Assignment #8: Binary Classification with Logistic Regression

Joel Riesen, MSDS 410

05/23/2021

1. Introduction

This report will be using the UniversalBank.csv data set and building a logistic regression model for binary classification. Logistic regression varies from linear regression. We will be evaluating the models-in-sample and out-of-sample using metrics based on a binary classification problem.

1.1 Sample data.

The original data contained 5000 data points and 14 variables. We will take out ID and Zip. Code right away since we will not be using them. Leaving us with 12 variables and 5000 observed.

Column names and description
ID: Customer ID
Age: Customer's age in completed years
Experience: Years of professional experience
Income Annual: Income of the customer ($000)
ZIPCode: Home Address ZIP code.
Family: The family size of the customer
CCAvg: Avg. spending on credit cards per month ($000)
Education: Education Level. 1: Undergrad; 2: Graduate; 3: Advanced/Professional
Mortgage: Value of house mortgage, if any. ($000)
Personal Loan: Did this customer accept the personal loan offered in the last campaign?
Securities Account: Does the customer have a securities account with the bank?
CD: Account Does the customer have a certificate of deposit (CD) account with the bank?
Online: Does the customer use internet banking facilities?
CreditCard: Does the customer uses a credit card issued by Universal Bank?
U: Info from seed.

1.2 Train/Test split

We will be using a 70/30 training split to check the performance for both in and out sample basic cross-validation. We added a random flag to the dataset, using 0 to 1. If the number is more significant than 0.70, it will go to the test data set. The Train data set will consist of 3492 observations or 70%, which leaves 1508 for the test data set.

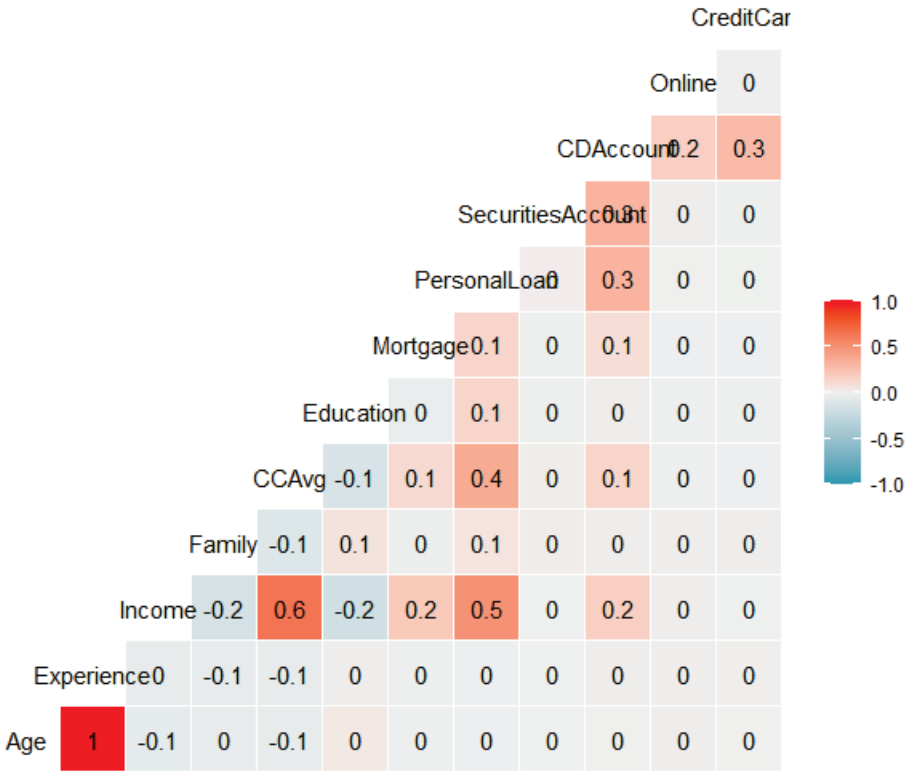
Table 1 Train/Test split

Nrow	Ncol	Complete
3492(Train)	13	3492
1508(Test)	13	1508
5000	13	5000

2. What does our data set look like

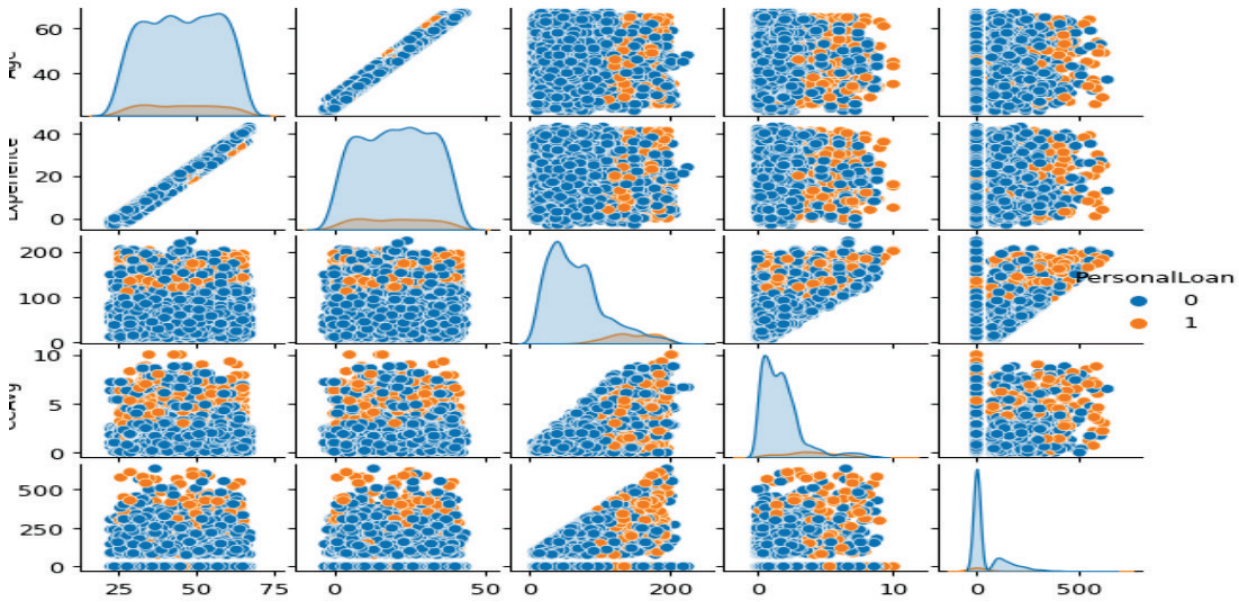
The response variable is PersonalLoan to find the best combination of variables to increase its chances of acceptance. Since this is logistic regression, we will need to transform a few of our variables into factors. Once that has been completed, we can check out the multi-colinearity of the data train data set.

Graph 1 -Correlation



Nothing shows up in the correlation matrix that I found surprising. I find the negative impact of Income as being prophetic.

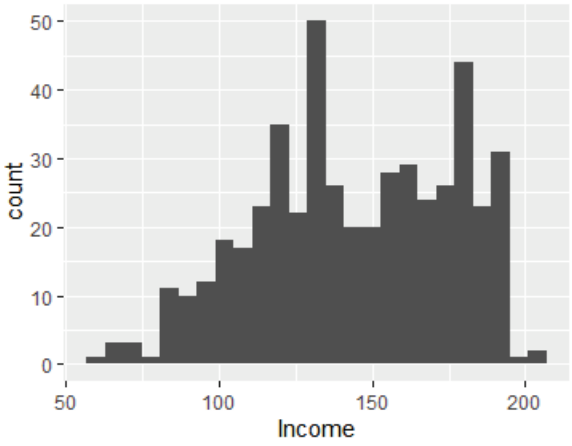
Graph 2 - Pairplot



Using the reticulate library for python and using the Seaborn package shows those who took the personal loans by each category. Notice the ones that get a PL are on the right side or with the most significant Income, experience, age, and others.

Since PersonalLoan is the response variable, how many customers purchased the loan. Just setting PL equal to one, we get 480 loans. How much was Income a factor? It seems like most are for customers with over 100k Income.

Graph 3 Income and PL



Looking at the statistics for the whole group, you can see that the average age is about 45, and the average income is about 145,000. If the average experience is 20 and the average age is 45, most people said their experience started when they were 25.

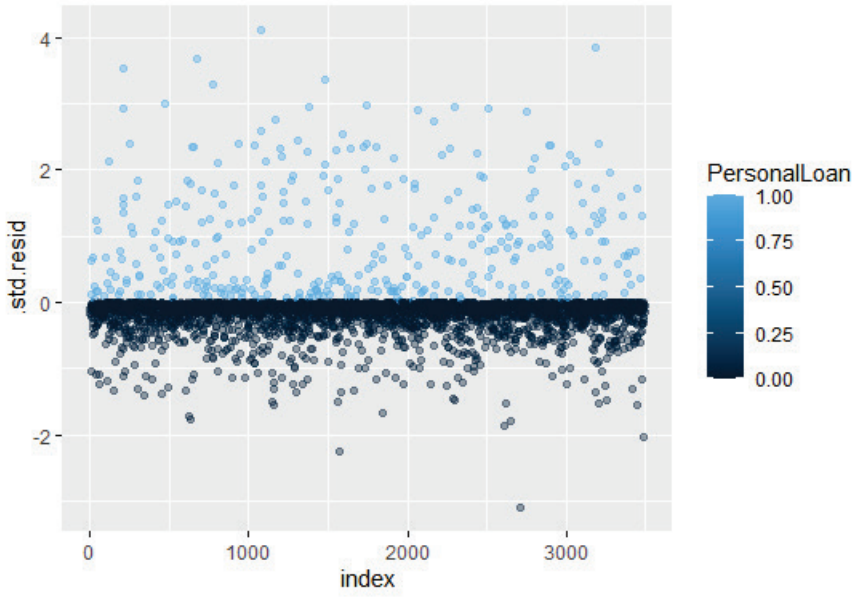
Table 2

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Age	480	45.067	11.591	26	35	55	65
Experience	480	19.844	11.582	0	9	30	41
Income	480	144.746	31.584	60	122	172	203
CCAvg	480	3.905	2.098	0.000	2.600	5.348	10.000
Mortgage	480	100.846	160.848	0	0	192.5	617
PersonalLoan	480	1.000	0.000	1	1	1	1

3. Fitting a Naïve Model as our Baseline model

First, we should check and see if we can find any outliers. We will check the residuals data from model.1

Graph 4 Residuals



Looks to be a few outliers that we can look into if needed.

Running the first model we used:

model.1 <- glm(PersonalLoan ~ Income+

CCAvg+

CDAccount+

factor(Education)+

Family+

SecuritiesAccount,

data=train.df, family=c('binomial'))

From that, the AIC for the individual data point was higher than the Models. So they all are having an impact on PL. Again, nothing surprising and Income was the largest number.

Start: AIC=895.0

Table 3 AIC

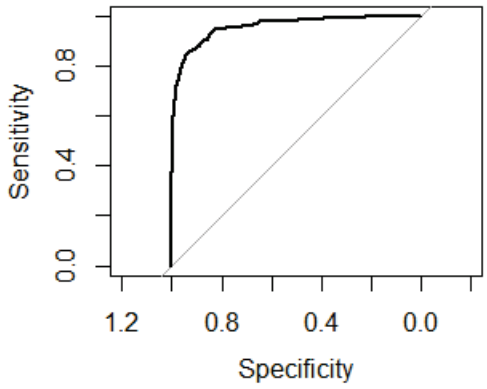
	Df	Deviance	AIC
		879.07	895.07
- SecuritiesAccount	1	881.97	895.97
- CCAvg	1	888.61	902.61
- Family	1	925.62	939.62
- CDAccount	1	940.26	954.26
- factor(Education)	2	1213.09	1225.09
- Income	1	1477.78	1491.78

ROC, for the model, was almost 0.96, which means virtually all the data points impact personalLoan.

```
  The ROC summary

  Data: model.1$fitted.values in 3148 controls (train.df$PersonalLoan 0)
< 344 cases (train.df$PersonalLoan 1).
The area under the curve: 0.9584
```

Graph 5 ROC



The AUC for this model is 0.9584.

The confusion matrix is also showing a very positive relationship

Confusion Matrix 1

	0 +	1 -
0 +	0.93996188	0.06003812
1-	0.14936692	0.85174419

4. **<u>Second model TOP 3 AIC of last model</u>**

The top three from the last model was:

Table 4

- CDAccount	1	940.26 - 954.26
- factor(Education)	2	1213.09 -1225.09
- Income	1	1477.78- 1491.78

When we run model2, we get for AIC:

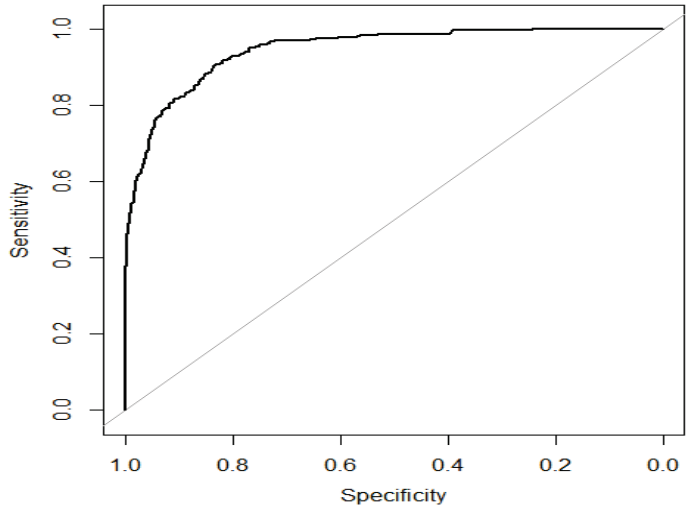
Table 5

Start: AIC=1061.45
PersonalLoan ~ Income + Education + CDAccount
Df Deviance AIC

		1053.5	1061.5
- CDAccount	1	1131.6	1137.6
- Education	1	1344.8	1350.8
- Income	1	2000.1	2006.1

The AUC for the second model2 was 0.9462, which is lower than the first model.

Graph7 Model2 AUC



Little surprised that there is even less having an impact on the model. However, it also used fewer data points. So, in this case, I would have to say, the more, the better.

The Third Model

For the third model, I would like to grab most data points and check them with their AIC to see if something was missed.

Start: AIC=2249.44

Table 6

PersonalLoan ~ 1			
	Df	Deviance	AIC
+ Income	1	1446.2	1450.2
+ CCAvg	1	1880.6	1884.6
+ CDAccount	1	2065.7	2069.7
+ Education	1	2184.2	2188.2
+ Mortgage	1	2197.1	2201.1
+ Family	1	2231.9	2235.9
<none>		2247.4	2249.4
+ SecuritiesAcc	1	2245.4	2249.4
+ Age	1	2246.6	2250.6
+ Experience	1	2246.7	2250.7
+ Online	1	2247.0	2251.0
+ CreditCard	1	2247.1	2251.1

Step: AIC=1450.24

PersonalLoan ~ Income

	Df	Deviance	AIC
+ Education	1	1131.6	1137.6
+ Family	1	1312.1	1318.1
+ CDAccount	1	1344.8	1350.8
+ CCAvg	1	1441.9	1447.9
+ SecuritiesAcco	1	1443.0	1449.0
<none>		1446.2	1450.2
+ Online	1	1444.8	1450.8
+ Age	1	1445.0	1451.0
+ Mortgage	1	1445.1	1451.1
+ CreditCard	1	1445.2	1451.2
+ Experience	1	1445.5	1451.5

Step: AIC=1137.57

PersonalLoan ~ Income + Education

	Df	Deviance	AIC
+ CDAccount	1	1053.5	1061.5
+ Family	1	1056.1	1064.1
+ CCAvg	1	1122.2	1130.2
+ SecuritiesAcc	1	1128.7	1136.7
+ Online	1	1129.4	1137.4
<none>		1131.6	1137.6
+ Mortgage	1	1130.2	1138.2
+ Experience	1	1130.3	1138.3
+ Age	1	1130.6	1138.6
+ CreditCard	1	1130.7	1138.7

Step: AIC=1061.45

PersonalLoan ~ Income + Education + CDAccount

	Df	Deviance	AIC
+ Family	1	985.0	995.0
+ Online	1	1039.6	1049.6
+ CreditCard	1	1044.6	1054.6
+ CCAvg	1	1046.7	1056.7
+ SecuritiesAc	1	1048.9	1058.9
<none>		1053.5	1061.5
+ Mortgage	1	1052.3	1062.3
+ Experience	1	1052.8	1062.8
+ Age	1	1053.0	1063.0

Step: AIC=995

PersonalLoan ~ Income + Education + CDAccount + Family

	Df	Deviance	AIC
+ Online	1	972.60	984.60

+ CreditCard	1	977.14	989.14
+ CCAvg	1	979.54	991.54
+ SecuritiesAc	1	980.98	992.98
<none>		985.00	995.00
+ Experience	1	983.01	995.01
+ Age	1	983.36	995.36
+ Mortgage	1	984.62	996.62

Step: AIC=984.6

PersonalLoan ~ Income + Education + CDAccount + Family + Online

	Df	Deviance	AIC
+ CreditCard	1	962.18	976.18
+ CCAvg	1	966.89	980.89
+ SecuritiesAcc	1	967.79	981.79
<none>		972.60	984.60
+ Experience	1	970.64	984.64
+ Age	1	970.96	984.96
+ Mortgage	1	972.36	986.36

Step: AIC=976.18

PersonalLoan ~ Income + Education + CDAccount + Family + Online + CreditCard

	Df	Deviance	AIC
+ SecuritiesAcc	1	954.86	970.86
+ CCAvg	1	956.44	972.44
<none>		962.18	976.18
+ Experience	1	960.33	976.33
+ Age	1	960.67	976.67
+ Mortgage	1	961.96	977.96

Step: AIC=970.86

PersonalLoan ~ Income + Education + CDAccount + Family + Online + CreditCard + SecuritiesAccount

	Df	Deviance	AIC
+ CCAvg	1	949.12	967.12
<none>		954.86	970.86
+ Experience	1	952.93	970.93
+ Age	1	953.30	971.30
+ Mortgage	1	954.71	972.71

Step: AIC=967.12

PersonalLoan ~ Income + Education + CDAccount + Family + Online + CreditCard + SecuritiesAccount + CCAvg

	Df	Deviance	AIC
+ Experience	1	946.21	966.21
+ Age	1	946.70	966.70

<none>		949.12	967.12
+ Mortgage	1	948.68	968.68

Step: AIC=964.85
PersonalLoan ~ Income + Education + CDAccount + Family + Online +
CreditCard + SecuritiesAccount + CCAvg + Experience

	Df	Deviance	AIC
<none>		942.85	964.85
+ Age	1	941.27	965.27
+ Mortgage	1	942.43	966.43

Call: glm(formula = PersonalLoan ~ Income + Education + CDAccount +
Family + Online + CreditCard + SecuritiesAccount + CCAvg +
Experience, family = "binomial", data = train.df)

Coefficients:

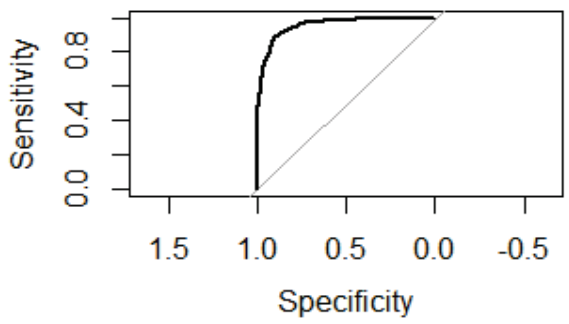
(Intercept)	Income	Education	CDAccount
-13.87135	0.05588	1.74819	3.35488
Family	**Online**	**CreditCard**	**SecuritiesAccount**
0.66790	-0.71831	-0.80503	-0.89475
CCAvg	**Experience**		
0.11674	0.01266		

Degrees of Freedom: 3491 Total (i.e. Null); 3481 Residual
Null Deviance: 2247
Residual Deviance: 942.9 AIC: 964

Based on the information above we chose to use Income + CCAvg + CDAccount +factor(Education) + Mortgage +factor(Family) . Since these look to have a positive impact on lowering the overall AIC for the models. Additionally added in factors for Education and Family which also had positive impact on AIC.

Area under the curve: 0.9609

Graph 8



Using the Caret package VarImp which is used for estimating the contribution of each variable to the model. The reason for this is I wanted to verify the impact using factors actually had. Looking at the tables it seems that factoring education was beneficial.

Caret VarImp Tables

	variable	Overall
1	Income	17.1281538
2	factor(Education)3	12.9661885
3	factor(Education)2	12.7788859
4	CDAccount	7.5365954
5	factor(Family)3	6.0585193
6	factor(Family)4	5.2217859
7	CCAvg	3.6904584
8	Online	3.5763086
9	SecuritiesAccount	1.5631477
10	Mortgage	1.3778573
11	factor(Family)2	1.0809669
12	Experience	1.0173587
13	Age	0.8763672

	variable	Overall
1	Income	17.6533598
2	Education	12.9764198
3	CDAccount	8.8031460
4	Family	7.6690543
5	Online	3.9930801
6	CreditCard	3.4752078
7	CCAvg	2.6579947
8	SecuritiesAccount	2.6129016
9	Experience	1.4205877
10	Age	1.2449881
11	Mortgage	0.7339022

Running the models on the test data frame for the first and last model.

Third Model

Confusion Matrix 2

	0	1
0	0.93658892	0.06341108
1	0.11029412	0.88970588

roc.specs

threshold	specificity	sensitivity
0.108253	0.9365889	0.8897059

0.93658892/(0.93658892 + 0.88970588)

[1] 0.5128356

First Model

Confusion Matrix 3

	0	1
0	0.96209913	0.03790087
1	0.15441176	0.84558824

roc.specs

threshold	specificity	sensitivity
0.1924083	0.9620991	0.8455882

0.96209913/(0.96209913 + 0.84558824)

[1] 0.5322265

From the confusion matrix the first model was a little more accurate in its predictions.