Summer 2019

Performance Validation Guide

Lauren Camero
NORTHWESTERN UNIVERSITY

Title:

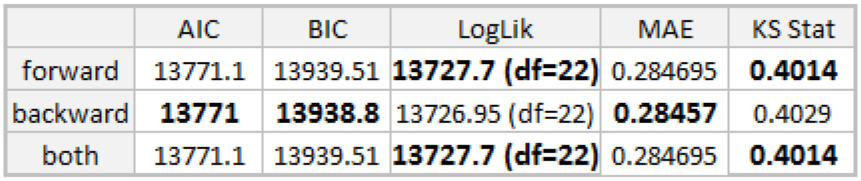
Model #101: Credit Card Default Model

Performance Validation Guide

1. The Production Model

After cleaning and transforming the raw and engineered variables, a simple logistic model is created. The model is executed using a stepwise selection process that drops a variable one at a time by checking to see if the significance has been reduced below tolerance. The stepwise model is executed in three directions: forward, backward, and both. Akaike information criterion (AIC), Bayesian information criterion (BIC), Log Likelihood, Mean Absolute Error, and the K-S Statistic were used to pick the best model. Log Likelihood Deviance represents the probability that the result is 0 or 1 in a logistic regression, and the AIC and BIC are information criteria methods to estimate the relative quality of each model.

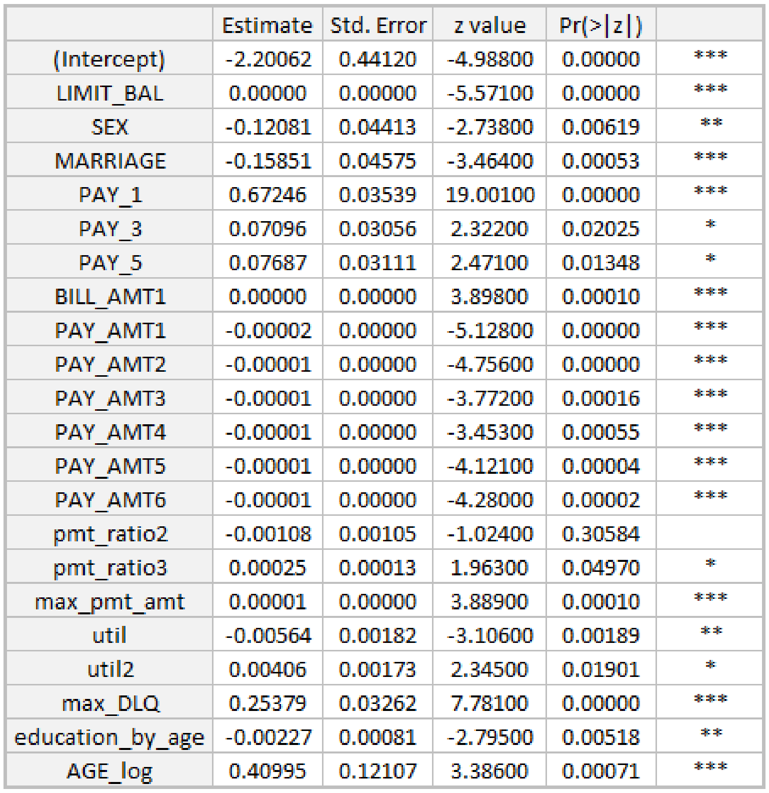
Figure 1: AIC and BIC Model Results



	AIC	BIC	LogLik	MAE	KS Stat
forward	13771.1	13939.51	**13727.7 (df=22)**	0.284695	**0.4014**
backward	**13771**	**13938.8**	13726.95 (df=22)	**0.28457**	0.4029
both	13771.1	13939.51	**13727.7 (df=22)**	0.284695	**0.4014**

Figure 1 highlights the best model based on these model performance metrics. Using this table, the backward selection model is chosen because it had the best results in three out of five metrics. The summary results illustrate that the remaining variables are all significant and should remain in the model for optimal performance.

Figure 2: Model Summary Results

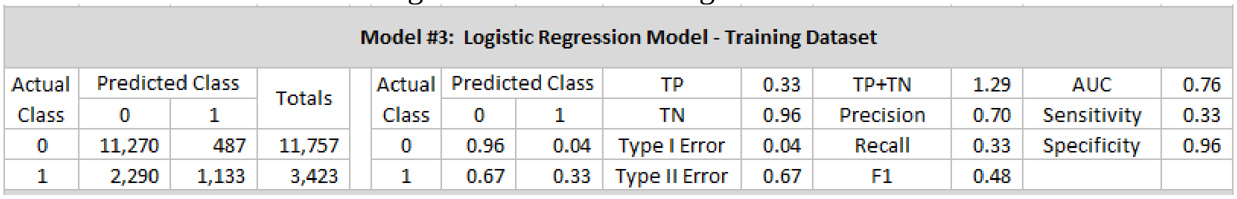


	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.20062	0.44120	-4.98800	0.00000	***
LIMIT_BAL	0.00000	0.00000	-5.57100	0.00000	***
SEX	-0.12081	0.04413	-2.73800	0.00619	**
MARRIAGE	-0.15851	0.04575	-3.46400	0.00053	***
PAY_1	0.67246	0.03539	19.00100	0.00000	***
PAY_3	0.07096	0.03056	2.32200	0.02025	*
PAY_5	0.07687	0.03111	2.47100	0.01348	*
BILL_AMT1	0.00000	0.00000	3.89800	0.00010	***
PAY_AMT1	-0.00002	0.00000	-5.12800	0.00000	***
PAY_AMT2	-0.00001	0.00000	-4.75600	0.00000	***
PAY_AMT3	-0.00001	0.00000	-3.77200	0.00016	***
PAY_AMT4	-0.00001	0.00000	-3.45300	0.00055	***
PAY_AMT5	-0.00001	0.00000	-4.12100	0.00004	***
PAY_AMT6	-0.00001	0.00000	-4.28000	0.00002	***
pmt_ratio2	-0.00108	0.00105	-1.02400	0.30584	
pmt_ratio3	0.00025	0.00013	1.96300	0.04970	*
max_pmt_amt	0.00001	0.00000	3.88900	0.00010	***
util	-0.00564	0.00182	-3.10600	0.00189	**
util2	0.00406	0.00173	2.34500	0.01901	*
max_DLQ	0.25379	0.03262	7.78100	0.00000	***
education_by_age	-0.00227	0.00081	-2.79500	0.00518	**
AGE_log	0.40995	0.12107	3.38600	0.00071	***

2. Model Development Performance

The results from the training data set are shown below. The AUC for the model using the training data is 0.756. The Type II error for this model is 0.67. The precision of the logistic model is 0.70. Precision is the positive prediction value meaning that about 70% of the estimated Default values of 1 were actually 1.

Figure 3: Train results logistic model

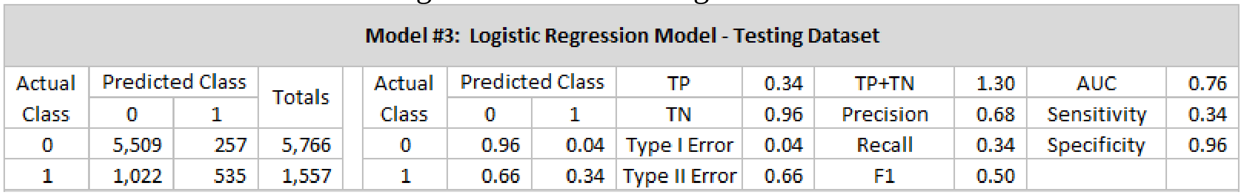


Model #3: Logistic Regression Model - Training Dataset

Actual	Predicted Class		Totals	Actual	Predicted Class		TP	0.33	TP+TN	1.29	AUC	0.76
Class	0	1		Class	0	1	TN	0.96	Precision	0.70	Sensitivity	0.33
0	11,270	487	11,757	0	0.96	0.04	Type I Error	0.04	Recall	0.33	Specificity	0.96
1	2,290	1,133	3,423	1	0.67	0.33	Type II Error	0.67	F1	0.48		

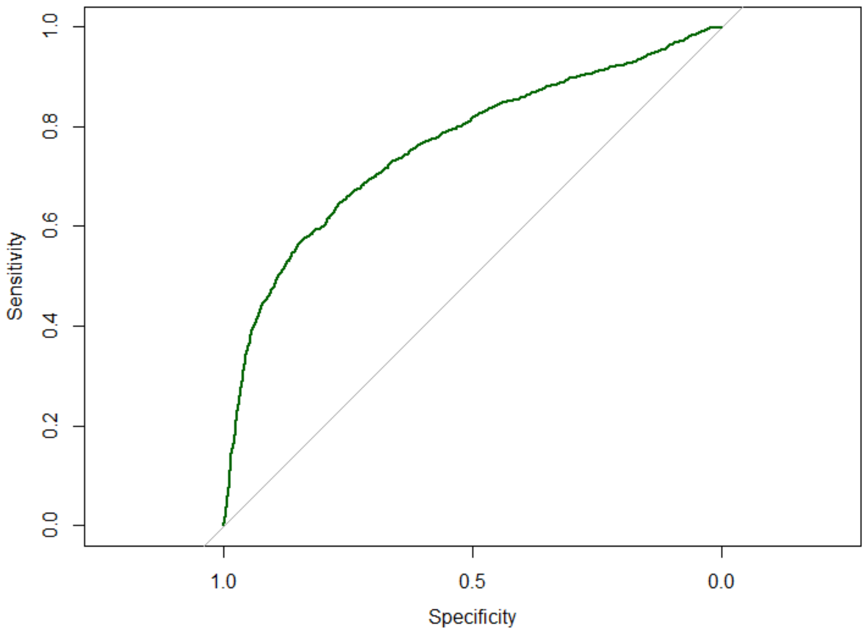
The precision of the logistic model decreased slightly when using the test data set. All metrics were very close between the two samples, meaning the model was not overfit for the training data. The specificity is the proportion of actual negatives that are correctly identified. This means that the default field was 0 and the predicted value was 0. Additionally, the ROC curve for the test data set is similar to the training data set which is unsurprising since the area under the curve was 0.76.

Figure 4: Test results logistic model



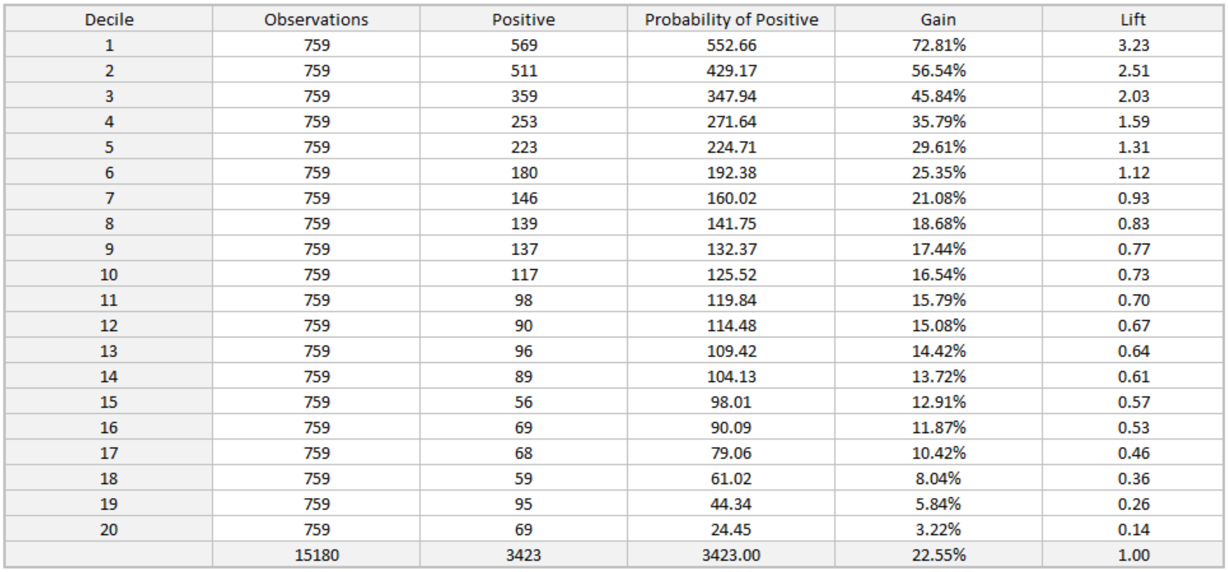
Model #3: Logistic Regression Model - Testing Dataset													
Actual	Predicted Class		Totals		Actual	Predicted Class		TP	0.34	TP+TN	1.30	AUC	0.76
Class	0	1			Class	0	1	TN	0.96	Precision	0.68	Sensitivity	0.34
0	5,509	257	5,766		0	0.96	0.04	Type I Error	0.04	Recall	0.34	Specificity	0.96
1	1,022	535	1,557		1	0.66	0.34	Type II Error	0.66	F1	0.50		

Figure 5: ROC Curve Test Data



Below is a lift chart showing the probability of positive results. Lift charts measure how much better one can expect to do with predictive modeling compared to without a model. The relative lift drops off significantly after five deciles.

Figure 6: Lift Table on Training Data



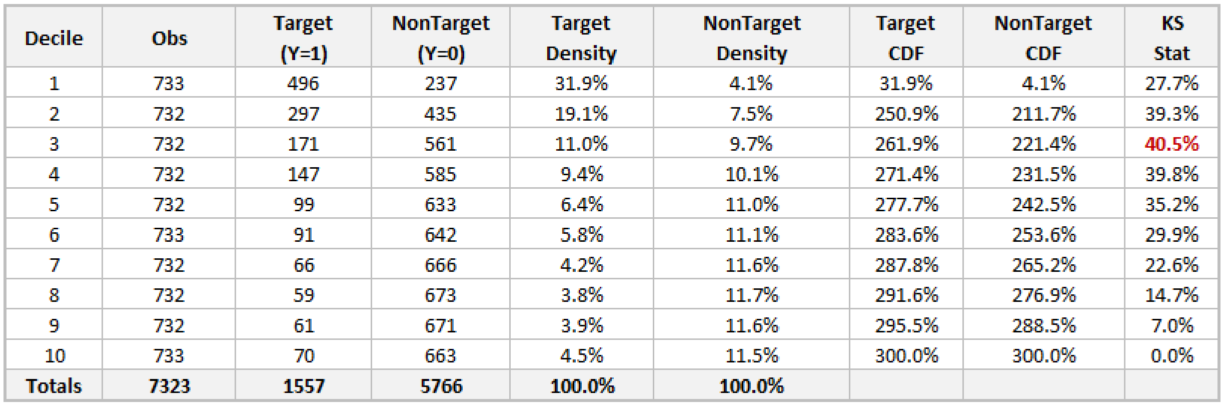
Decile	Observations	Positive	Probability of Positive	Gain	Lift
1	759	569	552.66	72.81%	3.23
2	759	511	429.17	56.54%	2.51
3	759	359	347.94	45.84%	2.03
4	759	253	271.64	35.79%	1.59
5	759	223	224.71	29.61%	1.31
6	759	180	192.38	25.35%	1.12
7	759	146	160.02	21.08%	0.93
8	759	139	141.75	18.68%	0.83
9	759	137	132.37	17.44%	0.77
10	759	117	125.52	16.54%	0.73
11	759	98	119.84	15.79%	0.70
12	759	90	114.48	15.08%	0.67
13	759	96	109.42	14.42%	0.64
14	759	89	104.13	13.72%	0.61
15	759	56	98.01	12.91%	0.57
16	759	69	90.09	11.87%	0.53
17	759	68	79.06	10.42%	0.46
18	759	59	61.02	8.04%	0.36
19	759	95	44.34	5.84%	0.26
20	759	69	24.45	3.22%	0.14
	15180	3423	3423.00	22.55%	1.00

3. Performance Monitoring Plan

When building a model, there is a potential of presenting risk based on inaccurate results. In order to prevent unnecessary risk, the model must be monitored in order to address necessary tweaks to improve performance. The KS – Statistic is used in logistic regressions to test the quality of two distribution functions. The bigger the KS - value, the better the model will perform delineating between the two binary outcomes. Since the score is the probability that the model outcome will be one, it can create an empirical cumulative distribution function. The table below shows results in semi - deciles. This table is helpful if the model is to be deployed in stages so that we know the probability of Y = 1 at each decile.

.

Figure 7: KS Stats on Testing Data

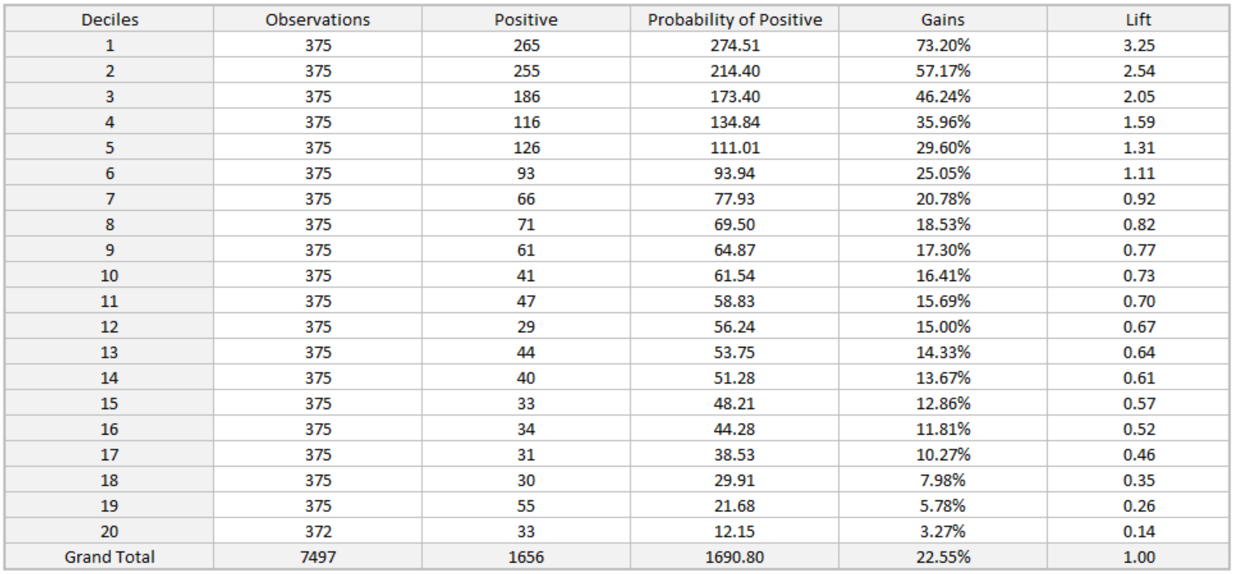


Decile	Obs	Target (Y=1)	NonTarget (Y=0)	Target Density	NonTarget Density	Target CDF	NonTarget CDF	KS Stat
1	733	496	237	31.9%	4.1%	31.9%	4.1%	27.7%
2	732	297	435	19.1%	7.5%	250.9%	211.7%	39.3%
3	732	171	561	11.0%	9.7%	261.9%	221.4%	40.5%
4	732	147	585	9.4%	10.1%	271.4%	231.5%	39.8%
5	732	99	633	6.4%	11.0%	277.7%	242.5%	35.2%
6	733	91	642	5.8%	11.1%	283.6%	253.6%	29.9%
7	732	66	666	4.2%	11.6%	287.8%	265.2%	22.6%
8	732	59	673	3.8%	11.7%	291.6%	276.9%	14.7%
9	732	61	671	3.9%	11.6%	295.5%	288.5%	7.0%
10	733	70	663	4.5%	11.5%	300.0%	300.0%	0.0%
Totals	**7323**	**1557**	**5766**	**100.0%**	**100.0%**			

4. Performance Monitoring Results

To monitor the performance of the model, the validation data set is used to predict and score the probability of a customer defaulting on a loan. A lift table will measure the model's ability to classify if the default field equals one. The first 6 deciles are performing well since they are above the average. These results align with the lift table outcomes from both the training and test data sets.

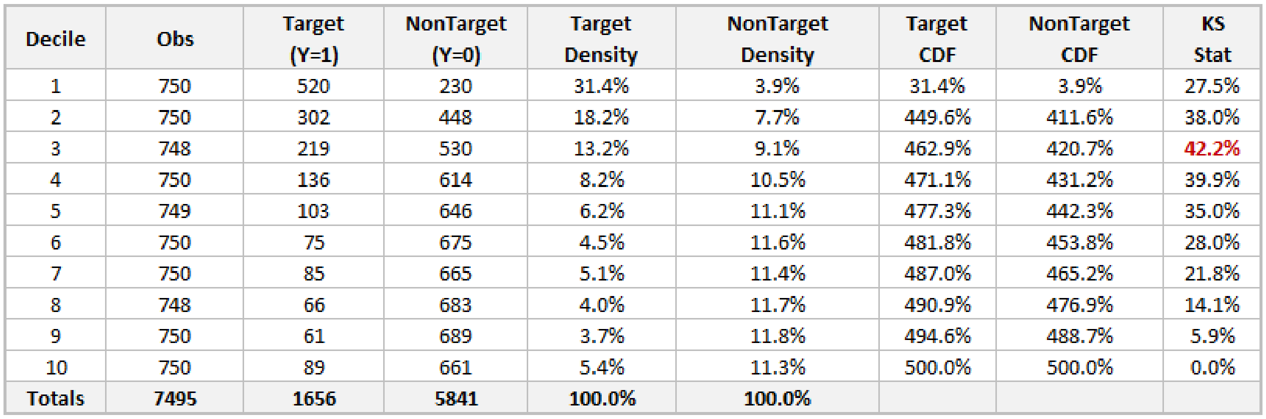
Figure 8: Lift Table on Validation Data



Deciles	Observations	Positive	Probability of Positive	Gains	Lift
1	375	265	274.51	73.20%	3.25
2	375	255	214.40	57.17%	2.54
3	375	186	173.40	46.24%	2.05
4	375	116	134.84	35.96%	1.59
5	375	126	111.01	29.60%	1.31
6	375	93	93.94	25.05%	1.11
7	375	66	77.93	20.78%	0.92
8	375	71	69.50	18.53%	0.82
9	375	61	64.87	17.30%	0.77
10	375	41	61.54	16.41%	0.73
11	375	47	58.83	15.69%	0.70
12	375	29	56.24	15.00%	0.67
13	375	44	53.75	14.33%	0.64
14	375	40	51.28	13.67%	0.61
15	375	33	48.21	12.86%	0.57
16	375	34	44.28	11.81%	0.52
17	375	31	38.53	10.27%	0.46
18	375	30	29.91	7.98%	0.35
19	375	55	21.68	5.78%	0.26
20	372	33	12.15	3.27%	0.14
Grand Total	7497	1656	1690.80	22.55%	1.00

Figure 9 illustrates the KS Statistic results for the validation data set. Similar to both the training and test data sets, the validation data set has the highest KS Stat in decile three at 42.2%. This is close to the testing data set at 40.5%. Based on the validation data, the model would need to be reviewed every six months to see if the KS Statistic is above 38%. This will be the “yellow” alert for model performance. If a KS Statistic rises above 45%, then the model must be reviewed for accuracy as soon as possible. All other metric results would be “green” in the red-amber-green performance validation and require no additional monitoring.

Figure 9: KS Stats on Validation Data



Decile	Obs	Target (Y=1)	NonTarget (Y=0)	Target Density	NonTarget Density	Target CDF	NonTarget CDF	KS Stat
1	750	520	230	31.4%	3.9%	31.4%	3.9%	27.5%
2	750	302	448	18.2%	7.7%	449.6%	411.6%	38.0%
3	748	219	530	13.2%	9.1%	462.9%	420.7%	42.2%
4	750	136	614	8.2%	10.5%	471.1%	431.2%	39.9%
5	749	103	646	6.2%	11.1%	477.3%	442.3%	35.0%
6	750	75	675	4.5%	11.6%	481.8%	453.8%	28.0%
7	750	85	665	5.1%	11.4%	487.0%	465.2%	21.8%
8	748	66	683	4.0%	11.7%	490.9%	476.9%	14.1%
9	750	61	689	3.7%	11.8%	494.6%	488.7%	5.9%
10	750	89	661	5.4%	11.3%	500.0%	500.0%	0.0%
Totals	**7495**	**1656**	**5841**	**100.0%**	**100.0%**			