**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:**

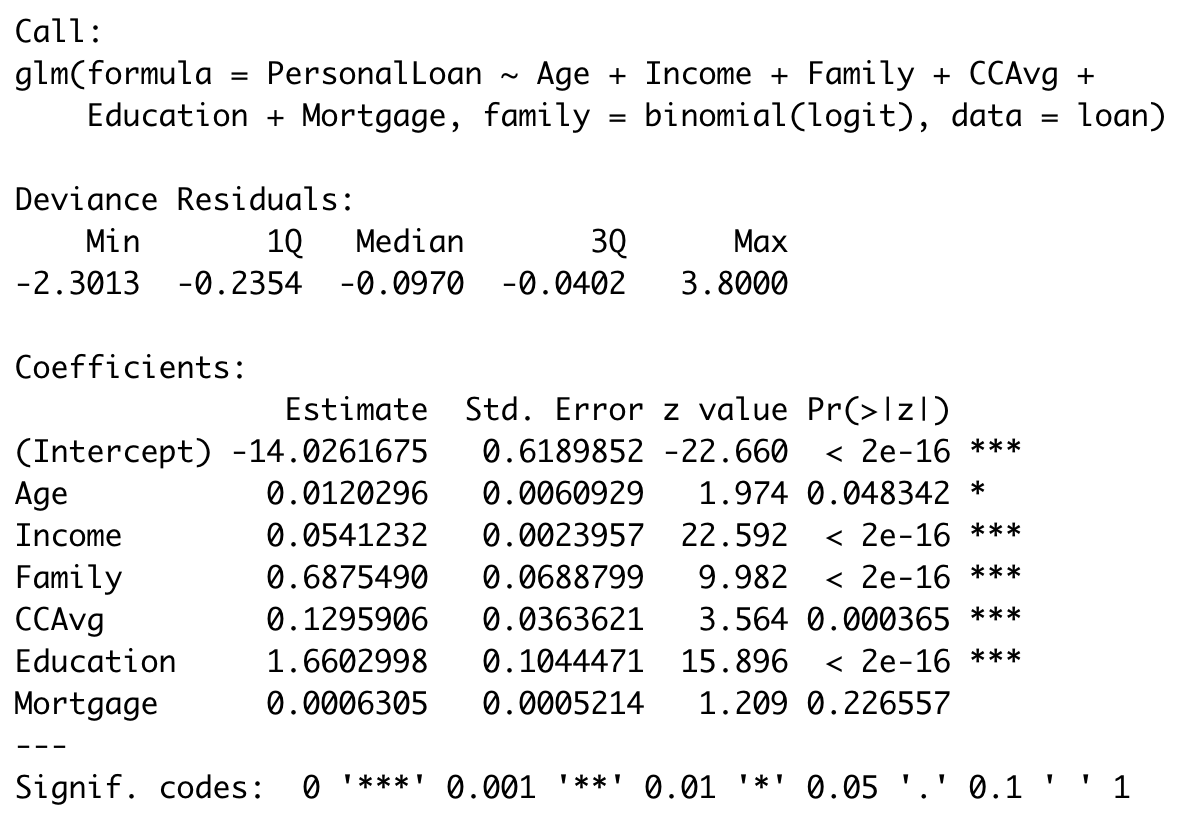
1. Perform a logit and probit analysis of the variables that affect whether a customer takes out a loan. Consider only main effects. Which variables are significant? How do the significant variables influence the likelihood of taking out a loan? Copy screen snapshots of your analysis in R to your report.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

LOGIT ANALYSIS

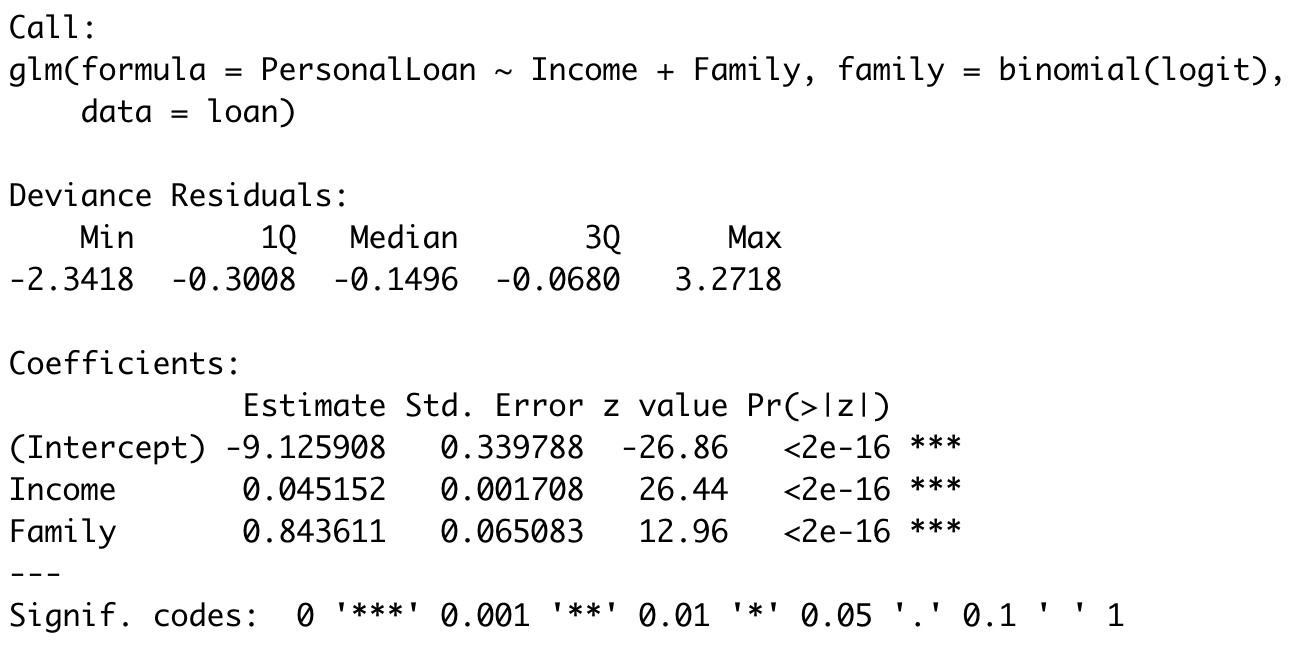
The first step in the analysis was to perform a logit and probit analysis for the variables with continuous or integer values (Age, Income, Family, CCAvg, Education, Mortgage). Binary data such as CDAccount was not included due to it being less granular. Experience was not included due to collinearity concerns with Age. Education was also eliminated because, like experience, there were concerns with collinearity. In addition, CCAvg was also eliminated from further analysis because it was determined that, while statistically significant, the Income and Family attributes had lower P-values and therefore were better options.

A logit and probit analysis will typically produce similar results, which has been illustrated with the data set. The logit analysis relies on logistic distribution and is better for detecting differences at extreme values while probit relies on normal distribution and is more sensitive to values near the mean. Below are the results of the analysis performed using the statistical package R.

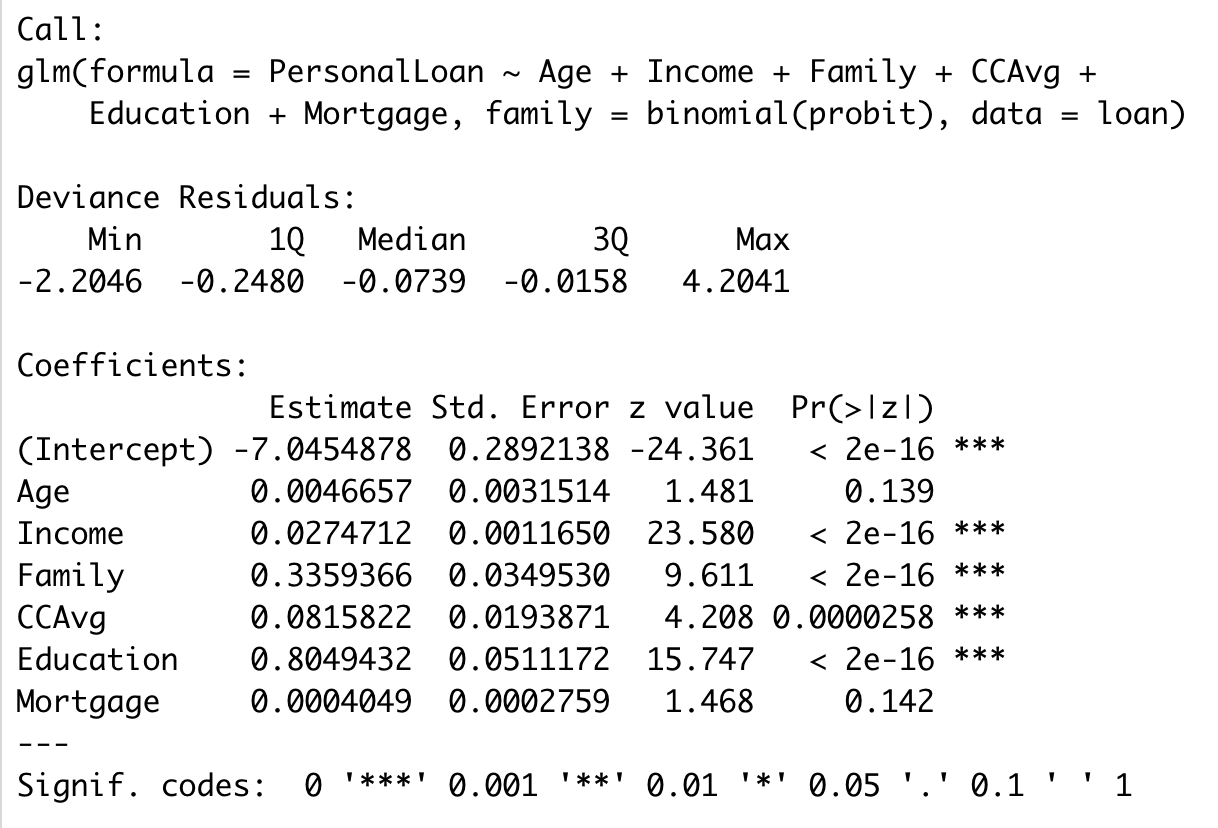
The logit analysis calculated low P-values for Income, Family, CCAvg and Education which indicates the variables are statistically significant and have a bearing on loan acquisition. Their coefficients are all positive, indicating that as each of their values increase, the likelihood of the customer securing a loan also increases. This is an intuitive result, as factors such as high income would increase loan acquisition.



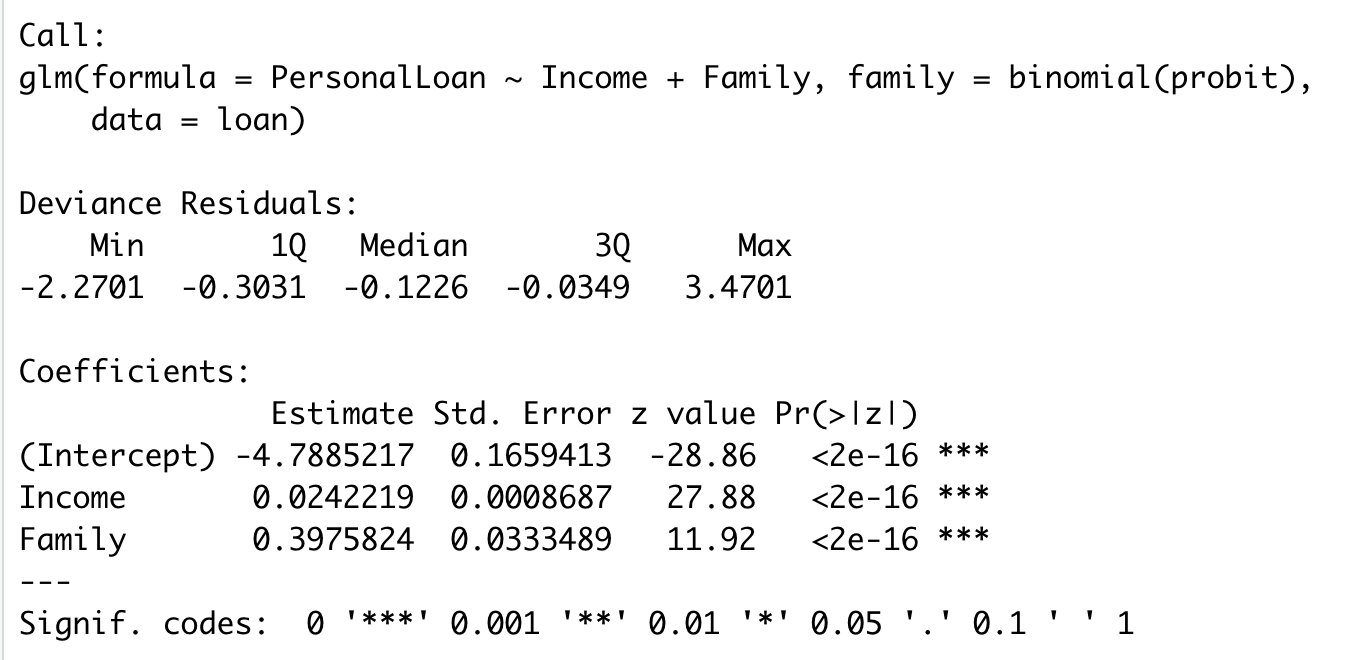
The logit analysis was then run again using only the Income and Family attributes. As indicated in the R output below, as Income or Family size increases so does the likelihood of loan acquisition.



The probit analysis also calculated low P-values for Income, Family, CCAvg and Education.

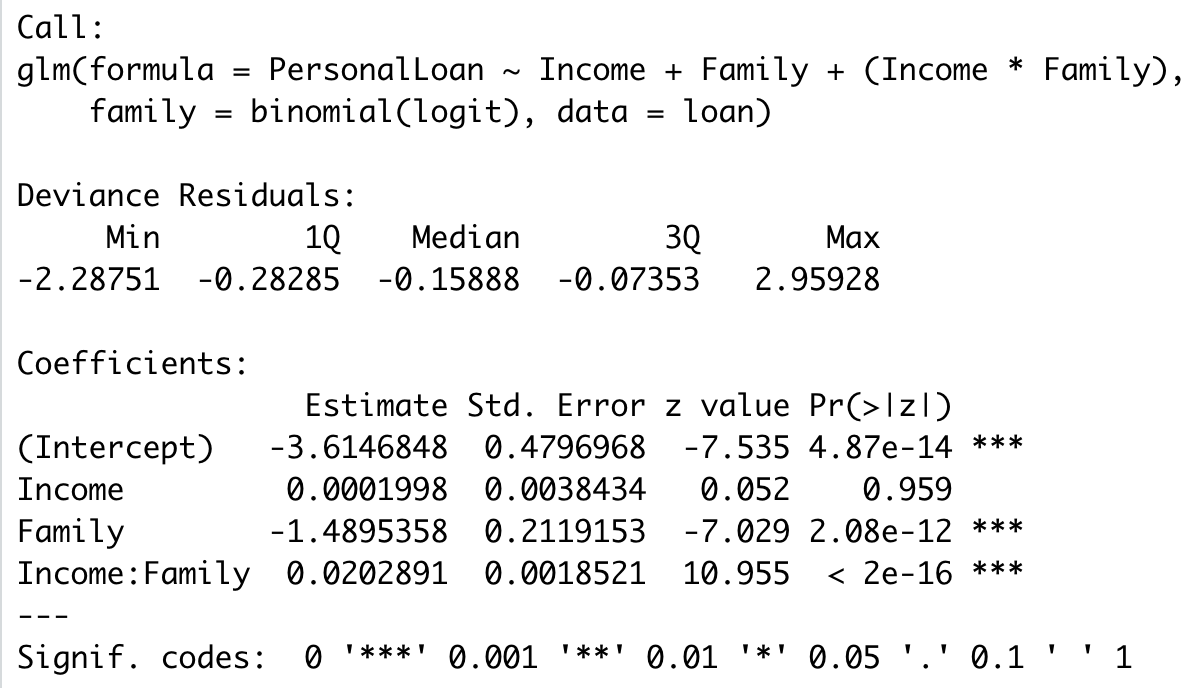


The probit analysis was run again only using the Income and Family attributes which were both statistically significant.



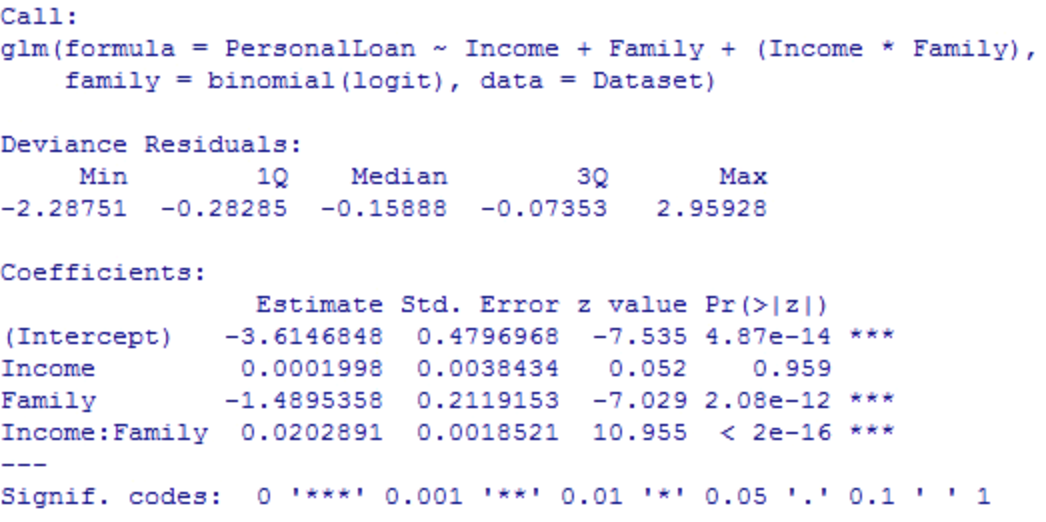
1. Add moderating effects (interactions of variables). Which interactions make sense conceptually? Which interactions are statistically significant? How do you interpret the coefficients on these variables? Copy screen snapshots of your analysis in R to your report. (20%)

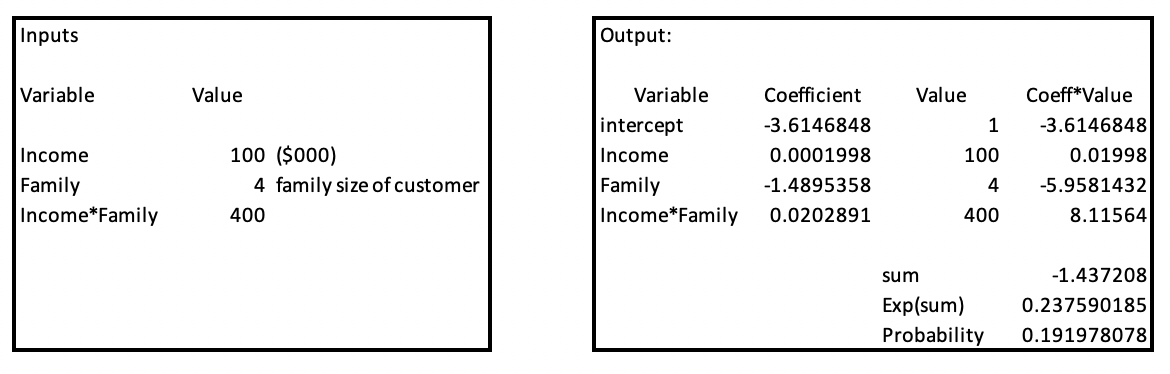
After the significant variables have been identified, the data set can be reviewed again for variables that could have a moderating effect on the key variables. A moderating effect allows the slope of the line to change allowing the primary variables to have a greater impact on the overall analysis. Moderating effects can help improve the accuracy of the overall analysis. The key variables of the analysis are Income and Family with a moderating effect of Family \* Income. The analysis shows that by using this combination of variables and moderating effects, Income becomes insignificant when Family \* Income is added (indicated by its very low coefficient). The negative coefficient for Family indicates that as its value goes up, loan acquisition goes down. However, the moderating effect must also be taken into account. With a positive value 100 times greater than Income, it is both significant and overcomes the low Income coefficient. Exactly how this affects the data is difficult to predict using coefficients alone, but becomes more apparent when a What-If Analysis is run (Q3).



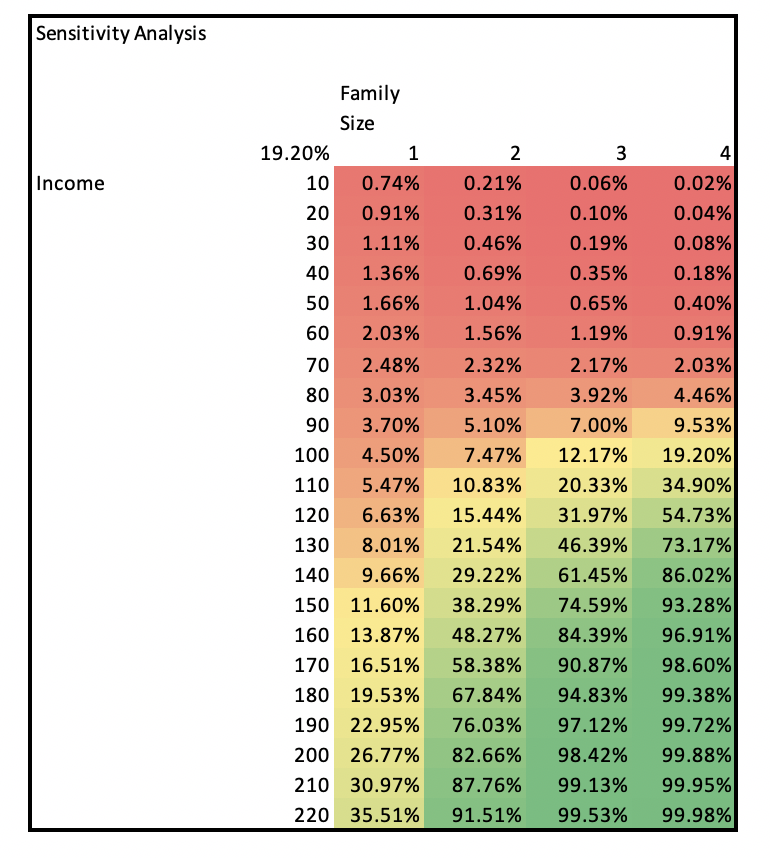
1. Create a final regression model with the variables that you feel are important (both main effects and interaction terms). Create a spreadsheet prediction of the model. Which variables have the greatest influence on the customers’ loan behavior (combined main effects and interaction effects)? Perform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

Income and Family were chosen as main effects, as they had the lowest p-values calculated in both the logit and probit models in question 1. Income\*Family was used as the moderating effect, as it was determined to have statistical significance in question 2.





Although the coefficient for Family is negative and the coefficient for Income is very small, the interaction of the two yields a coefficient that is both positive and much larger than either of the two main variables. This interaction isn’t obvious when using the spreadsheet prediction alone but becomes apparent when a What-If Analysis is used along with conditional formatting.

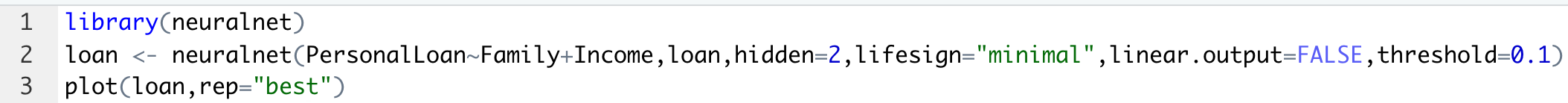


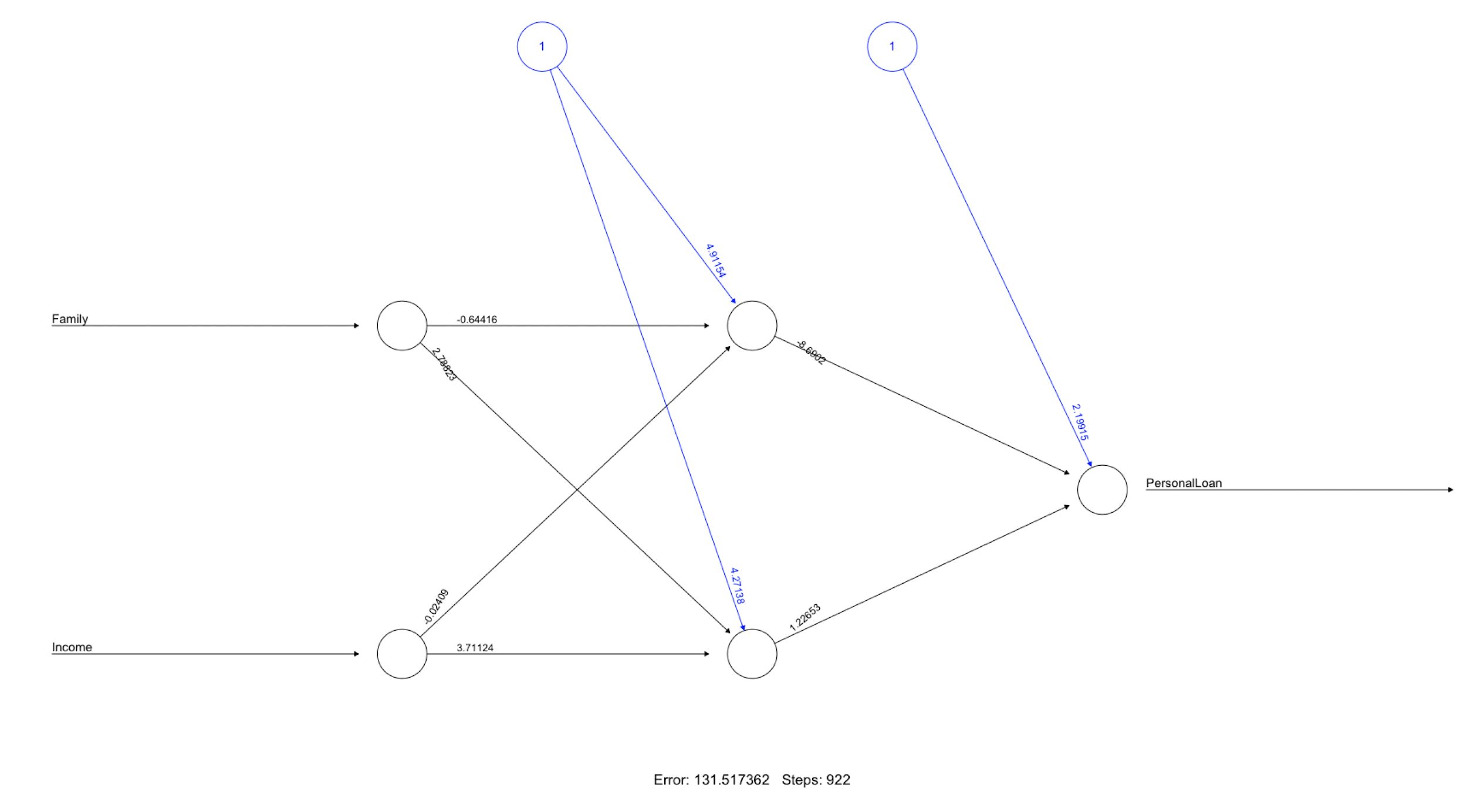
When examining the data table, a few patterns can be seen. First, among all family sizes the probability of loan acquisition goes up as income goes up. However, as family size increases, loan acquisition decreases if income is $80,000 or less. This may be because, from a business perspective, banks risk customers with larger families defaulting on their loan as they have more bills/expenses and are therefore less likely to make paying the loan back a priority.

At or above $80,000 income, loan acquisition increases as family size increases. For example, a family with an income of $220,000 and 4 family members is more likely to secure a loan than a family of 2 with the same income. This may be because a family with 4 members has more opportunities for tax deductions and for dependents to contribute to paying expenses and are therefore more likely to make paying the loan back a priority.

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%)

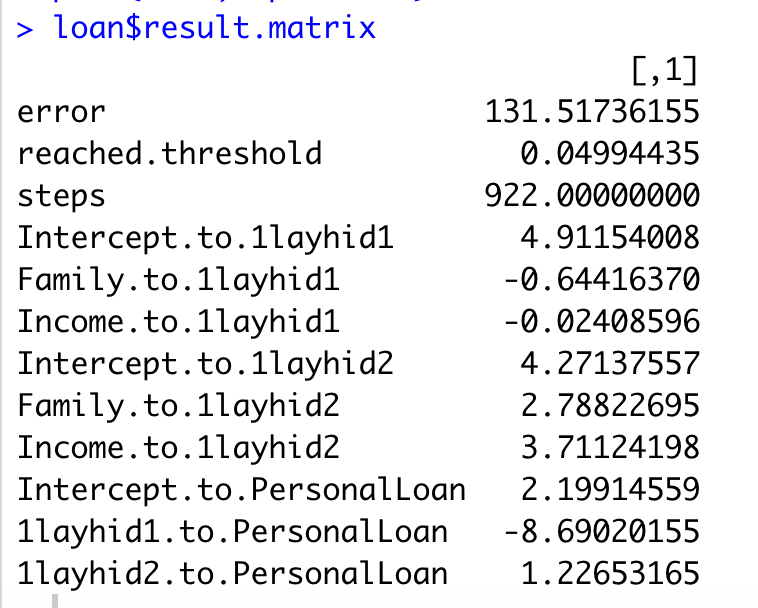
Neural networks allow for information to be processed similar to the way the human brain processes. It allows for the analysis of binary and nonbinary information and represents nonlinear behavior. The neural network creates the connections included the hidden ones between the different variables presented. The neural network below indicates the use of Income and Family as the input variables (X) and PersonalLoan as the output variable (Y). The number of hidden nodes is determined by the user, in this case 2. A larger number of nodes corresponds with an increased chance of finding a pattern, but it also increases run time and it becomes more complicated to interpret results.





One observation when comparing the neural net results for Family/Income with the logit/probit analysis is the difference in the coefficients. The logit/probit had negative intercepts (-9.125908 and -4.7885217 respectively) where the neural net had a positive intercept of 2.19914559 from the hidden layers. This difference is likely due to the more sophisticated analysis in the neural net algorithm and the differences in results will be explored in Step 5.

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%)



A screenshot of a cell phone

Description automatically generated

As mentioned in Step 4, the coefficients are significantly different in the neural net predictions compared to the logic/probit analysis. The predictions were also different where the neural net predicted a family of 4 with a $100k income to have a 32% probability of taking out a loan. This same family of 4 and income of $100k would only have a 19% probability in the logit analysis with moderating effects done on Step 3. In theory, the neural network is more sophisticated and should provide better results but to test this an analysis should be done on training and testing data to determine the actual accuracy of the different methods.

A screenshot of a cell phone

Description automatically generated

The final part of Step 5 was to create a sensitivity analysis based on the neural network analysis. As shown in the chart above, there is a much more gradual change in the values compared to the earlier sensitivity done in Step 3. Again, this is likely due to the neural net’s ability to provide more complex calculations to provide better predictions compared to the logit/probit techniques. One anomaly that did occur with the neural net was an error with any incomes greater than $187,000. The process was repeated several times by different team members, and we were unable to explain the reason for the error. It is possible the model predicted greater than a 100% probability which may not be supported in the sensitivity analysis in Microsoft Excel.

One final observation of the sensitivity analysis is the increase from red to yellow to green as both the income and family size increases. This makes logical sense for the income but not for the family size. As mentioned above, perhaps this is due to prioritization of home loans for large families or some other unexplained cause. We also did not see the values going up and down as was demonstrated in the Titanic data set demonstration. The neural net has this capability compared to logic/probit analysis, but we did not see this phenomenon in the bank data set.