MF004 – semestrální projekt

Marek Bryša, 323771

Brno, 9. ledna 2012

1 Data

1.1 Úvod

Zdrojem dat je soubor `MortgageDefaulters.xls`. Obsahuje údaje o 15153 klientech banky, kteří mají hypoteční úvěr. Cílem je identifikovat klienty s vysokou pravděpodobností, že nebudou splácet. K dispozici jsou tyto proměnné:

Bo_Age Věk klienta

Ln_Orig Výše půjčký v USD

Orig_LTV_Ratio_Pct Poměr půjcky a nákupní ceny domu

Credit_score Klientovo credit score

First_home První klientův nákup dům (Y/N)

Tot_mthly_debt_exp Klientův celkový měsíční výdaj na půjčku

Tot_mthly_incm Klinetův celkový měsíční příjem

orig_apprd_val_amt Odhad ceny domu v době žádosti

pur_prc_amt Kupní cena domu

DTI_ratio Poměr nákladu na dluh a příjmu klienta – měsíčně

StatusCurrent Stav půjčky

OUTCOME Binary binární verze stavu: 0=v pořádku, 1=default

StateUS Stát USA, kde se dům nachází

Median_state_inc Střední příjem domácnosti v daném státě (2002-2004)

UPB>Appraisal Je půjčka vyšší než odhad? 0=ne, 1=ano

1.2 Popisné statistiky

V následující tabulce jsou uvedeny popisné statistiky intervalových proměnných:

	Mean	Std.Err.	Median	Std.Dev.	Min	Max
Bo_Age	36.79	0.08	37	10.03	18	99
Ln_Orig	153467.57	555.42	141500	68370.61	19600	599000
Orig_LTV_Ratio_Pct	93.08	0.07	95	8.85	20	111
Credit_score	687.67	0.51	688	62.90	440	999
Tot_mthly_debt_exp	1745.46	8.85	1578	1089.20	0	17225
Tot_mthly_incm	5024.71	23.98	4632	2952.16	500	65000
orig_apprd_val_amt	170661.44	664.31	154000	81775.07	0	870000
pur_prc_amt	164681.56	647.62	148650	79719.84	20000	870000
DTI Ratio	0.37	0.00	0	0.18	0	3
Median_state_inc	44945.07	44.12	43988	5431.51	32589	57352

Je zřejmé, že data obsahují některé podezřelé hodnoty. Bude nunté porvést transformaci.

1.3 Transformace dat

U 903 klientů je proměnná `Tot_mthly_debt_exp` nulová. Tím je u stejného počtu nulová i proměnná `DTI_Ratio`. U ostatních klientů nabývá `Tot_mthly_debt_exp` v průměru 0.016045 násobku `Ln_Orig`. Touto hodnotou nulu nahradíme a dopočteme `DTI_Ratio`.

U 19 klientů je `orig_apprd_val_amt` nulová. Stejným postupem ji rekonstruujeme z `pur_prc_amt`, přičemž u ostatních klientů platí, že

$$\texttt{orig\_apprd\_val\_amt} = 1.059314 \cdot \texttt{pur\_prc\_amt}.$$

Dále budeme pracovat s daty kategorizovanými podle decilů.

2 Regrese

Model Information	
Data Set	WORK.CAT
Response Variable	OUTCOME
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	15153
Number of Observations Used	15153

Response Profile		
Ordered Value	**OUTCOME**	**Total Frequency**
1	0	14751
2	1	402

Probability modeled is OUTCOME='0'.

Step 5. Effect Tot_mthly_incm is removed:

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

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	3713.359	3557.623
SC	3720.985	3588.127
-2 Log L	3711.359	3549.623

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	161.7361	3	<.0001
Score	159.1847	3	<.0001
Wald	151.0764	3	<.0001

Residual Chi-Square Test		
Chi-Square	**DF**	**Pr > ChiSq**
8.3415	9	0.5001

Note: No effects for the model in Step 5 are removed.

Note: Model building terminates because the last effect entered is removed by the Wald statistic criterion.

Summary of Stepwise Selection								
Step	**Effect**		**DF**	**Number In**	**Score Chi-Square**	**Wald Chi-Square**	**Pr > ChiSq**	**Variable Label**
	Entered	**Removed**						
1	**Credit_score**		1	1	99.3280		<.0001	Values of Credit_score Were Replaced by Ranks
2	**Tot_mthly_debt_exp**		1	2	52.4367		<.0001	Values of Tot_mthly_debt_exp Were Replaced by Ranks
3	**First_home**		1	3	6.3314		0.0119	
4	**Tot_mthly_incm**		1	4	1.8461		0.1742	Values of Tot_mthly_incm Were Replaced by Ranks
5		**Tot_mthly_incm**	1	3		1.8449	0.1744	Values of Tot_mthly_incm Were Replaced by Ranks

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	2.2473	0.1136	391.3753	<.0001
First_home	1	0.2911	0.1160	6.2973	0.0121
Credit_score	1	0.1940	0.0197	97.3968	<.0001
Tot_mthly_debt_exp	1	0.1346	0.0184	53.6301	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
First_home	1.338	1.066	1.679
Credit_score	1.214	1.168	1.262
Tot_mthly_debt_exp	1.144	1.104	1.186

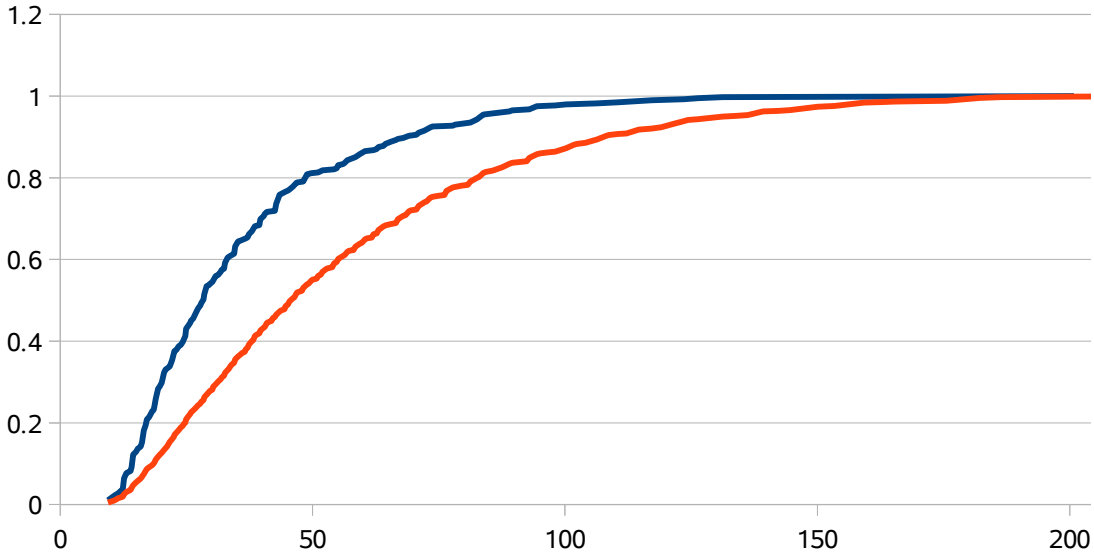
Association of Predicted Probabilities and Observed Responses

Percent Concordant	66.7	**Somers' D**	0.368
Percent Discordant	30.0	**Gamma**	0.380
Percent Tied	3.3	**Tau-a**	0.019
Pairs	5929902	**c**	0.684

Byla použita metoda stepwise, která přidáva a odebíra vysvětlující proměnné podle jejich statistické významnosti. Výsledkem je model se třemi proměnnými. Hypotézu o celkové nevýznamnosti modelu zamítáme.

Pokud klient žádá o půjčku na svůj první dům, snižuje se šance nesplácení. Stejně tak má pozitivní vliv credit score přidělené při žádosti o úvěr. Dále platí, že klient je tím lepší, čím vyšší jsou jeho měsíční výdaje na splátky dluhu. To lze vystvětlit tím, že vysoký úvěr dostanou jen solidní klienti.

Konkordantní páry tvoří 66.7%, diskordantní 30%. K porovnání vypovídací hodnoty s případným jiným modelem je možno použít hodnotu Somers' D. Kolmogorova-smirnovova statistika nabývá hodnoty 0.28647, její výpočet probíha pomocí skriptu `KS.py` v jazyce Python. Na následující straně je uveden graf distribučních funkcí pro dobré a špatné případy.


1.2
1
0.8
0.6
0.4
0.2
0
0
50
100
150
200