Loan Analysis

Homework 4

6/8/2021

<u>Team 71</u>
Toby Anderson
Garen Moghoyan
Daniel Caley
Michael Johnson

1. Logit and Probit

Variable	Significant	Influence
Age	No	
Experience	No	
Income	Yes	Positive
Family	Yes	Positive
CCAvg	Yes	Positive
Education	Yes	Positive
Mortgage	No	
SecuritiesAccount	Yes	Negative
CDAccount	Yes	Positive
Online	Yes	Negative
CreditCard	Yes	Negative

Table 1 - Summary of variable impacts to Logit and Probit models

We determined significance by looking for P values that were <0.05. To determine influence, we looked at the sign of the coefficient.

R code to produce models:

```
loan_logit <- glm(PersonalLoan ~
                    Age + Experience + Income + Family +
                    CCAvg + Education + Mortgage + SecuritiesAccount +
                    CDAccount + Online + CreditCard,
                  family=binomial(logit), data=bank_df)
summary(loan_logit)

loan_logit_sig_p <- glm(PersonalLoan ~
                          Income + Family + CCAvg + Education +
                          SecuritiesAccount + CDAccount + Online + CreditCard,
                        family=binomial(logit), data=bank_df)
summary(loan_logit_sig_p)


loan_probit <- glm(PersonalLoan ~
                    Age + Experience + Income + Family +
                    CCAvg + Education + Mortgage + SecuritiesAccount +
                    CDAccount + Online + CreditCard,
                  family=binomial(probit), data=bank_df)
summary(loan_probit)

loan_probit_sig_p <- glm(PersonalLoan ~
                          Income + Family + CCAvg + Education +
                          SecuritiesAccount + CDAccount + Online + CreditCard,
                        family=binomial(probit), data=bank_df)
summary(loan_probit_sig_p)
```

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.219e+01	1.645e+00	-7.411	1.25e-13	***
Age	-5.361e-02	6.131e-02	-0.874	0.38191	
Experience	6.376e-02	6.093e-02	1.046	0.29536	
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	***
Mortgage	4.745e-04	5.541e-04	0.856	0.39190	
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	***

Figure 1a – Results of Logit Model

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	0.0000205232	***
CreditCard	-1.123683	0.205003	-5.481	0.0000000422	***

Figure 1b – Results of Logit Model – Only P Values <= 0.05

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-6.0671118	0.8269708	-7.337	2.19e-13	***
Age	-0.0303628	0.0312820	-0.971	0.331740	
Experience	0.0337833	0.0311288	1.085	0.277800	
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	***
Mortgage	0.0002217	0.0002950	0.751	0.452395	
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	***

Figure 2a – Results of Probit Model

```
                    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 0.000015369 ***
CreditCard         -0.583261   0.104525  -5.580 0.000000024 ***
```

Figure 2b – Results of Probit Model – Only P Values <= 0.05

Question	Answer
Interpreting Logit Sensitivity Analysis	• As a person receives more income and education, the probability for accepting a personal loan increases based on the logit regression model. • See figure 3a
Interpreting Probit Sensitivity Analysis	• As a person receives more income and education, the probability for accepting a personal loan increases based on the probit regression model. • See figure 3b

Logit Sensitivity Analysis

		Education		
	97%	1	2	3
Income	0	0%	0%	0%
	20	0%	0%	0%
	40	0%	0%	1%
	60	0%	1%	4%
	80	0%	2%	12%
	100	1%	7%	28%
	120	4%	18%	54%
	140	10%	39%	78%
	160	26%	66%	91%
	180	51%	85%	97%
	200	76%	94%	99%
	220	90%	98%	100%
	240	96%	99%	100%

Figure 3a – Results of Logit Sensitivity Analysis

Probit Sensitivity Analysis				
		Education		
	96%	1	2	3
Income	0	0%	0%	0%
	20	0%	0%	0%
	40	0%	0%	2%
	60	0%	1%	6%
	80	0%	3%	15%
	100	2%	10%	32%
	120	6%	23%	54%
	140	15%	43%	74%
	160	32%	65%	89%
	180	54%	83%	96%
	200	74%	93%	99%
	220	89%	98%	100%
	240	96%	100%	100%

Figure 3b – Results of Probit Sensitivity Analysis

2. Moderating Effects

Question	Answer
Which interactions make sense conceptually?	• Education*Age and Education*Experience: These factors would combine training with experience and could be another way to find someone's potential expendable income. However, age and experience are not statistically significant. • Income*Family, Income*CCAvg and Income*Education: These factors would combine income with costs. For instance, income is less effective when costs are high. With high income and low costs, there would be more expendable income to pay for monthly interest.
Which interactions are statistically significant?	See figure 3 below.
How do you interpret coefficients on these variables?	As an example, the coefficient of Income*Education means that as Education increases and Income

	increases, there is a quadratic increase in the probability of a loan.

R code to test all pairs of variables in a logit model:

```r
library(gtools)

col_list = c("Age","Experience","Income","Family","CCAvg","Education",
               "Mortgage","SecuritiesAccount","CDAccount",
             "Online","CreditCard")
combos <- combinations(11, 2)

for (i in 1:55) {
  cols <- col_list[combos[i,]]
  loan_logit <- glm(PersonalLoan ~ bank_df[,cols[1]]:bank_df[,cols[2]],
                    family=binomial(probit), data=bank_df)
  p_val <- summary(loan_logit)$coefficients[,4][2]
  if (p_val < 0.05){
    print(paste(cols[1],"-",cols[2],"p-value:",p_val, sep=" "))
  }
}
```

```
"Age - CCAvg p-value: 0.0238707187689208"
"Age - Mortgage p-value: 0.0164754843807592"
"Experience - Mortgage p-value: 0.0241523839939576"
"Income - Family p-value: 7.083418831623e-22"
"Income - CCAvg p-value: 4.9509663259925e-21"
"Income - Education p-value: 1.02083507222787e-29"
"Family - CCAvg p-value: 6.64349625338396e-18"
"Family - Education p-value: 4.20548965534083e-06"
"Family - Mortgage p-value: 0.000428791491600724"
"Family - SecuritiesAccount p-value: 0.00166693212552682"
"CCAvg - Education p-value: 4.04308566849145e-33"
"CCAvg - Mortgage p-value: 0.00453541201854055"
"Mortgage - Online p-value: 0.0091368681635762"
```
Figure 4 – All statistically significant moderating combinations (when used as single factor in Logit model)

3. Final Regression Model

Question	Answer
Create final regression model	Final model was created using Income and Education. See figure 5.
Create a spreadsheet prediction of the model	See figure 6
Which variables have the greatest influence on the customers' loan behavior?	Since the values of the variables are not scaled, we cannot directly read the coefficient to determine influence. However, judging by the sensitivity analysis, Income appears to have the highest influence on loan behavior.
Perform a sensitivity analysis	See figure 7
Copy screenshots of your analysis in R to your report	See figure 8

Final model was created using Income and Education. We chose these factors because they had statistical significance as both main factors and as moderating effects. Additionally, the model produced a favorable Area Under the Curve (AUC) which is a measure of model accuracy (see section 6) and is easily interpretable in business terms.

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.905818   0.805047   2.367   0.0179 *
Income           -0.058799   0.007082  -8.303   <2e-16 ***
Education        -7.043853   0.676857 -10.407   <2e-16 ***
Income:Education  0.079411   0.006280  12.646   <2e-16 ***
```

Figure 5 – Final regression model

Output:

Variable	Coefficient	Value	Coeff*Value
Intercept	1.905818	1	1.905818
Education	-7.043853	3	-21.131559
Income	-0.058799	180	-10.58382
Education*Income	0.079411	540	42.88194
		Sum	13.07
		Exp(sum)	475,622.14
		Probability	100%

Figure 6 – Spreadsheet prediction model

```
train_df <- bank_df[0:4000,]
test_df <- bank_df[4001:5000,]

small_model <- glm(PersonalLoan ~ Income*Education,
                   family=binomial(logit), data=train_df)
summary(small_model)
```

Figure 7 – R Analysis

Question	Answer
Interpreting Logit Sensitivity Analysis with Moderating Effects	• As a person receives more income the probability for accepting a personal loan increases based on the logit regression model. • As a person receives more education the probability for accepting a personal loan increases at an income of 100,000 or higher. The probability for accepting a personal loan decreases below an income of 100,000. • See figure 8a
Interpreting Probit Sensitivity Analysis with Moderating Effects	• As a person receives more income the probability for accepting a personal loan increases based on the probit regression model. • As a person receives more education the probability for accepting a personal loan increases at an income of 100,000 or higher. The probability for accepting a personal loan decreases below an income of 100,000. • See figure 8b

Logit Sensitivity Analysis

Moderating Effects

		Education		
	100%	1	2	3
Income	0	1%	0%	0%
	20	1%	0%	0%
	40	1%	0%	0%
	60	2%	0%	0%
	80	3%	2%	1%
	100	4%	10%	22%
	120	7%	46%	91%
	140	10%	86%	100%
	160	14%	98%	100%
	180	19%	100%	100%
	200	27%	100%	100%
	220	35%	100%	100%
	240	45%	100%	100%

Figure 8a – Sensitivity Analysis

Probit Sensitivity Analysis
Moderating Effects

		Education		
	100%	1	2	3
Income	0	0%	0%	0%
	20	1%	0%	0%
	40	1%	0%	0%
	60	2%	0%	0%
	80	3%	1%	1%
	100	4%	12%	28%
	120	7%	45%	89%
	140	10%	81%	100%
	160	14%	97%	100%
	180	20%	100%	100%
	200	26%	100%	100%
	220	33%	100%	100%
	240	42%	100%	100%

Figure 8b – Sensitivity Analysis

4. Neural Network

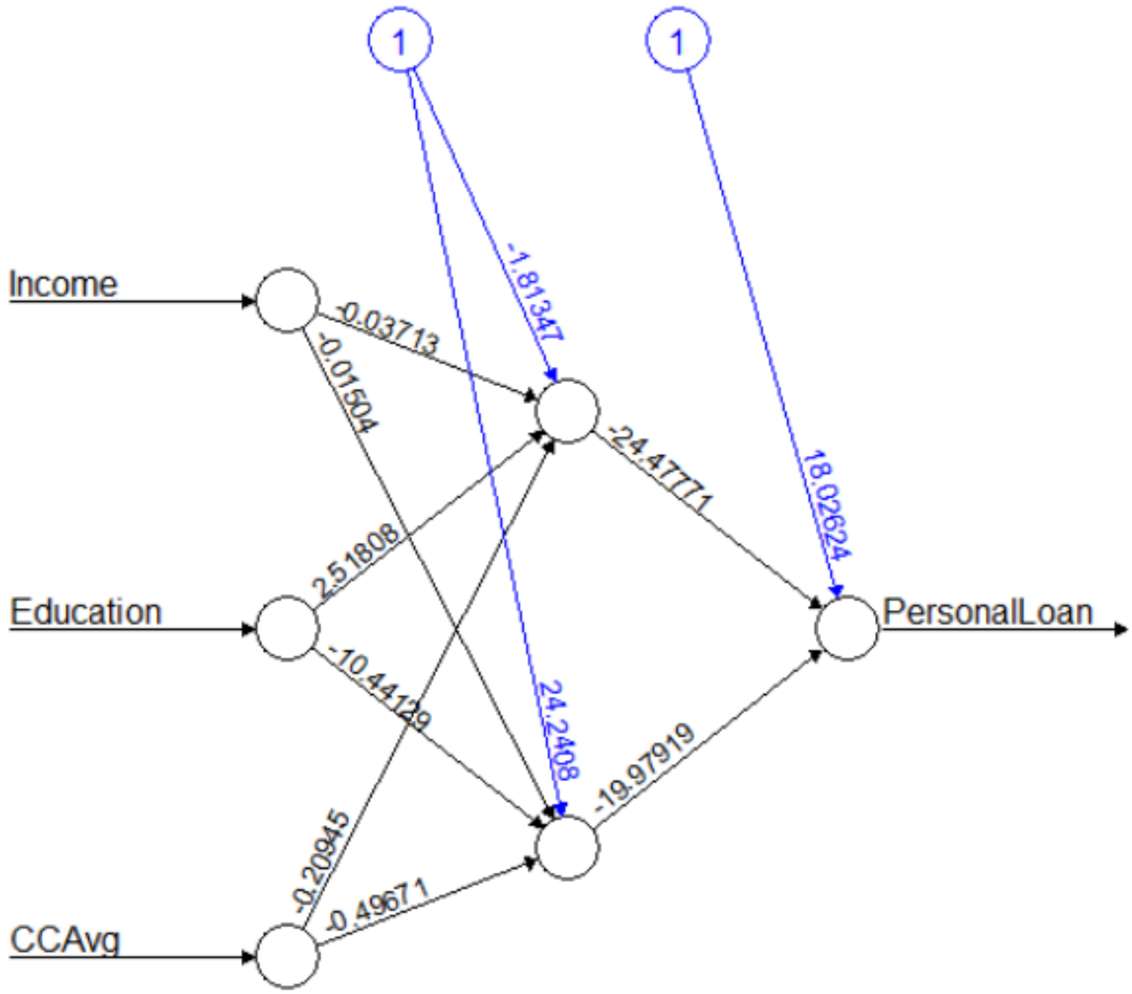


Figure 9 – Neural Network representation

Multiple variables were significant, per the Logit and Probit analysis above. Initially, we chose Income and Education as the only inputs to the Neural Network model, however, this was producing bad AUC and bad results in Excel analysis. After some experimentation, we found that including CCAvg greatly improved model results, and there is good statistical backing for using this variable based on its P value.

5. Neural Network Predictions

Neural Network

Inputs	
Variable	Value
Income	73.8
Education	1.9
CCAvg	8.9

Hidden node 1:			
Variable	Coefficient	Value	Coeff*Value
Intercept	-1.813	1	-1.813
Income	-0.0371	73.8	-2.73798
Education	2.518	1.9	4.7842
CCAvg	-0.209	8.9	-1.8601
		sum	-1.62688
		Exp(sum)	0.196541829
		Probability	0.164258218

Hidden node 2:			
Variable	Coefficient	Value	Coeff*Value
Intercept	24.24	1	24.24
Income	-0.015	73.8	-1.107
Education	-10.44	1.9	-19.836
CCAvg	-0.497	8.9	-4.4233
		sum	-1.1263
		Exp(sum)	0.324230693
		Probability	0.244844569

Output:			
Variable	Coefficient	Value	Coeff*Value
Intercept	18.03	1	18.03
Hidden1	-24.48	0.164258218	-4.021041185
Hidden2	-19.98	0.244844569	-4.891994488
		sum	9.116964326
		Exp(sum)	9108.509143
		Probability	100%

Figure 10 – Neural Network model in Excel
The Excel model was created using the coefficients from the R Neural Network model.

Sensitivity Analysis				
		Education		
	100%	1	2	3
Income	0	0%	0%	0%
	15	0%	4%	0%
	30	1%	66%	1%
	45	3%	99%	2%
	60	6%	100%	7%
	75	8%	100%	42%
	90	10%	100%	93%
	105	11%	100%	100%
	120	12%	100%	100%
	135	12%	100%	100%
	150	12%	100%	100%
	165	12%	100%	100%
	180	13%	100%	100%
	195	13%	100%	100%
	210	13%	100%	100%

Figure 11 – Neural Network sensitivity analysis

The sensitivity analysis in the Neural Network model still shows that increasing Income will increase your probability of getting a Personal Loan. However, the Education effect is no longer monotonic and only increases the likelihood of accepting a loan when going from an undergraduate degree to a graduate degree. Obtaining a professional or doctorate decreases the chance of accepting a loan below an income of 105,000.

6. Model Justification

Classification model performance is typically measured using some combination of Precision and Recall.
From the book Advanced Data Mining Techniques written by Olson, David L.; and Delen, Dursun (2008):

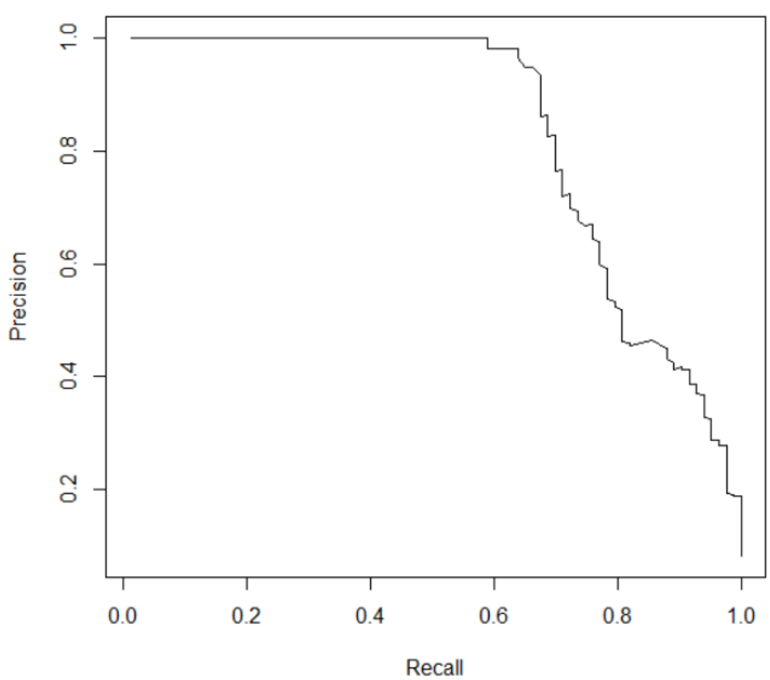
$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

Where:

- tp = True Positives
- fp = False Positives
- fn = False Negatives

However, since classification predictions are often given as probabilities, the Precision and Recall are a design tradeoff. Increasing our cutoff value will improve Precision but reduce Recall, and vice versa.
To show this, we can plot the change in Precision vs. Recall as we decrease this cutoff value.



We can convert this into a single model performance metric by measuring the Area Under the Curve (AUC). This measure tells us how robust the model is, by giving accurate predictions

whether we want to conservatively increase Precision or more liberally capture all positive cases by increasing Recall.
Below, we have measured the AUC for various models created in this homework.

Model	Description	AUC
Baseline	Constant prediction with no variables	0.5
Worst	Use the 2 worst variables, Online and CreditCard	0.485
Small	Using the 2 best variables, Income and Education	0.964
Main	Using main effects from all variables	0.965
Manual	Manually selecting multiple variables with moderating effects	0.989
All	Using all significant factors and moderating effects	0.991
NN	NN using Income, Education, CCAvg, with 2 hidden nodes	0.972

Summary:
The Small model, using only Income and Education is highly accurate. Only a small improvement is made when throwing all possible factors at the model, increasing the AUC by only 0.027.
The Neural Network gave inconsistent results. After multiple runs with different settings, the description in the table above was the simplest that gave a decent AUC.