**HW4\_Team 2**

**Loan Analysis**

Katherine Hurtado-Da Silva

Sahil Nanavaty

Victoria Shu

Gregory J Horth

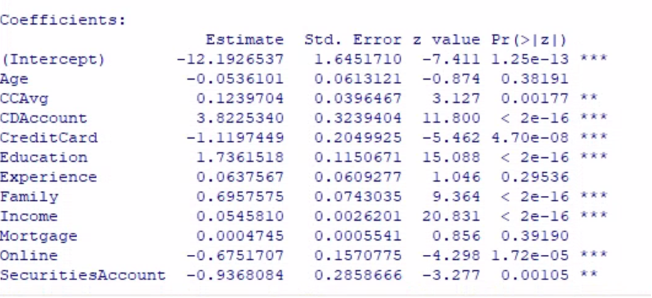
Michael McLean

**Universal Bank - 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?

The Logit Regression below includes all variables except customerID and ZIP.code. No further assumptions regarding the “main effect” variables had on probability of success were made. Given the initial results below, it was evident the p-values of Age, Experience, and Mortgage exceeded the assumed alpha level of .05, meaning these variables were not statistically significant and would be omitted. Therefore, the variables that have a “main effect” and were initially considered for the analyses that follow were CCAvg, CDAccount, CreditCard, Education, Family, Income, Online, and SecuritiesAccount.



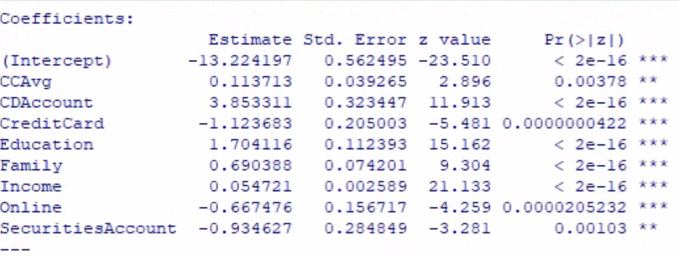
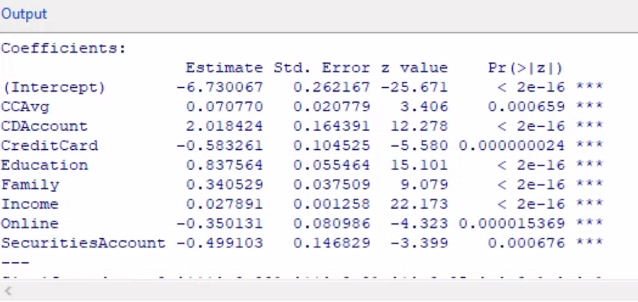
Logit and Probit Analysis

When considering the weights of the coefficients, having a CD account with the bank contributed the most in favor of a customer taking a loan. On the other hand, having a credit card with the bank had the most adverse effect on predicting a customer taking out a loan. The y-intercept being negative means a customer is expected to not take a loan when all of these variables are 0. These main takeaways make sense due to the following:

* Having a CD account means they have assets invested that they cannot access until a fixed period of time, therefore increasing the probability of them taking a loan
* Having a credit card is a form of a loan in itself and may be enough for an applicant, therefore reducing the probability of them needing to take out a loan
* A negative y-intercept insinuating a 0% probability of taking out a loan is expected since it means the customer does not exist (family size=0).

Since Probit is based on the normal distribution, the coefficients differed from the Logit results. However, the trends explained for each of the predictors remained the same. All variables had p-values below .05 and were therefore statistically significant. CDAccount and CreditCard also remained as the greatest contributing factors to the predicted success, for and against it, respectively

**Logit Analysis** **Probit Analysis**

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

After removing the non-significant variables, the predictors below remained statistically significant with p-values below the alpha level of .05. Each coefficient’s influence on the probability of a customer taking out a personal bank loan is summarized below:

· CCAvg: As the monthly average spending on credit cards increases, the likelihood of taking out a loan also increases.

· CDAccount: If a customer has a certificate of deposit with the bank, the likelihood of taking out a loan increases.

· CreditCard: If a customer has a credit card with the bank, the likelihood of taking out a loan decreases.

· Education: Higher levels of educational achievement increase the likelihood of taking out a loan.

· Family: The larger a family size is, the higher the likelihood is of the customer taking out a loan.

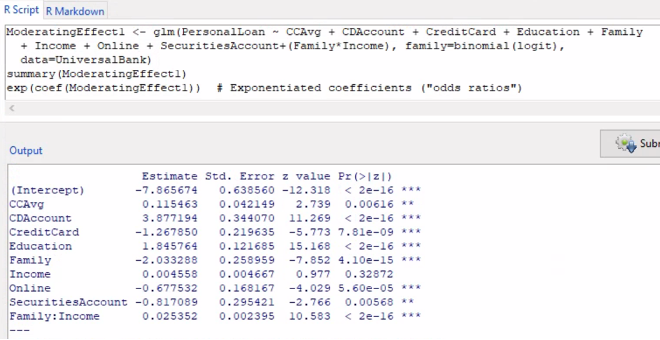
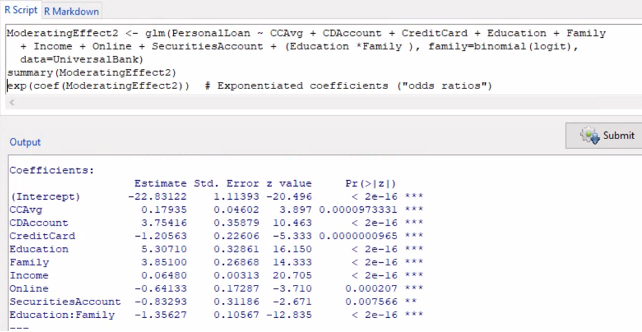
· Income: The higher the annual income of a customer, the higher the likelihood of taking out a loan.

· Online: If a customer uses the internet banking facilities, the likelihood of taking out a loan decreases.

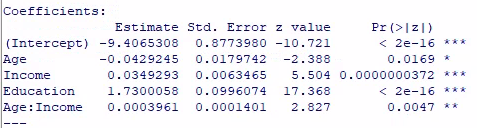
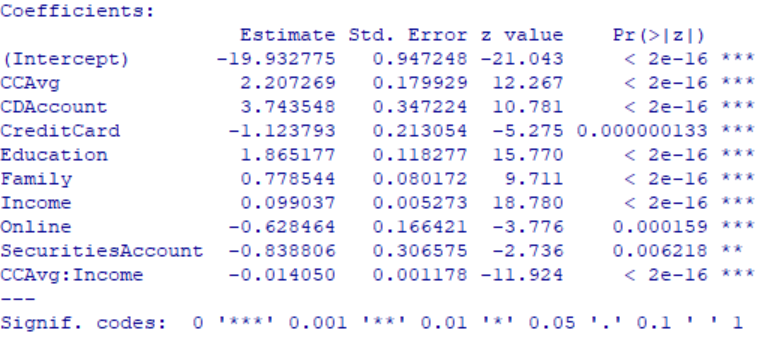
· SecuritiesAccount: If a customer has a securities account with the bank, the likelihood of taking out a loan decreases.

2. Add moderating effects (interactions of variables).

**Interaction #1 Family and Income** **Interaction #2 Family and Education**

**Interaction #3 CCAvg and Income** **Interaction #4 Age and Income**



Which interactions make sense conceptually? How do you interpret the coefficients on these variables? Which interactions are statistically significant?

Most personal loans are used for home renovations, consolidating debt, or other major life changes. All these events tend to be associated with the working family demographic. Working families have family size and income as interrelating predictors and were explored to see how the interaction of these factors catalyzed a decision to take out a loan. When both family size and income levels were chosen for Interaction #1, the results showed its coefficient to be positive, meaning it increased the chances of success (taking the loan). Its associated p-value was below .05, indicating it was statistically significant. However, income was no longer significant as it was in the original analysis. Therefore, this moderating effect was not explored further.

Since family and income did not yield a reliable model, Interaction #2 used education in place of income. Higher levels of education often correlate to higher income levels. When both education level and family size were chosen as moderating effects, its coefficient was negative and had an associated p-value below .05, making it statistically significant. Furthermore, all predictors remained statistically significant with p-values below .05. Therefore, the context of the moderating effect was explored further:

* Main Effects: Like in our preliminary analysis, the variables of CCAvg, CDAccount, CreditCard, Education, Family, Income, Online, and SecuritiesAccount retained the same direction of influence over a customer taking out a loan.
* Education:Family: The moderating effect decreased the likelihood of taking out a loan the higher the educational level and the larger the family size. This makes conceptual sense since it is implied that customers with advanced/professional degrees have higher levels of income, therefore reducing the need to take out a personal loan. On the other hand, customers with undergraduate degrees will have a higher likelihood of taking the loan, presumably due to the lower income levels associated with not having a graduate/advanced degree.

Interaction #3 highlights a meaningful interaction between the variables of CCAvg and Income. This makes sense conceptually, as one would expect an individual to spend more money on average as his or her income increases. On the contrary, an individual with a low income and high spending has a higher probability of taking out a loan. The p-values from the logit outputs indicated that Interactions #2 and #3 were statistically significant based on an alpha value of .05.

Interaction #4 was also statistically significant with the caveat that we are forgoing the fact that Age was initially observed to be a non-significant variable in our preliminary logit analysis (this moderating effect was selected for further exploration due to the interaction making intuitive business sense). It was interesting to observe how an unreliable predictor could become statistically significant, given a specific selection of variables. This highlighted how business intuition resulted in a statistically significant interaction that could be useful for a company.

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.

The final regression model includes the following variables and the rationale for the decision:

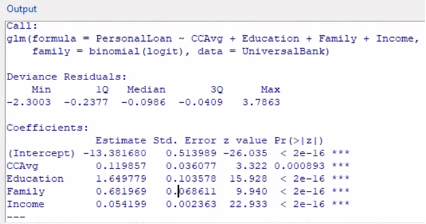
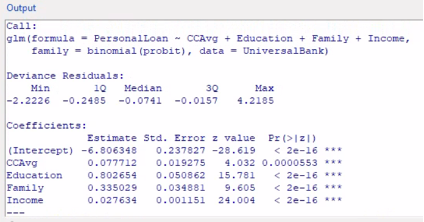
* CCAvg: offers insight on the household average expenses that can be linked to household size and potential debt size. This was also statistically significant in the moderating effects analysis.
* Education: higher levels of education correlate to higher earning potential, therefore reducing the need of taking out a loan. This was also statistically significant in the moderating effects analysis.
* Family: family size is important in properly scaling expenses and income a customer has. This was also statistically significant in the moderating effects analysis.
* Income: needed to determine debt to income ratio, which could also reveal the need for a personal loan. This was also statistically significant in the moderating effects analysis.
* Family\*Education: the interaction of family size and educational level was statistically significant in the moderating effects analysis and diminishes the likelihood of taking out a loan that family and Education provided as individual variables.

The following are the variables that were omitted and the rationale for the decision:

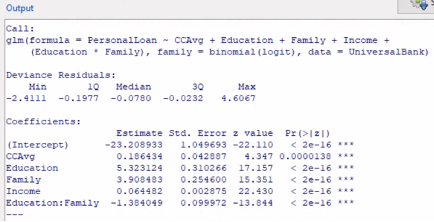
* CDAccount: declining rates on certificates of deposits questions the relevancy of this variable.
* CreditCard: this only shares whether a customer has a credit card from the bank. This variable would have been valuable if it shared debt to available credit data instead.
* Online: accessing internet banking services does not offer context on a customer’s banking history, much less if it makes them more likely to take out a loan.
* SecuritiesAccount: this was originally included in the final regression model due to current popularity in stock investments and the assumption that having enough liquid assets to invest reduces the likelihood of taking out a personal loan. However, the results showed it was the least statistically significant variable, so it was omitted.

1. Final Regression Models

**Logit**   **Probit**

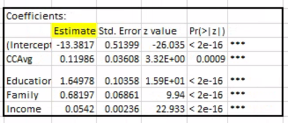
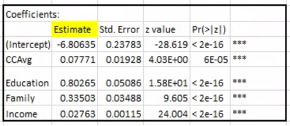
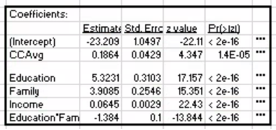
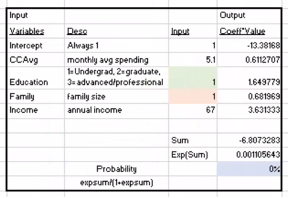
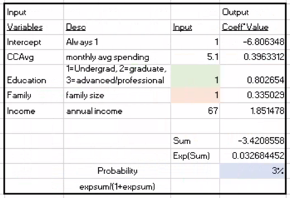
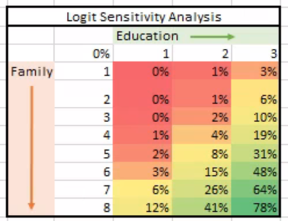
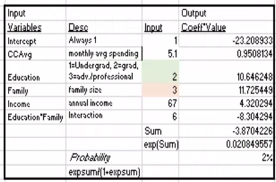
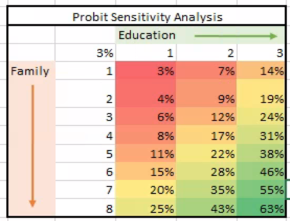
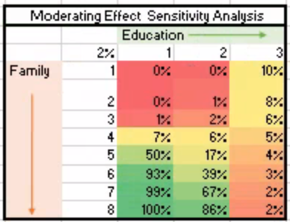
 

**Moderating Effect**



1. Prediction Models & Sensitivity Analyses

**Logit Probit Moderating Effect**

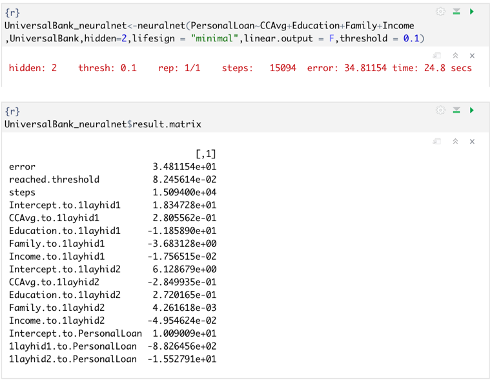
       . 

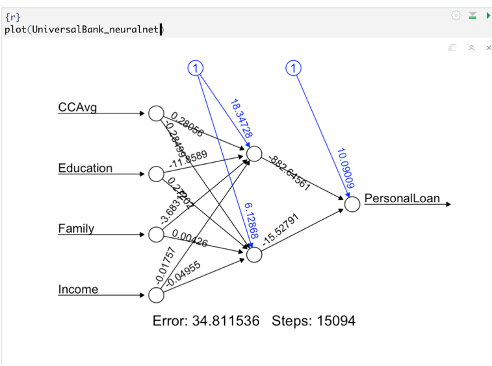
1. Variable Influence in Final Regression Models

Since the focus was on the probability of working families taking out personal loans, the U.S Bureau of Labor Statistics and U.S. Census Bureau were referred to when determining which values to use for the CCAvg and Income variables. The sources reported the average monthly expenses for American households are $5,111, and that the Median household income was $67,521 in 2020. Therefore, all models had an input of 5.1 for CCAvg and 67 for Income. The sensitivity analysis was used to explore the influence education level and family size had on the likelihood of taking out a loan given a CCAvg and Income that is indicative of a common working-class family.

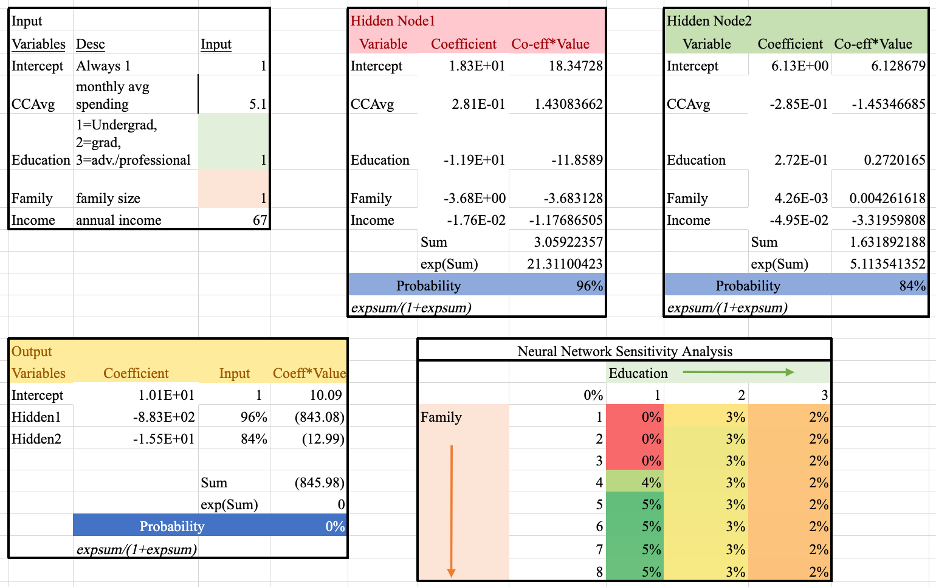
All variables in all models continued to have p-values below .05, making them statistically significant. In both the Logit and Probit models, the probability of taking out a loan increased as family size grew and educational levels increased. Though the probabilities vary due to the different coefficients, they both show a small window of having a 50% chance or greater of taking a loan and it is determined by educational level. Given a typical American family, applicants with advanced degrees had the highest probabilities of taking out a loan due to an exceptionally large family size. Conversely, a customer who is an undergraduate has the lowest probability of taking out a loan, even if they have the same large family size. This does not make conceptual sense since the assumed lower income associated with not having degrees is linked to a higher need to take out a loan.

4. Perform a neural network analysis of the variables found to be significant in the Logit and Probit analysis above.





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.



Compared to the previous Logit and Probit analyses, the neural network’s sensitivity analysis significantly scaled back the influence of education and family on the probability of a customer taking out a loan. This indicated CCAvg and/or Income may have a higher contribution to the decision made by a customer.

The results of this sensitivity analysis further indicated that a higher level of education combined with a large family size may not be an optimal interaction of predicting the probability of taking out a loan. This output suggested that the Family variable had a greater influence, since the probability of success remained 0% until the family size was at least 4 or above. In addition, an individual with a higher level of education was the 2nd least likely to take out a loan regardless of family size. Since the analysis was based on the same U.S. median income and average monthly spending, this insinuates the customer may have a higher level of financial literacy. Thus, a highly educated individual will avoid making common financial errors and, in turn, be less probable to take out a loan.