Spring 2019

Matthew Dobbin

Northwestern University

Default of Credit Card Clients

Performance Validation Results

Contents

1.0 The Production Model

This report documents the performance monitoring plan for a Logistic Regression model used to predict the probability of credit card default. The data set that was used to train the model was obtained from the UCI Machine Learning Repository (UCI, 2016). It contains 30,000 observations and 23 predictor variables and the binary variable, DEFAULT, is the response variable. A value of one represents a default payment. Table 1 shows the split between on time and default payment observations. A detailed description of the exploratory data analysis and feature engineering process can be found in the Model Development Guide. A data dictionary and a description of the engineered features can be found in Appendix 1.

Table 1 - Observation count of the values for the default payments variable.

Data Set	0	1
Training	11757	3423
Test	5766	1557
Validation	5841	1656

The Logistic Regression model was fitted using a backward variable selection algorithm. The full model contained 21 variables that had importance in a Gradient Boosted model that was fitted using all of the predictors. The backward selection method then reduced the full model down to 15 predictor variables. The variables in the final model and their logit coefficients are shown in Table 2. The predictor with the largest z-value is PAY_R1_G1. This variable indicates that the client has delayed payment by one or more months. The positive associate of the PAY_R1_G1 coefficient also makes logical sense. Delayed payments increase the probability of defaulting.

Table 2 - Logit coefficient estimates for the backward variable selection model.

```
Coefficients:
                           Estimate Std. Error z value              Pr(>|z|)
(Intercept)                -2.25251    0.09481 -23.758 < 0.0000000000000002
PAY_R1_G1                   1.35710    0.08183  16.585 < 0.0000000000000002
MAX_DLQ_G1                  0.63702    0.07075   9.004 < 0.0000000000000002
PAY_R1_LE0                 -0.34898    0.07392  -4.721   0.0000023446534336
AVG_PAY_AMT_LE2000          0.31176    0.05242   5.948   0.0000000027192309
PAY_R5_G0                   0.45818    0.07480   6.125   0.0000000009067502
PAY_R3_G0                   0.31164    0.07411   4.205   0.0000260742623733
PAY_R2_G0                   0.11886    0.08328   1.427              0.15354
MAX_BILL_AMT_LE600          1.11345    0.10837  10.274 < 0.0000000000000002
BAL_GROWTH_6MO_N10172_923   0.11074    0.05161   2.146              0.03191
MARRIAGE_MD                 0.15834    0.04420   3.583              0.00034
MAX_BILL_AMT_600_4079       0.63606    0.08743   7.276   0.0000000000003451
UTIL_AUG_G24                0.44047    0.05701   7.726   0.0000000000000111
AVG_PMT_RATIO_46_1         -0.10205    0.05348  -1.908              0.05636
MAX_BILL_AMT_4079_18400     0.31606    0.06670   4.738   0.0000021562288779
SEX_M                       0.14318    0.04478   3.197              0.00139
```

The variables in the final model are all indicator variables. The Weight of Evidence supervised binning algorithm was used to identify optimal bin sizes for the continuous variables which were then discretised into indicator variables. A description of each indicator variable in the model is provided in Table 3.

Table 3 - Description of the variables in the Logistic Regression model.

Variable	Description
PAY_R1_G1	PAY_R1 greater than one
MAX_DLQ_G1	MAX_DLQ greater than one
PAY_R1_LE0	PAY_R1 less or equal to zero
AVG_PAY_AMT_LE2000	AVG_PAY_AMT less than $2000
PAY_R5_G0	PAY_R5 greater than zero
PAY_R3_G0	PAY_R3 greater than zero
PAY_R2_G0	PAY_R2 greater than zero
MAX_BILL_AMT_LE600	MAX_BILL_AMT less than or equal to $600
BAL_GROWTH_6M0_N10172_923	BAL_GROWTH_6MO between -$10172 to $923
MARRIAGE_MD	Married indicator variable
MAX_BILL_AMT_600_4079	MAX_BILL_AMT between $600-$4079
UTIL_AUG_G24	UTIL_AUG greater than 24.4
AVG_PMT_RATIO_46_1	AVG_PMT_RATIO between 46.0-46.1
MAX_BILL_AMT_4079_18400	MAX_BILL_AMT between $4079-$18400
SEX_M	Male indicator variable

2.0 Model Development Performance

The in sample (training) and out of sample (test) performance of the model was analysed using several metrics. These metrics were accuracy, true positive rate (TPR), false positive rate (FPR) and AUC. The Logistic Regression model in sample and out of sample performance metrics are shown in Table 4. The AUC score for in sample and out of sample performance was 0.78 and the ROC curves shown in Figure 1 appear to be almost identical. The classes for the test data set were assigned based on the cutoff threshold from the training data set.

Table 4 - Logistic Regression model with backward selection performance metrics.

	In Sample	Out of Sample
AUC	0.778	0.778
Cut off	0.201	-
TPR	0.654	0.645
FPR	0.230	0.233
Accuracy	0.744	0.742

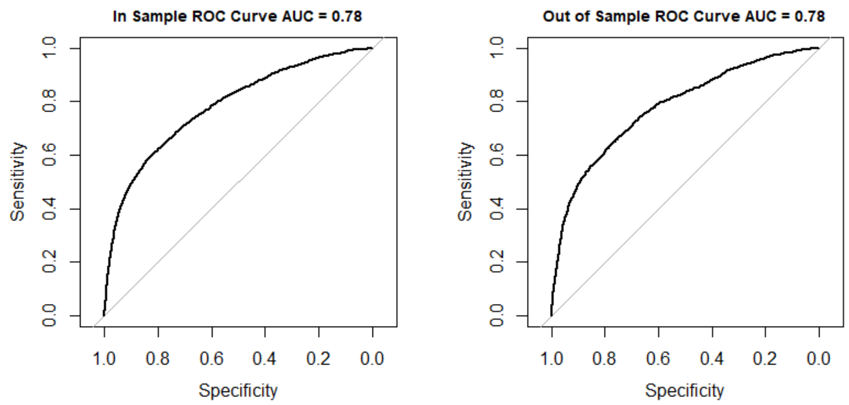

Figure 1 - ROC curve for Logistic Regression model with backward selection training and test sets.

The Kolmogorov-Smirnov (KS) statistic can be used as a relative sense of how well the model is performing. It calculates the maximum difference between two cumulative distributions. In this case, the two cumulative distributions of interest are the goods (Y=1, default payment) and the bads (Y=0, non-default payment). Table 5 and Table 6 show the lift charts in table format for the training and test data sets.

Table 5 - Lift chart for training data set.

Semi Decile	Obs	Target (Y=1)	NonTarget (Y=0)	Target Density	NonTarget Density	Target CDF	NonTarget CDF	KS Stat
1	753	566	187	16.5%	1.6%	16.5%	1.6%	14.9%
2	760	494	266	14.4%	2.3%	31.0%	3.9%	27.1%
3	764	391	373	11.4%	3.2%	42.4%	7.0%	35.4%
4	696	256	440	7.5%	3.7%	49.9%	10.8%	39.1%
5	804	239	565	7.0%	4.8%	56.9%	15.6%	41.3%
6	775	189	586	5.5%	5.0%	62.4%	20.6%	**41.8%**
7	741	166	575	4.8%	4.9%	67.2%	25.4%	**41.8%**
8	748	138	610	4.0%	5.2%	71.3%	30.6%	40.6%
9	737	117	620	3.4%	5.3%	74.7%	35.9%	38.8%
10	784	145	639	4.2%	5.4%	78.9%	41.3%	37.6%
11	732	117	615	3.4%	5.2%	82.3%	46.6%	35.7%
12	695	92	603	2.7%	5.1%	85.0%	51.7%	33.3%
13	530	74	456	2.2%	3.9%	87.2%	55.6%	31.6%
14	1,023	111	912	3.2%	7.8%	90.4%	63.3%	27.1%
15	528	45	483	1.3%	4.1%	91.7%	67.4%	24.3%
16	939	76	863	2.2%	7.3%	94.0%	74.8%	19.2%
17	874	72	802	2.1%	6.8%	96.1%	81.6%	14.4%
18	775	52	723	1.5%	6.1%	97.6%	87.8%	9.8%
19	700	36	664	1.1%	5.6%	98.6%	93.4%	5.2%
20	822	47	775	1.4%	6.6%	100.0%	100.0%	0.0%
Totals	**15,180**	**3423**	**11,757**	**100.0%**	**100.0%**			

Table 6 - Lift chart for test data set.

Semi Decile	Obs	Target (Y=1)	NonTarget (Y=0)	Target Density	NonTarget Density	Target CDF	NonTarget CDF	KS Stat
1	365	261	104	16.8%	1.8%	16.8%	1.8%	15.0%
2	368	233	135	15.0%	2.3%	31.7%	4.1%	27.6%
3	343	165	178	10.6%	3.1%	42.3%	7.2%	35.1%
4	346	129	217	8.3%	3.8%	50.6%	11.0%	39.6%
5	405	104	301	6.7%	5.2%	57.3%	16.2%	41.1%
6	368	87	281	5.6%	4.9%	62.9%	21.1%	**41.8%**
7	363	77	286	4.9%	5.0%	67.8%	26.0%	**41.8%**
8	352	63	289	4.0%	5.0%	71.9%	31.1%	40.8%
9	371	65	306	4.2%	5.3%	76.0%	36.4%	39.7%
10	373	58	315	3.7%	5.5%	79.8%	41.8%	37.9%
11	338	45	293	2.9%	5.1%	82.7%	46.9%	35.7%
12	360	40	320	2.6%	5.5%	85.2%	52.5%	32.8%
13	254	23	231	1.5%	4.0%	86.7%	56.5%	30.2%
14	485	49	436	3.1%	7.6%	89.9%	64.0%	25.8%
15	243	25	218	1.6%	3.8%	91.5%	67.8%	23.6%
16	520	53	467	3.4%	8.1%	94.9%	75.9%	19.0%
17	310	17	293	1.1%	5.1%	96.0%	81.0%	15.0%
18	419	30	389	1.9%	6.7%	97.9%	87.7%	10.1%
19	346	12	334	0.8%	5.8%	98.7%	93.5%	5.1%
20	394	21	373	1.3%	6.5%	100.0%	100.0%	0.0%
Totals	**7,323**	**1557**	**5,766**	**100.0%**	**100.0%**			

The KS statistic for both the training and test data set was 41.8. Thomas (2009) stated that "it is difficult to give limits on what are acceptable levels of discrimination because the discrimination depends on the population of borrowers, the lending product, and the definition of bad. However, a useful rule of thumb is that KS statistics of 0.4 suggest good discrimination" (p.112).

3.0 Performance Monitoring Plan

Credit card default prediction models cannot be left in production unmonitored. This section outlines how the model will be tracked using the KS statistic. The tolerances which define acceptable model performance have been defined using a RAG (Red-Amber-Green) operational status approach. For each status an action has been specified as shown in Table 7. The KS statistic from the training data set of 41.8 represents a baseline value. The Red status represents a degradation of the model performance greater than 20%.

Table 7 - Performance monitoring plan model status and action table.

Model Status	KS Statistic	Action
Green	$KS > 37.5$	Model is performing as expected. Model will be re-validated at the standard interval of six months.
Amber	$33.5 \leq KS \geq 37.5$	Model needs to be re-validated in three months.
Red	$KS < 33.5$	Model needs redevelopment.

4.0 Performance Monitoring Results

A KS statistic of 43.9 was calculated for the validation data set. A lift chart has been produced and is shown in Table 8. The KS score of 43.9 means the model status is Green and the model is performing as expected. The model will be re-validated at the standard interval of six months.

Table 8 - Lift chart for validation data set.

Semi Decile	Obs	Target (Y=1)	NonTarget (Y=0)	Target Density	NonTarget Density	Target CDF	NonTarget CDF	KS Stat
1	372	265	107	16.0%	1.8%	16.0%	1.8%	14.2%
2	378	262	116	15.8%	2.0%	31.8%	3.8%	28.0%
3	375	180	195	10.9%	3.3%	42.7%	7.2%	35.5%
4	374	144	230	8.7%	3.9%	51.4%	11.1%	40.3%
5	372	108	264	6.5%	4.5%	57.9%	15.6%	42.3%
6	378	104	274	6.3%	4.7%	64.2%	20.3%	**43.9%**
7	361	61	300	3.7%	5.1%	67.9%	25.4%	42.4%
8	389	70	319	4.2%	5.5%	72.1%	30.9%	41.2%
9	330	61	269	3.7%	4.6%	75.8%	35.5%	40.3%
10	402	57	345	3.4%	5.9%	79.2%	41.4%	37.8%
11	392	46	346	2.8%	5.9%	82.0%	47.3%	34.7%
12	301	44	257	2.7%	4.4%	84.7%	51.7%	32.9%
13	446	50	396	3.0%	6.8%	87.7%	58.5%	29.2%
14	378	41	337	2.5%	5.8%	90.2%	64.3%	25.9%
15	203	18	185	1.1%	3.2%	91.2%	67.5%	23.8%
16	464	40	424	2.4%	7.3%	93.7%	74.7%	18.9%
17	431	35	396	2.1%	6.8%	95.8%	81.5%	14.3%
18	355	27	328	1.6%	5.6%	97.4%	87.1%	10.3%
19	392	24	368	1.4%	6.3%	98.9%	93.4%	5.4%
20	404	19	385	1.1%	6.6%	100.0%	100.0%	0.0%
Totals	**7,497**	**1656**	**5,841**	**100.0%**	**100.0%**			

References

Thomas, L. C. (2009). Consumer Credit Models Pricing, Profit, and Portfolios. Oxford: Oxford University Press.

UCI. (2016). Default of Credit Card Clients Data Set. Retrieved April 10, 2019, from https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

Appendix 1

Table 9 - Data dictionary for the credit card default data set.

Variable	Description
ID	Unique ID number for each row of data.
LIMIT_BAL	Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family supplementary credit.
SEX	Gender: Male = 1, Female = 2
EDUCATION	Education: Graduate School = 1, University = 2, High school = 3, Others = 4
MARRIAGE	Marital Status: Married = 1, Single = 2, Others = 3
AGE	Age in years.
PAY_0	Repayment status in September 2005
PAY_2	Repayment status in August 2005
PAY_3	Repayment status in July 2005
PAY_4	Repayment status in June 2005
PAY_5	Repayment status in May 2005
PAY_6	Repayment status in April 2005
BILL_AMT1	Amount of bill statement (NT dollar) in September 2005.
BILL_AMT2	Amount of bill statement (NT dollar) in August 2005.
BILL_AMT3	Amount of bill statement (NT dollar) in July 2005.
BILL_AMT4	Amount of bill statement (NT dollar) in June 2005.
BILL_AMT5	Amount of bill statement (NT dollar) in May 2005.
BILL_AMT6	Amount of bill statement (NT dollar) in April 2005.
PAY_AMT1	Amount of previous payment (NT dollar) in September 2005
PAY_AMT2	Amount of previous payment (NT dollar) in August 2005
PAY_AMT3	Amount of previous payment (NT dollar) in July 2005
PAY_AMT4	Amount of previous payment (NT dollar) in June 2005
PAY_AMT5	Amount of previous payment (NT dollar) in May 2005
PAY_AMT6	Amount of previous payment (NT dollar) in April 2005
DEFAULT	Response Variable. Default Payment: Yes = 1, No = 0
u	A random number generated for splitting into training/test/validate set.
train	A value of 1 indicates the observation is used in the training data set.
test	A value of 1 indicates the observation is used in the test data set
validate	A value of 1 indicates the observation is used in the validation data set.
data.group	Constructed to partition the data set in a single dimension. data.group <- 1*train + 2*test + 3*validate

Table 10 - Measurement scale for PAY_0 to PAY6 variables.

-1 = Pay duly	5 = payment delay for five months
1 = payment delay for one month	6 = payment delay for six months
2 = payment delay for two months	7 = payment delay for seven months
3 = payment delay for three months	8 = payment delay for eight months
4 = payment delay for four months	9 = payment delay for nine months and above

Based on the information from the data dictionary the PAY_0 variable was renamed to PAY_1 so that it aligned with the BILL_AMT1 and PAY_AMT1 September variables. The values of -2 and -1 were recoded as 0, which indicates it was paid on time. The recoded variables are PAY_R1, PAY_R2, PAY_R3, PAY_R4, PAY_R5 and PAY_R6.

Table 11 - Engineered features for billing and payment history variables.

Feature	Description
AVG_BILL_AMT	The average monthly bill amount over the six months.
AVG_PAY_AMT	The average monthly payment amount over the six months.
PMT_RATIO_APR PMT_RATIO_MAY PMT_RATIO_JUN PMT_RATIO_JUL PMT_RATIO_AUG	This feature looks at the amount the customer repays compared to the bill amount. Note there is a time delay as PAY_AMT is previous month payment. Payment Ratio April = Previous Month Payment May / Bill Amount April PMT_RATIO_APR = PMT_AMT5 / BILL_AMT6 The PMT_RATIO value is then scaled between [0,100].
AVG_PMT_RATIO	The average payment ratio over the five months it can be calculated for.
UTIL_APR UTIL_MAY UTIL_JUN UTIL_JUL UTIL_AUG UTIL_SEP	Determines the amount of the credit line the customer is using each month. Utilization = Current Balance / Credit Limit UTIL_APR = BILL_AMT6 / LIMIT_BAL The utilization value is then scaled between [0,100].
AVG_UTIL	The average utilization over the six months.
BAL_GROWTH_6MO	The balance growth over the six months. BAL_GROWTH_6MO = BILL_AMT1 - BILL_AMT6
UTIL_GROWTH_6MO	The utilization growth over the six months. UTIL_GROWTH_60 = UTIL_SEP – UTIL_APR
MAX_DLQ	Maximum delinquency calculated by taking the maximum of the PAY_R# variables.
MAX_BILL_AMT	Maximum billed amount over the six months.
MAX_PAY_AMT	Maximum payment amount over the six months.