Credit Card Default Model

Performance Validation Results

Jonah Muniz

1.0 The Production Model

The production model that will be analyzed in the performance monitoring plan is the logistic regression model with variable selection. The logistic regression model was selected out of the four models analyzed in the model development plan because in practice generalized linear models are the models that are pushed to production. Linear models are often chosen because they are easy to implement across different platforms. A description of the logistic regression model can be found below.

1.1 Model Description

The logistic regression algorithm is a statistical method that predicts a binary outcome, such as yes or no or 0 or 1 for the current case. The addition of a variable selection method to logistic regression enables the “best” variables to be selected to create the “best” performing logistic regression model. Variable selection methods are implemented when there are a high number of features because it reduces the iterative process of selecting different variations of feature combinations to see which produces the best model. There are three variable selection processes that can be used to select the features: forward, backward, and stepwise variable selection.

Forward selection starts with only the intercept in the model and additional features are added to the model until the model no longer improves. Backward selection starts with all features included then removes one feature at a time. The process continues until the model performance no longer improves. Lastly, stepwise variable selection is a combination of the forward and backward methodology where a variable can be added or removed from the model until model performance no longer improves.

1.2 Model Development and Performance

The original list of variables that were available for variable selection can be seen below in Table 1.

Table 1: Summary of Initial Feature List

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Max_Deliquency	15,180	0.68	1.07	0	0	2	8
Max_Utilization	15,180	0.49	0.43	-0.10	0.07	0.92	6.46
Avg_Pay_Amt	15,180	5,255.11	10,150.14	0.00	1,111.75	5,554.42	627,344.30
Max_Pay_Amt	15,180	15,620.73	35,279.24	0	2,195.8	12,200.8	1,215,471
Max_Bill_Amt	15,180	60,425.45	77,746.88	-2,900	10,050.8	79,119.5	823,540
Six_Month_Bal_Grwth	15,180	-12,495.35	44,045.77	-497,231	-20,094.2	2,962	399,983
Avg_Payment_Ratio	15,180	11.65	35.78	-605.46	0.05	1.00	2,687.00
Avg_Utilization	15,180	0.37	0.35	-0.20	0.03	0.69	4.47
Six_Month_Utilization_Grwth	15,180	-0.11	0.30	-5.31	-0.18	0.03	1.83
Avg_Bill_Amt	15,180	44,934.91	63,150.01	-56,043	4,789.1	56,880.5	592,432
AGE_41_100	15,180	0.28	0.45	0	0	1	1
AGE_26_40	15,180	0.60	0.49	0	0	1	1
AGE_18_25	15,180	0.13	0.34	0	0	0	1

A logistic model was trained on the above features and then stepwise variable selection was used on that model to select the best features to improve overall performance. Below in Table 2 is a summary table of the final logistic model with the final features after stepwise selection.

Table 2: Model #3

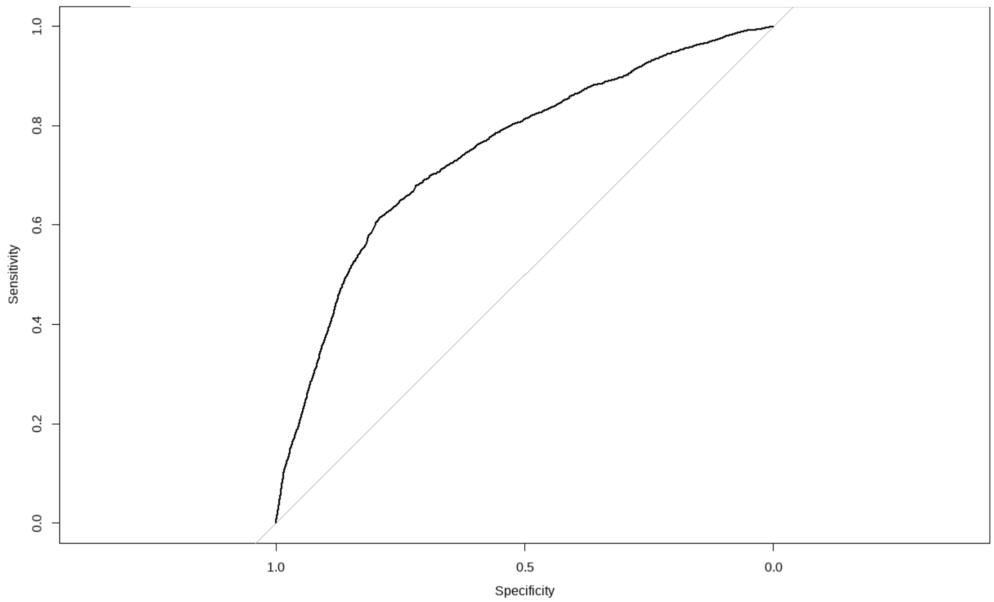
	Dependent variable:
	DEFAULT
Max_Deliquency	0.73***
	(0.02)
Max_Utilization	0.16
	(0.17)
Avg_Pay_Amt	-0.0001***
	(0.0000)
Max_Pay_Amt	0.0000***
	(0.0000)
Max_Bill_Amt	-0.0000
	(0.0000)
Avg_Payment_Ratio	0.004***
	(0.001)
Avg_Utilization	0.23
	(0.23)
Avg_Bill_Amt	0.0000*
	(0.0000)
AGE_41_100	0.13***
	(0.05)
Constant	-1.99***
	(0.05)
Observations	15,180
Log Likelihood	-7,052.76
Akaike Inf. Crit.	14,125.52
Note:	*p**p***p<0.01

The coefficients of the logistic regression model can be seen above in Table 2. The coefficients' magnitude and sign indicate how much an increase in each variable will affect the probability of the observation being a default customer if all other variables are held constant.

2.0 **Model Development Performance**

Once variable selection was performed and the final logistic regression model has been determined, the ROC can be generated to understand the threshold value to classify customers. The ROC curve for the logistic regression model can be seen below as Figure 1.

Figure 1: Logistic Regression ROC Curve



The threshold value is the point on the curve that is closest to the top left corner. The threshold value that was used to classify customers was 0.2864228. This means that all customers who had a predicted probability of default greater than 0.2864228 were assigned 1 and the rest were assigned 0. The performance of the predictions using the train and test datasets can be seen below in Table 3.

Model #3: Logistic Regression With Stepwise Feature Selection (Train)													
Actual	Predicted Class		Totals		Actual	Predicted Class		TP	0.46	TP+TN	1.34	AUC	0.75
Class	0	1			Class	0	1	TN	0.88	Precision	0.61	Sensitivity	0.46
0	9,306	1,318	10,624		0	0.88	0.12	Type I Error	0.12	Recall	0.46	Specificity	0.88
1	2,451	2,105	4,556		1	0.54	0.46	Type II Error	0.54	F1	0.58	Accuracy	62%

Model #3: Logistic Regression With Stepwise Feature Selection (Test)													
Actual	Predicted Class		Totals		Actual	Predicted Class		TP	0.43	TP+TN	1.32	AUC	0.71
Class	0	1			Class	0	1	TN	0.89	Precision	0.66	Sensitivity	0.43
0	4,385	531	4,916		0	0.89	0.11	Type I Error	0.11	Recall	0.43	Specificity	0.89
1	1,381	1,026	2,407		1	0.57	0.43	Type II Error	0.57	F1	0.56	Accuracy	74%

Table 3: Confusion matrix and classification metrics for Model #3: Logistic Regression With Stepwise Feature Selection Model.

As can be seen above, the Test dataset outperforms the Train dataset in overall accuracy which is a bit surprising given that the Train dataset was used to train the model and the Test dataset was never seen by the model. Diving deeper it can be seen that the Train dataset has a lower True Positive Rate compared to Test. The Train dataset has a lower False Positive Rate though compared to Test. Test outperforms Train slightly when comparing F1 Scores. Lastly, Train outperforms Test dataset in AUC. Overall, the Test dataset metrics will be used to evaluate the models as the Test dataset is a net new dataset to the model which models its performance in production better than the Train dataset.

After the confusion matrix was produced a lift chart was produced computing the Kolmogorov-Smirnov (KS) statistic for the train and test datasets. Twenty groups were used to generate the lift chart. Below you will see the lift chart for the Train dataset as Table 4.

Table 4: KS Lift Chart (Train Dataset)

Decile	Obs	Target (Y=1)	NonTarget (Y=0)	Target Density	NonTarget Density	Target CDF	NonTarget CDF	KS Stat
1	759	460	299	13.4%	2.5%	13.4%	2.5%	10.9%
2	759	371	388	10.8%	3.3%	24.3%	5.8%	18.4%
3	759	372	387	10.9%	3.3%	35.1%	9.1%	26.0%
4	759	360	399	10.5%	3.4%	45.7%	12.5%	33.1%
5	759	286	473	8.4%	4.0%	54.0%	16.6%	37.5%
6	759	254	505	7.4%	4.3%	61.4%	20.8%	40.6%
7	759	146	613	4.3%	5.2%	65.7%	26.1%	39.6%
8	759	148	611	4.3%	5.2%	70.0%	31.3%	38.8%
9	759	115	644	3.4%	5.5%	73.4%	36.7%	36.7%
10	759	120	639	3.5%	5.4%	76.9%	42.2%	34.7%
11	759	116	643	3.4%	5.5%	80.3%	47.6%	32.6%
12	759	83	676	2.4%	5.7%	82.7%	53.4%	29.3%
13	759	102	657	3.0%	5.6%	85.7%	59.0%	26.7%
14	759	89	670	2.6%	5.7%	88.3%	64.7%	23.6%
15	759	64	695	1.9%	5.9%	90.2%	70.6%	19.6%
16	759	108	651	3.2%	5.5%	93.3%	76.1%	17.2%
17	759	71	688	2.1%	5.9%	95.4%	82.0%	13.4%
18	759	58	701	1.7%	6.0%	97.1%	87.9%	9.1%
19	759	63	696	1.8%	5.9%	98.9%	93.9%	5.1%
20	759	37	722	1.1%	6.1%	100.0%	100.0%	0.0%
Totals	**15,180**	**3423**	**11,757**	**100.0%**	**100.0%**			

As can be seen above in Table 4, the KS statistic for the Train dataset is 40.6. A KS statistic of 100 would mean that there is no distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution so a KS of 40.6 shows there is a distance between the too.

The lift chart for the Test dataset can be seen below as Table 5.

Table 5: KS Lift Chart (Test Dataset)								
Decile	Obs	Target (Y=1)	NonTarget (Y=0)	Target Density	NonTarget Density	Target CDF	NonTarget CDF	KS Stat
1	367	215	152	13.8%	2.6%	13.8%	2.6%	11.2%
2	366	169	197	10.9%	3.4%	24.7%	6.1%	18.6%
3	366	175	191	11.2%	3.3%	35.9%	9.4%	26.5%
4	366	154	212	9.9%	3.7%	45.8%	13.0%	32.8%
5	366	149	217	9.6%	3.8%	55.4%	16.8%	38.6%
6	366	108	258	6.9%	4.5%	62.3%	21.3%	41.0%
7	366	83	283	5.3%	4.9%	67.6%	26.2%	41.4%
8	366	59	307	3.8%	5.3%	71.4%	31.5%	39.9%
9	366	51	315	3.3%	5.5%	74.7%	37.0%	37.7%
10	366	45	321	2.9%	5.6%	77.6%	42.5%	35.0%
11	367	55	312	3.5%	5.4%	81.1%	48.0%	33.2%
12	366	34	332	2.2%	5.8%	83.3%	53.7%	29.6%
13	366	44	322	2.8%	5.6%	86.1%	59.3%	26.8%
14	366	35	331	2.2%	5.7%	88.4%	65.0%	23.3%
15	366	31	335	2.0%	5.8%	90.4%	70.8%	19.5%
16	366	45	321	2.9%	5.6%	93.3%	76.4%	16.8%
17	366	30	336	1.9%	5.8%	95.2%	82.2%	12.9%
18	366	30	336	1.9%	5.8%	97.1%	88.1%	9.0%
19	366	29	337	1.9%	5.8%	99.0%	93.9%	5.1%
20	367	16	351	1.0%	6.1%	100.0%	100.0%	0.0%
Totals	7,323	1557	5,766	100.0%	100.0%			

As can be seen above in Table 5, the KS statistic for the Test dataset is 41.4. Comparing the KS statistic for the Train and Test datasets it can be seen that the Test dataset has a higher KS statistic than the Train dataset. This means that the model is performing slightly better when comparing KS statistic values.

3.0 Performance Monitoring Plan

After the KS statistic was determined for the Train and Test datasets a RAG status (Red-Amber-Green) can be determined to structure the monitoring plan. When a RAG status is Green it means that the model is performing as expected. When a RAG status is Amber, it means that the KS statistic of model has dropped below the original KS statistic and is within 10% of the original value. If a model is Amber, it should be reevaluated in three-months.

Lastly, if a RAG status is Red, it means that the KS statistic of the model has dropped further than 10% of the original value. If a model is Red, it needs to be remodeled. Now that the high-level definitions of each status have been discussed, the levels for each status for the logistic regression model can be seen below in Table 6.

Table 6: Logisitic Regression RAG Status			
	Status		
Baseline Test KS Statistic	Green	Amber	Red
41.4	41.4 >=	37.26 >=	< 37.26

As can be seen above in Table 6, 41.4 is the baseline KS statistic for the logistic regression model on the Test dataset. The model should be considered in Green status if the KS statistic is greater than or equal to 41.4. The model should be considered in Amber status if the KS statistic is greater than or equal to 37.26 and does not meet the Green criteria. Lastly, the model should be considered in Red status if the KS statistic is less than 37.26.

4.0 Performance Monitoring Results

The Kolmogorov-Smirnov (KS) statistic for the validation dataset was produced. Twenty groups were used to generate the lift chart. Below you will see the lift chart for the Validate dataset as Table 7.

Table 7: KS Lift Chart (Validate Dataset)								
Decile	Obs	Target (Y=1)	NonTarget (Y=0)	Target Density	NonTarget Density	Target CDF	NonTarget CDF	KS Stat
1	375	217	158	13.1%	2.7%	13.1%	2.7%	10.4%
2	375	213	162	12.9%	2.8%	26.0%	5.5%	20.5%
3	375	182	193	11.0%	3.3%	37.0%	8.8%	28.2%
4	375	181	194	10.9%	3.3%	47.9%	12.1%	35.8%
5	374	151	223	9.1%	3.8%	57.0%	15.9%	41.1%
6	375	103	272	6.2%	4.7%	63.2%	20.6%	42.6%
7	375	66	309	4.0%	5.3%	67.2%	25.9%	41.3%
8	375	62	313	3.7%	5.4%	71.0%	31.2%	39.7%
9	375	62	313	3.7%	5.4%	74.7%	36.6%	38.1%
10	374	43	331	2.6%	5.7%	77.3%	42.3%	35.0%
11	375	41	334	2.5%	5.7%	79.8%	48.0%	31.8%
12	375	49	326	3.0%	5.6%	82.7%	53.6%	29.2%
13	375	45	330	2.7%	5.6%	85.4%	59.2%	26.2%
14	375	42	333	2.5%	5.7%	88.0%	64.9%	23.1%
15	374	32	342	1.9%	5.9%	89.9%	70.8%	19.2%
16	375	46	329	2.8%	5.6%	92.7%	76.4%	16.3%
17	375	37	338	2.2%	5.8%	94.9%	82.2%	12.7%
18	375	32	343	1.9%	5.9%	96.9%	88.0%	8.8%
19	375	30	345	1.8%	5.9%	98.7%	94.0%	4.7%
20	375	22	353	1.3%	6.0%	100.0%	100.0%	0.0%
Totals	**7,497**	**1,656**	**5,841**	**100.0%**	**100.0%**			

As can be seen above in Table 7, the KS statistic for the Validate dataset is42.6. Comparing the KS statistic for the Validate dataset to the Train and Test datasets it can be seen that the Validate KS statistic is greater than the Train and Test KS statistic. The RAG status table, Table 6, was then used to determine which status the model should be given. Based off of the RAG status table the model received a RAG status of Green due to 42.6 being greater than 41.4. The recommendation is to leave the model alone and to revisit the performance of the model in 6 months.