Universal Bank Loan (Team 2)

(Homework Assignment 4)

Ahmed Sobi, Douglas Reed, Jason Tompkins,
Joe Shaxted, Michael Farrell, Timothy Hulak

SCM 651 (Business Analytics)

03/16/2021

Background and Introduction

Universal Bank tasked our analytics team to develop a model that can predict whether a customer will take out a loan with the bank. Using the provided dataset containing 5000 observations of 14 variables, the team will use multiple analysis tools, such as regression analysis and neural networks, to develop the predictive model, showing the sensitivity of each significant variable.

Logit & Probit Analysis

As part of the analyses, the team used both logit and probit models to perform initial regression to determine possible significant variables. The team removed any insignificant variables from the model and reran the regression until all remaining variables remained significant. The final regressions provided estimates for each variable that can be used to develop the predictive model.

Logit Analysis

For the Logit Analysis, the team excluded the *Zip Code* variable because it is categorical and sparse, and not relevant without further research into each zip code region and/or more occurrences. After the initial run of logit regression with all variables considered, *Mortgage* returned a P-value of 0.39190. Removing *Mortgage* and rerunning the model, *Age* had a P-Value of 0.398185. Removing *Age* and rerunning the model, *Experience* had a P-Value of 0.09458 and was removed due to its low P-value. In the final regression model, the P-Values of the remaining variables were less than 5%, which indicates that they were significant and retained.

Initial Logit

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

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)      -1.219e+01  1.645e+00  -7.411 1.25e-13 ***
Income            5.458e-02  2.620e-03  20.831  < 2e-16 ***
Family            6.958e-01  7.430e-02   9.364  < 2e-16 ***
CCAvg             1.240e-01  3.965e-02   3.127  0.00177 **
Education         1.736e+00  1.151e-01  15.088  < 2e-16 ***
SecuritiesAccount -9.368e-01  2.859e-01  -3.277  0.00105 **
CDAccount         3.823e+00  3.239e-01  11.800  < 2e-16 ***
Online           -6.752e-01  1.571e-01  -4.298 1.72e-05 ***
CreditCard       -1.120e+00  2.050e-01  -5.462 4.70e-08 ***
Mortgage          4.745e-04  5.541e-04   0.856  0.39190
Experience        6.376e-02  6.093e-02   1.046  0.29536
Age              -5.361e-02  6.131e-02  -0.874  0.38191
---
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
```

Second Logit

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

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 ***
Income              0.054721   0.002589  21.133  < 2e-16 ***
Family              0.690388   0.074201   9.304  < 2e-16 ***
CCAvg               0.113713   0.039265   2.896  0.00378 **
Education           1.704116   0.112393  15.162  < 2e-16 ***
SecuritiesAccount  -0.934627   0.284849  -3.281  0.00103 **
CDAccount           3.853311   0.323447  11.913  < 2e-16 ***
Online             -0.667476   0.156717  -4.259 2.05e-05 ***
CreditCard         -1.123683   0.205003  -5.481 4.22e-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: 1288.6  on 4991  degrees of freedom
AIC: 1306.6

Number of Fisher Scoring iterations: 8
```

Once the less significant variables were removed from the Logit analysis, 8 significant variables remained that contribute meaningful data to the prediction of taking out a loan. The coefficient values were then used in the prediction formula.

[Y = I + (V1*X1) + … + (Vn*Xn)]

Using *Personal Loan* as the Y-variable in the logit regression model, the remaining eight significant variables provided values that were used to develop a predictive model. Initial analysis returned a negative intercept, which indicated that the positive factors multiplied by their corresponding variables, such as income, must be large enough to offset the initial deficit to contribute to a significant probability that the customer will apply for a personal loan. The individual variables will affect the final prediction as follows:

Positively affects model: *Income*, *Family*, *CCAvg*, *Education*, *CDAcct*

Negatively affects model: *SecuritiesAccount*, *Online Banking*, and *CreditCard*.

Note: While it appears that Income does not impact the decision as much as some of the other variables, the variable is in the thousands which affects the model significantly.

Probit Analysis

As with the Logit analysis, for the Probit Analysis, the team excluded *Zip Code* for the same reasons previously explained above. The initial probit regression calculated the *Mortgage* variable's P-value as 0.452395. Removing *Mortgage* and rerunning the model, *Age* had a P-Value of 0.345023 and was removed. Once *Age* was removed, *Experience* had a P-Value of 0.26798 and was also removed. This left the same eight variables as significant for the final regression (P-Value below0.05): *Income*, *Family*, *CCAvg*, *Education*, *SecuritiesAccount*, *CDAccount*, *Online*, and *CreditCard*.

Initial Probit

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

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 ***
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 ***
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 ***
Mortgage           0.0002217  0.0002950   0.751 0.452395
Experience         0.0337833  0.0311288   1.085 0.277800
Age               -0.0303628  0.0312820  -0.971 0.331740
---
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
```

Second Probit

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

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 ***
Income             0.027891   0.001258  22.173  < 2e-16 ***
Family             0.340529   0.037509   9.079  < 2e-16 ***
CCAvg              0.070770   0.020779   3.406 0.000659 ***
Education          0.837564   0.055464  15.101  < 2e-16 ***
SecuritiesAccount -0.499103   0.146829  -3.399 0.000676 ***
CDAccount          2.018424   0.164391  12.278  < 2e-16 ***
Online            -0.350131   0.080986  -4.323 1.54e-05 ***
CreditCard        -0.583261   0.104525  -5.580 2.40e-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: 1305.8  on 4991  degrees of freedom
AIC: 1323.8

Number of Fisher Scoring iterations: 8
```

Overall, the eight variables appeared more significant than in logit regression (lower P-Values). Additionally, both the Intercept and estimates for the individual variables were smaller, so may affect the final model differently. The individual variables appear to have the same positive or negative effects as in the logit regression. *SecuritiesAccount*, *Online Banking*, and *CreditCard* each had a negative impact on the probability of taking out a personal loan.

Looking at the individual variables further, a few of the variables had very large ranges. For instance, Income had a range from 8 ($8,000) to 224 ($224,000). Because Probit works best with data that is normally distributed, and the provided distribution of incomes does not fit that model, we chose to use Logit to determine the viability of moderating effects and conduct our final predictive analysis.

Moderating Effect

The team added a new independent variable as a moderating effect to explore and determine which variables provide better results when run together. Looking for two independent variables that have commonalities, Family and Education each have small integers in their range. *SecuritiesAccount*, *CDAccount*, *Online Banking*, and *CreditCard* all have binary factors (YES or NO represented by 0 or 1). We chose the last two available variables: *Income* and *CCAvg*.

Both *Income* and *CCAvg* are significant, both use the same scale and data type ($1,000s), and both have a larger range of values than any other variable. Further, many believe that the interaction between these two variables is intuitive. Because the ratio of an applicant's income to their average spending on credit cards per month are indirectly correlated; it can safely be assumed that an individual with a larger salary would be paying more towards their credit card balances. This debt-to-income ratio is used by the lending industry as a factor in determining loan approval.

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3527  -0.1430  -0.0336  -0.0061   3.8983

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -19.932775   0.947248 -21.043  < 2e-16 ***
Income              0.099037   0.005273  18.780  < 2e-16 ***
Family              0.778544   0.080172   9.711  < 2e-16 ***
CCAvg               2.207269   0.179929  12.267  < 2e-16 ***
Education           1.865177   0.118277  15.770  < 2e-16 ***
SecuritiesAccount  -0.838806   0.306575  -2.736 0.006218 **
CDAccount           3.743548   0.347224  10.781  < 2e-16 ***
Online             -0.628464   0.166421  -3.776 0.000159 ***
CreditCard         -1.123793   0.213054  -5.275 1.33e-07 ***
Income:CCAvg       -0.014050   0.001178 -11.924  < 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: 1105.1  on 4990  degrees of freedom
AIC: 1125.1

Number of Fisher Scoring iterations: 8
```

Running the logit regression model, the new ‘*Income:CCAvg*’ variable produced a significant impact on the Y-variable (P-value virtually zero). Both *Income* and *CreditCard* Average negatively impacted the chances of taking out a loan. With the moderating effect, the estimated impact for each of the two separate variables is reduced (*Income* from 0.054 to 0.099 for example). Further, the combined moderating variable negatively impacted the total chances of obtaining a loan. The remaining variables also changed slightly.

<u>Final Regression Model</u>

Determining if the moderating effect is significant, the team used a logit regression with the moderating effect in the final model as shown below.

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3527  -0.1430  -0.0336  -0.0061   3.8983

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -19.932775   0.947248 -21.043  < 2e-16 ***
Income              0.099037   0.005273  18.780  < 2e-16 ***
Family              0.778544   0.080172   9.711  < 2e-16 ***
CCAvg               2.207269   0.179929  12.267  < 2e-16 ***
Education           1.865177   0.118277  15.770  < 2e-16 ***
SecuritiesAccount  -0.838806   0.306575  -2.736 0.006218 **
CDAccount           3.743548   0.347224  10.781  < 2e-16 ***
Online             -0.628464   0.166421  -3.776 0.000159 ***
CreditCard         -1.123793   0.213054  -5.275 1.33e-07 ***
Income:CCAvg       -0.014050   0.001178 -11.924  < 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: 1105.1  on 4990  degrees of freedom
AIC: 1125.1

Number of Fisher Scoring iterations: 8
```

To determine the sample value for each variable, the mean, median, or mode of each variable was chosen, depending on the variable type and range. The specific values chosen for each variable are outlined below in the Logit Mod Effects (Interactions) table.

Logit Mod Effects (Interactions) - Personal Loan

Inputs

Variable	Value		
Income	73	MEAN	income of customer
Family	2	MEDIAN	family size of customer
CCAvg	1.5	MEDIAN	avg monthly CC spend
Education	2	MEDIAN	three categories
SecuritiesAccount	0	MODE	No/Yes (0,1)
CDAccount	0	MODE	No/Yes (0,1)
Online	1	MODE	No/Yes (0,1)
CreditCard	0	MODE	No/Yes (0,1)
Income:CCAvg	1		Interaction

Output: Personal Loan 0 Did Not Take, 1 Did Take

Variable	Coefficient	Value	Coeff*Value
intercept	-19.932775	1	-19.932775
Income	0.099037	73	7.229701
Family	0.778544	2	1.557088
CCAvg	2.207269	1.5	3.3109035
Education	1.865177	2	3.730354
SecuritiesAccount	-0.838806	0	0
CDAccount	3.743548	0	0
Online	-0.628464	1	-0.628464
CreditCard	-1.123793	0	0
Income:CCAvg	-0.01405	1	-0.01405
		sum	-4.7331925
		Exp(sum)	0.008798337
		Probability	1%

These values provided inputs for the predictive model above (Output: Personal Loan) table. Multiplying these values by the coefficients for each variable in the regression model and summing the results, the exponential of the sum provides the predicted probability of someone with the same factors (values for each variable) taking out a loan. Using the two variables chosen for moderating effects, we can conduct a sensitivity analysis for further study, shown below.

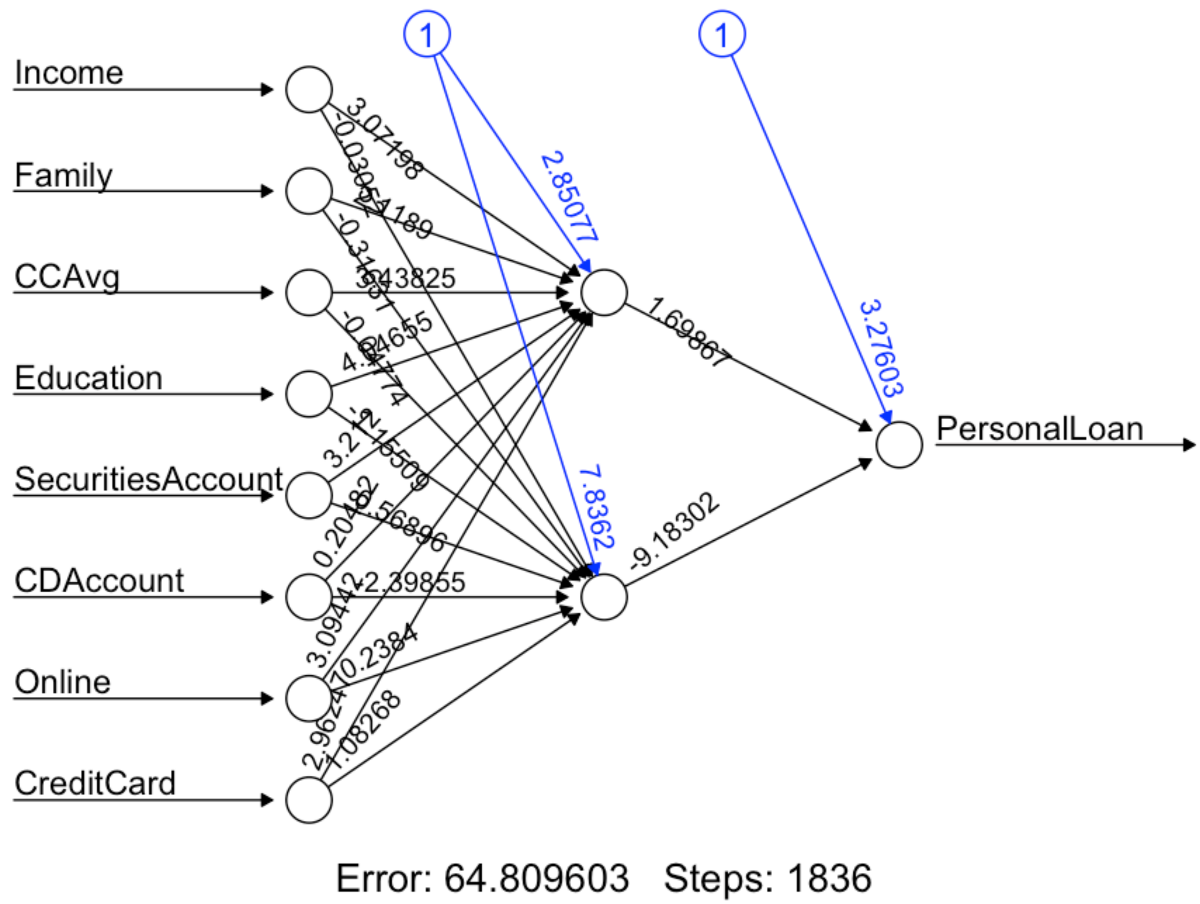
Sensitivity Analysis

CCAvg

		0	1	2	3	4	5	6	7	8	9	10
	1%											
Income	10	0%	0%	0%	0%	0%	4%	26%	76%	97%	100%	100%
	20	0%	0%	0%	0%	1%	9%	48%	90%	99%	100%	100%
	40	0%	0%	0%	1%	8%	43%	87%	98%	100%	100%	100%
	60	0%	0%	1%	6%	37%	84%	98%	100%	100%	100%	100%
	80	0%	1%	5%	32%	81%	98%	100%	100%	100%	100%	100%
	100	0%	4%	27%	78%	97%	100%	100%	100%	100%	100%	100%
	120	3%	23%	73%	96%	100%	100%	100%	100%	100%	100%	100%
	140	19%	69%	95%	99%	100%	100%	100%	100%	100%	100%	100%
	160	64%	94%	99%	100%	100%	100%	100%	100%	100%	100%	100%
	180	93%	99%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	200	99%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	220	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

The sensitivity analysis above shows how different values for the *Income* and *CCAvg* variables affect the probability of a customer applying for a personal loan. The higher the income the more likely the client is to apply for the loan. Additionally, the higher the average amount spent on credit cards per month, the higher the probability of a customer applying for a personal loan. Our interpretation of this data is that a customer with high credit card debt and a high income can use the personal loan to consolidate their credit card debt and can pay off the loan. Conversely, a customer with low credit card debt and low income, neither needs a personal loan nor has the capacity to afford it.

Neural Network



	actual	prediction
3501	0	0
3502	0	0
3503	0	0
3504	0	0
3505	0	0
3506	0	0
3507	0	0
3508	0	0
3509	1	1
3510	0	0
3511	0	0
3512	0	0
3513	0	0
3514	0	0
3515	0	0
3516	0	0
3517	0	0
3518	1	0
3519	0	0
3520	0	0

For the neural network test/train data sets, the team did a 70/30 split. This was due to the complete dataset being only 5000 rows (a relatively small dataset). A comparison between the actual *PersonalLoan* output from the testing data and a prediction using the model above shows the accuracy of the model. The neural network model correctly predicted 19 out of the 20 observations. Observation 3518 shows an inaccurate prediction, which is expected given that there will be some error.

Prediction Model: Neural Network

Neural Network

Inputs

Variable	Value	Descriptoion
Income	73.77	
Family	1	
CCAvg	1.938	
Education	1	
SecuritiesAccount	0	
CDAccount	0	
Online	1	
CreditCard	0	

Hidden node 1:

Variable	Coefficient	Value	Coeff*Value
Intercept	2.85076584	1	2.85076584
Income	3.07198068	73.77	226.6200148
Family	2.11889906	1	2.11889906
CCAvg	3.43825269	1.938	6.663333713
Education	4.04655349	1	4.04655349
SecuritiesAccount	3.21200303	0	0
CDAccount	0.20482073	0	0
Online	3.09441875	1	3.09441875
CreditCard	2.96246523	0	0
		sum	245.3939856
		Exp(sum)	3.7433E+106
		Probability	100.00%

Hidden node 2:

Variable	Coefficient	Value	Coeff*Value
Intercept	7.83619793	1	7.83619793
Income	-0.03052189	73.77	-2.251599825
Family	-0.31630889	1	-0.31630889
CCAvg	-0.04773753	1.938	-0.092515333
Education	-1.15509016	1	-1.15509016
SecuritiesAccount	0.56896011	0	0
CDAccount	-2.39854944	0	0
Online	0.23840352	1	0.23840352
CreditCard	1.08268436	0	0
		sum	4.259087242
		Exp(sum)	70.74538053
		Probability	98.61%

Output:

Variable	Coefficient	Value	Coeff*Value
Intercept	3.27602966	1	3.27602966
Hidden1	1.69866551	1	1.69866551
Hidden2	-9.18302492	0.98606182	-9.055030269
		sum	-4.080335099
		Exp(sum)	1.6902E-02
		Probability	1.66%

Sensitivity Analysis

		CCAvg										
	2%	0	1	2	3	4	5	6	7	8	9	10
Income	0	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%
	10	1%	1%	1%	1%	1%	1%	1%	1%	2%	2%	2%
	20	1%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%
	30	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%
	40	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%
	50	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%
	60	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%
	70	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%
	80	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%
	90	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%
	100	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%
	110	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%
	120	2%	2%	2%	2%	3%	3%	3%	3%	3%	3%	3%
	130	3%	3%	3%	3%	3%	3%	3%	3%	3%	4%	4%
	140	3%	3%	3%	4%	4%	4%	4%	4%	4%	5%	5%
	150	4%	4%	5%	5%	5%	5%	6%	6%	6%	7%	7%
	160	6%	6%	6%	7%	7%	8%	8%	9%	9%	10%	10%
	170	8%	9%	9%	10%	11%	11%	12%	13%	14%	15%	16%
	180	13%	14%	15%	16%	17%	18%	20%	21%	23%	25%	26%
	190	20%	22%	24%	25%	27%	29%	31%	34%	36%	39%	41%
	200	32%	35%	37%	40%	42%	45%	47%	50%	53%	56%	58%
	210	49%	51%	54%	57%	59%	62%	64%	67%	69%	72%	74%
	220	65%	68%	70%	72%	75%	76%	78%	#NUM!	#NUM!	#NUM!	#NUM!
	230	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!

The input values for the neural network model were chosen based on the unique values within each variable. The mean, instead of the median, was taken from the *Income* and *CCAvg* variables because of the extreme range of the *Income* variable. The mode was taken from the remaining *Family, Education, SecuritiesAccount, CDAccount, Online,* and *CreditCard* because there were a limited number of unique values (1, 2, 3, and 4 for *Family* and 0 or 1, representing a false or true, for *Education, SecuritiesAccount, CDAccount, Online,* and *CreditCard).* The mode was chosen in an effort to represent the majority of the possible attributes for a loan application and appeal to as many applicants as possible.

Summary

Upon completion of the analysis, we found that the neural network had an interesting relationship between the error and the hidden node coefficients. The team ran the neural network several times, to generate a lower error. As the error shrunk, the hidden node coefficients also decreased drastically (as low as -500 coefficient for an error or 22). This resulted in an infinitesimally small probability. The team determined that an error between 50 - 65 would be an exceptional error so that a probability could be calculated, and the sensitivity table could be populated.

The *Education* and *CDAccount* variables did have a drastic effect on the models. When those values were increased, and the model took those changes into account, the model predicted a much higher likelihood of taking out a loan for higher-income earners.