Northwestern MSDS-498

Model #101: Credit Card Default Model

Performance Monitoring Plan

Andrew Stevens

September 1, 2022

1. The Production Model

To develop a logistic regression model, the feature importance from two of the other models developed for this study, Random Forest and XGboost, were calculated and the top three were taken from each. The resulting explanatory variables are: pay_max, payment_avg, payment_max, ratio_avg, balance_growth_6mo. Logistic regression is one of the simplest data science approaches to fitting a model – using an algebraic expression with the number of variables chosen as a hyperparameter. The coefficient for each variable is provided in table 1. A review of the p-values tells that all the input variables are likely to be relevant, as the p-values are very low.

Table 1: Logistic Regression Model Summary

Variable Name	coef	std err	z	P-values
payment_avg	.55	3.45E-06	0.715	1.66834366e-03
ratio_avg	-1.20 e-04	0.003	-0.62	0.00000000e+00
balance_growth_6mo	1.59 e-05	0.088	-3.351	0.00000000e+00
payment_max	-.11	0.018	-5.356	0.00000000e+00
pay_max	-4.88 e-04	0.021	34.966	2.44573664e-04

2. Model Development Performance

The logistic regression model has proven to generalize well, as it performs similarly on the test data set in comparison to the training set. The overall performance, however, may not be sufficient to ensure business profitability, dependent on the cost function.

a. Training Data Model Performance

It is expected that the model will perform best on the training data set, as this was where the optimization was performed for selecting the coefficients. The Kolmogorov-Smirnov (KS) test statistic measures "goodness of fit" through the degree of separation between the positive and negative distributions (Darling, 1957). A perfect score would be 100, with complete separation; a zero would be a completely useless model providing no separation. For the variable selection-based logistic regression model, the KS is 16.2% at decile 6. This indicates poor performance. This is confirmed with the ROC-AUC curve, displaying an AUC of only thirty-eight percent.

Table 2: Logistic Regression Model Kolmogorov-Smirnov chart: Training Data

Decile	min_prob	max_prob	events	nonevents	event_rate	nonevent_rate	Cum eventrate	Cum noneventrate	Abs(KS)
1	0.7	1	72	1118	0.81%	7.51%	0.81%	7.51%	6.7
2	0.63	0.7	172	1017	1.93%	6.83%	2.74%	14.34%	11.6
3	0.59	0.63	260	930	2.92%	6.25%	5.66%	20.59%	14.9
4	0.57	0.59	410	779	4.61%	5.23%	10.27%	25.82%	15.6
5	0.55	0.57	415	770	4.66%	5.17%	14.93%	31.00%	16.1
6	0.54	0.55	437	756	4.91%	5.08%	19.84%	36.07%	16.2
7	0.53	0.54	529	661	5.94%	4.44%	25.78%	40.51%	14.7
8	0.53	0.53	462	728	5.19%	4.89%	30.97%	45.41%	14.4

|									
9	0.52	0.53	483	706	5.43%	4.74%	36.39%	50.15%	13.8
10	0.52	0.52	512	677	5.75%	4.55%	42.14%	54.70%	12.6
11	0.52	0.52	479	707	5.38%	4.75%	47.52%	59.45%	11.9
12	0.51	0.52	459	734	5.16%	4.93%	52.68%	64.38%	11.7
13	0.51	0.51	424	766	4.76%	5.15%	57.44%	69.52%	12.1
14	0.51	0.51	480	708	5.39%	4.76%	62.83%	74.28%	11.4
15	0.51	0.51	467	723	5.25%	4.86%	68.08%	79.13%	11.1
16	0.5	0.51	437	753	4.91%	5.06%	72.99%	84.19%	11.2
17	0.5	0.5	544	644	6.11%	4.33%	79.10%	88.52%	9.4
18	0.5	0.5	461	727	5.18%	4.88%	84.27%	93.40%	9.1
19	0.5	0.5	420	772	4.72%	5.19%	88.99%	98.59%	9.6
20	0.5	0.5	980	210	11.01%	1.41%	100.00%	100.00%	0
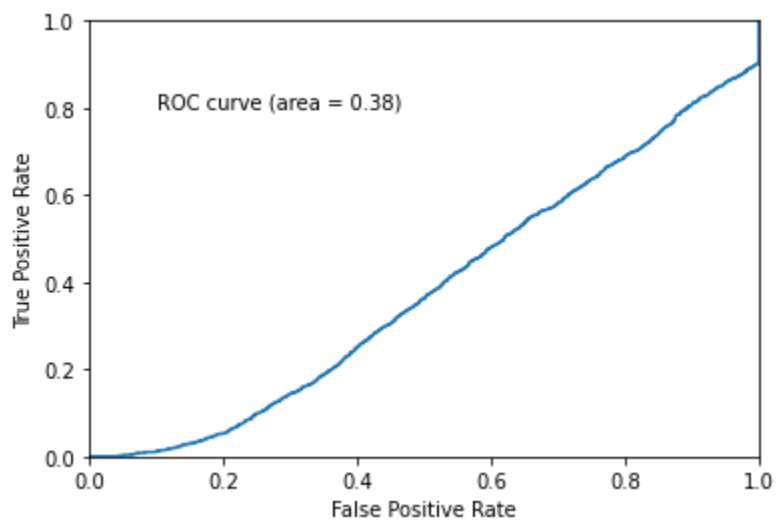
Figure 1: ROC: Training Performance

b. Test Data Model Performance

When the model was executed using the test data set, the KS is 22.0% at decile 5 – slightly improved from the training set, though the inclusion of the ROC-AUC plot reveals that overall performance is in fact slightly worse.

Table 3: Logistic Regression Kolmogorov-Smirnov chart

Decile	min_prob	max_prob	events	nonevents	event_rate	nonevent_rate	Cum eventrate	Cum noneventrate	Abs(KS)
1	0.74	1	5	359	0.22%	7.21%	0.22%	7.21%	7
2	0.65	0.74	13	351	0.57%	7.05%	0.78%	14.26%	13.5
3	0.61	0.65	48	316	2.09%	6.35%	2.87%	20.61%	17.7
4	0.58	0.61	73	290	3.18%	5.82%	6.05%	26.43%	20.4
5	0.56	0.58	90	274	3.92%	5.50%	9.97%	31.93%	22
6	0.55	0.56	126	238	5.49%	4.78%	15.45%	36.71%	21.3
7	0.54	0.55	114	250	4.96%	5.02%	20.42%	41.74%	21.3
8	0.53	0.54	131	232	5.70%	4.66%	26.12%	46.39%	20.3
9	0.53	0.53	134	230	5.83%	4.62%	31.95%	51.01%	19.1
10	0.52	0.53	127	237	5.53%	4.76%	37.48%	55.77%	18.3
11	0.52	0.52	141	223	6.14%	4.48%	43.62%	60.25%	16.6
12	0.52	0.52	144	219	6.27%	4.40%	49.89%	64.65%	14.8
13	0.51	0.52	132	232	5.75%	4.66%	55.64%	69.31%	13.7
14	0.51	0.51	125	239	5.44%	4.80%	61.08%	74.11%	13

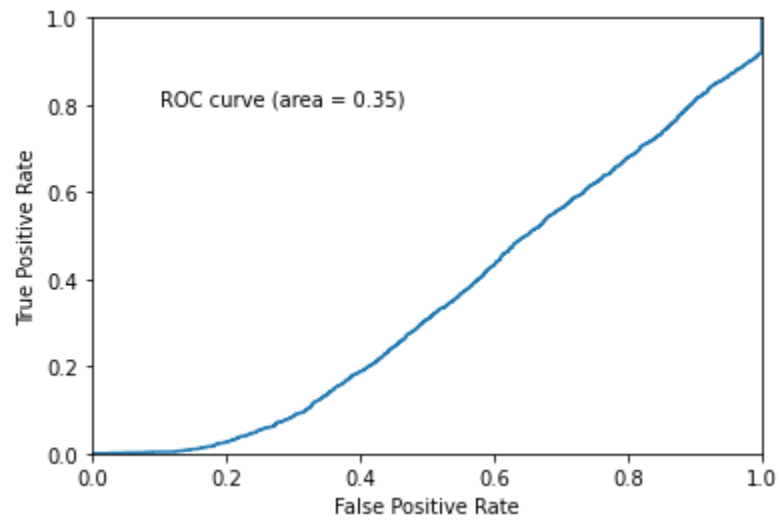|									
15	0.51	0.51	123	241	5.35%	4.84%	66.43%	78.95%	12.5
16	0.51	0.51	126	237	5.49%	4.76%	71.92%	83.71%	11.8
17	0.5	0.51	143	221	6.23%	4.44%	78.15%	88.15%	10
18	0.5	0.5	143	221	6.23%	4.44%	84.37%	92.59%	8.2
19	0.5	0.5	115	249	5.01%	5.00%	89.38%	97.59%	8.2
20	0.5	0.5	244	120	10.62%	2.41%	100.00%	100.00%	0

Figure 2: ROC: Test Performance

3. Performance Monitoring Plan

In order to ensure continued positive results, the model performance must be monitored through test statistics and thresholds must be established for when action is required. Specific performance values are not as relevant for a business as the impacts they produce. If a model accuracy reduction of five-percent results in twenty-percent increase in costs, 50-percent decrease in profits, five-percent is much worse in reality than it initially seems. A more relevant performance plan may require a more complex model. Classification of defaulting customers binarily may be insufficient. The model may also need to monitor and predict severity of that default, with high loss (either graduated or threshold of dollar amount) being recognized as another risk factor. To begin, a simple plan can be devised.

Table 4: RAG Thresholds

Status	KS threshold	AUC threshold
Green	>15%	>30%
Amber	10-15%	25-30%
Red	<10%	<25%

These thresholds are set assuming that current performance is sufficient – though that is not likely the case with the low initial performance values. It can also be helpful to monitor multiple measures if one is insufficient for recognizing all areas of potential loss. While the

model continues to score in the green domain no action is necessary, and the model is considered healthy. Performance falling into the amber region would require more active monitoring and preparation for response within three months with re-validation. When KS or AUC falls below the red threshold, immediate redevelopment is required. The thresholds will also be periodically re-evaluated annually to determine if business impacts have changed and require corresponding adjustment of the monitoring plan.

4. Bibliography

1.   Model Evaluation. https://www.saedsayad.com/model_evaluation_c.htm. Accessed 20 Aug. 2022.

2.   Darling, D. A. "The Kolmogorov-Smirnov, Cramer-von Mises Tests." The Annals of Mathematical Statistics, vol. 28, no. 4, Dec. 1957, pp. 823–38. DOI.org (Crossref), https://doi.org/10.1214/aoms/1177706788.