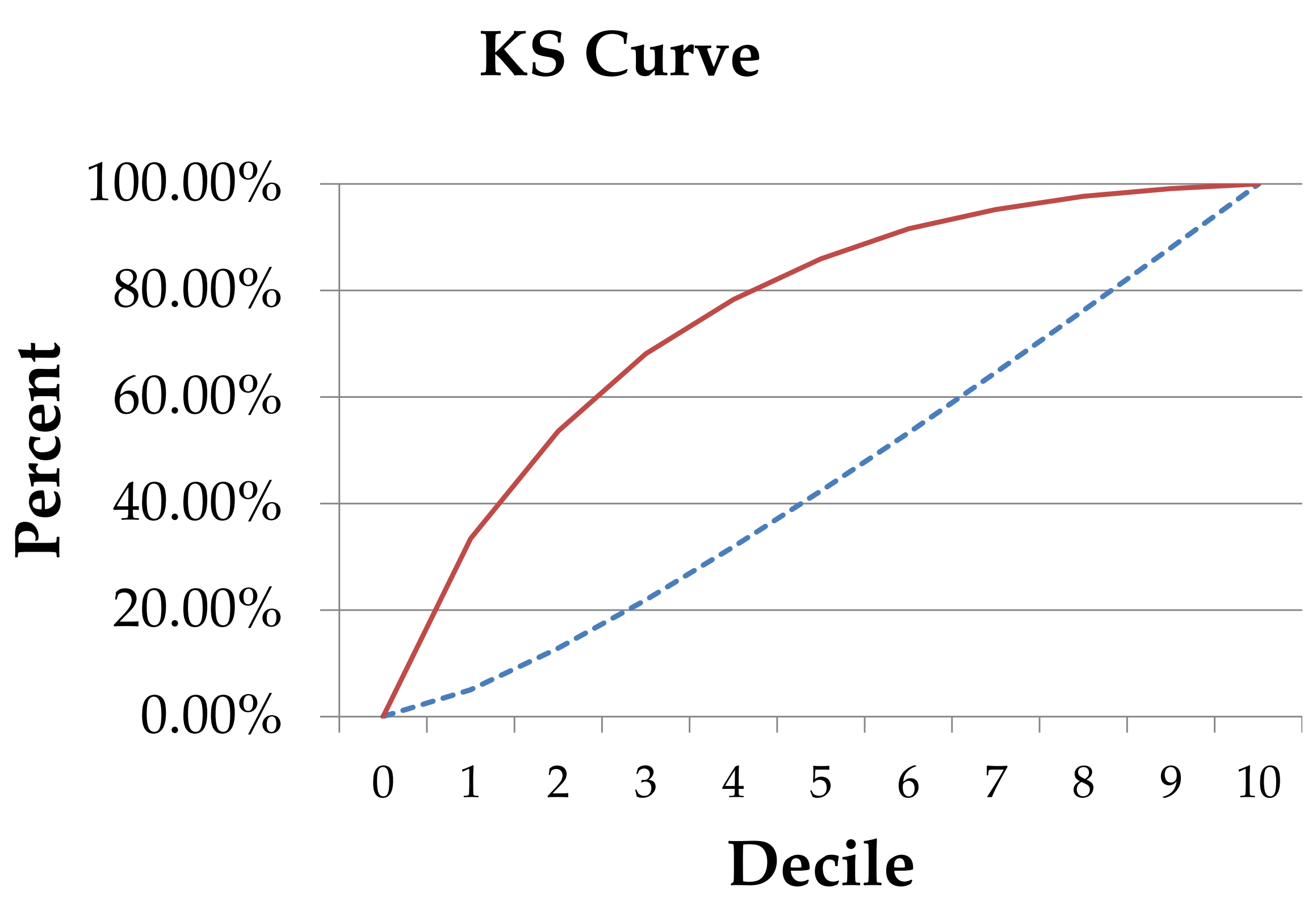


KS Curve and Cluster Analysis

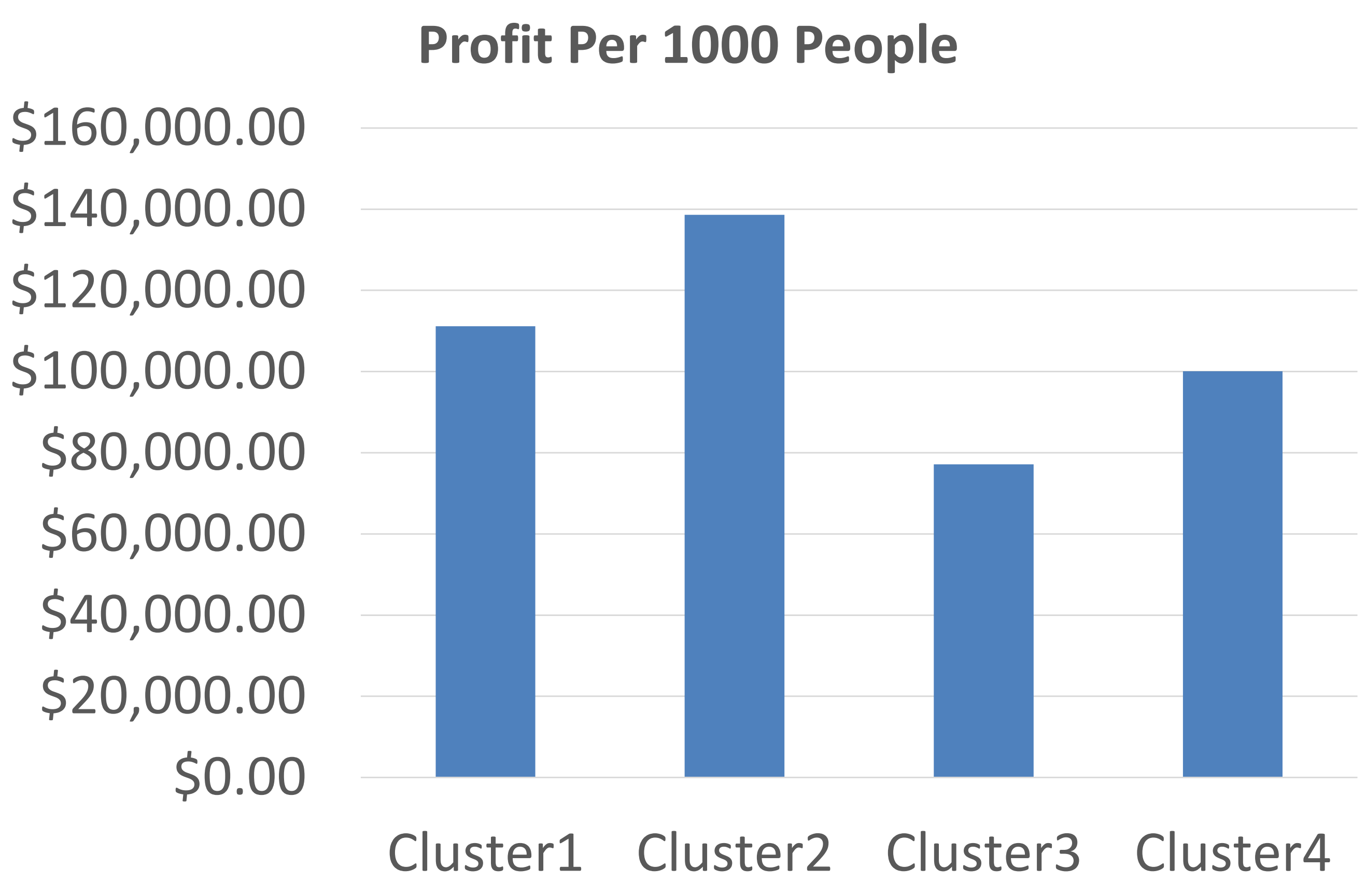
A KS test is predominantly used in a marketing context but can be used in the financial market as well. The idea for a KS test is if a list of *x* amount of potential customers existed and stretched out over some domain then how deep into this list should solicitation attempt to acquire in order to optimize the profit.

A clustering analysis determines how many, if any, groups or clusters exists within a dataset. In order to decide how many clusters the data set has three tests are used; Cubic Clustering Criterion, Pseudo-F Statistics, and Pseudo-T Statistic. The inflection points within the testing occurs when the number of clusters is four.

KS Curve



Clusters 1 - 4



Conclusion

After several procedures (cleansing the data, eliminating variables that were over coded, transforming the remaining, and running proc logistic) the model had a C-stat of .8122. The profitability function maxed out at approximately 25%. In other words, the probability for someone to default is expectable at or below .25 to receive a credit loan. This function showed an average profit per costumer of $117.11. KS testing showed that by targeting 31-40% percent of customers, the greatest difference between actual good and bad credit observations will be found. Each cluster was scored based on the validation file used for the model and the profit for 1000 people varies from $70,000 to $140,000. Based on the clustering analysis, cluster 2 yielded the largest profit and cluster 3 yielded the lowest.