HW4_Team #92

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%)

The following variables are significant in the logit and probit analysis and their influence in the likelihood of taking out a loan:

1) CCAvg
    a. As average credit card spend per month increases, the likelihood increases by 0.114 (logit), 0.071 (probit). This could be because an individual is more likely to take out a loan to pay off credit card spend.
2) CDAccount
    a. If the individual has a certificate of deposit with the bank, the likelihood increases by 3.853 (logit), 2.018 (probit). This could be because if an individual has an existing relationship with the bank via Certificate of Deposit, they are more likely to continue that relationship with the bank.
3) Credit Card
    a. The likelihood of taking a loan decreases if the individual has a credit card issued by the Universal Bank by -1.124 (logit), -0.583 (probit). This could be because the bank is less likely to issue a loan to an individual with a line of credit/money owed to the bank from the credit card.
4) Education
    a. As education levels increase from undergraduate, graduate, and/or advanced/professional, the likelihood of taking out a loan increases by 1.704 (logit), 0.838 (probit). This could be because individuals that have more education typically need a method to pay for that education.
5) Family
    a. As family grows, the likelihood of taking out a loan increases by 0..690 (logit), 0.341 (probit) This could be because there are more people to take care of which requires more money.
6) Income
    a. As an income increases, the likelihood of an individual taking out a loan increases by 0.055 (logit), 0.028 (probit). This could be because individuals that make more income have a greater chance of paying the loan back.
7) Online
    a. The likelihood of taking out a loan decreases by -0.667 (logit), and -0.350 (probit) if the customer uses internet banking facilities.
8) SecuritiesAccount
    a. The likelihood of taking out a loan decreases negatively by -0.935 (logit), and -0.499 (probit) if the customer has a securities account with the bank.

R screenshot of Logit analysis with all variables and then logit analysis with only the significant variables:

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

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
```

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1451  -0.2054  -0.0804  -0.0310   3.8986

Coefficients:
                   Estimate Std. Error z value     Pr(>|z|)
(Intercept)      -13.224197   0.562495 -23.510      < 2e-16 ***
CCAvg              0.113713   0.039265   2.896      0.00378 **
CDAccount          3.853311   0.323447  11.913      < 2e-16 ***
CreditCard        -1.123683   0.205003  -5.481 0.0000000422 ***
Education          1.704116   0.112393  15.162      < 2e-16 ***
Family             0.690388   0.074201   9.304      < 2e-16 ***
Income             0.054721   0.002589  21.133      < 2e-16 ***
Online            -0.667476   0.156717  -4.259 0.0000205232 ***
SecuritiesAccount -0.934627   0.284849  -3.281      0.00103 **
---
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: 1288.6  on 4991  degrees of freedom
```

R screenshot of Probit analysis with all variables and then Probit analysis with only the significant variables:

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2759  -0.2065  -0.0524  -0.0090   4.4706

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -6.0671118  0.8269708  -7.337 2.19e-13 ***
Age               -0.0303628  0.0312820  -0.971 0.331740
Experience         0.0337833  0.0311288   1.085 0.277800
Income             0.0277314  0.0012705  21.828  < 2e-16 ***
Family             0.3417417  0.0375270   9.107  < 2e-16 ***
CCAvg              0.0743382  0.0209287   3.552 0.000382 ***
Education          0.8509102  0.0567310  14.999  < 2e-16 ***
Mortgage           0.0002217  0.0002950   0.751 0.452395
SecuritiesAccount -0.4991692  0.1470525  -3.394 0.000688 ***
CDAccount          2.0049036  0.1646493  12.177  < 2e-16 ***
Online            -0.3515799  0.0810717  -4.337 1.45e-05 ***
CreditCard        -0.5825612  0.1045810  -5.570 2.54e-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: 1303.2  on 4988  degrees of freedom
AIC: 1327.2

Number of Fisher Scoring iterations: 8
```

```
Call:
glm(formula = PersonalLoan ~ CCAvg + CDAccount + CreditCard +
    Education + Family + Income + Online + SecuritiesAccount,
    family = binomial(probit), data = UniversalBank)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2706  -0.2067  -0.0532  -0.0089   4.4378

Coefficients:
                   Estimate Std. Error z value    Pr(>|z|)
(Intercept)       -6.730067   0.262167 -25.671     < 2e-16 ***
CCAvg              0.070770   0.020779   3.406    0.000659 ***
CDAccount          2.018424   0.164391  12.278     < 2e-16 ***
CreditCard        -0.583261   0.104525  -5.580 0.000000024 ***
Education          0.837564   0.055464  15.101     < 2e-16 ***
Family             0.340529   0.037509   9.079     < 2e-16 ***
Income             0.027891   0.001258  22.173     < 2e-16 ***
Online            -0.350131   0.080986  -4.323 0.000015369 ***
SecuritiesAccount -0.499103   0.146829  -3.399    0.000676 ***
---
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: 1305.8  on 4991  degrees of freedom
```

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%)

There are four variables (outside of the binary variables) that are statistically significant which are Family, Income, CCAvg, and Education.

Using these four variables, there are three interactions that make sense conceptually as moderating effects, these are:

1) Family*Income
2) CCAvg*Income
3) Education*Income

When we plotted these moderating effects using the Logit analysis, we see the following analysis:

Family*Income

The moderating effect of Family*Income is significant. Family as a main effect decreases the likelihood of a personal loan by -1.489. Income as a main effect from this analysis becomes insignificant. However, when combine together Family*Income, the effect is that it increases the likelihood of a personal loan by 0.020.

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

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.28751  -0.28285  -0.15888  -0.07353   2.95928

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.6146848  0.4796968  -7.535 4.87e-14 ***
Family        -1.4895358  0.2119153  -7.029 2.08e-12 ***
Income         0.0001998  0.0038434   0.052    0.959
Family:Income  0.0202891  0.0018521  10.955  < 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: 1667.2  on 4996  degrees of freedom
AIC: 1675.2

Number of Fisher Scoring iterations: 7


> exp(coef(GLM.9))  # Exponentiated coefficients ("odds ratios")
  (Intercept)          Family         Income Family:Income
```

CCAvg*Income

The moderating effect of CCAvg*Income is significant. In this model, all three variables are significant. As average credit card spend per month increases, the likelihood of a personal loan increases by 1.340. The same goes for income, as it increases, the likelihood of a personal loan increases by 0.060. However, when combining the two variables, the interaction is significant but going in a negative direction, which means as CCAvg*Income goes up, the likelihood of a personal loan decreases by -0.009.

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0995  -0.2660  -0.1155  -0.0416   2.6515

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)  -9.5556476  0.4309632  -22.17   <2e-16 ***
CCAvg         1.3398815  0.1215377   11.02   <2e-16 ***
Income        0.0600142  0.0031175   19.25   <2e-16 ***
CCAvg:Income -0.0086107  0.0008001  -10.76   <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  on 4999  degrees of freedom
Residual deviance: 1880  on 4996  degrees of freedom
AIC: 1888

Number of Fisher Scoring iterations: 7
```

Education*Income

The moderating effect of Education*Income is significant. In this model, all three variables are significant. As education increases, it decreases the likelihood of taking out a loan by -7.044. Same for Income, as income increases, it decreases the likelihood of taking out a loan by - 0.0589. However, when combining the two variables as a moderating effect, the interaction increases the likelihood of personal loan by 0.079.

```
Call:
glm(formula = PersonalLoan ~ Education + Income + Education *
    Income, family = binomial(logit), data = UniversalBank)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8043  -0.2403  -0.1236  -0.0091   3.9351

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.905818   0.805047   2.367   0.0179 *
Education        -7.043853   0.676857 -10.407   <2e-16 ***
Income           -0.058799   0.007082  -8.303   <2e-16 ***
Education:Income  0.079411   0.006280  12.646   <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: 1157.2  on 4996  degrees of freedom
AIC: 1165.2

Number of Fisher Scoring iterations: 9
```

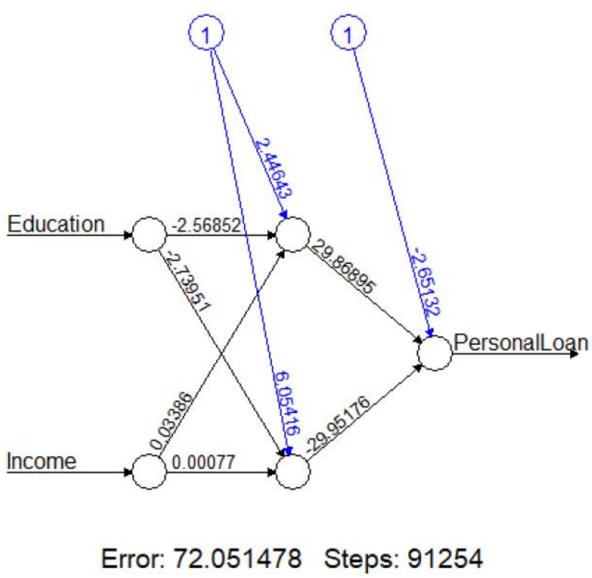
3. Create a final regression model with the variables that you feel are important (both main effects and interaction terms). Create a spreadsheet prediction of the model. Which variables have the greatest influence on the customers' loan behavior (combined main effects and interaction effects)? Perform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

Income seems to be the greatest influence in customers' loan behavior, if the customers' income is high, even with an undergraduate degree, the probability is still higher than if they have low income and high education level. With the moderating effect of income and education, it becomes probability becomes linear as both variables increase.


```
Call:
glm(formula = PersonalLoan ~ Education + Income + Education *
    Income, family = binomial(logit), data = UniversalBank)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8043  -0.2403  -0.1236  -0.0091   3.9351

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.905818   0.805047   2.367   0.0179 *
Education        -7.043853   0.676857 -10.407   <2e-16 ***
Income           -0.058799   0.007082  -8.303   <2e-16 ***
Education:Income  0.079411   0.006280  12.646   <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: 1157.2  on 4996  degrees of freedom
AIC: 1165.2

Number of Fisher Scoring iterations: 9
```

Sensitivity Analysis				
		Education		
		Undergraduate	Graduate	Advanced/Professional
	22%	1	2	3
Income	10	1%	0%	0%
	20	1%	0%	0%
	30	1%	0%	0%
	40	1%	0%	0%
	50	2%	0%	0%
	60	2%	0%	0%
	70	2%	1%	0%
	80	3%	2%	1%
	90	4%	4%	4%
	100	4%	10%	22%
	110	5%	24%	63%
	120	7%	46%	91%
	130	8%	69%	98%
	140	10%	86%	100%
	150	11%	94%	100%
	160	14%	98%	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%)

Sensitivity Analysis				
		Education		
		Undergraduate	Graduate	Advanced/Professional
	12%	1	2	3
Income	10	0%	0%	0%
	20	0%	0%	0%
	30	0%	0%	0%
	40	0%	0%	1%
	50	0%	0%	1%
	60	0%	0%	1%
	70	1%	0%	1%
	80	2%	0%	2%
	90	4%	1%	5%
	100	5%	8%	12%
	110	7%	41%	34%
	120	9%	80%	72%
	130	10%	94%	95%
	140	11%	98%	99%
	150	12%	99%	100%
	160	12%	100%	100%