**Customer choice:**

**Logit, probit, and neural network analysis of loan acquisition behavior**

**Team 2:**

**BB, RS, NK, Jennifer Mead**

Contents

[Background 1](#_Toc509206780)

[**1. Perform a logit and probit analysis of the variables that affect whether a customer takes out a loan.** 2](#_Toc509206781)

[**2. Add moderating effects (interactions of variables).** 5](#_Toc509206782)

[**3. Create a final regression model with the variables that you feel are important (both main effects and interaction terms).** 7](#_Toc509206783)

[**4. Perform a neural network analysis of the variables found to be significant in the logit and probit analysis above.** 9](#_Toc509206784)

[**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.** 10](#_Toc509206785)

# Background

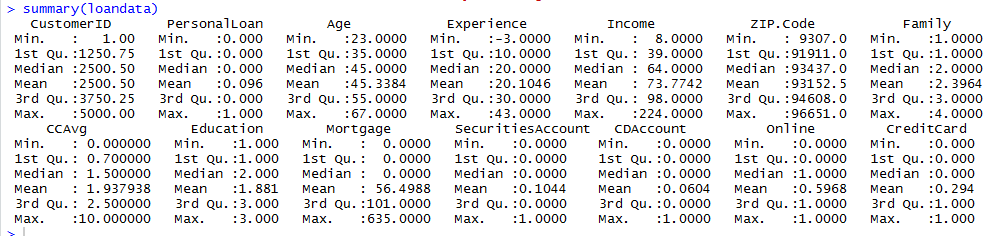
Using the Universal Bank data, determine the factors which influence whether a customer takes out a loan. Compare logit, probit, and neural network models on this data.

## **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.

Before we started the analysis, we wanted to have basic understanding of our data. Below is the result from a summary function, showing basic statistics on our variables.



Universal Bank Data Fields

* **ID** unique identifier
* **Personal Loan** did the customer accept the personal loan offered (1=Yes, 0=No)
* **Age** customer’s age
* **Experience** number of years of profession experience
* **Income** annual income of the customer ($000)
* **Zip code** home address zip code
* **Family** family size of customer
* **CCAvg** average spending on credit cards per month ($000)
* **Education** education level (1) undergraduate, (2) graduate, (3) advanced/professional
* **Mortgage** value of house mortgage ($000)
* **Securities** does the customer have a securities account with the bank? (1=Yes, 0=No)
* **CDAccount** does the customer have a certificate of deposit with the bank? (1=Yes, 0=No)
* **Online** does the customer use Internet banking facilities (1=Yes, 0=No)
* **CreditCard** does the customer use a credit card issued by Universal Bank? (1=Yes, 0=No)

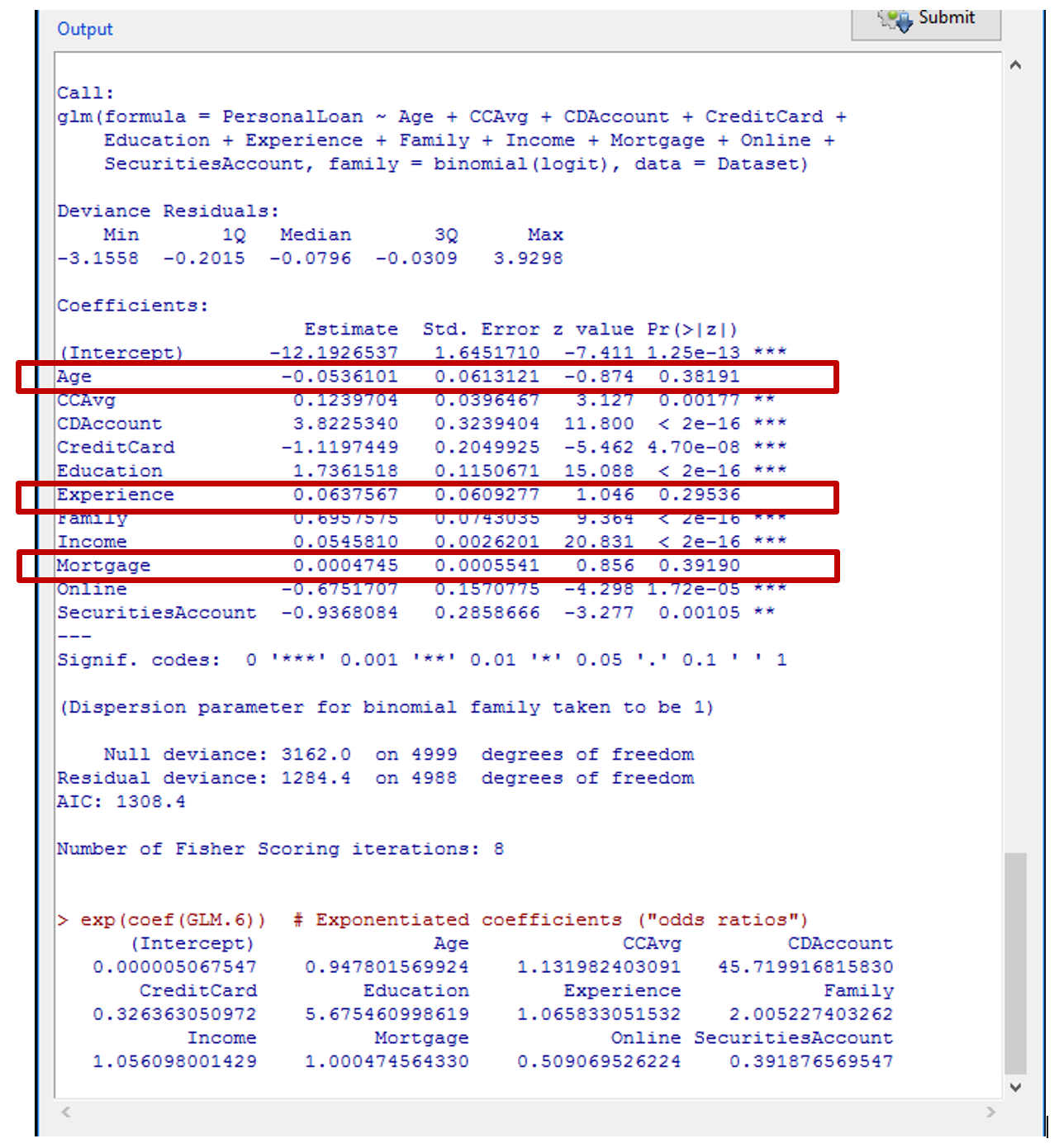
**Logit Analysis**

All variables except Customer ID and Zip Code have been included in the model selection process. Customer ID is a variable which doesn’t include any useful information for deciding if a customer will take a loan and is just a sequential number. Zip codes were mapped by the team to “regions” with the highest percentage of customers who took the loan compared to others and were found to be statistically significant. While the effort was made to transform the variable and test the model it was out of scope of this homework assignment and was excluded from the analysis.

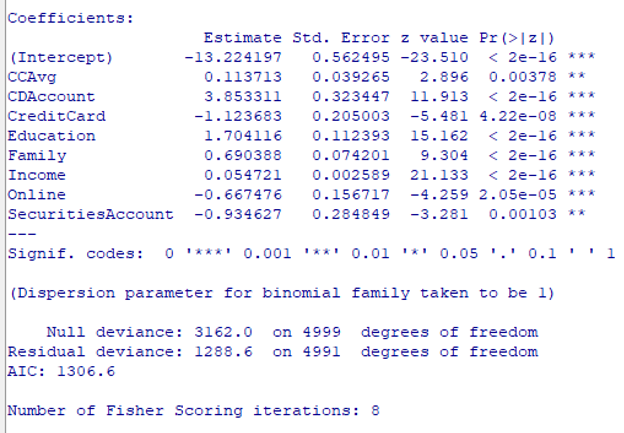
**Logit including all variables**

The screenshot below shows an output from R (Rmcdr library).

This preliminary Logit analysis resulted in Age, Experience and Mortgage independent variables to be statistically insignificant with a p-value above 0.05 as shown below.



Once statistically insignificant variables have been excluded, a resulted Logit model was the following with Akaike Information Criterion of 1306.6. The team decided to use AIC to evaluate all models so that we had a fair metric to compare them; Rcmdr results include AIC automatically.

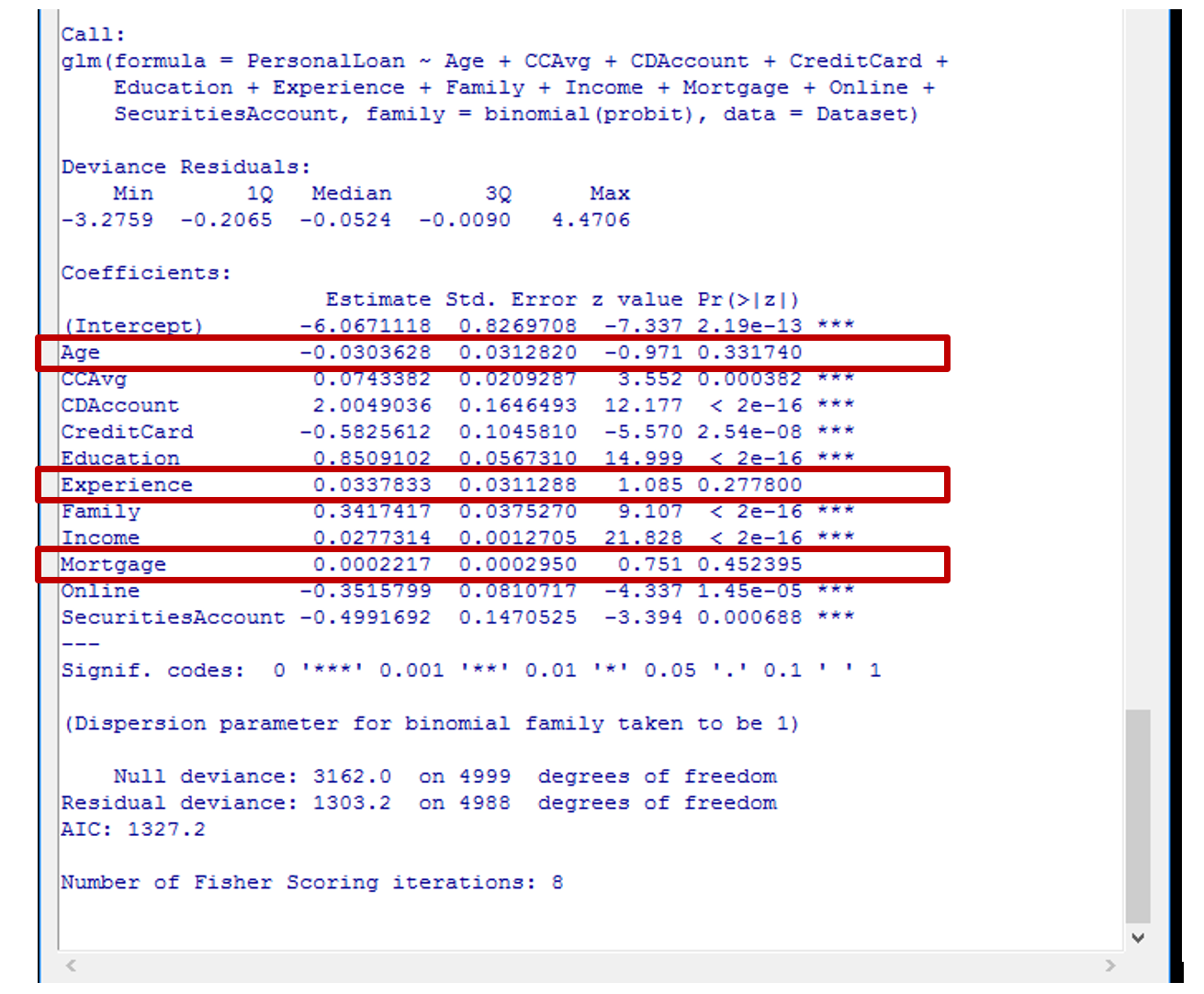


**Probit Analysis**

Similar to a Logit analysis all variables except Customer ID and Zip Code have been included in the model selection process. (Customer ID and Zip codes have been excluded.)

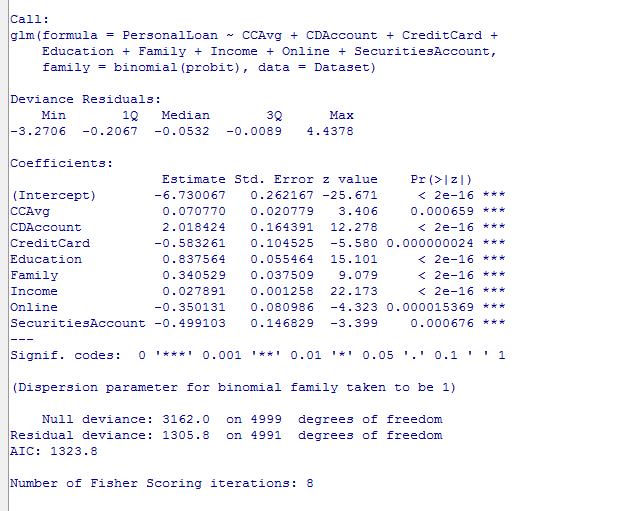
**Probit including all variables**

The screenshot below shows an output from R (Rmcdr library).



Age, Experience and Mortgage, similar to a logit model described above, remained statistically insignificant with a p-value above 0.05.

Once statistically insignificant variables have been excluded, the resulted Probit model was the following with Akaike Information Criterion of 1323.8.



|  |  |  |
| --- | --- | --- |
| How do significant variables influence the likelihood of taking out a loan? | | |
| CCAvg | Average spending on credit cards per month | The average spending per month **increases** the likelihood of a customer to take a loan. This is intuitive as high credit card utilization usually means the customer has less cash available and most likely will need external financing. |
| CDAccount | Does the customer have a certificate of deposit with the bank? | The certificate of deposit with the bank **increases** likelihood for a customer to take out a loan. This is intuitive from a convenience perspective. A customer has a CD with the bank which in turn provides easier approval process for a loan. |
| CreditCard | Does the customer use a credit card issued by Universal Bank? | The existing credit card with the bank **decreases** the likelihood of taking out another loan. This is intuitive as the customer would rather increase existing credit line than apply for a second card. If the loan is a mortgage for a house purchase, the customer most likely will go to another bank if the existing credit card limit is fully utilized. |
| Education | Education levels from undergraduate to professional | The higher education level translates to an **increased** likelihood of taking a loan. This is intuitive if a customer takes a Home Equity Line loan or a Residential mortgage (this is not specified in the instructions and affects our response). If we look at a Home Equity lines, fresh out of college graduates most likely will not need a home equity loan or residential loan as this is a population who usually rents. |
| Family | Family size | A bigger family size **increases** the likelihood of taking a loan. |
| Income | Annual income of a customer | As an income increases there is a **higher** likelihood a customer would take a loan. This is intuitive especially if a loan type is financing the mortgage/home. |
| Online | Does the customer use Internet banking? | The likelihood of taking a loan **decreases** when a client uses Internet banking facilities. |
| SecuritiesAccount | Does the customer have securities or deposit account with the bank? | The likelihood of taking a loan **decreases** when a client has securities or deposit accounts with the bank. |

## **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.

When selecting moderating effect, we are looking for variables that influence outcome (taking out a loan), and the combination of those variables which would influence the likelihood of taking a loan even more. Overall, four moderating effect variables were tested. The team selected a p-value of 0.05 to decide which variables are statistically significant.

**Statistically significant:**

**Family\*CCAvg** moderating combination is conceptually sound as the family size and credit card spending per month are related and one potentially triggers the other. The sign is positive and intuitive. The larger the family, the larger credit card spending month the less disposable income the customer might have, hence the larger the need to look for other sources of cash and external financing.

**Education\*Income** moderating combination is conceptually sound as the education level affects and potentially increases the income as well. The sign is positive and intuitive. Customer with high income and customer with higher education are most likely have more luxurious demands (a larger house, a top of the line car) and potentially will be looking to take a loan.

**Not statistically significant:**

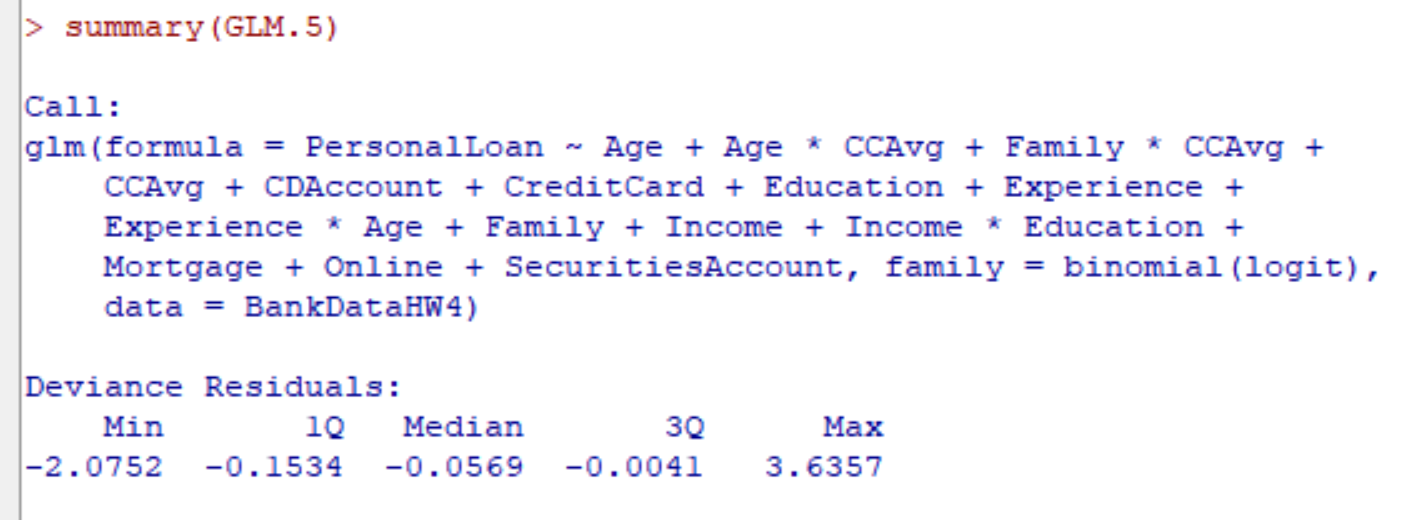
In addition, **Age\*CCAvg** and **Age\*Experience** were included to see whether with age credit card usage increases and whether experience is relevant. Neither of these two variables were statistically significant and were eliminated from the models considered.

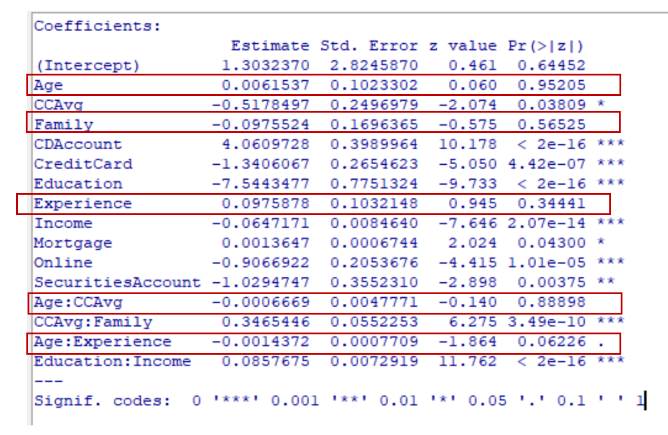
**Impact of variables and coefficients:**

Negative coefficients mean the variables reduce the likelihood of choosing to take out a loan. Positive coefficients mean the variables increase the likelihood of choosing to take out a loan.

The following screenshots are from R which are a logit (GLM.5) and probit (GLM.6) models with 4 moderating effects (probit model included same variables as a logit model and resulted in similar statistically insignificant variables shown on the following page).

**Logit with moderating effects**





**Probit with moderating effects**

## 

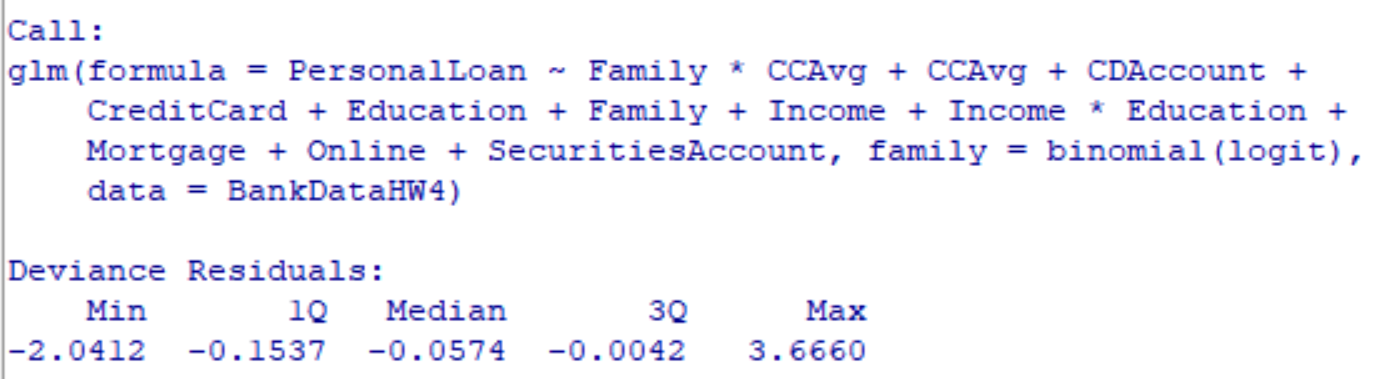
## **3. Create a final regression model with the variables that you feel are important (both main effects and interaction terms).**

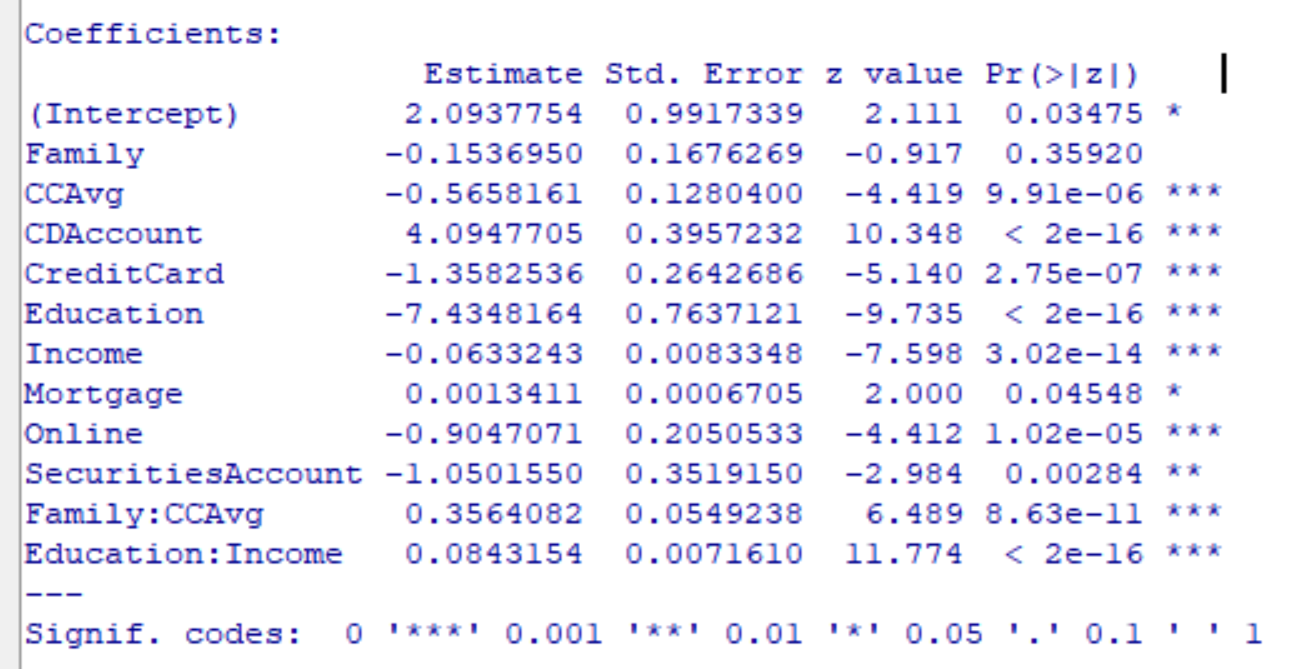
* Create a spreadsheet prediction of the model.
* Which variables have the greatest influence on the customers’ loan behavior (combined main effects and interaction effects)?
* Perform a sensitivity analysis as seen earlier in the semester.
* Copy screen snapshots of your analysis in R to your report.

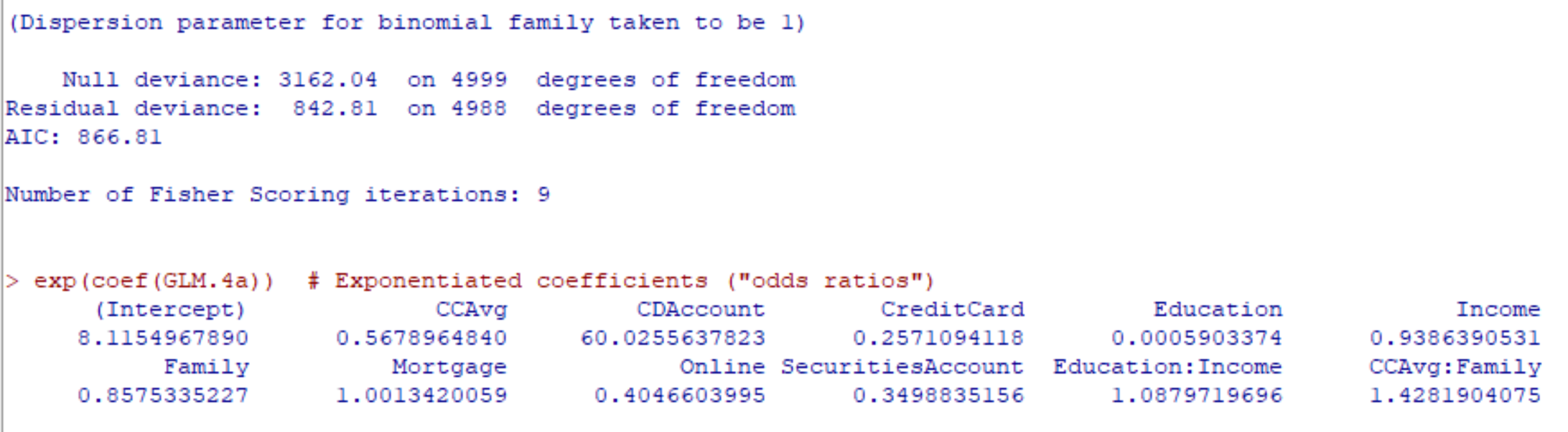
Based on the analysis from the above sections 1 and 2, once variables with insignificant moderating effects were excluded from both logit and probit models the team selected a final logit model with moderating effects using Akaike Information Criterion. This model had the lowest AIC value of 866.81 among other alternative models as shown below.

We chose the following logit model based on its lower AIC, inclusion of all statistically significant variables and variables with moderating effects as the team believed all these variables belong to the final model and should be included.

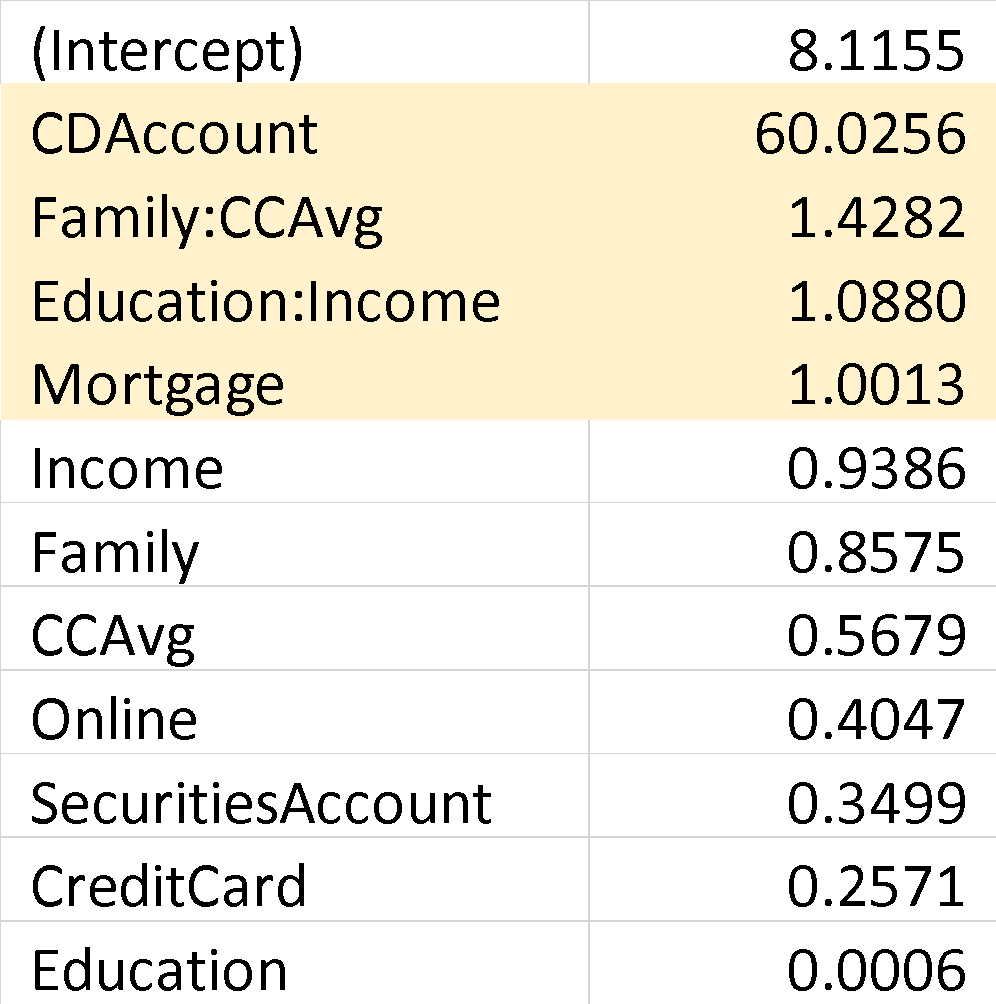
**Logit model with moderating effects**



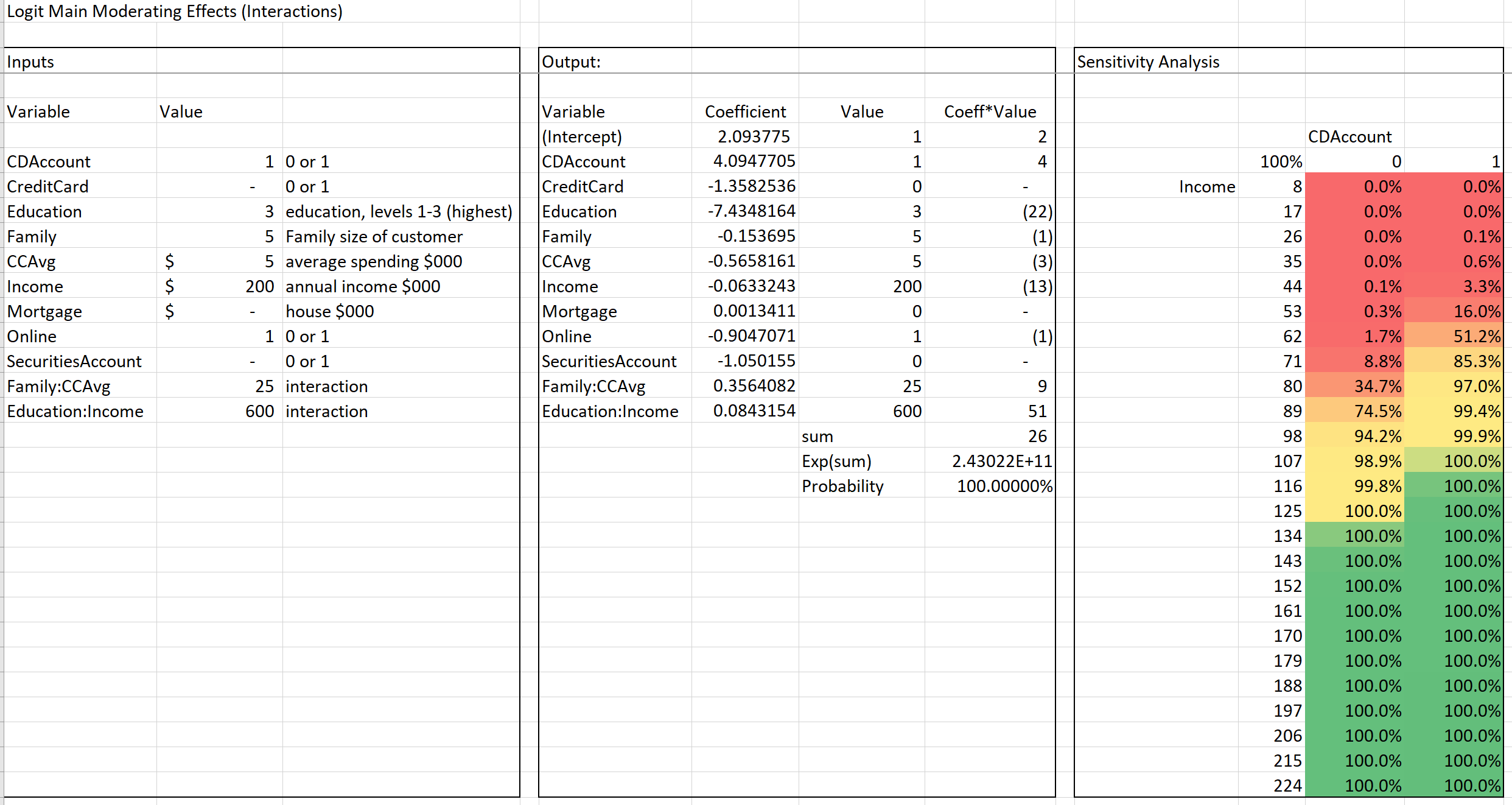




The variables with greatest influence (the odds ratio above 1 as shown below) are CD Account, interactions of Family size and Credit Card Averages, Education and Income, and a size of a Mortgage using exponentiated coefficients. For example, the odds ratio of 60.02 for a CD Account implies that a customer who has a CD Account results in increasing odds of taking a loan by a factor of 60.02 and so forth.

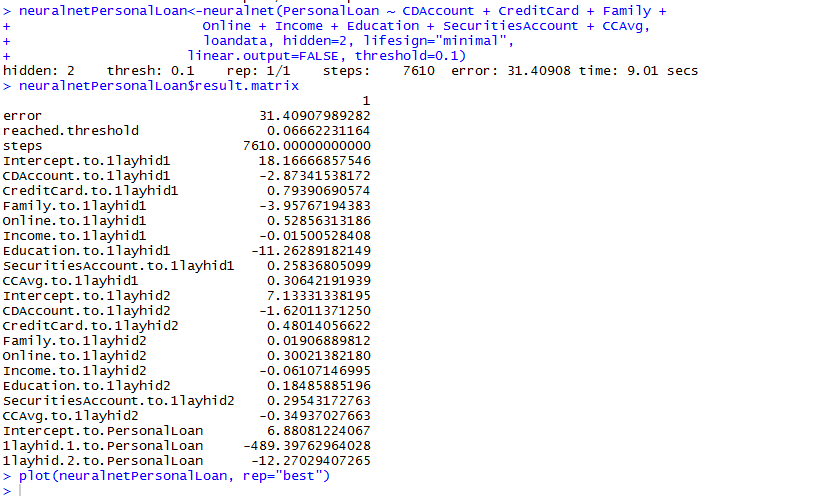


The sensitivity analysis is presented below. A customer with a CD Account has a higher likelihood of taking a loan as the Income increases. The conditional formatting highlights this trend.

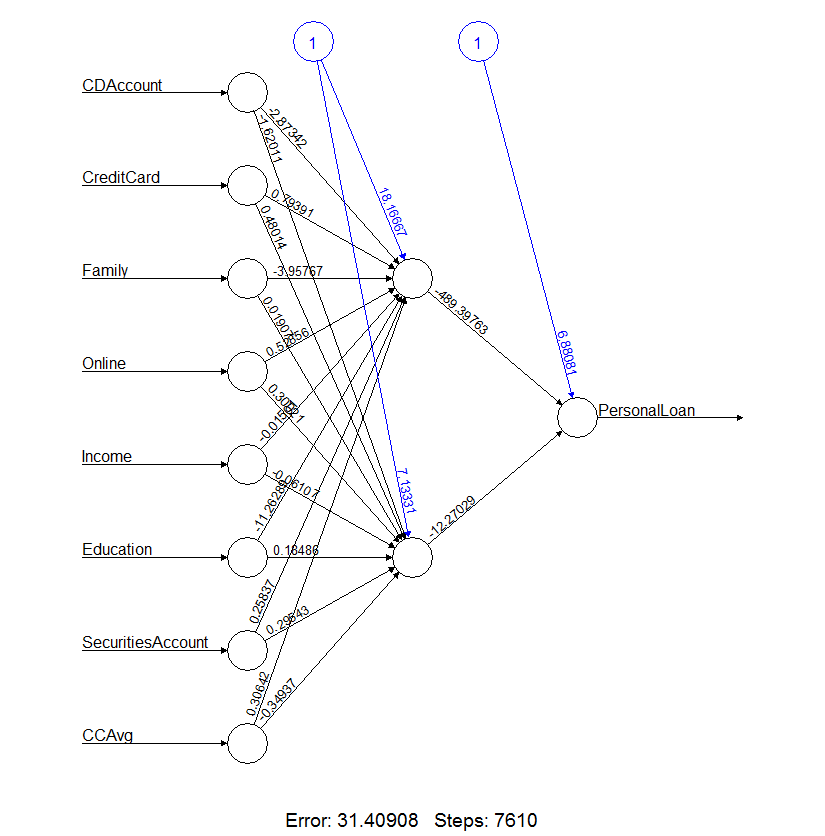


## **4. Perform a neural network analysis of the variables found to be significant in the logit and probit analysis above.**

We ran several dozen tries of our model, with the error ranging from 31.4 to 134.60 and the number of steps between 1,500 to 7610.

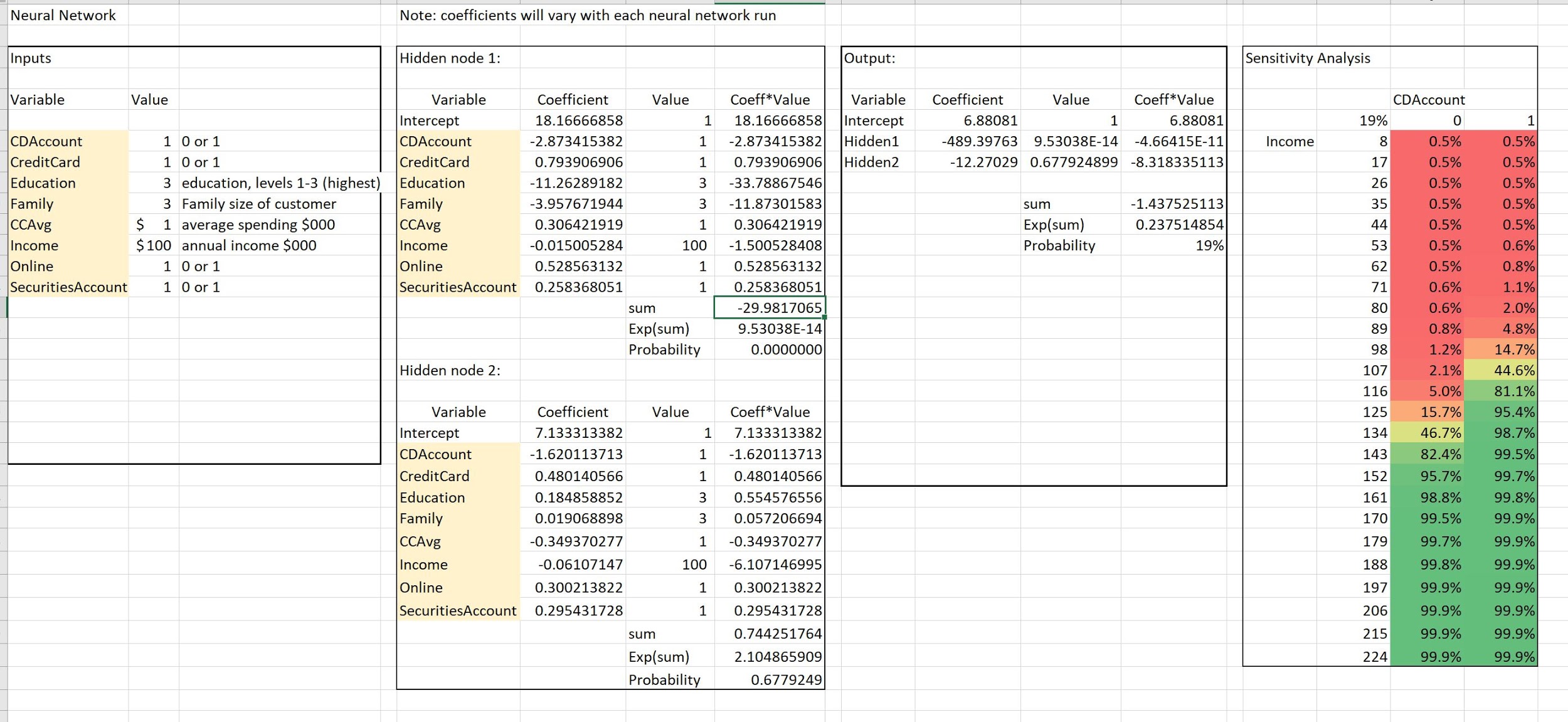


The best model among alternative models is shown below. Our significant variables are the inputs; there are two hidden nodes; PersonalLoan is the output.



## **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.**

The prediction model of neural network is presented below. While results from sensitivity testing are incomparable due to the fact that the Mortgage variable is not included in Neural network model and moderating effects variables are included in the Logit with moderating effects model, the neural network model gradually assigns a higher likelihood of a customer taking a loan with higher Income and the existence of CD Account.



**Our final choice in models:**

Both the logit with moderating effects and the neural network models have good qualities. If we wanted to decide between them, we would consider running tests for accuracy using these models and comparing the results.