Logistic Regression – Credit Score Model

Case Summary

The company has data on customers involving the details about the demographic and credit bureau variables. We need to understand the data and a model to help the company identify potential good customers.

Note – The SAS code and writeup about project are given in other files in project. This presentation explains the observation, results and conclusion of the project

https://github.com/Vicky-Crasto/Credit-Scoring-Model-SAS-

Data Exploration and Initial Finding from the data

In the given data set out of 150,000 observations, 6.68% observations related to customers who have defaulted.

Age

About 75% of the defaulting customers are in the age group of 30 to 60

Credit lines

About 40% of the defaulted customers have at least 1 credit lines

Credit Utilization

Credit utilization is high in defaulted customers as compared to non defaulted. About 44% of defaulted customers have used their entire credit limit.

Debt to income ratio

Debt to income ratio is relatively higher for defaulted customers.

Region

Defaulted customers seem to be concentrated in West(48%) and North (23%)

60 -90 days past due

Out of the defaulted customers, in last two years 28% have not repaid the loan at least once for a period of 60-90 days. 18% have not repaid only once.

90 + days past due

Out of the defaulted customers, 35% have not paid their loan at least once for more than 90 days in the last two years. 17.5% not paid only once.

Income

93% of the defaulters have a income less than 10000. Income level of 2500 to 5000 have about 34% defaulters.

https://github.com/Vicky-Crasto/Credit-Scoring-Model-SAS-

Understanding factors affecting default

Cross table – Income, region and NPA Status

NPA Status	dummy_region	incomegroup: Upto 2500 N	2500 to 5000 N	5000 to 7500 N	7500 to 10000 N	10000 to12500 N	12500 to 15000 N	15000 to 20000 N	20000 to 25000 N	25000 to high N
0	Centre	4654	11129	13710	8324	2993	1228	966	323	377
	East	2117	4905	5955	4199	1302	559	480	142	188
	North	3432	7993	9747	6043	2130	916	718	196	286
	South	2404	5328	6471	4669	1583	627	509	152	180
	West	2504	5796	7262	4375	1564	611	563	153	211
1	Centre	40	77	83	34	6	9	3	.	.
	East	91	227	209	115	34	8	11	2	7
	North	374	886	774	354	120	50	43	16	21
	South	224	530	442	250	62	31	19	3	11
	West	720	1699	1336	705	183	88	68	32	29

- Default rate is high across income group in the West Region, the company must investigate if loans approval procedures are correctly followed.
- On the other loan default rate are relatively lower in the central region, company must learn how they are able to maintain this.
- The northern region also seem to have high loan defaults in the income groups of 2500 to 7500.

https://github.com/Vicky-Crasto/Credit-Scoring-Model-SAS-

Understanding factors affecting default

Cross table – Income, age group and NPA Status

		incomegroup								
		Upto 2500	2500 to 5000	5000 to 7500	7500 to 10000	10000 to12500	12500 to 15000	15000 to 20000	20000 to 25000	25000 to high
		N	N	N	N	N	N	N	N	N
NPA Status	agegroup									
0	Upto 30	3051	3731	783	136	41	17	14	5	8
	30 to 40	2698	6688	7716	2163	875	276	225	87	120
	40 to 50	2427	7083	11779	5257	2619	967	755	248	364
	50 to 60	2319	6585	7102	10774	2964	1477	1147	302	353
	60 to 70	2278	5723	7410	7916	2098	980	881	262	307
	70 to 80	1613	3346	6061	1036	724	198	177	45	70
	80 to 90	661	1750	2025	295	197	25	35	15	17
	Above 90	64	245	269	33	54	1	2	2	3
1	Upto 30	404	545	62	15	2	3	1	2	1
	30 to 40	347	915	793	166	49	19	23	5	18
	40 to 50	287	849	1036	405	151	61	48	22	19
	50 to 60	216	656	515	605	131	65	48	17	25
	60 to 70	130	310	270	233	51	27	19	5	5
	70 to 80	49	101	122	27	19	9	3	1	.
	80 to 90	14	41	39	7	2	1	1	.	.
	Above 90	2	2	7	.	.	1	1	1	.

Default in the age group of 30 to 50 with income 2500 to 7500 is on the higher side

Default in the age group of 60 and above must be checked and prevented, as repayment ability decrease post retirement

https://github.com/Vicky-Crasto/Credit-Scoring-Model-SAS-

Credit Scoring Model

The model has been constructed based on the following parameters

- Region
- Rented/ owned house
- Income group
- Age group
- Gender
- Education
- Occupation
- Credit utilization
- Default behavior in 30-60 days,60-90 days and more than 90 days.

Some interesting observations from the model

- One unit increase in the credit utilization, increases the odds of default by about 424%.
- If a person defaults on a loan repayment for more than 90 days, then the odds of default increases by 90%.
- If a debt to income ratio increase by one unit, then the odds of default increases by 70%
- If a person move up to the next income group than the odds of default decrease by 6%.
- If a person education is just matric than the odds of default increases by 307%

https://github.com/Vicky-Crasto/Credit-Scoring-Model-SAS-

SAS code Used for Modelling

```
proc logistic data = gd3.train_data descending outmodel = gd3.train_out;
model NPA_Status =
credit_lines
incomegroup
credit_utiliznew
age_new
n30_59pastdue_new
N90pastdue_new
N60_90pastdue_new
debtratio_new
gender_dummy
dummy_house
edu_matric
edu_phd
edu_postgrad
dummy_occup1
dummy_occup3
dummy_centre
dummy_east
dummy_north
dummy_south / ctable lackfit outroc = gd3.train_roc ;
output out = gd3.train_predicted  p = pred;
score out = gd3.train_score;
run;
```

```
*** ploting ROC curve***;

symbol2 i=join v=none c=blue;
proc gplot data = gd3.train_roc;
title "ROC plot";
plot  _SENSIT_ *_1MSPEC_=1/cframe = ligr;
run;

*** creating lift chart ***;

proc sort data = gd3.train_score;
by P_0;
run;
```

```
proc rank data = gd3.train_score
out =gd3.train_gain
groups = 10
ties = mean;
var P_1 ;
ranks decile;
run;
```

```
proc rank data = gd3.train_score
out =gd3.train_gain
groups = 10
ties = mean;
var P_1 ;
ranks decile;
run;

*** export train_gain to csv**;

proc export data = gd3.train_gain
outfile = "Y:\Vickydec2016\2.1 Graded Assignment -Regression (reat
dbms = csv replace;
run;


*****running the model on the validation dataset****;

proc logistic inmodel = gd3.train_out;
score data = gd3.valid_data out =gd3.valid_score fitstat;
run;
```

```
***checking accuracy**;

data gd3.valid_testaccu;
set gd3.valid_score;
if F_NPA_Status = 1 and I_NPA_Status = 1 then result = "True Positive";
if F_NPA_Status = 0 and I_NPA_Status = 0 then result = "True Negative";
if F_NPA_Status = 1 and I_NPA_Status = 0 then result = "False Negative";
if F_NPA_Status = 0 and I_NPA_Status = 1 then result = "False Positive";
run;

proc freq data = gd3.valid_testaccu;
tables result;
run;
```

https://github.com/Vicky-Crasto/Credit-Scoring-Model-SAS-

Logistic Regression output

Response Profile		
Ordered Value	**NPA_Status**	**Total Frequency**
1	1	7120
2	0	97804

Probability modeled is NPA_Status='1'.

The training data set consists of 104924 observations out of which 2159 are observations of churned members

Model Convergence Status
Convergence criterion (GCONV=1e-008) satisfied.

Convergence criteria is satisfied

Model Fit Statistics		
Criterion	**Intercept only**	**Intercept and Covariates**
AIC	52057.810	34829.020
SC	52067.371	35020.240
-2 Log L	52055.810	34789.020

The model fit statistics are the lowest for this iteration compared to the other iterations. In simple words the data lost in building the model is the lowest in this model iterations.

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > Chi-Square**
Likelihood Ratio	17266.7892	19	<.0001
Score	23649.1188	19	<.0001
Wald	9691.1539	19	<.0001

The Global Null Hypothesis test for the condition that the independent variables have not effect on the dependent variable by showing that one of the Beta variable is zero. But in this case p value are less than 0.05 and hence we reject the hypothesis.

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.2748	0.0774	863.0601	<.0001
credit_lines	1	0.0233	0.00320	52.8454	<.0001
incomegroup	1	-0.0605	0.0111	29.4248	<.0001
credit_utiliznew	1	1.6568	0.0424	1525.0313	<.0001
age_new	1	-0.0197	0.00115	294.7980	<.0001
n30_59pastdue_new	1	0.4730	0.0147	1042.1033	<.0001
N90pastdue_new	1	0.6409	0.0206	972.0299	<.0001
N60_90pastdue_new	1	0.5641	0.0292	374.2654	<.0001
debtratio_new	1	0.5314	0.0525	102.5381	<.0001
gender_dummy	1	0.4290	0.0373	132.3116	<.0001
dummy_house	1	-0.4778	0.0350	186.2048	<.0001
edu_matric	1	1.4041	0.0538	680.2726	<.0001
edu_phd	1	1.1702	0.0688	289.4369	<.0001
edu_postgrad	1	0.4353	0.0537	65.8384	<.0001
dummy_occup1	1	0.4615	0.0433	113.7110	<.0001
dummy_occup3	1	0.7211	0.0612	138.6229	<.0001
dummy_centre	1	-3.8643	0.0893	1871.8454	<.0001
dummy_east	1	-2.0401	0.0644	1003.7060	<.0001
dummy_north	1	-0.9897	0.0399	613.8028	<.0001
dummy_south	1	-1.5386	0.0495	968.0418	<.0001

The p values for the variables is less than 0.05 making them significant.

https://github.com/Vicky-Crasto/Credit-Scoring-Model-SAS-

Logistic Regression output

Effect	Point Estimate	Lower 95% Wald Confidence Limit	Upper 95% Wald Confidence Limit
credit_lines	1.024	1.017	1.030
incomegroup	0.941	0.921	0.962
credit_utiliznew	5.242	4.824	5.697
age_new	0.981	0.978	0.983
n30_59pastdue_new	1.605	1.559	1.652
N90pastdue_new	1.898	1.823	1.976
N60_90pastdue_new	1.758	1.660	1.861
debtratio_new	1.701	1.535	1.886
gender_dummy	1.536	1.427	1.652
dummy_house	0.620	0.579	0.664
edu_matric	4.072	3.664	4.525
edu_phd	3.223	2.816	3.688
edu_postgrad	1.545	1.391	1.717
dummy_occup1	1.586	1.457	1.727
dummy_occup3	2.057	1.824	2.319
dummy_centre	0.021	0.018	0.025
dummy_east	0.130	0.115	0.148
dummy_north	0.372	0.344	0.402
dummy_south	0.215	0.195	0.237

The odds ratio also give us an indication if the coefficient are significant. In this case, they are significant as their range is less than or greater than 1. Also they help in calculating the percentage change in the odds with one unit increase in the variable, with all others variable remaining constant

Concordant % of 89.2% indicated that a high % of predicated values of 1 and 0 are in sink with actual responses.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	89.2	Somer's D	0.788
Percent Discordant	10.4	Gamma	0.791
Percent Tied	0.4	Tau-a	0.1
Pairs	696364480	c	0.894

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > Chi-Square
74.6159	8	<.0001

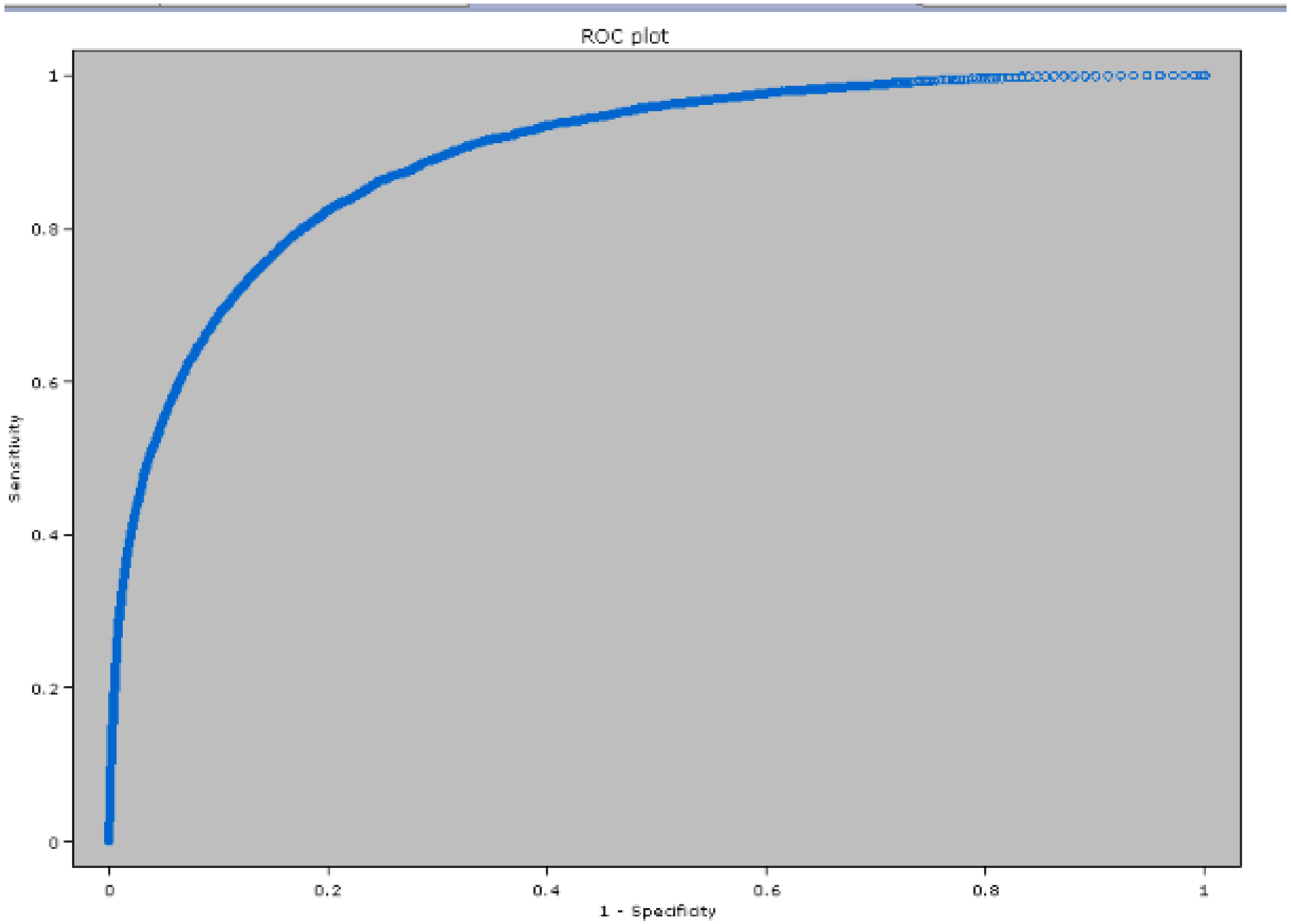
The Hosmer and Lemeshow test tell you how well your data fits in them model. In this case the low p-value indicates it a not a very good fit.

https://github.com/Vicky-Crasto/Credit-Scoring-Model-SAS-

Logistic Regression output – ROC curve

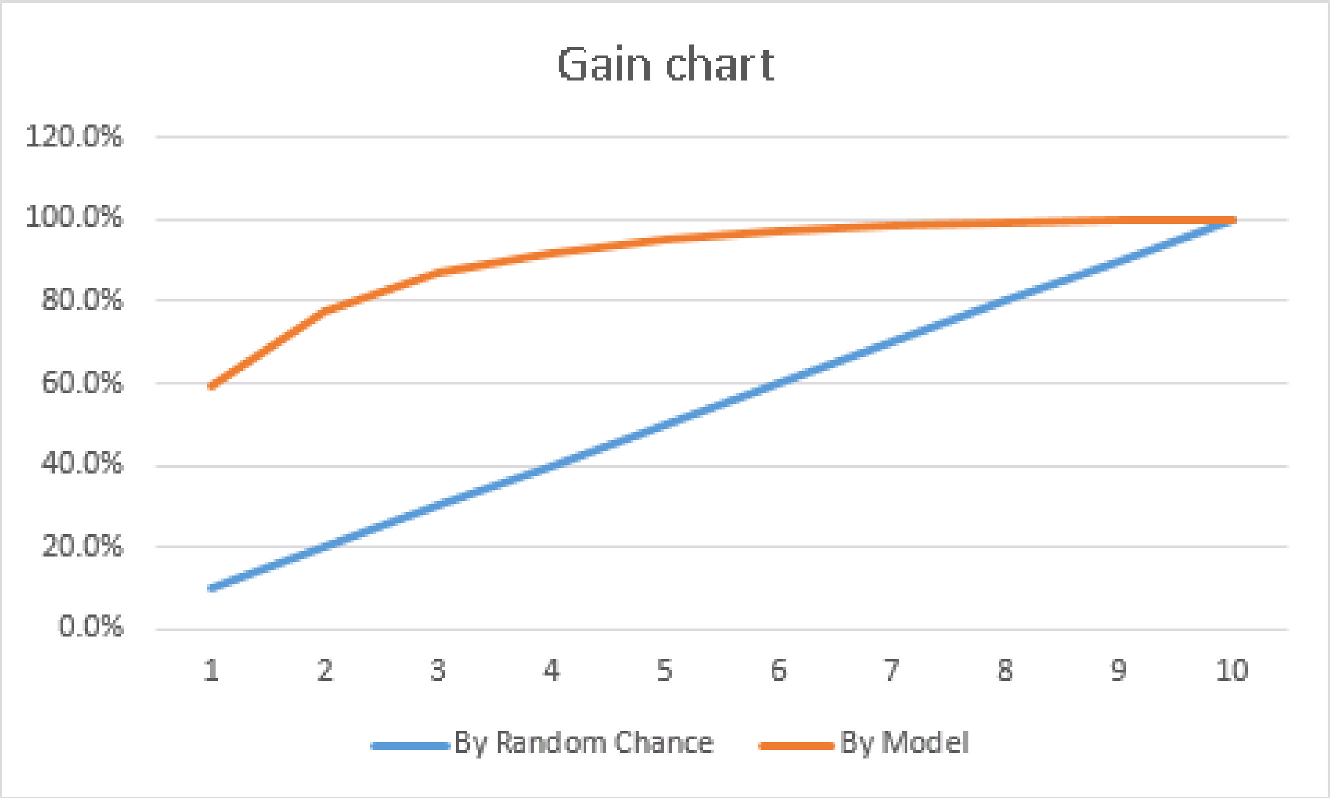
ROC curve is constructed using sensitivity vs (1- Specificity). Sensitivity is the proportion of observation correctly predicted by the model as true positive when they are actually so in the data. Similarly, Specificity is the proportion of true negative correctly predicted by the model when they are actually so in the data.

A good model must have high sensitivity and specificity, which is indicated by a curve moving closer to the top left corner , as shown above.. Hence this model is good model based on the ROC output.



https://github.com/Vicky-Crasto/Credit-Scoring-Model-SAS-

Logistic Regression output – Gain Chart



Gain chart helps to evaluated the performance of the model. It helps us to understand how much better the model is able to predict the values as against if there was no model but only a random chance probability.
The orange straight line indicate the event (churned) predicted using random chance probability where as the blue curve line indicate the event predicted by the model. Hence the model performs better than if there was no model at all.

https://github.com/Vicky-Crasto/Credit-Scoring-Model-SAS-

Logistic Regression output – Predicted Value

The model was applied to the validation data to check the accuracy, which is given below.

Fit Statistics for SCORE Data			
Data Set	Total Frequency	Log Likelihood	Misclassification Rate
GD3.valid_data	45076	-7069.3	0.0540

We see that about 94% of the event are correctly predicted by the model.

result	Frequency	Percent	Cumulative Frequency	Cumulative Percent
False Negativ	2113	4.69	2113	4.69
False Positiv	320	0.71	2433	5.40
True Negative	41850	92.84	44283	98.24
True Positive	793	1.76	45076	100.00

Here we see that the model has classified about 0.7% events wrongly that is Non churned are indicated as churned (False Positives) .
And in 4.69% events it has wrongly classified non churned members as churned (False negative).