

Loan Analysis

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?

Stepwise Logit Results

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	0.000205232 ***
SecuritiesAccount	-0.934627	0.284849	-3.281	0.00103 **

Stepwise Probit Results

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 ***

- The variables of CCAvg, CDAccount, CreditCard, Education, Family, Income, Online, and SecuritiesAccount are all statistically significant. Of these variables, CCAvg, CDAccount, Education, Family, and Income are all positively correlated with the likelihood of taking out a loan. While CreditCard, Online, and SecuritiesAccount are all negatively correlated with the likelihood of taking out a loan. Meaning that having the following traits results in a lower likelihood of taking out a loan; having a Universal Bank credit card, having a Universal Bank securities account, and/or utilizing online banking. While the following traits are associated with a higher likelihood of taking out a loan; spending more on credit cards, having a Universal Bank CD account, higher levels of education, having larger families, having higher incomes.

R Code:

```

UniversalBank <-
readXL("/Users/Dyllan/Desktop/scm651_homework_4_universal_bank.xls",
        rownames=FALSE, header=TRUE, na="",
        sheet="scm651_homework_4_universal_ban",
        stringsAsFactors=TRUE)

#Logit

Logit <- glm(PersonalLoan ~ CCAvg + CDAccount + CreditCard + Education +
              Family + Income + Online + SecuritiesAccount,
              family=binomial(logit),
              data=UniversalBank)
summary(Logit)
exp(coef(Logit)) # Exponentiated coefficients ("odds ratios")

#Probit

Probit <- glm(PersonalLoan ~ CCAvg + CDAccount + CreditCard + Education +
              Family + Income + Online + SecuritiesAccount,
              family=binomial(probit),
              data=UniversalBank)
summary(Probit)

```

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%)

```

              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 ***

```

```

      mean      sd IQR 0% 25% 50% 75% 100%    n
Education 1.8810 0.8398691  2  1  1  2  3  3 5000
Income   73.7742 46.0337293 59  8 39 64 98 224 5000

```

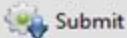
Fig 1: Numerical summaries of education and income data

The following correlation tests in R indicated conceptually that interactions between income and level of education make the most sense and it is highly significant ($p < 2.2e-16$). Interaction between age and income is also significant ($p < 0.00009226$). Correlation between credit card : education and credit card : family were computed and it was not statistically significant.

Fig 2: Correlation matrix of selected data: Education, income, credit card, and family

```
> cor(hw4[,c("CreditCard", "Education", "Family", "Income")], use="complete")
```

	CreditCard	Education	Family	Income
CreditCard	1.000000000	-0.01101413	0.01158807	-0.002385008
Education	-0.011014134	1.000000000	0.06492891	-0.187524257
Family	0.011588066	0.06492891	1.000000000	-0.157500785
Income	-0.002385008	-0.18752426	-0.15750079	1.000000000

Output 

```
Pearson's product-moment correlation

data: Education and Income
t = -13.497, df = 4998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2141305 -0.1606400
sample estimates:
      cor
-0.1875243

> with(hw4, cor.test(Age, Income, alternative="two.sided", method="pearson"))

Pearson's product-moment correlation

data: Age and Income
t = -3.9133, df = 4998, p-value = 0.00009226
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.08286097 -0.02759160
sample estimates:
      cor
-0.05526862
```

Output Submit

```

Pearson's product-moment correlation

data: CreditCard and Education
t = -0.77871, df = 4998, p-value = 0.4362
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.03872161  0.01671026
sample estimates:
      cor
-0.01101413

> with(hw4, cor.test(CreditCard, Family, alternative="two.sided",
+   method="pearson"))

Pearson's product-moment correlation

data: CreditCard and Family
t = 0.81929, df = 4998, p-value = 0.4127
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.01613641  0.03929474
sample estimates:
      cor
0.01158807

```

Fig 3: Screen snapshots from R statistics

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.

PersonalLoan = 1.905818 - 7.043853*Education - .058799*Income + .079411*(Education*Income)						
Inputs			Output:			
Variable	Value		Variable	Coefficient	Value	Coeff*Value
			intercept	1.905818	1	1.905818
Education	2	1 to 3	Education	-7.043853	2	-14.087706
Income	150	8 to 200	Income	-0.058799	150	-8.81985
Educ*Inc	300	moderating	Educ*Inc	0.079411	300	23.8233
					sum	2.821562
					Exp(sum)	16.80307659
					Probability	94.4%

Which variables have the greatest influence on the customers' loan behavior (combined main effects and interaction effects)?

- 'Income' and 'Education' present as the most logical variables affecting whether a loan is initiated. While both variables were positively correlated during the initial logit and probit analysis, introducing a moderating effect by multiplying them together creates negative correlation for both variables individually leaving the moderating effect as the only variable with a positive correlation. The bank should focus their loan product offerings towards customers with 'Education' > 1 AND 'Income' > 100.

Perform a sensitivity analysis as seen earlier in the semester. Copy screen snapshots of your analysis in R to your report.

```

              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 ***

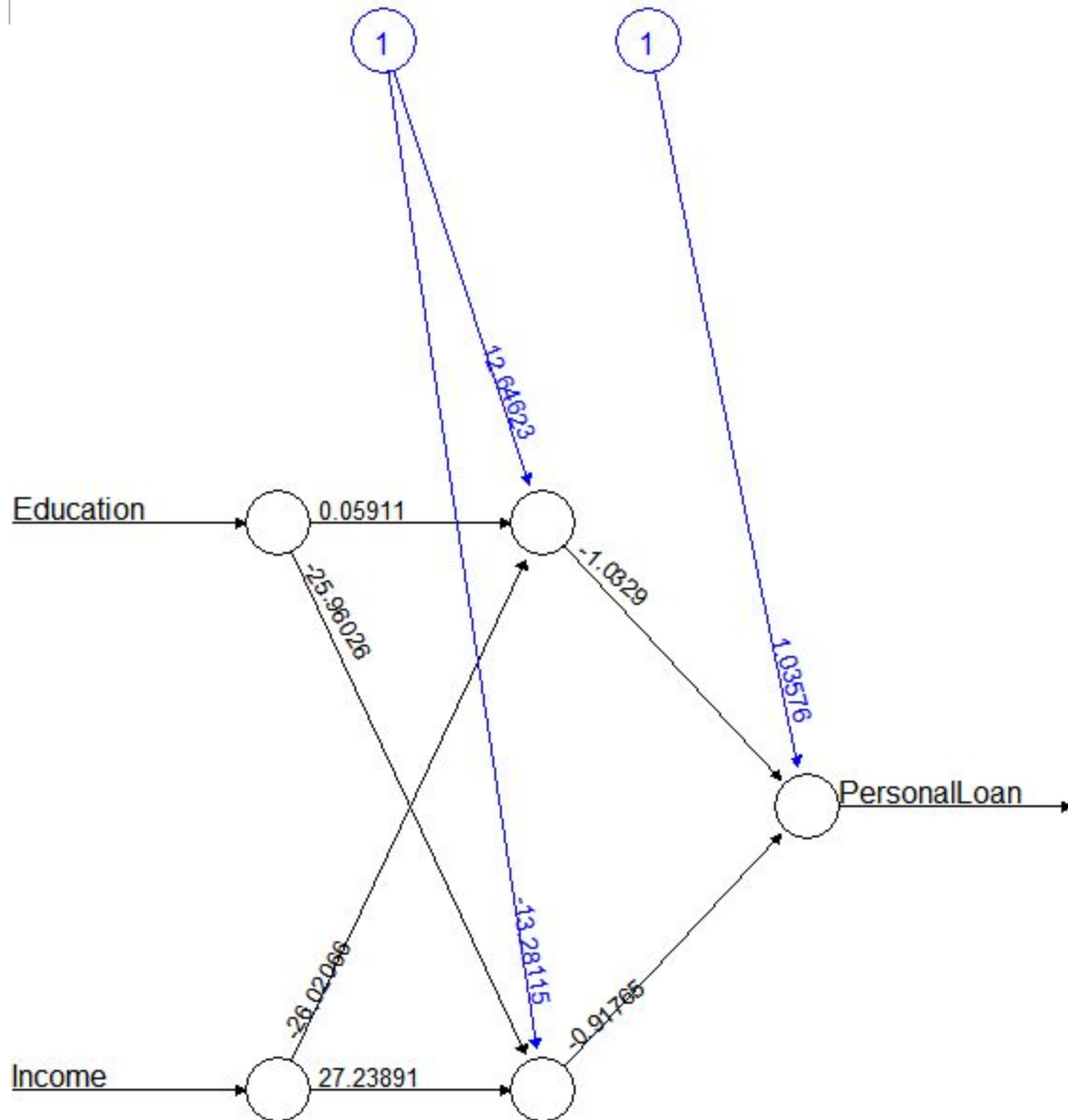
```

Sensitivity Analysis		Education		
	94%	1	2	3
Income	10	0.7%	0.0%	0.0%
	20	0.9%	0.0%	0.0%
	30	1.1%	0.0%	0.0%
	40	1.3%	0.0%	0.0%
	50	1.6%	0.1%	0.0%
	60	2.0%	0.2%	0.0%
	70	2.4%	0.6%	0.1%
	80	3.0%	1.5%	0.8%
	90	3.6%	4.0%	4.4%
	100	4.4%	10.2%	21.7%
	110	5.4%	23.5%	62.5%
	120	6.5%	45.5%	90.9%
	130	7.9%	69.4%	98.4%
	140	9.5%	86.1%	99.7%
	150	11.4%	94.4%	100.0%
	160	13.7%	97.9%	100.0%
	170	16.3%	99.2%	100.0%
	180	19.3%	99.7%	100.0%
	190	22.8%	99.9%	100.0%
	200	26.6%	100.0%	100.0%

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%)

The neural network trained originally could not predict any customers who selected a loan, so feature scaling was used on Education and Income to scale them between zero and one. The neural network resulted in two hidden layers across five repetitions at a threshold of 0.15, this model performed with an accuracy of 79.92% with 251 errors; the test set was developed from a random sample of 25% of the data. The network was trained on the significant variables found in our Logit and Probit analysis, being Education and Income; the confusion matrix is as follows:

	Predicted: 0	Predicted: 1
Actual: 0	915	215
Actual: 1	36	84



Error: 53.477126 Steps: 3962

```
library(neuralnet)
library(caTools)

# Import the dataset
fp <- redacted
dataset <- read.csv(paste(fp, 'scm651_homework_4_universal_bank.csv', sep=''))

# Apply feature scaling [0:1]
normalize <- function(x) { (x-min(x)) / (max(x)-min(x)) }
dataset$Education <- normalize(dataset$Education)
dataset$Income <- normalize(dataset$Income)

# Education, Income
dataset <- dataset[,c(2,5,9)]

# Train/ Test Split
split <- sample.split(dataset$PersonalLoan, SplitRatio = 0.75)
test_set <- subset(dataset, split == FALSE)
training_set <- subset(dataset, split == TRUE)

# Train the NN
loan_net <- neuralnet(PersonalLoan ~ Education + Income
                      , training_set
                      , rep = 5
                      , hidden = 2
                      , lifesign = "minimal"
                      , linear.output = TRUE
                      , threshold = 0.15)

# View the NN
plot(loan_net, rep = "best")

# Report results in CM and Accuracy %
loan_net.results <- compute(loan_net, test_set[, -1])
results <- data.frame(actual = test_set$PersonalLoan
                      , prediction = loan_net.results$net.result)
results$prediction <- round(results$prediction)
results$actual <- factor(results$actual, levels = c(0,1), labels = c("0","1"))
results$prediction <- factor(results$prediction, levels = c(0,1), labels =
c("0","1"))

table(results$actual, results$prediction)
cm <- table(results$actual, results$prediction)
print(paste('Accuracy: ', 100*(cm[1]+cm[4])/sum(cm), '%', sep=''))
```

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%)

Inputs			Hidden node 1:				Output:			
Variable	Value		Variable	Coefficient	Value	Coeff*Value	Variable	Coefficient	Value	Coeff*Value
Education	2	1 to 3	Intercept	12.64623	1	12.64623	Intercept	1.03576	1	1.03576
Income	150	8 to 224	Education	0.05911	0.5	0.029555	Hidden1	-1.0329	0.011769673	-0.012156896
			Income	-26.02066	0.6574074	-17.10617463	Hidden2	-0.91765	0.000235346	-0.000215965
					sum	-4.43038963				
					Exp(sum)	0.011909848				
					Probability	0.011769673				
			Hidden node 2:						sum	1.023387139
			Variable	Coefficient	Value	Coeff*Value			Exp(sum)	2.782603887
			Intercept	-13.28115	1	-13.28115			Probability	73.56%
			Education	-25.96026	0.5	-12.98013				
			Income	27.23891	0.6574074	17.9070612				
					sum	-8.354218796				
					Exp(sum)	0.000235401				
					Probability	0.000235346				

Sensitivity analysis:

		Education					
	74%	1	2	3			
Income	10	50.22%	50.12%	50.02%			
	20	50.17%	50.07%	49.96%			
	30	50.12%	50.02%	49.91%			
	40	50.08%	49.98%	49.88%			
	50	50.06%	49.97%	49.88%			
	60	50.12%	50.07%	50.03%			
	70	50.41%	50.47%	50.56%			
	80	51.24%	51.66%	52.01%			
	90	52.22%	54.27%	54.67%			
	100	51.54%	61.29%	59.30%			
	110	52.14%	71.59%	70.21%			
	120	52.55%	73.24%	73.13%			
	130	52.75%	73.43%	73.36%			
	140	52.79%	73.45%	73.36%			
	150	52.76%	73.42%	73.32%			
	160	52.71%	73.37%	73.28%			
	170	52.65%	73.29%	73.23%			
	180	52.59%	72.80%	73.18%			
	190	52.53%	67.82%	73.12%			
	200	52.46%	55.63%	73.07%			

Since the inputs to neural network - Education and Income are normalised while training the model, same normalisation is applied to the inputs for building prediction model.

From sensitivity plot, as income and education increase, the probability for person taking loan increases. For Income < 100K, the effect of education is not significant and the probability of person taking loan is lower regardless of education. But as Income increases greater than 100K, the higher education plays significant role in increasing the probability of person taking loan.