Logistic Regression Model

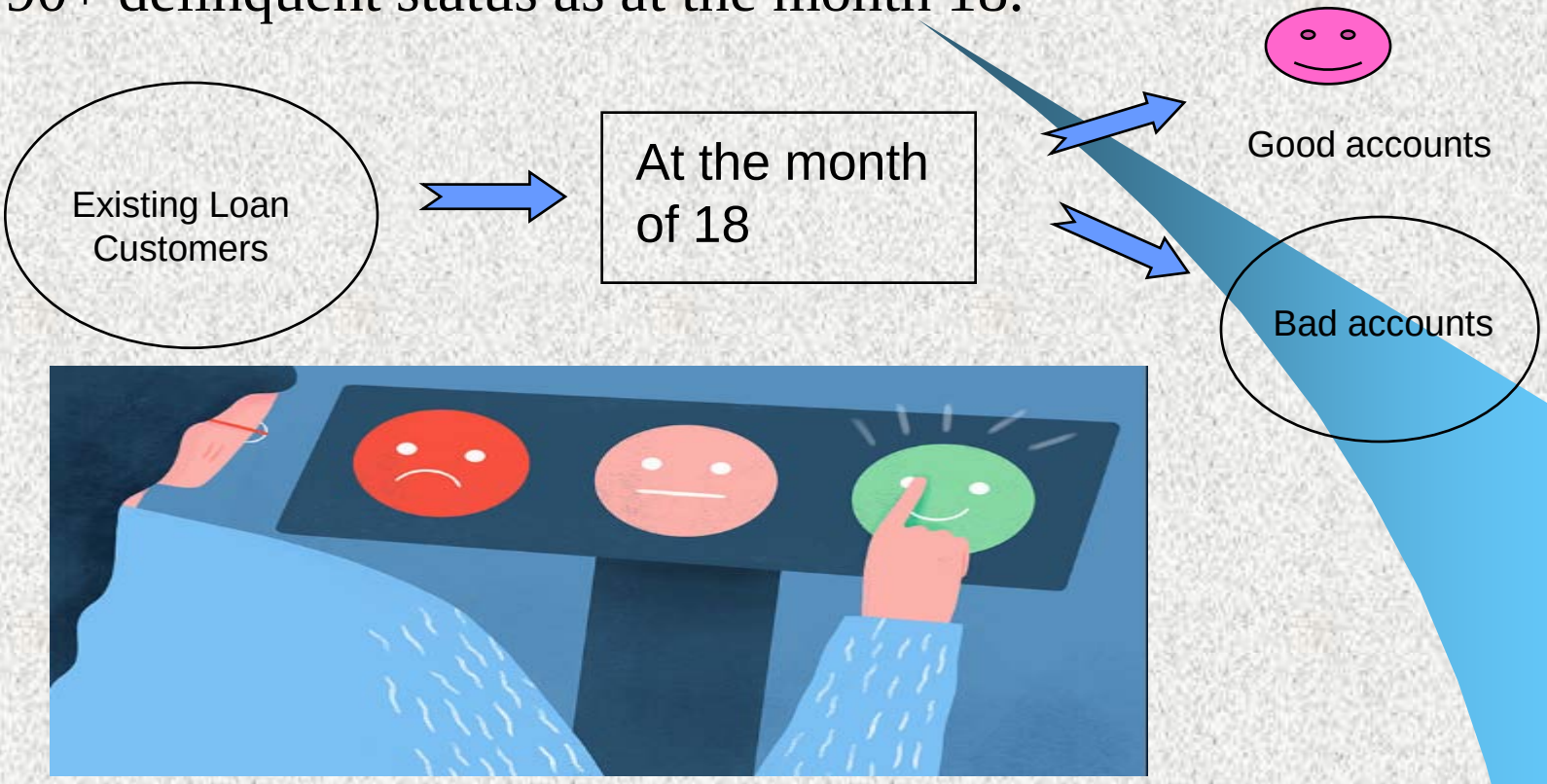
Who, amongst existing loan customers, are more likely to be default (bad)?

By: Ramzi Abdel.

SAS Project

Definition of Default (bad):

An account that was in NPNA or DWO or 90+ delinquent status as at the month 18.



- Who, amongst existing loan customers, are more likely to be default (bad)?

Modeling Process

- **Modeling data set has**
  - **more than 50 independent variables, and**
  - **one dependent binary variable: bad=1, non-bad=0.**
- **Preliminary variable transformations through Rank & Plot.**
- **Logistic Regression**

  logit(P / (1-P))= $\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_N x_N$, using stepwise selection technique
- **Correlation between variables, model concordance and goodness-of-fit are used to select the best-fit model**

Modeling Procedure

- **Importing data HMEQ into SAS Environment**
- **Feature Engineering: creating new variables with new labels from old variables**
- **Grouping (10 groups)**
- **Data Cleaning: running macros to find, replace and clean missing values**
- **Models improving: running several models with different Independent variables to compare and select the best performance.**

Modeling Procedure

- **Importing data HMEQ into SAS Environment**
- **Feature Engineering: creating new variables with new labels from old variables**
- **Grouping (10 groups)**
- **Data Cleaning: running macros to find, replace and clean missing values**
- **Models improving: running several models with different Independent variables to compare and select the best performance.**

Summary of Model 1

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-0.0937	0.1303	0.5167	0.4722	
JOB_Office	1	-0.6410	0.1222	27.5044	<.0001	-0.1293
JOB_Sales	1	0.9519	0.2407	15.6379	<.0001	0.0703
JOB_Self	1	0.6214	0.1952	10.1290	0.0015	0.0606
JOB_miss	1	-2.0284	0.3252	38.8975	<.0001	-0.2362
REASON_DebtCon	1	-0.1986	0.0855	5.3963	0.0202	-0.0519
VALUE_MISS	1	4.5119	0.4407	104.8030	<.0001	0.3378
CLAGE	1	-0.00768	0.000591	169.2115	<.0001	-0.3098
DELINQ	1	0.7087	0.0373	361.4254	<.0001	0.4217
DEROG	1	0.5604	0.0486	132.8586	<.0001	0.2467
LOAN	1	-0.00003	5.518E-6	38.8602	<.0001	-0.1458
NINQ	1	0.1822	0.0205	78.9922	<.0001	0.1693
YOJ	1	-0.0135	0.00506	7.0626	0.0079	-0.0633

Partition for the Hosmer and Lemeshow Test					
		BAD = 1		BAD = 0	
Group	Total	Observed	Expected	Observed	Expected
1	596	17	15.65	579	580.35
2	596	19	27.10	577	568.90
3	596	42	37.34	554	558.66
4	596	45	49.40	551	546.60
5	596	59	64.08	537	531.92
6	596	93	83.31	503	512.69
7	596	91	106.26	505	489.74
8	596	151	140.86	445	455.14
9	596	246	216.32	350	379.68
10	596	426	448.70	170	147.30

Summary of Model 2

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-0.8538	0.1214	49.4475	<.0001	
JOB_Office	1	-0.4877	0.1270	14.7522	0.0001	-0.0984
JOB_Other	1	0.1932	0.0862	5.0181	0.0251	0.0522
JOB_Sales	1	1.1671	0.2411	23.4294	<.0001	0.0862
JOB_Self	1	0.5474	0.1929	8.0550	0.0045	0.0534
JOB_miss	1	-0.9808	0.2522	15.1228	0.0001	-0.1142
REASON_HomeImp	1	0.3462	0.0809	18.3210	<.0001	0.0874
CLAGE	1	-0.00808	0.000575	197.3439	<.0001	-0.3257
DELINQ	1	0.7512	0.0368	416.1331	<.0001	0.4470
DEROG	1	0.5501	0.0474	134.5686	<.0001	0.2422
NINQ	1	0.1599	0.0200	64.1470	<.0001	0.1486
YOJ	1	-0.0151	0.00485	9.7444	0.0018	-0.0712

Partition for the Hosmer and Lemeshow Test

Group	Total	BAD = 1		BAD = 0	
		Observed	Expected	Observed	Expected
1	596	20	21.42	576	574.58
2	596	25	31.86	571	564.14
3	596	22	42.06	574	553.94
4	596	62	55.03	534	540.97
5	596	74	70.44	522	525.56
6	596	107	88.54	489	507.46
7	596	112	113.43	484	482.57
8	596	129	141.40	467	454.60
9	596	247	209.25	349	386.75
10	596	391	415.58	205	180.42

Summary of Model 3

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-0.7169	0.1082	43.9088	<.0001	
JOB_Office	1	-0.6134	0.1173	27.3331	<.0001	-0.1237
JOB_Sales	1	1.0444	0.2366	19.4906	<.0001	0.0772
JOB_miss	1	-1.1161	0.2462	20.5490	<.0001	-0.1300
REASON_HomeImp	1	0.3698	0.0802	21.2654	<.0001	0.0933
CLAGE	1	-0.00823	0.000571	207.4370	<.0001	-0.3318
DELINQ	1	0.7464	0.0367	414.4562	<.0001	0.4441
DEROG	1	0.5542	0.0474	136.4671	<.0001	0.2440
NINQ	1	0.1614	0.0199	66.0284	<.0001	0.1500
YOJ	1	-0.0144	0.00482	8.9575	0.0028	-0.0678

Partition for the Hosmer and Lemeshow Test					
		BAD = 1		BAD = 0	
Group	Total	Observed	Expected	Observed	Expected
1	596	19	21.34	577	574.66
2	596	25	32.02	571	563.98
3	596	30	42.29	566	553.71
4	596	53	55.35	543	540.65
5	596	82	71.43	514	524.57
6	596	104	89.91	492	506.09
7	596	117	112.19	479	483.81
8	596	119	141.69	477	454.31
9	596	249	208.27	347	387.73
10	596	391	414.50	205	181.50

Model Result: Logit Equation

Logit=

-1.5204			
+**2.6277**	*	DEBTINC_MISS	/*Debt to income ratio IS MISSING */
-0.6524	*	JOB_Office	
+**1.1876**	*	JOB_Sales	
-1.933	*	JOB_miss	
+**0.2527**	*	REASON_HomeImp	/*Home improvement */
+**4.5662**	*	VALUE_MISS	/*Value of current property IS MISSING */
-0.00616	*	clage	/*Age of oldest trade line in months*/
-0.0148	*	clno	/*Number of trade (credit) lines*/
+**0.6875**	*	delinq	/*Number of delinquent trade lines*/
+**0.5215**	*	derog	/*Number of major derogatory reports*/
+**1.32E-01**	*	ninq	/*Number of recent credit inquiries*/
-0.0168	*	yoj	/*Years on current job*/

```
prob=1/(1+exp(-logit));
```

Plot of Logit(Response) by debtinc
Plot of Response by debtinc
Table of Response by Grouped debtinc

prob=0.1076091367

Obs	BAD	LOAN	MORTDUE	VALUE	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC	Value_grp	Value_grp2
1	1	1100	25860	39025	10.5	0	0	94.3667	1	15	.	.	1

JOB_Mgr	JOB_Office	JOB_Other	JOB_ProfExe	JOB_Sales	JOB_Self	JOB_miss	REASON_DebtCon	REASON_HomeImp	REASON_Miss	DEBTINC_MISS	VALUE_MISS	Logit
0	0	1	0	0	0	0	0	1	0	0	0	-2.11540

Task1: Data Manipulation

```
libname TT "D:\SAS\SAS Capstone Project\data";
proc contents data=tt.HMEQ
;
run;
proc freq data=tt.hmeq;
   tables bad reason job;
run;

data Hmeq(drop=job reason);
  set tt.Hmeq;
  JOB_Mgr=(JOB='Mgr');
  JOB_Office=(JOB='Office');
  JOB_Other=(JOB='Other');
  JOB_ProfExe=(JOB='ProfExe');
  JOB_Sales=(JOB='Sales');
  JOB_Self=(JOB='Self');
  JOB_miss=(JOB=' ');
  REASON_DebtCon=(REASON='DebtCon');
  REASON_HomeImp=(REASON='HomeImp');
  REASON_Miss=(REASON=' ');
  /*
  if CLAGE=. then CLAGE=179.7662752;
  if CLNO=. then CLNO=21.2960962;
  if DEBTINC=. then DEBTINC=33.7799153;
  if DELINQ=. then DELINQ=0.4494424;
  if DEROG=. then DEROG=0.2545697;
  if LOAN=. then LOAN= 18607.97;
  if MORTDUE=. then MORTDUE=73760.82;
  if NINQ=. then NINQ= 1.1860550;
  if VALUE=. then VALUE=101776.05;
  if YOJ=. then YOJ= 8.9222681;
  */
run;

proc contents data=Hmeq;
run;
```

```
%let inter_var=

clage
clno
debtinc
delinq
derog
loan
mortdue
ninq
value
yoj
;

%LET DSN=Hmeq;
%LET RESP=BAD;
%LET GROUPS=10;

%MACRO LOGTCONT                    ;
   OPTIONS CENTER PAGENO=1 DATE;
                        data test;
                          set &DSN;
                        run;
                        %do i=1 %to 10;
                        %LET VBLE=%scan(&inter_var, &i);
   PROC RANK DATA =TEST (KEEP=&RESP &VBLE)
        GROUPS = &GROUPS
          OUT = JUNK1     ;
      RANKS NEWVBLE        ;
      VAR &VBLE            ;
   RUN                     ;

   PROC SUMMARY DATA = JUNK1 NWAY ;
      CLASS NEWVBLE           ;
      VAR &RESP &VBLE         ;
      OUTPUT OUT = JUNK2
          MEAN =
          MIN(&VBLE)=MIN
          MAX(&VBLE)=MAX
            N = NOBS        ;
   RUN                       ;
```

```
DATA JUNK2                  ;
      SET JUNK2             ;
      IF &RESP NE 0 THEN
        LOGIT = LOG ( &RESP / (1- &RESP) ) ;
      ELSE IF &RESP = 0 THEN LOGIT = .     ;
   RUN                      ;

   PROC SQL NOPRINT;
    CREATE TABLE JUNK3 AS
    SELECT 99 AS NEWVBLE, COUNT(*) AS NOBS,
MEAN(&RESP) AS &RESP
    FROM test
    WHERE &VBLE=.
    ;

   DATA JUNK3;
    SET JUNK3;
    LOGIT=LOG(&RESP/(1-&RESP));
    RUN;


   DATA JUNK4;
    SET JUNK2 JUNK3;
    RUN;


   PROC PLOT DATA = JUNK4       ;
      TITLE1 "Plot of Logit(Response) by &&VBLE" ;
      PLOT  LOGIT* &VBLE       ;
   RUN  ;


    proc plot data=junk4;
    plot &resp*&vble;
    plot _freq_*&vble;
    TITLE2 "Plot of Response by &&VBLE" ;
    run;

   PROC PRINT DATA = JUNK4 LABEL SPLIT = '*'
NOOBS ;
      TITLE3 "Table of Response by Grouped &&VBLE" ;
      VAR NEWVBLE NOBS &VBLE MIN MAX &RESP ;
      LABEL NEWVBLE = "&&VBLE Grouping"
            NOBS = '# of*Records'
            LOGIT = "Logit of Response"
            MIN   ='MIN'
            MAX   ='MAX'                ;
   RUN                                  ;

                        %end;


%MEND LOGTCONT     ;
%LOGTCONT     ;
```

Task1: Modeling Code

```
data Hmeq1;
 set Hmeq;
 /*
 JOB_Mgr=(JOB='Mgr');
 JOB_Office=(JOB='Office');
 JOB_Other=(JOB='Other');
 JOB_ProfExe=(JOB='ProfExe');
 JOB_Sales=(JOB='Sales');
 JOB_Self=(JOB='Self');
 JOB_miss=(JOB=' ');
 REASON_DebtCon=(REASON='DebtCon');
 REASON_HomeImp=(REASON='HomeImp');
 REASON_Miss=(REASON=' ');
 */
 if CLAGE=. then CLAGE=95.205;
 if CLAGE>295 then CLAGE=295;
 if CLNO<10 then CLNO=0;
 if CLNO=. then CLNO= 42.2;
 if CLNO<15 then CLNO= 15;
 DEBTINC_MISS=(DEBTINC=45.8);
 if DELINQ=. then DELINQ=0;
 if DEROG=. then DEROG=0;
 if LOAN>30500 then LOAN=30500;
 if MORTDUE=. then MORTDUE= 46141.88;
 if NINQ=. then NINQ=0;
  VALUE_MISS=(VALUE=37502.57);
 if YOJ=. then YOJ=25;
run;
```

```
%LET INPUT2=

JOB_Mgr
JOB_Office
JOB_Other
JOB_ProfExe
JOB_Sales
JOB_Self
JOB_miss
REASON_DebtCon
REASON_HomeImp
REASON_Miss
VALUE_MISS
clage
clno
delinq
derog
loan
mortdue
ninq
yoj
;

proc logistic data=Hmeq1 descending;
model bad=&input2
  /selection=stepwise fast lackfit rsquare corrb stb;
run;
;
```

```
%LET INPUT3=
DEBTINC_MISS
JOB_Office
JOB_Other
JOB_Sales
JOB_Self
JOB_miss
REASON_HomeImp
VALUE_MISS
clage
clno
delinq
derog
ninq
yoj
;

proc logistic data=Hmeq1 descending;
model bad=&input3
  /selection=stepwise fast lackfit rsquare corrb stb;
run;

%LET INPUT4=
DEBTINC_MISS
JOB_Office
JOB_Sales
JOB_miss
REASON_HomeImp
VALUE_MISS
clage
clno
delinq
derog
ninq
yoj
;

proc logistic data=Hmeq1 descending;
model bad=&input4
  /selection=stepwise fast lackfit rsquare corrb stb;
run;
```

Task1: Modeling Code

```
data val;
  set Hmeq1;
Logit=
-1.5204
+2.6277     *     DEBTINC_MISS     /*Debt to income ratio IS MISSING */
-0.6524     *     JOB_Office
+1.1876     *     JOB_Sales
-1.933      *     JOB_miss
+0.2527     *     REASON_HomeImp   /*Home improvement */
+4.5662     *     VALUE_MISS       /*Value of current property IS MISSING */
-0.00616    *     clage                  /*Age of oldest trade line in months*/
-0.0148     *     clno                   /*Number of trade (credit) lines*/
+0.6875     *     delinq                 /*Number of delinquent trade lines*/
+0.5215     *     derog                  /*Number of major derogatory reports*/
+1.32E-01   *     ninq                   /*Number of recent credit inquiries*/
-0.0168     *     yoj                    /*Years on current job*/
;

prob=1/(1+exp(-logit));
run;
```

Task 2: SQL Code

Supervisors of crew flying to Copenhagen on March 4, 1990.

FNAME	LNAME
KAREN	CARTER
SHARON	DEAN
ROGER	DENNIS
KATRINA	FERNANDEZ
ANNE	KIRBY
JACKSON	KRAMER
RUSSELL	LONG
JASPER	MARSHBURN
MILTON	RAYNOR
SIMON	RIVERS
ALAN	TUCKER
ANNE	WALTERS
DEBORAH	YOUNG

```
libname s1 "D:\SAS\SAS Capstone Project\data";

proc sql;
select fname, lname
from s1.staff a
where a.idnum in
(select supid
from s1.superv,
(select substr(jobcode, 1,2) as jobct,
state as sstate
from s1.staff s,
s1.payroll p
where s.idnum=p.idnum
and s.idnum in
(select w.idnum
from s1.schedule w
where date='04mar90'd
and dest='CPH'))
where superv.jobcat=jobcat and sstate=superv.state);
quit;
```