Model #101: Credit Card Default Model

Performance Monitoring Plan

Tim Crnkovic

1. The Production Model

The production model is a logistic regression, created with the "glm" R package. Before training this model, feature selection was performed with L1 (Lasso) regularization. The results of the regularization are shown in Table 1 below. The variables whose weights were zeroed-out by regularization were excluded from the subsequent logistic regression training. The eleven predictor features that were used to train the model are those shown with non-zero regularized weights in the Table 1.

Variable	L1 Regularized Weight
(Intercept)	1.53E-01
LIMIT_BAL	-1.71E-07
SEX	-1.58E-02
EDUCATION	0
MARRIAGE	-2.22E-02
Avg_Util	1.68E-03
Avg_Pmt_Ratio	0
Pmt_Ratio_Weighted_Mean	0
Age_21_25	0
Age_26_40	-9.82E-03
Age_41_79	0
Bal_Growth_6Mo	-6.79E-07
Tot_Neg_Balances	0
Util_Growth_6Mo	-1.29E-02
Util_Incr_From_Min	0
Util_Weighted_Mean	0
Avg_Bill_Amt	7.06E-08
Avg_Pmt_Amt	-1.41E-06
Max_Bill_Amt	0
Max_Pmt_Amt	0
Max_DLQ	5.99E-02
Tot_DLQ	2.63E-02

Table 1. Features Selected for Logistic Regression by L1 Regularization.

Table 2 below is the data dictionary of the features that were included in the training of the logistic regression model.

Element	Description	Data type	Acceptable values
LIMIT_BAL	The amount of credit given.	Integer, continuous	> 0
SEX	Gender.	Integer, categorical	1 = male, 2 = female
EDUCATION	Level of education.	Integer, categorical	1 = graduate school, 2 = university, 3 = high school, 4 = others
MARRIAGE	Marital status.	Integer, categorical	1 = married, 2 = single, 3 = others
Max_DLQ	Maximum delinquency value over six consecutive pay periods.	Integer, categorical	0-9
Tot_DLQ	Total of delinquency values over six consecutive pay periods.	Integer, categorical	0-54
Util_Growth_6Mo	Growth of utlization over six consectutive pay periods, where utilization is the ratio of the anount billed to the LIMIT_BAL for a given pay period.	Real, continuous	[-1,1]
Age_26_40	Age bin of 26 to 40, inclusive.	Integer, categorical	0 = not in age group, 1 = in age group
Avg_Util	Arithmetic mean of utlization over six consectutive pay periods, where utilization is the ratio of the amount billed to the LIMIT_BAL for a given pay period.	Real, continuous	[0,1]
Bal_Growth_6Mo	Percentage change in balance over six consectutive pay periods.	Real, continuous	Theoretically [-1,∞) but to avoid values of ∞, limited to [-1,b] where b = max value of this feature other than ∞ in the overall data set
Avg_Pmt_Amt	Arithmetic mean of payments made over six consectutive pay periods.	Real, continuous	>= 0
Avg_Bill_Amt	Arithmetic mean of billing amounts over six consectutive pay period.	Real, continuous	>= 0

Table 2. Data Dictionary of Features Included in Logistic Regression.

Table 3 below shows coefficients, z scores, P values and significance for the features that were included in model training. This table itemizes the categorical features by level. By convention, the first level of each categorical variable is defined to have an estimate of zero and that level is not displayed.

Feature	Estimate	z value	Pr(>|z|)	Significance
(Intercept)	-1.39	-17.663	< 2.00E-16	***
LIMIT_BAL	0.00	-5.209	1.90E-07	***
Max_DLQ1	13.30	0.058	0.953784	
Max_DLQ2	15.03	0.065	0.94781	
Max_DLQ3	15.23	0.066	0.947105	
Max_DLQ4	14.82	0.065	0.948511	
Max_DLQ5	13.51	0.059	0.953064	
Max_DLQ6	14.38	0.063	0.950036	
Max_DLQ7	27.54	0.069	0.9448	
Max_DLQ8	14.01	0.043	0.96559	
Tot_DLQ1	-12.26	-0.053	0.957421	
Tot_DLQ2	-13.83	-0.06	0.951965	
Tot_DLQ3	-14.23	-0.062	0.950585	
Tot_DLQ4	-13.44	-0.059	0.953297	
Tot_DLQ5	-13.48	-0.059	0.953178	
Tot_DLQ6	-12.84	-0.056	0.955383	
Tot_DLQ7	-13.27	-0.058	0.953913	
Tot_DLQ8	-12.63	-0.055	0.95614	
Tot_DLQ9	-13.06	-0.057	0.954619	
Tot_DLQ10	-12.80	-0.056	0.955545	
Tot_DLQ11	-12.95	-0.056	0.955021	
Tot_DLQ12	-11.95	-0.052	0.958488	
Tot_DLQ13	-12.27	-0.053	0.957363	
Tot_DLQ14	-12.54	-0.055	0.956441	
Tot_DLQ15	-11.77	-0.051	0.959101	
Tot_DLQ16	-12.23	-0.053	0.957511	
Tot_DLQ17	-10.90	-0.047	0.962131	
Tot_DLQ18	-10.91	-0.048	0.962105	
Tot_DLQ19	-11.98	-0.052	0.958385	
Tot_DLQ20	-11.73	-0.051	0.959261	
Tot_DLQ21	-13.57	-0.059	0.952861	
Tot_DLQ22	-13.38	-0.058	0.953509	
Tot_DLQ24	-11.61	-0.051	0.959647	
Tot_DLQ27	-13.31	-0.032	0.974676	
Tot_DLQ28	-23.22	-0.058	0.953443	
Tot_DLQ31	-13.27	-0.041	0.967407	
Tot_DLQ32	-24.36	-0.061	0.951166	
Tot_DLQ33	-11.63	-0.036	0.971429	
MARRIAGE2	-0.18	-3.968	7.25E-05	***
MARRIAGE3	-0.11	-0.566	0.571414	
SEX2	-0.16	-3.665	0.000248	***
Util_Growth_6Mo	-0.25	-3.318	0.000906	***
Age_26_40	-0.09	-1.902	0.057234	.
Avg_Util	-0.18	-1.8	0.07185	.
Bal_Growth_6Mo	0.00	-2.103	0.035434	*
Avg_Pmt_Amt	0.00	-5.733	9.86E-09	***
Avg_Bill_Amt	0.00	3.373	0.000743	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table 3. Coefficients, z Scores and P Values for Logistic Regression Features.

Table 3 shows that the only feature significantly correlated with an increased likelihood of default is Avg_Bill_Amt, meaning that an increase in its value corresponds to increased probability of default with very high confidence. Note that this is true even though its estimate shows as 0. This is a result of the great range in values for this feature. Even a small increase in Avg_Bill_Amt will likely lead to a greater chance of default.

Table 3 also shows several features significantly related to a decreased probability of default: Avg_Pmt_Amt, Util_Growth_6Mo, LIMIT_BAL, MARRIAGE2 (Single) and SEX2 (Female). Increases in these should lead to a smaller chance of default. This is true for LIMIT_BAL and Avg_Pmt_Amt even with their estimates of 0. Again, that is a result of their large ranges of values. Even small changes in their value could lead to significant effects on chance of default.

2. Model Development Performance

The full results of the training and testing with the logistic regression model are summarized in Table 4 below. The model produced a true positive rate of 0.37, a false positive rate of 0.07 and an overall accuracy of 0.80, an F1 score of 0.51 and an AUC of 0.63 against the training set. The model then produced a true positive rate of 0.36, a false positive rate of 0.07 and an overall accuracy of 0.81, an F1 score of 0.51 and an AUC of 0.65 against the test set.

Logistic Regression Model - Training Set													
Actual Class	**Predicted Class**		**Totals**		**Actual Class**	**Predicted Class**		**TP**	**0.37**	**TP+TN**	**1.30**	**AUC**	**0.63**
	0	**1**				**0**	**1**	**TN**	**0.93**	**Precision**	**0.61**	**Sensitivity**	**0.37**
0	**10,762**	**813**	**11,575**		**0**	**0.93**	**0.07**	**Type I Error**	**0.07**	**Recall**	**0.37**	**Specificity**	**0.93**
1	**2,144**	**1,263**	**3,407**		**1**	**0.63**	**0.37**	**Type II Error**	**0.63**	**F1**	**0.51**		
Logistic Regression Model - Test Set													
Actual Class	**Predicted Class**		**Totals**		**Actual Class**	**Predicted Class**		**TP**	**0.36**	**TP+TN**	**1.29**	**AUC**	**0.65**
	0	**1**				**0**	**1**	**TN**	**0.93**	**Precision**	**0.57**	**Sensitivity**	**0.36**
0	**5,253**	**421**	**5,674**		**0**	**0.93**	**0.07**	**Type I Error**	**0.07**	**Recall**	**0.36**	**Specificity**	**0.93**
1	**983**	**565**	**1,548**		**1**	**0.64**	**0.36**	**Type II Error**	**0.64**	**F1**	**0.51**		

Table 4. Confusion Matrix and Classification Metrics for Logistic Regression Model.

The ROC (Receiver Operating Characteristic) curve of the results of the logistic regression model on the training set is shown in Figure 1 below. As mentioned, the area under the ROC curve for the training set is 0.63.

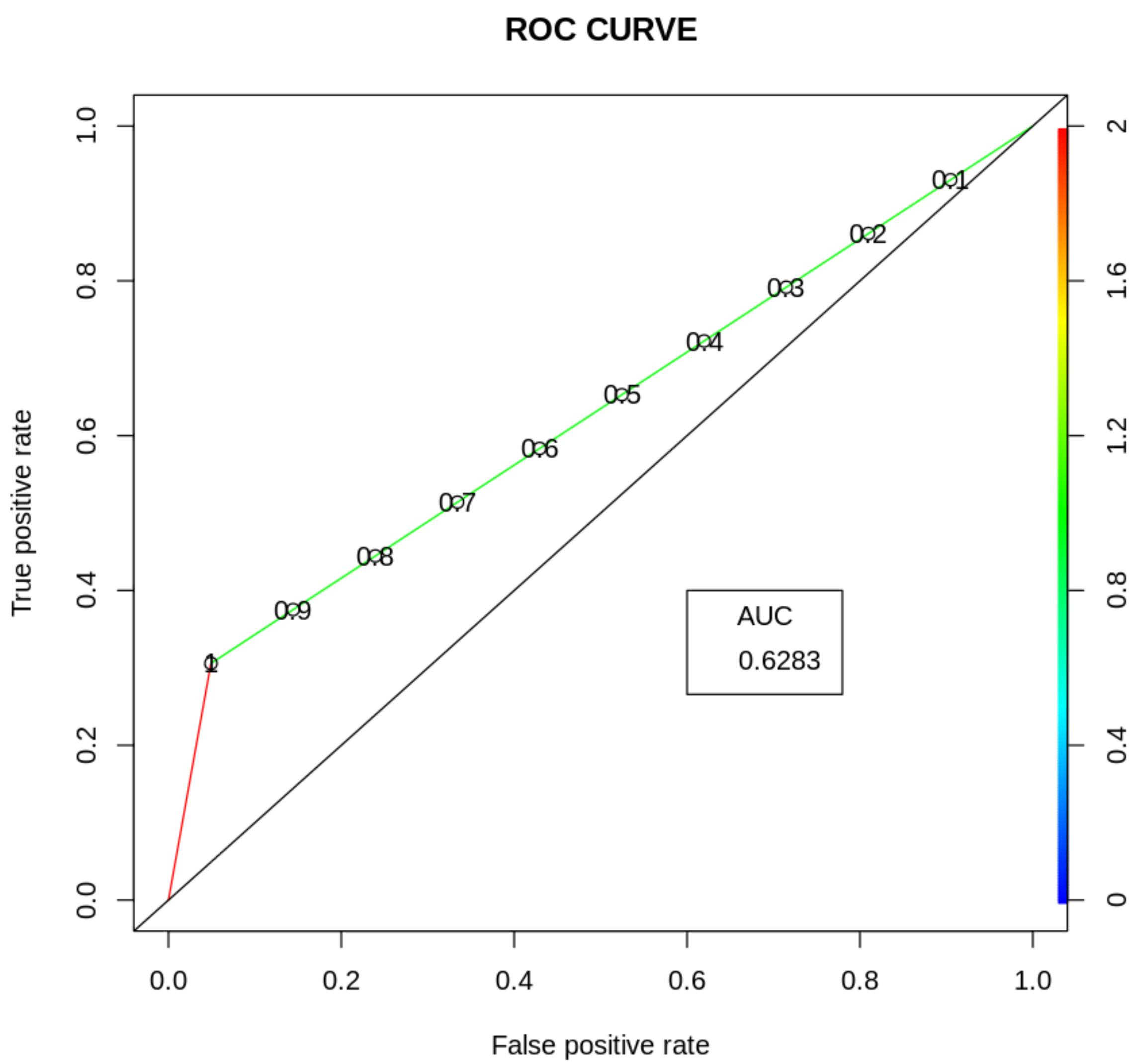

Figure 1. ROC Curve for Logistic Regression on Training Set.

The ROC (Receiver Operating Characteristic) curve of the results of the logistic regression model on the test set is shown in Figure 2 below. As mentioned, the area under the ROC curve for the training set is 0.65.

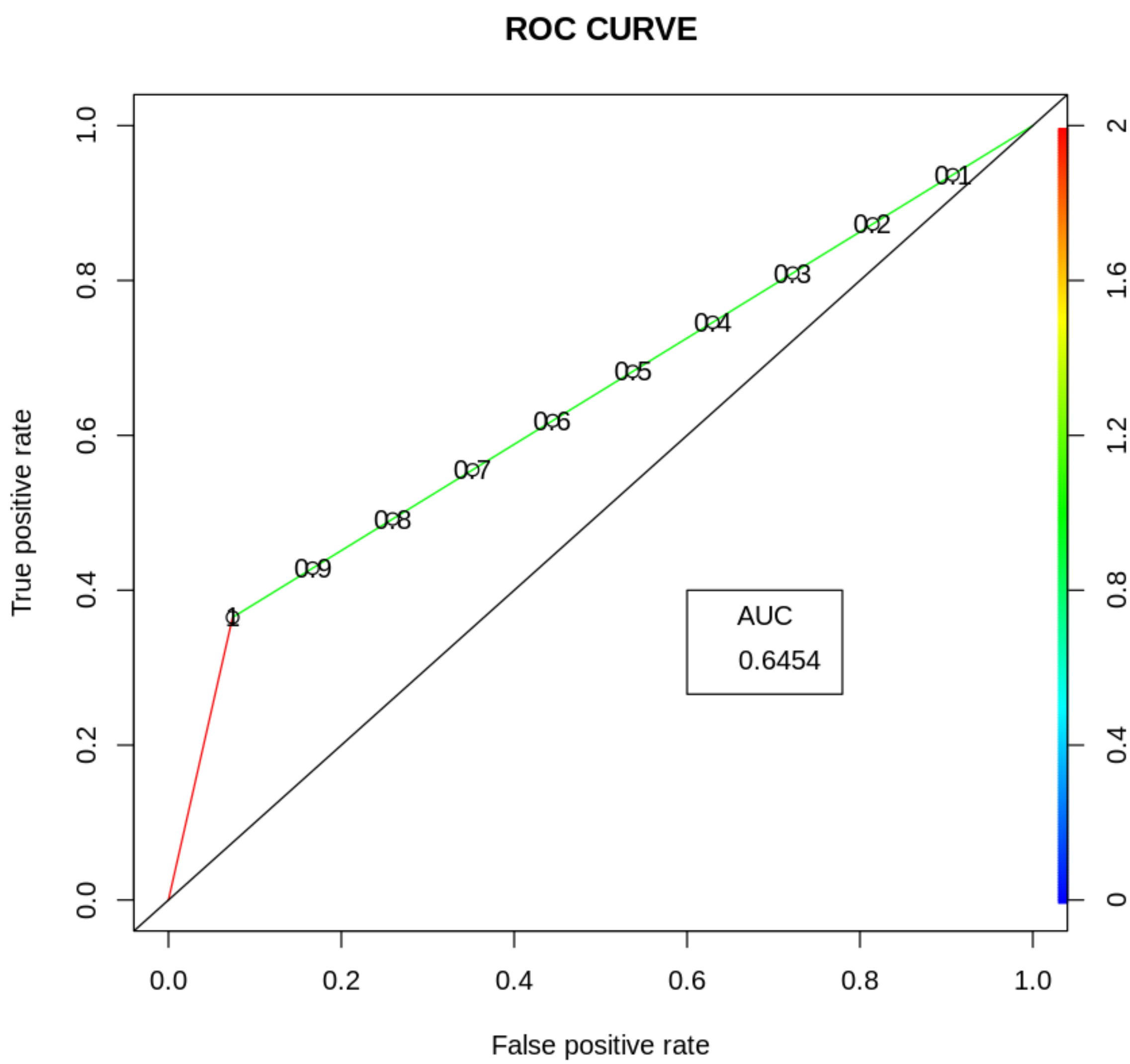

Figure 2. ROC Curve for Logistic Regression on Test Set.

Lift charts were created for the results of this model against both the training and the test sets. The chart contains 20 semi-deciles and it shows the Kolmogorov-Smirnov (KS) statistic. The lift chart for the training set is shown in Table 5 below. The maximum KS statistic on this set is 39.2%.

Semi Decile	Obs	Target (Y=1)	NonTarget (Y=0)	Target Density	NonTarget Density	Target CDF	NonTarget CDF	KS Stat
1	750	552	198	16.2%	1.7%	16.2%	1.7%	14.5%
2	749	427	322	12.5%	2.8%	28.7%	4.5%	24.2%
3	749	359	390	10.5%	3.4%	39.3%	7.9%	31.4%
4	749	279	470	8.2%	4.1%	47.5%	11.9%	35.5%
5	749	231	518	6.8%	4.5%	54.2%	16.4%	37.8%
6	749	202	547	5.9%	4.7%	60.2%	21.1%	39.0%
7	749	175	574	5.1%	5.0%	65.3%	26.1%	**39.2%**
8	749	142	607	4.2%	5.2%	69.5%	31.3%	38.1%
9	749	120	629	3.5%	5.4%	73.0%	36.8%	36.2%
10	749	98	651	2.9%	5.6%	75.9%	42.4%	33.5%
11	749	108	641	3.2%	5.5%	79.0%	47.9%	31.1%
12	749	88	661	2.6%	5.7%	81.6%	53.6%	28.0%
13	749	98	651	2.9%	5.6%	84.5%	59.3%	25.2%
14	749	88	661	2.6%	5.7%	87.1%	65.0%	22.1%
15	749	75	674	2.2%	5.8%	89.3%	70.8%	18.5%
16	749	87	662	2.6%	5.7%	91.8%	76.5%	15.3%
17	749	69	680	2.0%	5.9%	93.9%	82.4%	11.5%
18	749	74	675	2.2%	5.8%	96.0%	88.2%	7.8%
19	749	78	671	2.3%	5.8%	98.3%	94.0%	4.3%
20	750	57	693	1.7%	6.0%	100.0%	100.0%	0.0%
Totals	**14,982**	**3407**	**11,575**	**100.0%**	**100.0%**			

Table 5. Lift Chart Including KS Statistic for Training Set.

The lift chart for the test set is shown in Table 6 below. Its maximum KS statistic is 40.7%.

Semi Decile	Obs	Target (Y=1)	NonTarget (Y=0)	Target Density	NonTarget Density	Target CDF	NonTarget CDF	KS Stat
1	362	246	116	15.9%	2.0%	15.9%	2.0%	13.8%
2	361	191	170	12.3%	3.0%	28.2%	5.0%	23.2%
3	361	165	196	10.7%	3.5%	38.9%	8.5%	30.4%
4	361	124	237	8.0%	4.2%	46.9%	12.7%	34.2%
5	361	102	259	6.6%	4.6%	53.5%	17.2%	36.3%
6	361	103	258	6.7%	4.5%	60.1%	21.8%	38.4%
7	361	106	255	6.8%	4.5%	67.0%	26.3%	**40.7%**
8	361	61	300	3.9%	5.3%	70.9%	31.6%	39.4%
9	361	53	308	3.4%	5.4%	74.4%	37.0%	37.4%
10	361	53	308	3.4%	5.4%	77.8%	42.4%	35.4%
11	361	43	318	2.8%	5.6%	80.6%	48.0%	32.5%
12	361	38	323	2.5%	5.7%	83.0%	53.7%	29.3%
13	361	41	320	2.6%	5.6%	85.7%	59.4%	26.3%
14	361	33	328	2.1%	5.8%	87.8%	65.1%	22.7%
15	361	36	325	2.3%	5.7%	90.1%	70.9%	19.2%
16	361	34	327	2.2%	5.8%	92.3%	76.6%	15.7%
17	361	31	330	2.0%	5.8%	94.3%	82.4%	11.9%
18	361	27	334	1.7%	5.9%	96.1%	88.3%	7.7%
19	361	27	334	1.7%	5.9%	97.8%	94.2%	3.6%
20	362	34	328	2.2%	5.8%	100.0%	100.0%	0.0%
Totals	**7,222**	**1548**	**5,674**	**100.0%**	**100.0%**			

Table 6. Lift Chart Including KS Statistic for Test Set.

3. Performance Monitoring Plan

Table 7 below defines the performance monitoring plan for the logistic regression model .

Status	Kologormov-Smirnov Statisitic Threshhold	Prescribed Action
Red	< 37%	Model needs redevelopment
Amber	>= 37% and < 39%	Model needs to be re-validated in three months
Green	>= 39%	Model is performing as expected.

Table 7. Performance Monitoring Plan Definition.

If the KS statistic for the model in Production remains at a value of at least 39%, then the model is considered to be in Green status - performing as expected. In this status, no action is needed. If the Production model produces a KS statistic less than 39% but at least 37%, then it is considered to be in Amber status. In this case, the model needs to be re0validated in three months. If the KS statistic falls below 37% in Production, then the model is in Red status, meaning that it needs redevelopment.