HW 4 GROUP 5

Christopher Webster, Ryan Timbrook, David Eboh

Resources

Use the dataset SCM 651 Homework 4 Universal Bank.csv.

Assignment What's due:

Submit a logit, probit, and neural network analysis of loan acquisition behavior **before the live class in week 10**. Suggested length is five pages, but should not exceed ten pages, single- spaced, 12-point font.

This is a group assignment; each student should upload a copy of the assignment to the Learning Management System. The paper must be a Microsoft Word document. You should also submit the Excel spreadsheet with the prediction models and sensitivity analyses. Name the file HW4_Team# where # is your team number. Be sure to include the names of everyone on the team on the first page of the paper. Late assignments will not be accepted. Failure to follow directions will be penalized.

Outline and grading criteria:

1. Perform a logit and probit analysis of the variables that affect whether a customer takes out a loan. Consider only main effects. Which variables are significant? How do the significant variables influence the likelihood of taking out a loan? Copy screen snapshots of your analysis in R to your report. (20%)

Preliminary Model to see which variables are significant:

```
Call:
glm(formula = PersonalLoan ~ Age + Experience + Income + Family +
    CCAvg + Education + Mortgage + SecuritiesAccount + CDAccount +
    Online + CreditCard, family = binomial(logit), data = HW4Data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1558  -0.2015  -0.0796  -0.0309   3.9298

Coefficients:
                     Estimate  Std. Error z value Pr(>|z|)
(Intercept)       -12.1926537   1.6451710  -7.411 1.25e-13 ***
Age                -0.0536101   0.0613121  -0.874  0.38191
Experience          0.0637567   0.0609277   1.046  0.29536
Income              0.0545810   0.0026201  20.831  < 2e-16 ***
Family              0.6957575   0.0743035   9.364  < 2e-16 ***
CCAvg               0.1239704   0.0396467   3.127  0.00177 **
Education           1.7361518   0.1150671  15.088  < 2e-16 ***
Mortgage            0.0004745   0.0005541   0.856  0.39190
SecuritiesAccount  -0.9368084   0.2858666  -3.277  0.00105 **
CDAccount           3.8225340   0.3239404  11.800  < 2e-16 ***
Online             -0.6751707   0.1570775  -4.298 1.72e-05 ***
CreditCard         -1.1197449   0.2049925  -5.462 4.70e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3162.0  on 4999  degrees of freedom
Residual deviance: 1284.4  on 4988  degrees of freedom
AIC: 1308.4

Number of Fisher Scoring iterations: 8
```

After this, the team also looked at a correlation matrix to further investigate which 2-4 variables are the most correlated with taking a personal loan:

	Age	CCAvg	CDAccount	CreditCard
Age	1.0000000000	-0.052012179	0.008042552	0.007681037
CCAvg	-0.0520121791	1.000000000	0.136533655	-0.006689494
CDAccount	0.0080425521	0.136533655	1.000000000	0.278644365
CreditCard	0.0076810368	-0.006689494	0.278644365	1.000000000
Education	0.0413343834	-0.136123922	0.013933888	-0.011014134
Experience	0.9942148570	-0.050076511	0.010353331	0.008967447
Family	-0.0464176636	-0.109274506	0.014110365	0.011588066
Income	-0.0552686182	0.645983670	0.169738080	-0.002385008
Mortgage	-0.0125385869	0.109904723	0.089311058	-0.007230919
Online	0.0137024021	-0.003611009	0.175880016	0.004209656
PersonalLoan	-0.0077256172	0.366888736	0.316354829	0.002801509
SecuritiesAccount	-0.0004362422	0.015086311	0.317034416	-0.015028319

	Education	Experience	Family	Income
Age	0.04133438	0.994214857	-0.04641766	-0.055268618
CCAvg	-0.13612392	-0.050076511	-0.10927451	0.645983670
CDAccount	0.01393389	0.010353331	0.01411036	0.169738080
CreditCard	-0.01101413	0.008967447	0.01158807	-0.002385008
Education	1.00000000	0.013151813	0.06492891	-0.187524257
Experience	0.01315181	1.000000000	-0.05256315	-0.046574178
Family	0.06492891	-0.052563147	1.00000000	-0.157500785
Income	-0.18752426	-0.046574178	-0.15750079	1.000000000
Mortgage	-0.03332712	-0.010581552	-0.02044493	0.206806228
Online	-0.01500382	0.013897900	0.01035404	0.014205919
PersonalLoan	0.13672155	-0.007413098	0.06136704	0.502462292
SecuritiesAccount	-0.01081201	-0.001232134	0.01999408	-0.002616497

	Mortgage	Online	PersonalLoan	SecuritiesAccount
Age	-0.012538587	0.013702402	-0.007725617	-0.0004362422
CCAvg	0.109904723	-0.003611009	0.366888736	0.0150863114
CDAccount	0.089311058	0.175880016	0.316354829	0.3170344157
CreditCard	-0.007230919	0.004209656	0.002801509	-0.0150283189
Education	-0.033327125	-0.015003821	0.136721550	-0.0108120136
Experience	-0.010581552	0.013897900	-0.007413098	-0.0012321344
Family	-0.020444931	0.010354036	0.061367044	0.0199940798
Income	0.206806228	0.014205919	0.502462292	-0.0026164967
Mortgage	1.000000000	-0.005994898	0.142095236	-0.0054109700
Online	-0.005994898	1.000000000	0.006277815	0.0126274704
PersonalLoan	0.142095236	0.006277815	1.000000000	0.0219538822
SecuritiesAccount	-0.005410970	0.012627470	0.021953882	1.0000000000

Looking at the correlation matrix, it's clear that Income, CCAvg and CDAccount are very highly correlated with Personal Loan. Unfortunately, after running a Logit with these three variables, CCAvg is no longer significant. As such, the team removed the CCAvg variable and replaced it with the Family variable. This variable is statistically significant in the first Logit model.

Logit:

```
Call:
glm(formula = PersonalLoan ~ Income + Family + CDAccount, family = binomial(logit),

    data = HW4Data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0629  -0.2752  -0.1381  -0.0618   3.3554

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.38530    0.36016  -26.06   <2e-16 ***
Income       0.04479    0.00178   25.17   <2e-16 ***
Family       0.84825    0.06878   12.33   <2e-16 ***
CDAccount    2.33748    0.19156   12.20   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3162.0  on 4999  degrees of freedom
Residual deviance: 1665.6  on 4996  degrees of freedom
AIC: 1673.6

Number of Fisher Scoring iterations: 7
```

Coefficients							
	Estimate	Std. Error	z value	Pr(>|z|)		Input	Input*Coeff
(Intercept)	-9.3853	0.36016	-26.06	<2e-16	***	1	-9.3853
Income	0.04479	0.00178	25.17	<2e-16	***	100	4.479
Family	0.84825	0.06878	12.33	<2e-16	***	4	3.393
CDAccount	2.33748	0.19156	12.2	<2e-16	***	1	2.33748
						Sum	0.82418
						Exp(Sum)	2.28001039
						Probability	69.5%

Probit:

```
Call:
glm(formula = PersonalLoan ~ Income + Family + CDAccount, family = binomial(probit)
,
    data = HW4Data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1917  -0.2778  -0.1055  -0.0286   3.5996

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.9173319  0.1752108  -28.07   <2e-16 ***
Income       0.0239108  0.0009048   26.43   <2e-16 ***
Family       0.4027290  0.0349849   11.51   <2e-16 ***
CDAccount    1.2600334  0.1020413   12.35   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3162.0  on 4999  degrees of freedom
Residual deviance: 1657.7  on 4996  degrees of freedom
AIC: 1665.7

Number of Fisher Scoring iterations: 8
```

Coefficients:							
	Estimate	Std. Error	z value	Pr(>|z|)		Input	Input*Coeff
(Intercept)	-4.91733	0.175211	-28.07	<2e-16	***	1	-4.91733
Income	0.023911	0.000905	26.43	<2e-16	***	100	2.39108
Family	0.402729	0.034985	11.51	<2e-16	***	4	1.610916
CDAccount	1.260033	0.102041	12.35	<2e-16	***	1	1.260033
						Sum	0.344698
						Probability	63.5%

In both of these models, the significant variables are: Income, Family, CDAccount. As the customer's income increases, the probability of taking a personal loan out increases. As the customer's family increases, the probability of taking out a personal loan increases. As the customer's Average spending on Credit Cards per month increases, the probability of taking out a personal loan increases.

2. Add moderating effects (interactions of variables). Which interactions make sense conceptually? Which interactions are statistically significant? How do you interpret the coefficients on these variables? Copy screen snapshots of your analysis in R to your report. (20%)

Looking at the correlation matrix, it's clear that both Income and CCAvg are related with each other as well as statistically significant in the logit. Also, if one's Income is higher, you would expect them to spend more. However, it one's income is lower and the credit card average is higher, you would expect the probability of them taking out a loan to increase.

The final model with moderating effects is below:

```
Call:
glm(formula = PersonalLoan ~ CCAvg * Income + Income + Family,
    family = binomial(logit), data = HW4Data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3362  -0.2391  -0.0780  -0.0221   3.1918

Coefficients:
                Estimate  Std. Error z value Pr(>|z|)
(Intercept)  -13.7771831   0.6031461  -22.84   <2e-16 ***
CCAvg          1.6469211   0.1361439   12.10   <2e-16 ***
Income         0.0750491   0.0037406   20.06   <2e-16 ***
Family         0.9395873   0.0679152   13.84   <2e-16 ***
CCAvg:Income  -0.0106102   0.0008969  -11.83   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3162.0  on 4999  degrees of freedom
Residual deviance: 1647.2  on 4995  degrees of freedom
AIC: 1657.2

Number of Fisher Scoring iterations: 8
```

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)		Input	Coeff*Value
(Intercept)	-13.7772	0.603146	-22.84	<2e-16	***	1	-13.7771831
CCAvg	1.646921	0.136144	12.1	<2e-16	***	2.5	4.11730275
Income	0.075049	0.003741	20.06	<2e-16	***	100	7.50491
Family	0.939587	0.067915	13.84	<2e-16	***	4	3.7583492
CCAvg:Income	-0.01061	0.000897	-11.83	<2e-16	***	250	-2.65255
						Sum	-1.04917115
						Exp(Sum)	0.350227915
						Probability	25.9%

As CC Avg spending increase, the probability of taking out a personal loan increases. As income increases, the probability of taking out a personal loan increases. As the family size increases, the probability of taking out a loan increases. As the combination of CCAvg and Income increases, the probability of taking out a personal loan decreases.

3. Create a final regression model with the variables that you feel are important (both main effects and interaction terms). Create a spreadsheet prediction of the model. Which variables have the greatest influence on the customers' loan behavior (combined main effects and interaction effects)? Perform a sensitivity analysis as seen earlier in the semester. Copy screen snapshots of your analysis in R to your report. (20%)

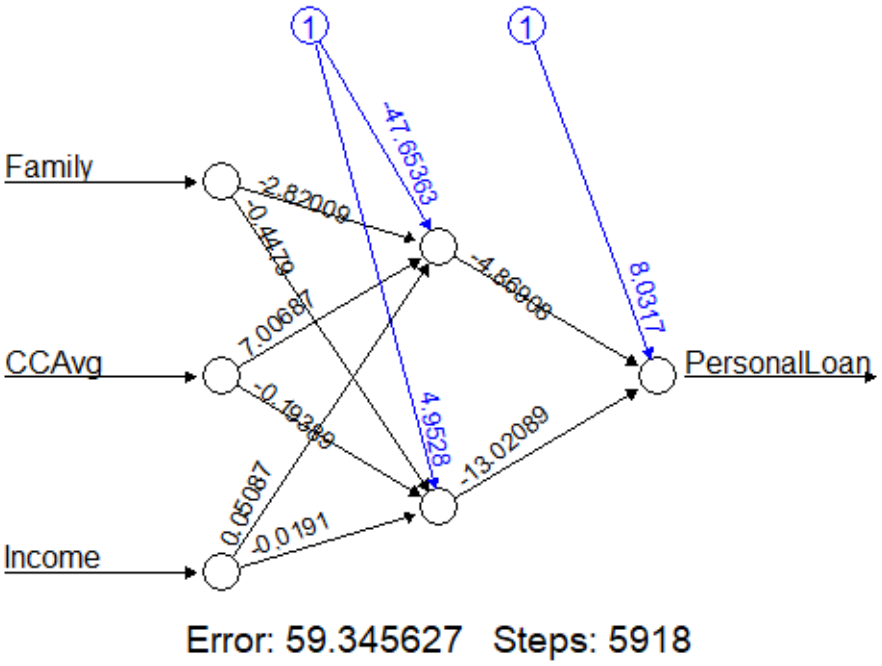
```
Call:
glm(formula = PersonalLoan ~ CCAvg * Income + Income + Family,
    family = binomial(logit), data = HW4Data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3362  -0.2391  -0.0780  -0.0221   3.1918

Coefficients:
                Estimate  Std. Error z value Pr(>|z|)
(Intercept)  -13.7771831   0.6031461  -22.84   <2e-16 ***
CCAvg          1.6469211   0.1361439   12.10   <2e-16 ***
Income         0.0750491   0.0037406   20.06   <2e-16 ***
Family         0.9395873   0.0679152   13.84   <2e-16 ***
CCAvg:Income  -0.0106102   0.0008969  -11.83   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3162.0  on 4999  degrees of freedom
Residual deviance: 1647.2  on 4995  degrees of freedom
AIC: 1657.2

Number of Fisher Scoring iterations: 8
```

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)		Input	Coeff*Value		Wald Test
(Intercept)	-13.7772	0.603146	-22.84	<2e-16	***	1	-13.7772		
CCAvg	1.646921	0.136144	12.1	<2e-16	***	2.5	4.117303		146.3353
Income	0.075049	0.003741	20.06	<2e-16	***	100	7.50491		402.5394
Family	0.939587	0.067915	13.84	<2e-16	***	4	3.758349		191.3993
CCAvg:Income	-0.01061	0.000897	-11.83	<2e-16	***	250	-2.65255		139.9455
						Sum	-1.04917		
						Exp(Sum)	0.350228		
						Probability	25.9%		

Looking at the Wald Test, it is clear that Income and Family have the highest impact on the model. As such, the 2-way sensitivity analysis between Income and Family is below:

		Income									
	25.9%	25	50	75	100	125	150	175	200	225	250
Family	1	0%	0%	1%	2%	7%	19%	44%	73%	90%	97%
	2	0%	0%	2%	5%	15%	38%	67%	87%	96%	99%
	3	0%	1%	4%	12%	32%	61%	84%	95%	98%	99%
	4	1%	3%	9%	26%	54%	80%	93%	98%	99%	100%
	5	2%	7%	21%	47%	75%	91%	97%	99%	100%	100%
	6	6%	17%	41%	70%	89%	96%	99%	100%	100%	100%
	7	13%	34%	64%	85%	95%	99%	100%	100%	100%	100%
	8	28%	57%	82%	94%	98%	99%	100%	100%	100%	100%
	9	50%	77%	92%	97%	99%	100%	100%	100%	100%	100%
	10	72%	90%	97%	99%	100%	100%	100%	100%	100%	100%

4. Perform a neural network analysis of the variables found to be significant in the logit and probit analysis above. Copy screen snapshots of your final neural network model in R to your report. (20%)



5. Create a prediction model of the neural network. Using the prediction model, perform a sensitivity analysis for the neural network model similar to the logit and probit sensitivity analysis. (20%)

error	59.34562728
reached.threshold	0.07654892
steps	5918
Intercept.to.1layhid1	-47.65362907
Family.to.1layhid1	-2.82009272
CCAvg.to.1layhid1	7.00686758
Income.to.1layhid1	0.05086553
Intercept.to.1layhid2	4.95279997
Family.to.1layhid2	-0.44790067
CCAvg.to.1layhid2	-0.19388949
Income.to.1layhid2	-0.01909625
Intercept.to.PersonalLoan	8.03170461
1layhid1.to.PersonalLoan	-4.86907574
1layhid2.to.PersonalLoan	-13.02088921

Variable	Input
Intercept	1
Family	4
CCAvg	3.5
Income	125

Hidden Node1		
Variable	Coefficient	Coeff*Value
Intercept	-47.6536291	-47.6536291
Family	-2.82009272	-11.2803709
CCAvg	7.00686758	24.5240365
Income	0.05086553	6.35819125
	Sum	-28.0517722
	exp(Sum)	6.5655E-13
	Probability	0%

Hidden Node2		
Variable	Coefficient	Co-eff*Value
Intercept	4.95279997	4.95279997
Family	-0.44790067	-1.79160268
CCAvg	-0.19388949	-0.67861322
Income	-0.01909625	-2.38703125
	Sum	0.09555283
	exp(Sum)	1.10026694
	Probability	52%

Output			
Variables	Coefficient	Value	Coeff*Value
Intercept	8.03170461	1	8.03170461
Hidden1	-4.86907574	0%	-2.0989E-24
Hidden2	-13.0208892	52%	-6.82125385
		Sum	1.21045076
		Exp(Sum)	3.35499663
		Probability	77%

As the neural network prediction above shows, a customer with a family of 4, CC average spend of 3.5 thousand per month, and annual income of 125 thousand is predicted to take a personal loan 77% of the time.

		Income									
	77%	25	50	75	100	125	150	175	200	225	250
Family	1	0.01048404	0.01349964	0.01983453	0.03490962	0.0763087	0.19959915	0.49360693	0.81732916	0.95368201	0.9878061
	2	0.01324595	0.01927697	0.03350638	0.07225393	0.18768089	0.47106339	0.8025521	0.94948124	0.986832	0.99570936
	3	0.01874783	0.03218662	0.06846745	0.17646662	0.44880239	0.7869	0.944891	0.98576728	0.99543768	0.99807841
	4	0.03094466	0.0649307	0.16592926	0.42691556	0.7703787	0.93987894	0.98460303	0.99514362	0.99798852	0.99892674
	5	0.06162634	0.15604	0.40548691	0.75300371	0.93441109	0.98332942	0.99482509	0.99789237	0.9988914	0.99928426
	6	0.14676887	0.38459171	0.73480099	0.92845196	0.98193571	0.9944798	0.99778944	0.99885401	0.99926821	0.9994557
	7	0.36429571	0.71580735	0.92196484	0.98041011	0.99410523	0.99767917	0.9988144	0.99925137	0.99944757	0.99954604
	8	0.69607071	0.91491198	0.97873973	0.99369859	0.99756095	0.99877244	0.99923369	0.9994391	0.99954159	0.99959688
	9	0.90725506	0.97691047	0.99325682	0.99743409	0.99872794	0.99921512	0.99943025	0.99953697	0.99959431	0.99962678
	10	0.97490694	0.99277656	0.99729788	0.99868071	0.9991956	0.99942102	0.99953217	0.99959165	0.99962524	0.99964488

Justify your answers. Provide a snapshot of output from your analysis in your final paper.