Logistic Regression Graded assignment

Credit card Dataset

The objective of this case study is to build a credit scoring model with the aim of capturing all important factors that can be used to predict default.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.7499	0.1574	123.6050	<.0001
Ruti_1	1	-2.8326	0.1008	789.7325	<.0001
Ruti_2	1	-2.1957	0.1113	389.2983	<.0001
Ruti_3	1	-1.7055	0.1098	241.2897	<.0001
Ruti_4	1	-1.2235	0.1068	131.3297	<.0001
Ruti_5	1	-0.8328	0.0935	79.3297	<.0001
age	1	-0.0119	0.00187	40.5767	<.0001
Gender_F	1	0.4664	0.0600	60.3783	<.0001
Region_C	1	-1.5685	0.1593	96.9990	<.0001
Region_N	1	1.0416	0.1013	105.8166	<.0001
Region_S	1	0.4449	0.1214	13.4409	0.0002
Region_W	1	2.0918	0.1045	400.5069	<.0001
Income_new	1	-0.00004	6.178E-6	39.5211	<.0001
House_own	1	-0.4183	0.0565	54.8835	<.0001
Occ2	1	0.5358	0.0700	58.6416	<.0001
Occ4	1	0.8574	0.0993	74.6125	<.0001
Edu2	1	1.4897	0.0884	284.1116	<.0001
Edu3	1	1.2388	0.1112	124.1894	<.0001
Edu4	1	0.5150	0.0871	34.9270	<.0001
DuenW3059	1	0.3164	0.0243	169.8846	<.0001
Dayslate_90	1	0.2606	0.0301	74.8412	<.0001
Realestate_loan	1	0.1311	0.0179	53.8224	<.0001
DuenW6089	1	-0.5448	0.0366	221.1007	<.0001
Dependents_new	1	0.0836	0.0195	18.2836	<.0001

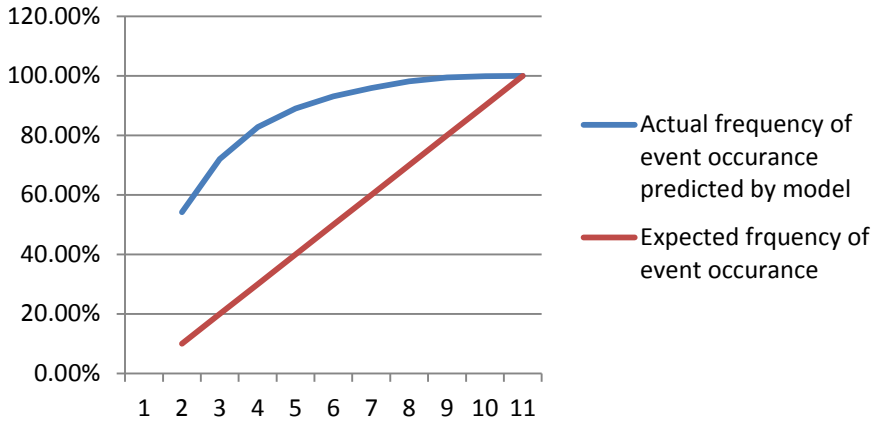
As per the credit scoring model, the parameters stated in the table have been determined to be the most important factors to predict default of customers.

The null hypothesis probability (Pr>Chisq) has been considered at a significance level of 5%

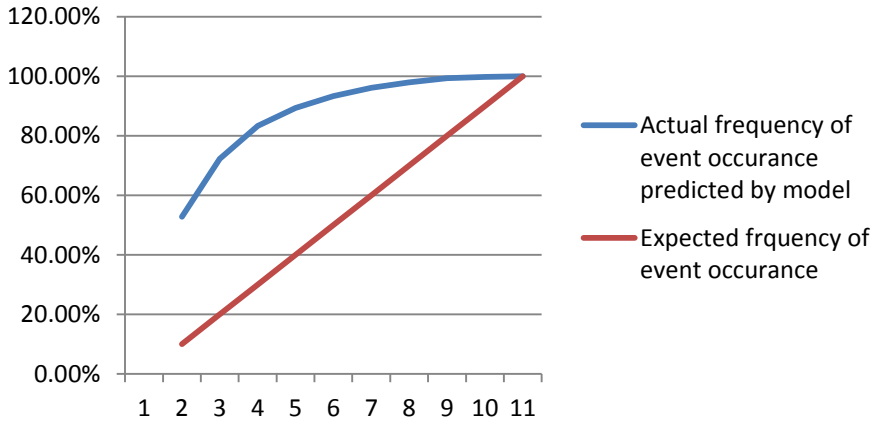
Logistic Regression Graded assignment

Lift curves

Training dataset



Validation dataset



Logistic Regression Graded assignment

SAS code

- proc logistic data=Graded.train descending outmodel = Graded.dmm;
  model NPA_Status=Ruti_1 Ruti_2 Ruti_3 Ruti_4 Ruti_5
  age Gender_F Region_C Region_N Region_S Region_W
  Income_new House_own Occ2 Occ4 Edu2 Edu3 Edu4 DuenW3059 Dayslate_90
  Realestate_loan DuenW6089 Dependents_new / ctable lackfit;
  score out=Graded.dmp;
  run;
- Proc rank data=Graded.dmp out=Graded.Decile_new groups=10 ties=mean;
  var p_1;
  ranks decile;
  run;

Model Usefulness

In the validation dataset, if the probability of default predicted by the model is considered to be 0.5, the following results are observed.

Model prediction	Actual Result	
	0	1
0	33465	233
1	2035	513

The total number of observations in the validation dataset are 36246.
As per the above results, the model accuracy is 93.74%

Logistic Regression Graded assignment

Model output

The LOGISTIC Procedure

Model Information	
Data Set	GRADED.VALIDATE
Response Variable	NPA_Status
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	36246
Number of Observations Used	36246

Response Profile		
Ordered Value	NPA_Status	Total Frequency
1	1	2548
2	0	33698

Probability modeled is NPA_Status='1'.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	18444.514	13138.822
SC	18453.012	13342.776
-2 Log L	18442.514	13090.822

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	5351.6926	23	<.0001
Score	6129.3860	23	<.0001
Wald	3397.0991	23	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Ruti_1	0.059	0.048	0.072
Ruti_2	0.111	0.089	0.138
Ruti_3	0.182	0.147	0.225
Ruti_4	0.294	0.239	0.363
Ruti_5	0.435	0.362	0.522
age	0.988	0.985	0.992
Gender_F	1.594	1.417	1.793
Region_C	0.208	0.152	0.285
Region_N	2.834	2.324	3.456
Region_S	1.560	1.230	1.979
Region_W	8.100	6.599	9.941
Income_new	1.000	1.000	1.000
House_own	0.658	0.589	0.735
Occ2	1.709	1.490	1.960
Occ4	2.357	1.940	2.863
Edu2	4.436	3.730	5.275
Edu3	3.452	2.776	4.292
Edu4	1.674	1.411	1.985
DuenW3059	1.372	1.308	1.439
Dayslate_90	1.298	1.223	1.377
Realestate_loan	1.140	1.101	1.181
DuenW6089	0.580	0.540	0.623
Dependents_new	1.087	1.046	1.130

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	87.0	Somers' D	0.739
Percent Discordant	13.0	Gamma	0.739
Percent Tied	0.0	Tau-a	0.097
Pairs	85862504	c	0.870

Partition for the Hosmer and Lemeshow Test					
		NPA_Status = 1		NPA_Status = 0	
Group	Total	Observed	Expected	Observed	Expected
1	3625	8	4.11	3617	3620.89
2	3626	10	9.65	3616	3616.35
3	3625	35	24.42	3590	3600.58
4	3624	48	47.83	3576	3576.17
5	3625	69	72.53	3556	3552.47
6	3625	103	106.94	3522	3518.06
7	3625	154	163.19	3471	3461.81
8	3625	283	267.94	3342	3357.06
9	3625	494	492.44	3131	3132.56
10	3621	1344	1358.99	2277	2262.01

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
10.3655	8	0.2403