Model Assessment with K-Fold Cross Validation

Harry Snart, SAS Institute

October 2024

This document shows how K-Fold Cross Validation can be used to assess model goodness of fit with few holdout samples. We start by loading the HMEQ dataset which has a binary target of BAD. After performing a brief exploratory analysis we then perform oversampling on the event class and then partition the dataset into Train, Test and Validate. We then train a logistic regression model with stepwise selection and perform k-fold sampling on the holdout dataset to score each of the partitions in order to generate a distribution of model assessment scores.

Load Dataset

Here we load the dataset using PROC IMPORT then print via PROC PRINT

BAD	LOAN	MORTDUE	VALUE	REASON	JOB	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
1	1100	25860	39025	HomeImp	Other	10.5	0	0	94.366666667	1	9	.
1	1300	70053	68400	HomeImp	Other	7	0	2	121.83333333	0	14	.
1	1500	13500	16700	HomeImp	Other	4	0	0	149.46666667	1	10	.
1	1500	.	.			.	.	.	.	.	.	.
0	1700	97800	112000	HomeImp	Office	3	0	0	93.333333333	0	14	.

Exploratory Data Analysis

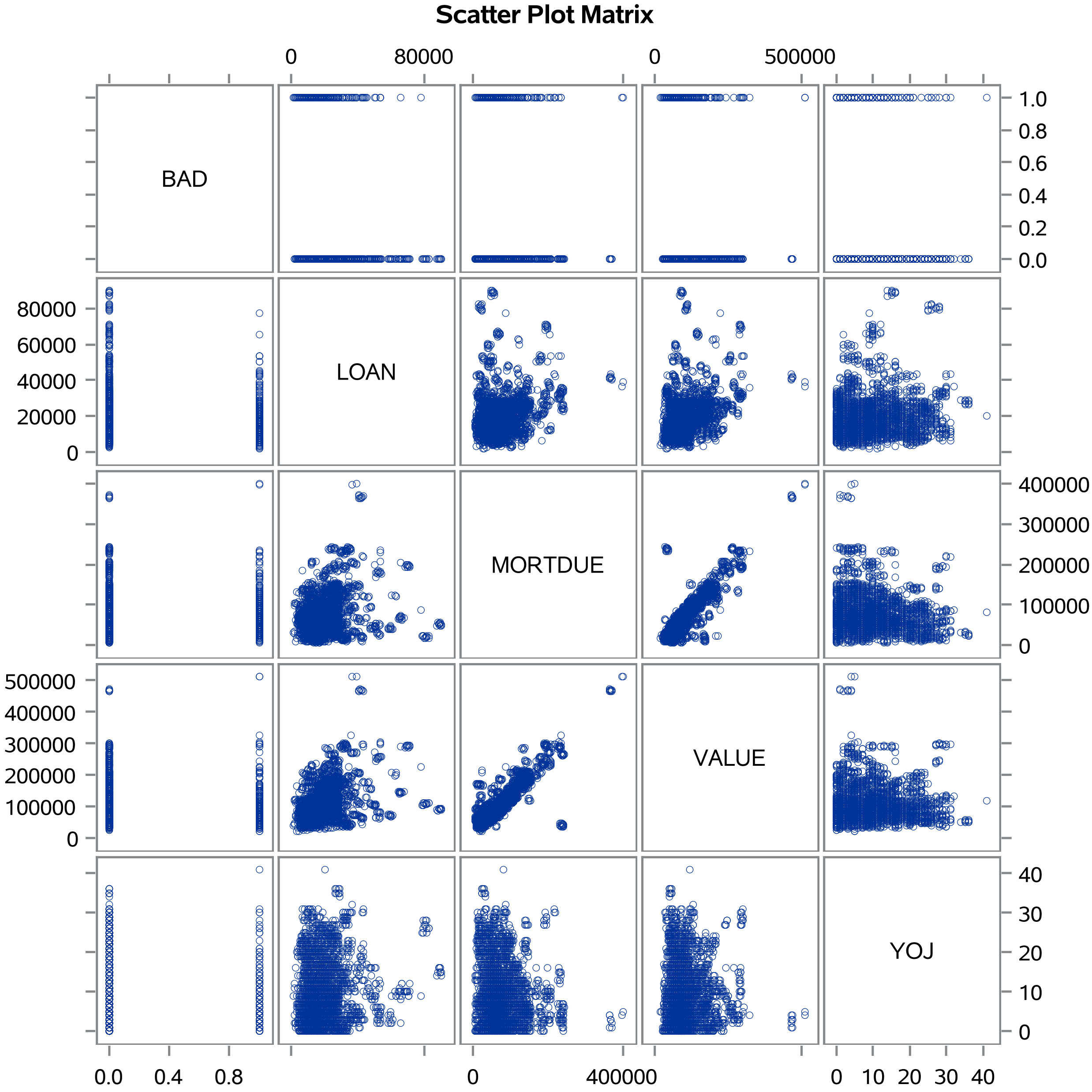
Here we perform an exploratory data analysis including variable correlation with PROC CORR, variable summary analysis with PROC CARDINALITY and visual analysis with PROC SGPLOT

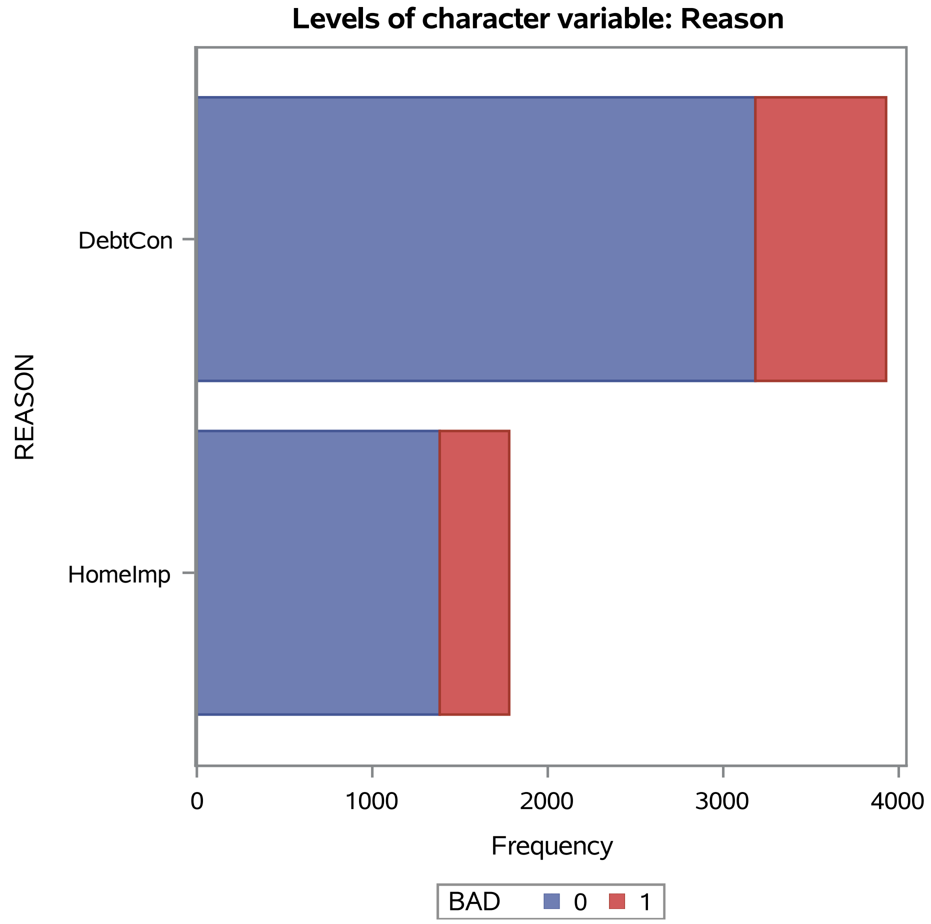
Bad	Number of Observations
0	4771
1	1189

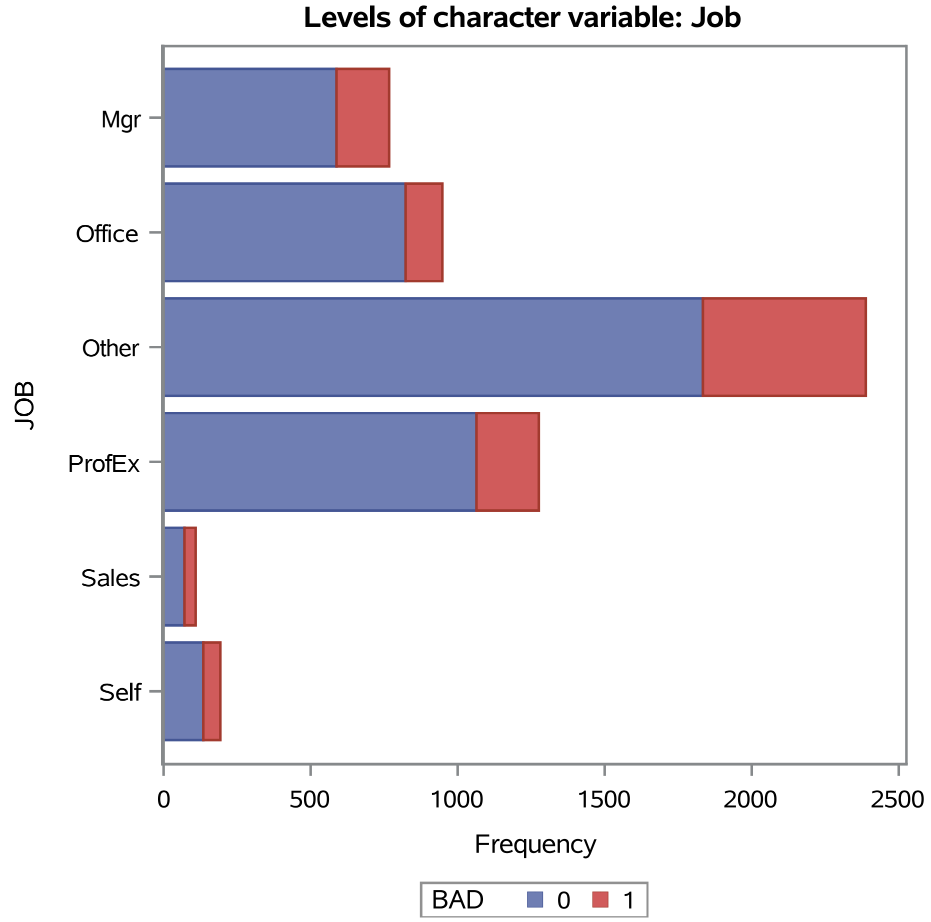
Variable name	Type of the raw values	Number of levels	Number of observations	Number of missing values	Mean	Standard deviation
LOAN	N	20	1189	0	16922.119428	11418.455152
MORTDUE	N	20	1189	106	69460.452973	47588.194467
VALUE	N	20	1189	105	98172.846227	74339.822506
REASON	C	2	1189	48	.	.
JOB	C	6	1189	23	.	.
YOJ	N	20	1189	65	8.0278024911	7.1007348316
DEROG	N	11	1189	87	0.7078039927	1.468380909
DELINQ	N	14	1189	72	1.2291853178	1.9029614156
CLAGE	N	20	1189	78	150.19018341	84.952286255
NINQ	N	16	1189	75	1.7827648115	2.2469764219
CLNO	N	20	1189	53	21.211267606	11.81298083
DEBTINC	N	20	1189	786	39.387644892	17.723586299

11 Variables:	BAD	LOAN	MORTDUE	VALUE	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC

Variable Correlation
Variables such as Loan amount appear to have an explanatory power for Bad.







Perform oversampling of event class

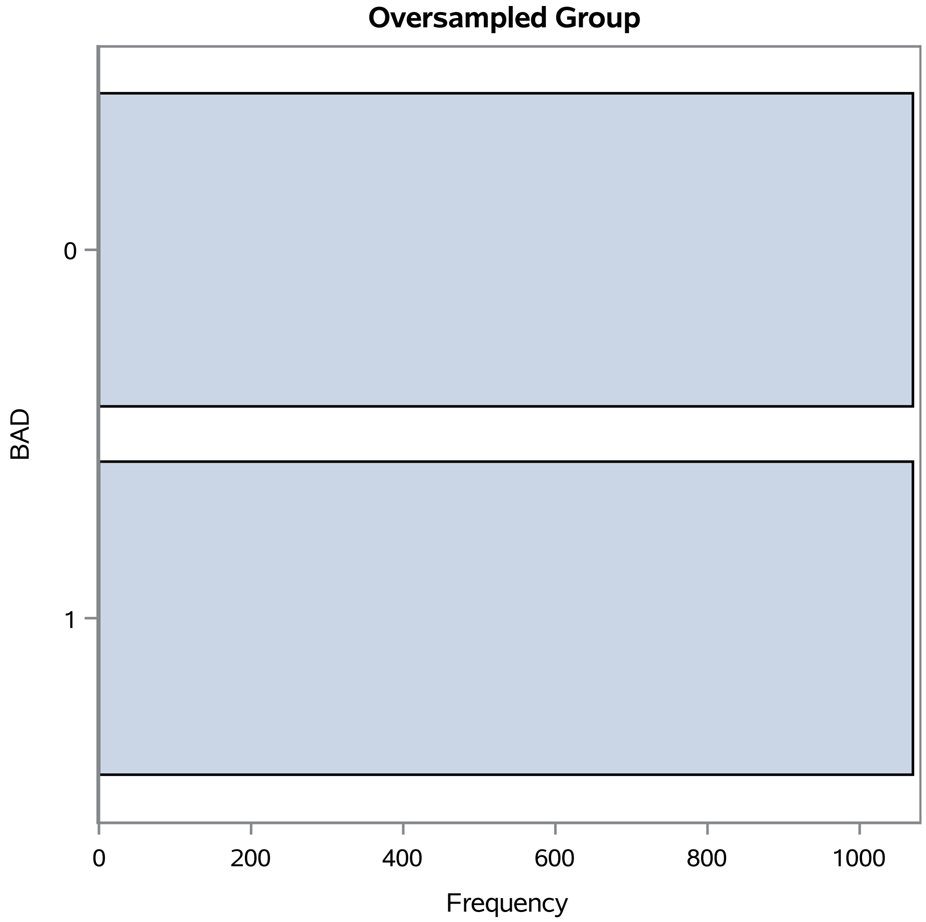
Here we oversample the event class, 1, given that the exploratory analysis shows there is a class imbalance we do this using PROC PARTITION

Levels of character variable: Job

The PARTITION Procedure

Oversampling Frequency			
Index	**BAD**	**Number of Obs**	**Number of Samples**
0	0	4771	1070
1	1	1189	1070

Output CAS Tables			
CAS Library	**Name**	**Number of Rows**	**Number of Columns**
CASUSER(sukhsn)	**SAMPLES**	2140	14

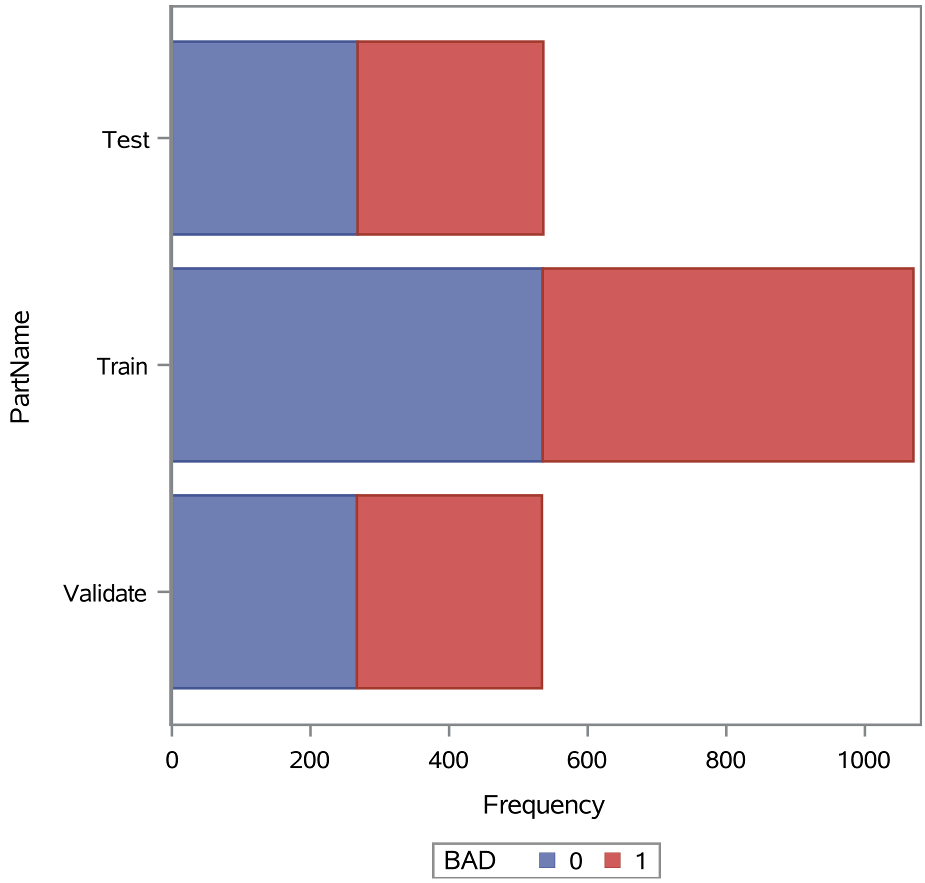


The PARTITION Procedure

Stratified Sampling Frequency				
Index	**BAD**	**Number of Obs**	**Sample Size 1**	**Sample Size 2**
0	0	1070	535	268
1	1	1070	535	268

Output CAS Tables			
CAS Library	**Name**	**Number of Rows**	**Number of Columns**
CASUSER(sukhsn)	**HMEQ_PART**	2140	15

Check number of observations by partition



Create Logistic Regression Model

Here we perform stepwise Logistic Regression using the Train and Test partitions using PROC LOGSELECT. The procedure prints summary statistics for both partitions.
We also save the scoring code to a SAS file that we can then use to score the kfold partitions later.

The LOGSELECT Procedure

Model Information	
Data Source	TRAIN_TEST
Response Variable	BAD
Distribution	Binary
Link Function	Logit
Optimization Technique	Newton-Raphson with Ridging
Predicted Response	P_BAD
Predicted Response Level	I_BAD

Number of Observations			
Description	**Total**	**Training**	**Testing**
Number of Observations Read	1606	1070	536
Number of Observations Used	723	478	245

The LOGSELECT Procedure

Response Profile				
Ordered Value	**BAD**	**Total Frequency**	**Training**	**Testing**
1	**0**	516	344	172
2	**1**	207	134	73

Probability modeled is BAD = 1.

Class Level Information		
Class	**Levels**	**Values**
REASON	2	DebtCon HomeImp
JOB	6	Mgr Office Other ProfEx Sales Self

Selection Information	
Selection Method	Stepwise
Select Criterion	SBC
Choose Criterion	SBC
Stop Criterion	SBC
Effect Hierarchy Enforced	None
Stop Horizon	3

Selection Details

Convergence criterion (GCONV=1E-8) satisfied.

Selection Summary				
Step	**Effect Entered**	**Effect Removed**	**Number Effects In**	**SBC**
0	**Intercept**		1	573.3349
1	**DELINQ**		2	527.4582
2	**DEBTINC**		3	517.9136
3	**NINQ**		4	508.6410
4	**CLAGE**		5	506.6004
5	**DEROG**		6	503.4530
6		**NINQ**	5	502.0618*
*** Optimal Value Of Criterion**				

Stepwise selection stopped because adding or removing an effect does not improve the SBC criterion.

The model at step 6 is selected where SBC is 502.0618.

Selected Effects:	Intercept DEROG DELINQ CLAGE DEBTINC

The LOGSELECT Procedure

Selected Model

Dimensions	
Columns in Design	5
Number of Effects	5
Max Effect Columns	1
Rank of Design	5
Parameters in Optimization	5

Testing Global Null Hypothesis: BETA=0			
Test	DF	Chi-Square	Pr > ChiSq
Likelihood Ratio	4	95.9033	<.0001

Fit Statistics		
Description	Training	Testing
-2 Log Likelihood	471.26203	238.86143
AIC (smaller is better)	481.26203	248.86143
AICC (smaller is better)	481.38914	249.11248
SBC (smaller is better)	502.11008	266.36772
Average Square Error	0.15653	0.15320
-2 Log L (Intercept-only)	567.16533	298.47134
R-Square	0.18179	0.21597
Max-rescaled R-Square	0.26167	0.30666
McFadden's R-Square	0.16909	0.19972
Misclassification Rate	0.21339	0.19184
Difference of Means	0.21381	0.23889

The LOGSELECT Procedure

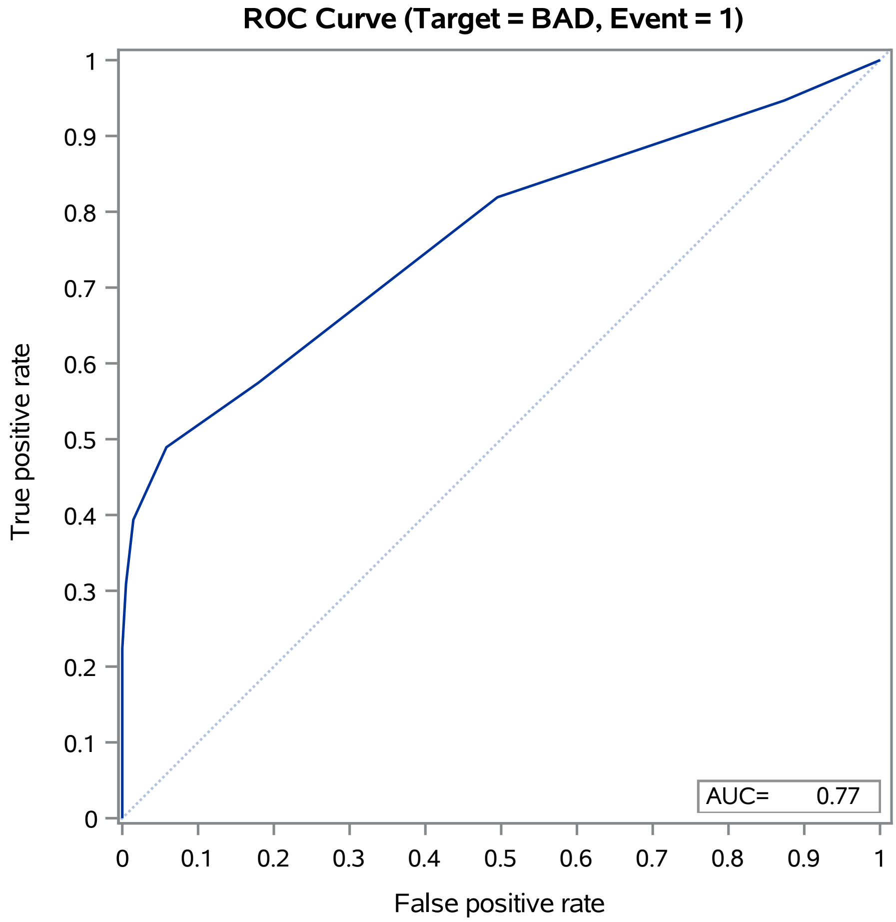
Selected Model

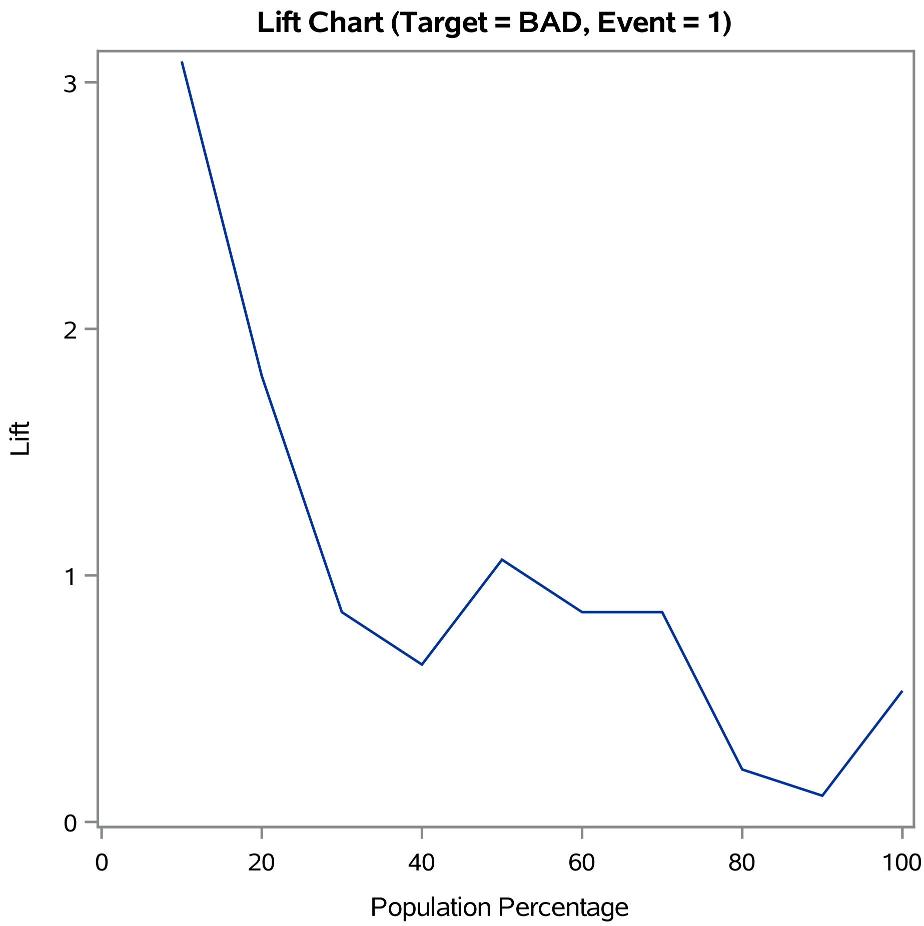
Parameter Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Chi-Square**	**Pr > ChiSq**
Intercept	1	-2.904953	0.642747	20.4267	<.0001
DEROG	1	0.601847	0.191464	9.8810	0.0017
DELINQ	1	0.587723	0.118384	24.6467	<.0001
CLAGE	1	-0.005247	0.001638	10.2570	0.0014
DEBTINC	1	0.067695	0.016116	17.6431	<.0001

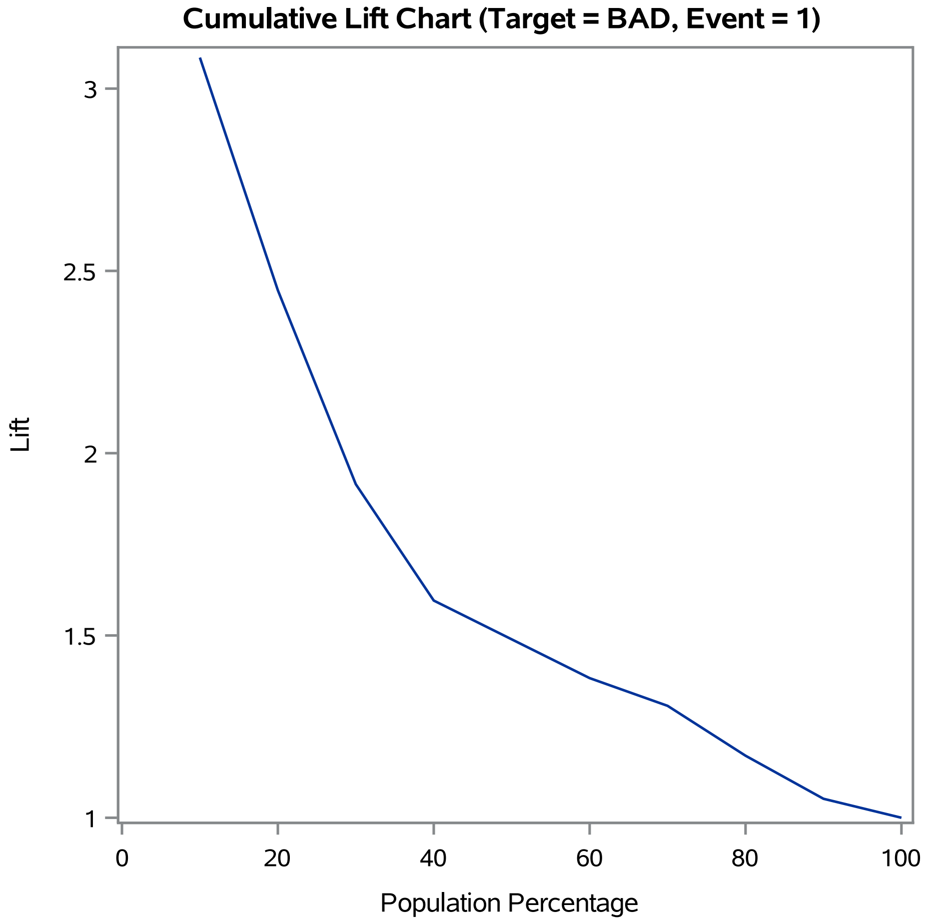
Task Timing		
Task	**Seconds**	**Percent**
Setup and Parsing	0.00	7.15%
Levelization	0.00	1.56%
Model Initialization	0.00	0.70%
SSCP Computation	0.00	3.93%
Model Selection	0.05	84.22%
Producing Score Code	0.00	1.56%
Display	0.00	0.69%
Cleanup	0.00	0.00%
Total	0.06	100.00%

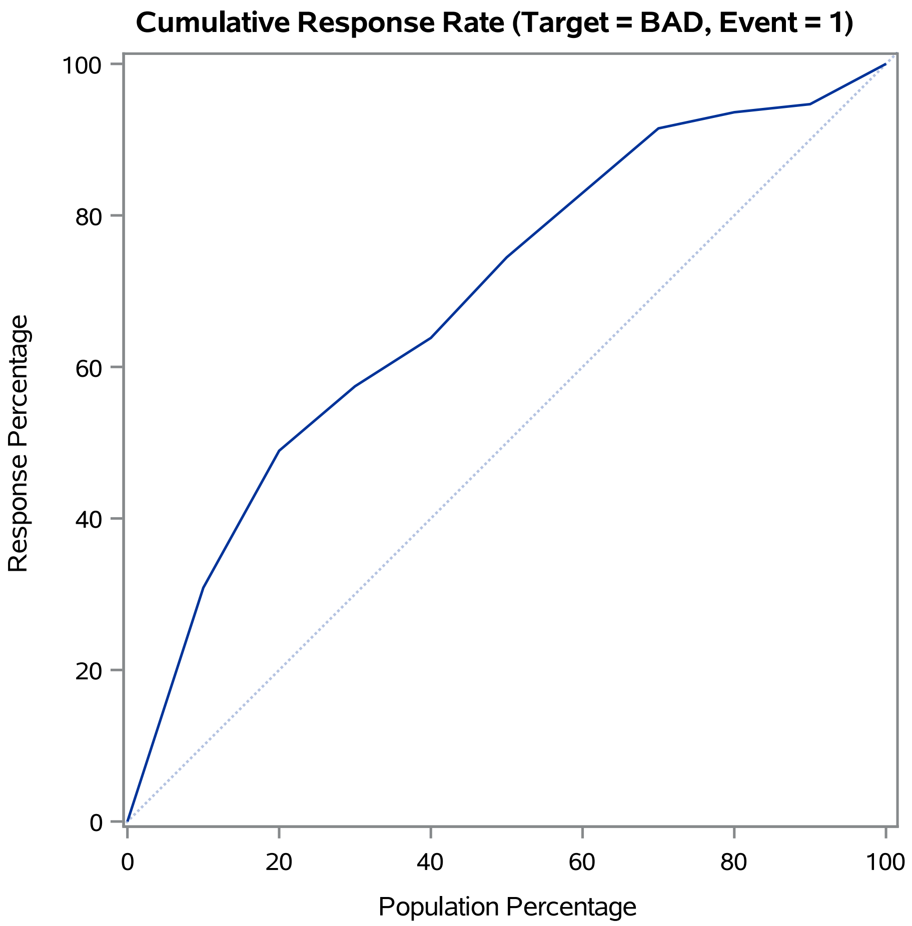
Visualise Model Fit on Test Dataset

Here we score the Test dataset using DataStep scorecode and visualise the ROC, Lift & Response charts.


ROC Curve (Target = BAD, Event = 1)
True positive rate
False positive rate
AUC= 0.77
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


Lift Chart (Target = BAD, Event = 1)
Lift
Population Percentage
0
1
2
3
0
20
40
60
80
100





Perform K-Fold Cross Validation

Here we define a macro, kFoldCV, which uses the CAS Sampling Actionset to perform k-fold partitioning stratified by BAD. We then score each dataset and append the results to a single table

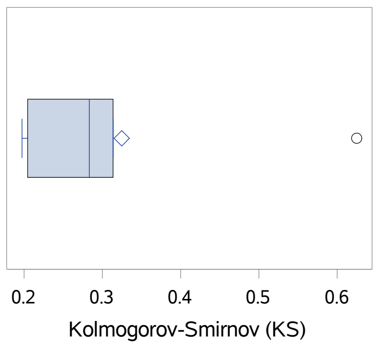
Cumulative Response Rate (Target = BAD, Event = 1)

including paritition identifier. Finally, we use PROC ASSESS which runs model assessment by Kfold partition.

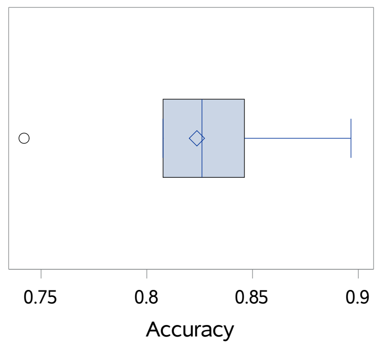
Visualise Estimated Fit Statistics by Kfold

Here we retain only values for the 0.5 cutoff from the ROC and visualise the estimated distributions for KS, Accuracy, F1, AUC, Gini and Misclassification rate from our k-fold partitions.

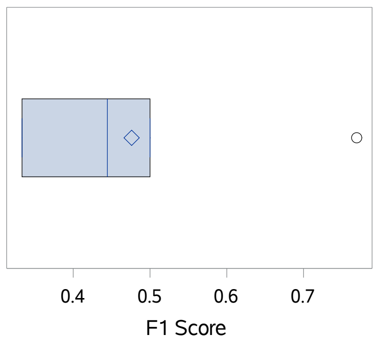
Distribution of KS Score



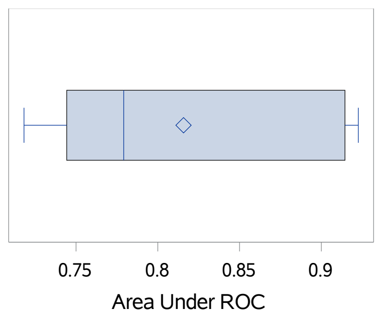
Distribution of Accuracy



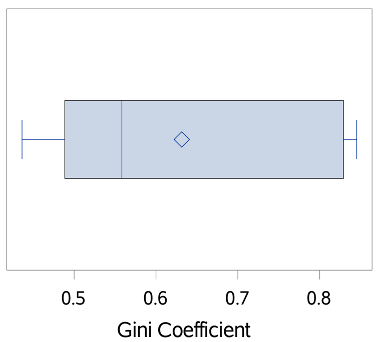
Distribution of F1 Score



Distribution of Area under Curve Score



Distribution of Gini



Distribution of Event Misclassification Rate