**Homework #4: Loan Analysis**

**Group #3**

Alter, Cassandra

Fjellberg, John

Parker, Victoria

Tamilmani, Tamilselvan

Wong, Steven

**Background**

Using the Universal Bank data, determine the factors which influence whether a customer takes out a loan

**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%)**

Logit with Main Effects

model.logit\_significant<-glm(formula = PersonalLoan ~ Age+Income+Family+CCAvg+Education+SecuritiesAccount+CDAccount+Online+CreditCard,data=bankData,family=binomial(logit))  
  
summary(model.logit\_significant)

##   
## Call:  
## glm(formula = PersonalLoan ~ Age + Income + Family + CCAvg +   
## Education + SecuritiesAccount + CDAccount + Online + CreditCard,   
## family = binomial(logit), data = bankData)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.1639 -0.2033 -0.0795 -0.0308 3.9351   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -13.757158 0.667122 -20.622 < 2e-16 \*\*\*  
## Age 0.010193 0.006492 1.570 0.11637   
## Income 0.054956 0.002602 21.121 < 2e-16 \*\*\*  
## Family 0.697989 0.074350 9.388 < 2e-16 \*\*\*  
## CCAvg 0.120317 0.039433 3.051 0.00228 \*\*   
## Education 1.710092 0.112884 15.149 < 2e-16 \*\*\*  
## SecuritiesAccount -0.936061 0.285341 -3.280 0.00104 \*\*   
## CDAccount 3.842066 0.323607 11.873 < 2e-16 \*\*\*  
## Online -0.672729 0.156881 -4.288 1.80e-05 \*\*\*  
## CreditCard -1.121904 0.204916 -5.475 4.38e-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: 1286.1 on 4990 degrees of freedom  
## AIC: 1306.1  
##   
## Number of Fisher Scoring iterations: 8

Logit with Only Significant Main Effects

|  |
| --- |
| model.logit\_significant<-**glm**(formula = PersonalLoan **~** Age**+**Income**+**Family**+**CCAvg**+**Education**+**SecuritiesAccount**+**CDAccount**+**Online**+**CreditCard,data=bankData,family=**binomial**(logit))  summary(model.logit\_significant)  ##  ## Call: ## glm(formula = PersonalLoan ~ Age + Income + Family + CCAvg +  ## Education + SecuritiesAccount + CDAccount + Online + CreditCard,  ## family = binomial(logit), data = bankData) ##  ## Deviance Residuals:  ## Min 1Q Median 3Q Max  ## -3.1639 -0.2033 -0.0795 -0.0308 3.9351  ##  ## Coefficients: ## Estimate Std. Error z value Pr(>|z|)  ## (Intercept) -13.757158 0.667122 -20.622 < 2e-16 \*\*\* ## Age 0.010193 0.006492 1.570 0.11637  ## Income 0.054956 0.002602 21.121 < 2e-16 \*\*\* ## Family 0.697989 0.074350 9.388 < 2e-16 \*\*\* ## CCAvg 0.120317 0.039433 3.051 0.00228 \*\*  ## Education 1.710092 0.112884 15.149 < 2e-16 \*\*\* ## SecuritiesAccount -0.936061 0.285341 -3.280 0.00104 \*\*  ## CDAccount 3.842066 0.323607 11.873 < 2e-16 \*\*\* ## Online -0.672729 0.156881 -4.288 1.80e-05 \*\*\* ## CreditCard -1.121904 0.204916 -5.475 4.38e-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: 1286.1 on 4990 degrees of freedom ## AIC: 1306.1 ##  ## Number of Fisher Scoring iterations: 8 |

The model above contains only the significant effects (and Age). This model is a lot more accurate than the first model because it is missing irrelevant variables. As you can see the AIC dropped by 5.

Probit with Main Effects

|  |
| --- |
| model.probit<-**glm**(formula = PersonalLoan **~** .,data=bankData,family=**binomial**(probit)) summary(model.probit)  ##  ## Call: ## glm(formula = PersonalLoan ~ ., family = binomial(probit), data = bankData) ##  ## Deviance Residuals:  ## Min 1Q Median 3Q Max  ## -3.2281 -0.2079 -0.0526 -0.0087 4.4415  ##  ## Coefficients: ## Estimate Std. Error z value Pr(>|z|)  ## (Intercept) -5.537e+00 2.079e+00 -2.664 0.007721 \*\*  ## CustomerID -4.033e-05 2.671e-05 -1.510 0.131090  ## Age -3.047e-02 3.133e-02 -0.973 0.330764  ## Experience 3.382e-02 3.118e-02 1.085 0.278095  ## Income 2.783e-02 1.275e-03 21.824 < 2e-16 \*\*\* ## ZIP.Code -4.777e-06 2.056e-05 -0.232 0.816295  ## Family 3.413e-01 3.757e-02 9.084 < 2e-16 \*\*\* ## CCAvg 7.300e-02 2.096e-02 3.483 0.000495 \*\*\* ## Education 8.562e-01 5.694e-02 15.036 < 2e-16 \*\*\* ## Mortgage 2.151e-04 2.953e-04 0.728 0.466367  ## SecuritiesAccount -5.035e-01 1.473e-01 -3.419 0.000629 \*\*\* ## CDAccount 2.006e+00 1.650e-01 12.157 < 2e-16 \*\*\* ## Online -3.492e-01 8.119e-02 -4.300 1.70e-05 \*\*\* ## CreditCard -5.779e-01 1.046e-01 -5.527 3.27e-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: 1300.9 on 4986 degrees of freedom ## AIC: 1328.9 ##  ## Number of Fisher Scoring iterations: 8 |

We also ran the probit analysis on the Bank Loan data. This shows that only Income, Family, CCAvg, Education, SecuritiesAccount, CDAccount, Online, and CreditCard (the same as the ones in the logit) are significant.

**Which Variables are Significant?**

In the first logit analysis, it looks like Income, Family, CCAvg, Education, SecuritiesAccount, CDAccount, Online, and CreditCard were all significant because their Pr(>|z|) scores are below 0.05. The other variables had higher p-values and should not be included in the model. So, we ran it again with only the significant effects.

**How do the significant variables influence the likelihood of taking out a loan?**

Education and CDAccount both add significant boot to the changes of getting loan. Each have a huge coefficient 1.71 and 3.84 respectively, the more the number of CD accounts we have , the higher the changes are getting loan.

On the other hand, Securities account and credit card is going to bring down the chances way more than other.

1. **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%)**

Logit ith Significant Main Effects and All Moderating Effects

|  |
| --- |
| formula\_sig <- PersonalLoan **~** Age**+**Income**+**Family**+**CCAvg**+**Education**+**SecuritiesAccount**+**CDAccount**+**Online**+**CreditCard model.logit\_significant\_2<-**glm**(formula = PersonalLoan **~** (Age**+**Income**+**Family**+**CCAvg**+**Education**+**SecuritiesAccount**+**CDAccount**+**Online**+**CreditCard)**^**2,data=bankData,family=**binomial**(logit))  ## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  **summary**(model.logit\_significant\_2)  ##  ## Call: ## glm(formula = PersonalLoan ~ (Age + Income + Family + CCAvg +  ## Education + SecuritiesAccount + CDAccount + Online + CreditCard)^2,  ## family = binomial(logit), data = bankData) ##  ## Deviance Residuals:  ## Min 1Q Median 3Q Max  ## -2.2459 -0.0760 -0.0069 -0.0001 5.1254  ##  ## Coefficients: ## Estimate Std. Error z value Pr(>|z|)  ## (Intercept) 8.343e+00 3.661e+00 2.279 0.02265 \*  ## Age -5.577e-02 6.892e-02 -0.809 0.41838  ## Income -1.682e-01 2.575e-02 -6.531 6.54e-11 \*\*\* ## Family -2.381e+00 7.340e-01 -3.244 0.00118 \*\*  ## CCAvg 1.406e+00 5.099e-01 2.758 0.00582 \*\*  ## Education -1.005e+01 1.362e+00 -7.381 1.57e-13 \*\*\* ## SecuritiesAccount 1.399e+00 4.486e+00 0.312 0.75513  ## CDAccount -3.874e+00 5.621e+00 -0.689 0.49069  ## Online 1.565e+00 1.916e+00 0.817 0.41399  ## CreditCard 1.964e+00 2.298e+00 0.855 0.39269  ## Age:Income 4.747e-04 4.257e-04 1.115 0.26485  ## Age:Family -9.874e-03 9.804e-03 -1.007 0.31385  ## Age:CCAvg -6.171e-03 6.889e-03 -0.896 0.37031  ## Age:Education 2.671e-02 1.299e-02 2.056 0.03978 \*  ## Age:SecuritiesAccount -1.026e-01 5.988e-02 -1.713 0.08673 .  ## Age:CDAccount 1.105e-01 5.589e-02 1.977 0.04804 \*  ## Age:Online -1.097e-02 2.504e-02 -0.438 0.66126  ## Age:CreditCard -1.518e-02 3.016e-02 -0.503 0.61469  ## Income:Family 4.629e-02 5.449e-03 8.495 < 2e-16 \*\*\* ## Income:CCAvg -1.390e-02 2.209e-03 -6.292 3.13e-10 \*\*\* ## Income:Education 1.041e-01 9.981e-03 10.432 < 2e-16 \*\*\* ## Income:SecuritiesAccount 3.105e-02 2.757e-02 1.127 0.25994  ## Income:CDAccount 6.296e-04 2.060e-02 0.031 0.97562  ## Income:Online 6.684e-03 9.157e-03 0.730 0.46540  ## Income:CreditCard -4.191e-03 1.231e-02 -0.340 0.73353  ## Family:CCAvg 1.954e-01 8.496e-02 2.299 0.02149 \*  ## Family:Education -8.339e-01 1.501e-01 -5.555 2.77e-08 \*\*\* ## Family:SecuritiesAccount -1.322e+00 6.906e-01 -1.914 0.05563 .  ## Family:CDAccount 1.008e+00 6.294e-01 1.602 0.10907  ## Family:Online -3.697e-01 2.780e-01 -1.330 0.18363  ## Family:CreditCard -2.158e-01 3.398e-01 -0.635 0.52544  ## CCAvg:Education 4.960e-01 1.076e-01 4.612 4.00e-06 \*\*\* ## CCAvg:SecuritiesAccount 3.806e-02 3.457e-01 0.110 0.91235  ## CCAvg:CDAccount -2.572e-01 3.335e-01 -0.771 0.44052  ## CCAvg:Online -2.268e-01 1.570e-01 -1.444 0.14869  ## CCAvg:CreditCard 1.825e-01 2.084e-01 0.876 0.38107  ## Education:SecuritiesAccount 3.517e-01 7.815e-01 0.450 0.65275  ## Education:CDAccount 3.169e-01 7.810e-01 0.406 0.68492  ## Education:Online -1.516e-01 3.550e-01 -0.427 0.66939  ## Education:CreditCard -5.282e-01 4.449e-01 -1.187 0.23508  ## SecuritiesAccount:CDAccount 4.830e+00 3.492e+00 1.383 0.16664  ## SecuritiesAccount:Online -3.575e+00 2.413e+00 -1.482 0.13836  ## SecuritiesAccount:CreditCard -2.437e+00 2.260e+00 -1.078 0.28088  ## CDAccount:Online 2.256e+00 2.774e+00 0.813 0.41605  ## CDAccount:CreditCard 5.049e-01 2.686e+00 0.188 0.85091  ## Online:CreditCard -4.139e+00 1.643e+00 -2.519 0.01176 \*  ## --- ## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1 ##  ## (Dispersion parameter for binomial family taken to be 1) ##  ## Null deviance: 3162.04 on 4999 degrees of freedom ## Residual deviance: 527.21 on 4954 degrees of freedom ## AIC: 619.21 ##  ## Number of Fisher Scoring iterations: 10 |

This analysis shows the significance of variables interacting. It looks like two very significant interactions are Family:Education and CCAvg:Education (with extremely low p-values). Additionally, the AIC dropped substantially (from 1306.1 to 619.21), indicating that this model is much more effective with the moderating effects. Also, the previously significant variable SecuritiesAccount is not significant any more

**Which interactions are statistically significant?**

Income interacting with Family, CCAvg and Education are significant

Education Interacting with CCAvg and Family are significant

And there are somewhat significant interactions like Age:CDAccount, Family:CCAvg, Online:Creditcard

**How do you interpret the coefficients on these variables?**

It is similar to other variables but the coefficients get a big boost due to the multiplication of the primary variables. For example, the 0.1 coefficient of Income and Education is going to be pronounced if we consider the income is in range of 8-224 and multiplied by 2 or 3 for education

1. **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%)**

|  |
| --- |
| interaction\_vars<- **c**("Income","Family","CCAvg","Education","Income:Family","Income:CCAvg","Income:Education","Family:CCAvg","Family:Education","Family:SecuritiesAccount","Family:CDAccount","CCAvg:Education","Online:CreditCard") interaction\_vars\_vif<-**c**(interaction\_vars,"Age:Education","Age:SecuritiesAccount","Age:CDAccount")  formula\_int<-**paste**("PersonalLoan", **paste**(interaction\_vars\_vif, collapse=" + "), sep=" ~ ") model.logit\_significant\_interaction<-**glm**(formula =formula\_int,data=bankData,family=**binomial**(logit))  ## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  **summary**(model.logit\_significant\_interaction)  ##  ## Call: ## glm(formula = formula\_int, family = binomial(logit), data = bankData) ##  ## Deviance Residuals:  ## Min 1Q Median 3Q Max  ## -2.2222 -0.0913 -0.0102 -0.0002 5.0347  ##  ## Coefficients: ## Estimate Std. Error z value Pr(>|z|)  ## (Intercept) 6.5985943 1.9349670 3.410 0.000649 \*\*\* ## Income -0.1425678 0.0170050 -8.384 < 2e-16 \*\*\* ## Family -2.8728071 0.6080344 -4.725 2.30e-06 \*\*\* ## CCAvg 0.8841448 0.3781676 2.338 0.019389 \*  ## Education -8.6969967 1.0895529 -7.982 1.44e-15 \*\*\* ## Income:Family 0.0444700 0.0052230 8.514 < 2e-16 \*\*\* ## Income:CCAvg -0.0123248 0.0019621 -6.282 3.35e-10 \*\*\* ## Income:Education 0.1022398 0.0092372 11.068 < 2e-16 \*\*\* ## Family:CCAvg 0.1831904 0.0801107 2.287 0.022213 \*  ## Family:Education -0.8567226 0.1425634 -6.009 1.86e-09 \*\*\* ## Family:SecuritiesAccount -0.4760582 0.4604861 -1.034 0.301222  ## Family:CDAccount 0.8021589 0.4315022 1.859 0.063028 .  ## CCAvg:Education 0.4776256 0.0976015 4.894 9.90e-07 \*\*\* ## Online:CreditCard -4.0474169 0.7391907 -5.475 4.36e-08 \*\*\* ## Education:Age -0.0006435 0.0050418 -0.128 0.898443  ## SecuritiesAccount:Age -0.0192758 0.0232358 -0.830 0.406781  ## CDAccount:Age 0.0831058 0.0229520 3.621 0.000294 \*\*\* ## --- ## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1 ##  ## (Dispersion parameter for binomial family taken to be 1) ##  ## Null deviance: 3162.04 on 4999 degrees of freedom ## Residual deviance: 554.98 on 4983 degrees of freedom ## AIC: 588.98 ##  ## Number of Fisher Scoring iterations: 10 |

Now we removed the insignificant variables from the model and ran it again. It looks like Income, Family, CCAvg, Education, Income:Family, Income:CCAvg, Income:Education, Family:CCAvg, Family:Education, Family:SecuritiesAccount, Family:CDAccount, CCAvg:Education, Online:CreditCard, Education:Age, SecuritiesAccount:Age, and CDAccount:Age are good indicators of whether someone has accepted a personal loan. Additionally, you can see the AIC is even lower now at 588.98!

**Which variables have the greatest influence on the customers’ loan behavior**

Family, Education, Online+Creditcard has the highest coefficient which will mainly drive the final outcome

Logit Sensitivity Analysis

Graphical user interface, application, table, Excel

Description automatically generated

Logit Sensitivity Analysis Graph

Probit Sensitivity Analysis

Graphical user interface, application, table, Excel

Description automatically generated

Probit Sensitivity Analysis Graph

1. **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%)**

|  |
| --- |
| library(neuralnet)  formula\_nn\_sig<-PersonalLoan ~ Age+Income+Family+CCAvg+Education+CDAccount+Online+CreditCard  model\_nn <- neuralnet(formula\_nn\_sig,bankData,hidden = 3,lifesign = "minimal", linear.output = FALSE,threshold = 0.1)  summary(model\_nn)  plot(model\_nn) |

![Diagram

Description automatically generated]()

1. **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%)**

Neural Net model

**Table

Description automatically generated**

Sensitivity Analysis of Neural Net model

Table

Description automatically generated

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

The neural net is much closer to the original data compared to login and probit models

Chart, line chart

Description automatically generated

Actual Data ( CCAvg over Personal Loan segregated by Education)

![Chart, line chart

Description automatically generated]()