LoanLogisticRegModel

Laura Ahumada, Erin McClure-Price, Duy Nguyen

EDA

Looking into the data

Statistics

- ID can be dropped since it is not a useful predictor and just a unique identifier and will affect the results of the model
- Age mean is 45 years old, 23 has is the lowest and 67 is the highest. There seems to be a normal distribution.
- ***There are negative values in Experience which is odd because there can't be negative years of experience. This will be looked at next***.
- The mean income is 73 while the median is 64 showing skewness which seems normal representation of society. The minimum income is 8,000 and the max is 224,000. These are common Salaries
- ZIP.Code should not be numberic they should be changed to categorical. There are 467 distinct ZIP codes
- There seems to be an equal distribution for families size 1,2,3,4. Each are compose of about 25%
- The CCAVG, Average Spending per 1000 goes from 0 to 10,000, however the median is 1,5000. Here we can see that there are outliers.
- For Education, 41% have have up to highschool, 28% up to under grad studies and interestingly 30% have up to grad school. That seems like a reasonable distribution.
- For mortgage we can see how skewed it is, the mean is 56.5 and the median is 0. Showing the extreme outliers. **This needs to be looked at**
- For the target variable, personal loan, we can see quite a difference, 90% without a loan and only 10% with loan. This makes sense because not many people take loan from banks
- Securities account also has a big difference where 90% does not have a security account while 10% does
- For CD.Account (Ceritificate deposit) once again we see the 94% does not have it while 6% has it.
- As for online (online banking capability), 40% doe not have it while 60% has it. That distribution is a little more balanced and makes sense.
- As for Credit card 70% does not have it while 30% has it.

```
PersonalLoan

 14  Variables      5000  Observations
--------------------------------------------------------------------------------
ID
       n  missing distinct     Info     Mean      Gmd      .05      .10
    5000        0     5000        1     2500     1667    251.0    500.9
     .25      .50      .75      .90      .95
  1250.8   2500.5   3750.2   4500.1   4750.1

lowest :    1    2    3    4    5, highest: 4996 4997 4998 4999 5000
```

```
--------------------------------------------------------------------------------
Age
       n  missing distinct     Info     Mean      Gmd      .05      .10
    5000        0       45    0.999    45.34    13.23       27       30
     .25      .50      .75      .90      .95
      35       45       55       61       63

lowest : 23 24 25 26 27, highest: 63 64 65 66 67
--------------------------------------------------------------------------------
Experience
       n  missing distinct     Info     Mean      Gmd      .05      .10
    5000        0       47    0.999     20.1    13.23        2        4
     .25      .50      .75      .90      .95
      10       20       30       36       38

lowest : -3 -2 -1  0  1, highest: 39 40 41 42 43
--------------------------------------------------------------------------------
Income
       n  missing distinct     Info     Mean      Gmd      .05      .10
    5000        0      162        1    73.77    50.91       18       22
     .25      .50      .75      .90      .95
      39       64       98      145      170

lowest :   8   9  10  11  12, highest: 203 204 205 218 224
--------------------------------------------------------------------------------
ZIP.Code
       n  missing distinct     Info     Mean      Gmd      .05      .10
    5000        0      467        1    93153     2042    90073    90275
     .25      .50      .75      .90      .95
   91911    93437    94608    95138    95670

lowest :  9307 90005 90007 90009 90011, highest: 96091 96094 96145 96150 96651

Value       9000 90000 91000 92000 93000 94000 95000 96000 97000
Frequency      1   573   472   837   626   940  1117   428     6
Proportion 0.000 0.115 0.094 0.167 0.125 0.188 0.223 0.086 0.001

For the frequency table, variable is rounded to the nearest 1000
--------------------------------------------------------------------------------
Family
       n  missing distinct     Info     Mean      Gmd
    5000        0        4    0.934    2.396    1.279

Value          1     2     3     4
Frequency   1472  1296  1010  1222
Proportion 0.294 0.259 0.202 0.244
--------------------------------------------------------------------------------
CCAvg
       n  missing distinct     Info     Mean      Gmd      .05      .10
    5000        0      108    0.999    1.938    1.794      0.1      0.3
     .25      .50      .75      .90      .95
     0.7      1.5      2.5      4.3      6.0

lowest :  0.0  0.1  0.2  0.3  0.4, highest:  8.8  8.9  9.0  9.3 10.0
```

```
--------------------------------------------------------------------------------
Education
       n  missing distinct     Info     Mean      Gmd
    5000        0        3    0.877    1.881   0.9073

Value          1     2     3
Frequency   2096  1403  1501
Proportion 0.419 0.281 0.300
--------------------------------------------------------------------------------
Mortgage
       n  missing distinct     Info     Mean      Gmd      .05      .10
    5000        0      347    0.668     56.5    88.16        0        0
     .25      .50      .75      .90      .95
       0        0      101      200      272

lowest :   0  75  76  77  78, highest: 590 601 612 617 635
--------------------------------------------------------------------------------
Personal.Loan
       n  missing distinct     Info      Sum     Mean      Gmd
    5000        0        2     0.26      480    0.096   0.1736

--------------------------------------------------------------------------------
Securities.Account
       n  missing distinct     Info      Sum     Mean      Gmd
    5000        0        2    0.281      522   0.1044    0.187

--------------------------------------------------------------------------------
CD.Account
       n  missing distinct     Info      Sum     Mean      Gmd
    5000        0        2     0.17      302   0.0604   0.1135

--------------------------------------------------------------------------------
Online
       n  missing distinct     Info      Sum     Mean      Gmd
    5000        0        2    0.722     2984   0.5968   0.4814

--------------------------------------------------------------------------------
CreditCard
       n  missing distinct     Info      Sum     Mean      Gmd
    5000        0        2    0.623     1470    0.294   0.4152

--------------------------------------------------------------------------------

       ID            Age          Experience       Income          ZIP.Code
 Min.   :   1   Min.   :23.00   Min.   :-3.0   Min.   :  8.00   Min.   : 9307
 1st Qu.:1251   1st Qu.:35.00   1st Qu.:10.0   1st Qu.: 39.00   1st Qu.:91911
 Median :2500   Median :45.00   Median :20.0   Median : 64.00   Median :93437
 Mean   :2500   Mean   :45.34   Mean   :20.1   Mean   : 73.77   Mean   :93152
 3rd Qu.:3750   3rd Qu.:55.00   3rd Qu.:30.0   3rd Qu.: 98.00   3rd Qu.:94608
 Max.   :5000   Max.   :67.00   Max.   :43.0   Max.   :224.00   Max.   :96651
     Family          CCAvg          Education        Mortgage
 Min.   :1.000   Min.   : 0.000   Min.   :1.000   Min.   :  0.0
 1st Qu.:1.000   1st Qu.: 0.700   1st Qu.:1.000   1st Qu.:  0.0
 Median :2.000   Median : 1.500   Median :2.000   Median :  0.0
```

```
 Mean   :2.396   Mean   : 1.938   Mean   :1.881   Mean   : 56.5  
 3rd Qu.:3.000   3rd Qu.: 2.500   3rd Qu.:3.000   3rd Qu.:101.0  
 Max.   :4.000   Max.   :10.000   Max.   :3.000   Max.   :635.0  
 Personal.Loan   Securities.Account   CD.Account         Online      
 Min.   :0.000   Min.   :0.0000     Min.   :0.0000   Min.   :0.0000  
 1st Qu.:0.000   1st Qu.:0.0000     1st Qu.:0.0000   1st Qu.:0.0000  
 Median :0.000   Median :0.0000     Median :0.0000   Median :1.0000  
 Mean   :0.096   Mean   :0.1044     Mean   :0.0604   Mean   :0.5968  
 3rd Qu.:0.000   3rd Qu.:0.0000     3rd Qu.:0.0000   3rd Qu.:1.0000  
 Max.   :1.000   Max.   :1.0000     Max.   :1.0000   Max.   :1.0000  
   CreditCard   
 Min.   :0.000  
 1st Qu.:0.000  
 Median :0.000  
 Mean   :0.294  
 3rd Qu.:1.000  
 Max.   :1.000  
```

Looking into questions obtained in the statistical analysis

- ***Looking at the entries with negative experience***
- The 52 people with negative total experience are between 23 to 29 years old and salary median of 65,000. Sound like these could just young people that just started working. We can change their negative values to 1 since they have an income, thus are are working.

```
      Age            Income       Personal.Loan
 Min.   :23.00   Min.   : 12.00   Min.   :0    
 1st Qu.:24.00   1st Qu.: 40.75   1st Qu.:0    
 Median :24.00   Median : 65.50   Median :0    
 Mean   :24.52   Mean   : 69.94   Mean   :0    
 3rd Qu.:25.00   3rd Qu.: 86.75   3rd Qu.:0    
 Max.   :29.00   Max.   :150.00   Max.   :0    
```

- Setting categories as factors
- Creating a new variable called Exprience 2 due to Age and experience having a correlation of 97%. That way we can get rid of Age and Experience

```r
# Identification Columns (ID and ZIP.Code)
# Droppingzip and ID as there are too many unique values and will affect the model
PersonalLoan = PersonalLoan[-c(1,5)]

# making sure yes is our targer variable
PersonalLoan$Personal.Loan<-factor(ifelse(PersonalLoan$Personal.Loan==1,"Yes","No"),levels=c("No","Yes"

# To categorical
factor_vars = c("Family", "Education",
                "Securities.Account", "CD.Account", "Online", "CreditCard")
PersonalLoan[factor_vars] = lapply(PersonalLoan[factor_vars], as.factor)

PersonalLoan = PersonalLoan %>% mutate(Experience2 = Experience/Age)

# getting rid of Age and Experience
```

```
PersonalLoan = PersonalLoan[-c(1,2)]
names(PersonalLoan)
```

```
 [1] "Income"             "Family"             "CCAvg"
 [4] "Education"          "Mortgage"           "Personal.Loan"
 [7] "Securities.Account" "CD.Account"         "Online"
[10] "CreditCard"         "Experience2"
```

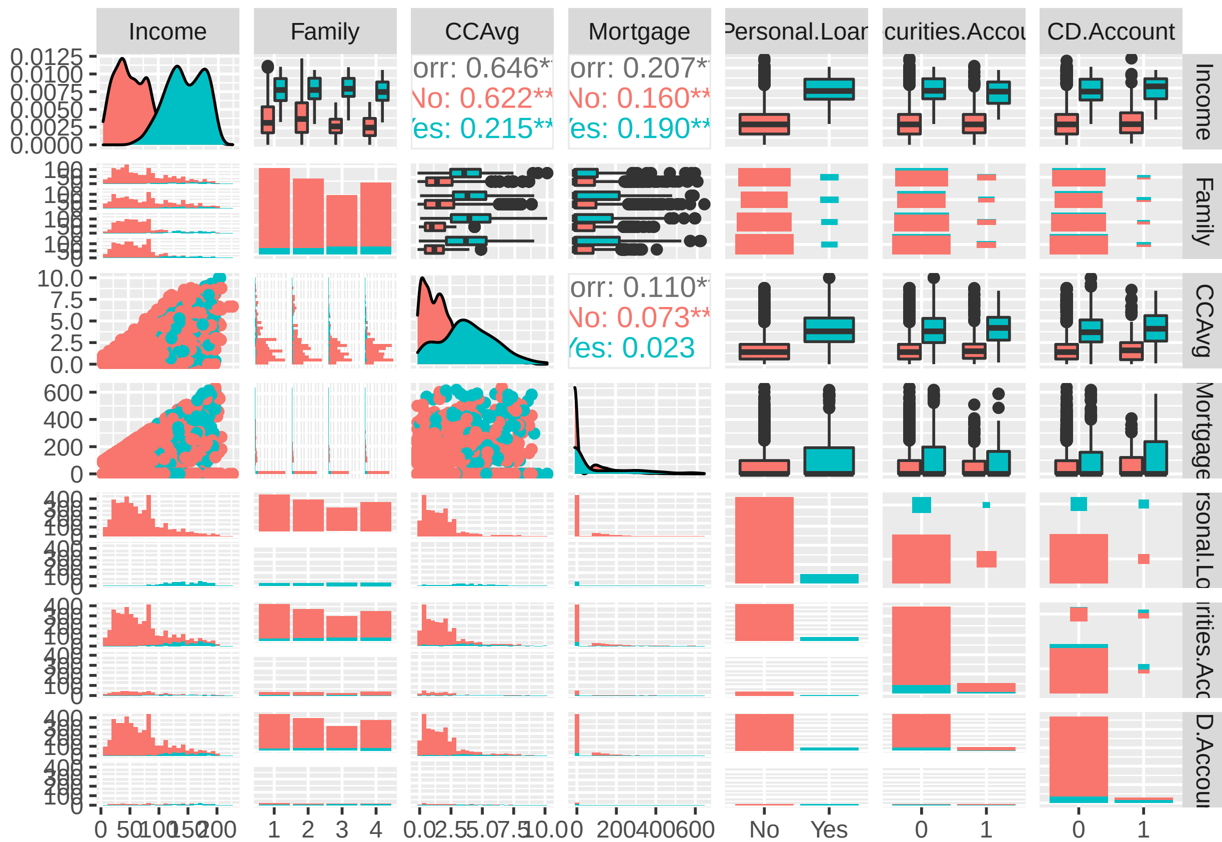
- ***Checking mortgage distribution***
- Mortgage may need to be logged as it is very skewed.

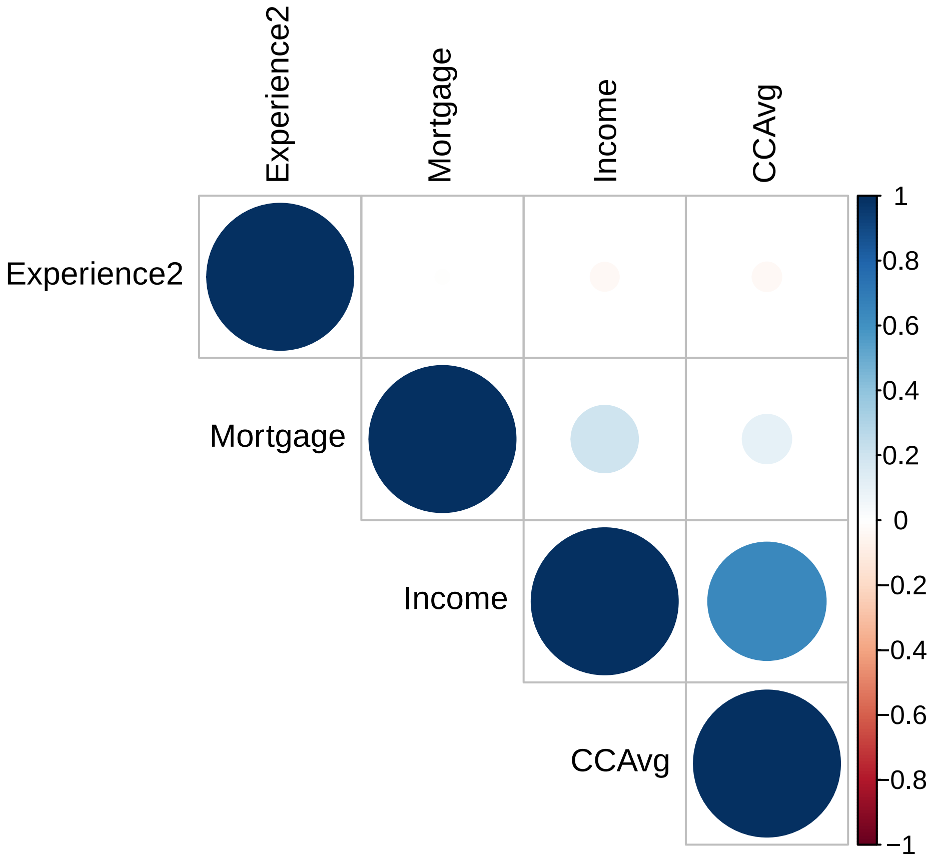
- Creating another data set with Mortgage logged since logging did improve the distribution. The zero's were replaced to with 1 in order to not get infitiy when logging. Also, removing age since it has a 99% correlation with Experience, therefore only one is needed

```
#Creating a new data set with the modified attributes
#PersonalLoan2=mutate(PersonalLoan)
#Updating values
#PersonalLoan2$Mortgage[PersonalLoan2$Mortgage==0]=1
#PersonalLoan2$MortgageLogged=log(PersonalLoan2$Mortgage)
#PersonalLoan2=PersonalLoan2 %>% select(-Mortgage)
#PersonalLoan2=PersonalLoan2 %>% select(-Age)
```

Relationships and correlations

- Experience and Age has a Correlation of 99%.Too high. However, they seem to have no relationship with Loan as both, Yes loan and no loan, have correlation of 0.99
- **Make a plot with just loan and experience and experience and loan**
- CCAvg and Income has a 64% correlations and it does seems to have a relationship with Loan since there is 0.62 with no loan and .02 with yes loan
- The rest of the explanatory variables do not seem to have relationship between each other





Checking realtionship between Loan (response variable) and the rest of the predictors

Relationship present

- ***Income***, the more income the more changes to get a loan
- ***CD.Account*** (certificate of deposit) seem to have a relationship with Personal Loan. Those with personal loan tend to have CD.Account more so than those with no CD.Account
- ***Mortgage***, people high mortgages seem to ask for loans more so thatn those with lower mortgages
- ***Personal education***, people with up to highschool tend do not get loans as much as does with an education of higher than highschool
- `!!**there seems to be relationship with loan and education but that of education 2 and 3 seem to be`

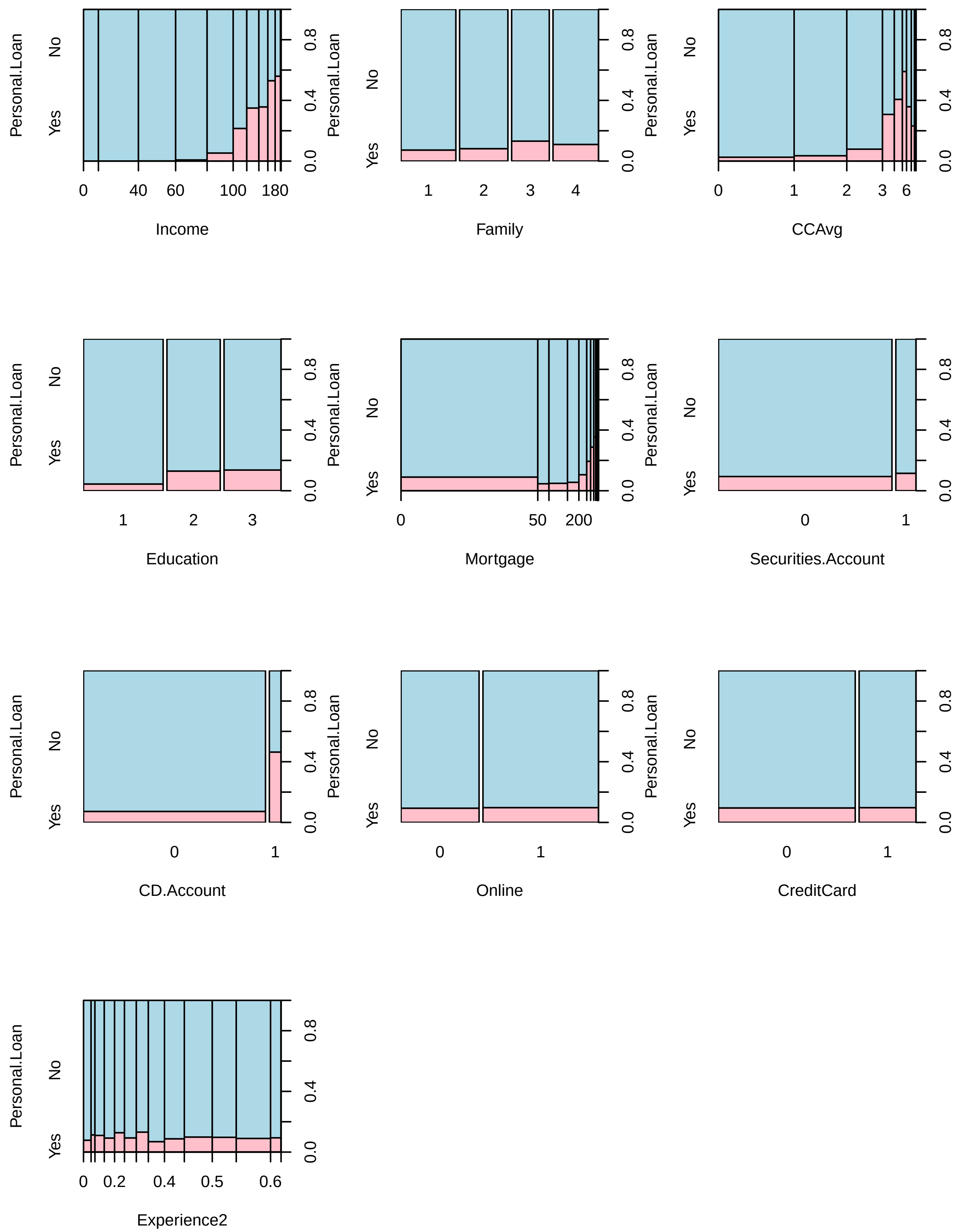
Very small relationship

- ***Security account*** and Personal Loan seem to have slight relationship

No relationship with response

- ***Online and Credit card*** don't seem to have a relationship

Looking closer at each realationship to see if anything was missed



Model: Objective 1

```
# Train Test Split
set.seed(123)
index<-sample(1:dim(PersonalLoan)[1],round(.70 * dim(PersonalLoan)[1]))
train<-PersonalLoan[index,]
test<-PersonalLoan[-index,]


# Split Predict for lasso
dat.test.x = model.matrix(Personal.Loan ~ Income + Family + CCAvg-1 + Education + Securities.Account-1 +

dat.train.x = model.matrix(Personal.Loan ~ Income + Family + CCAvg-1 + Education + Securities.Account-1
dat.train.y = train$Personal.Loan
```

Performing model with all variables, some feature selection methods (forward, stepwise, LASSO) and another based on EDA + With the full model with all attributes it showed that the only important were Income, Family,CCavg, Education,Securites.Account, CD.Account, Online, CreditCard

- Once Stepwise was added to the full model it selected all of those that appeared as significant in the full model:Income, Family, CCAvg, Education, Securities.Account, CD.Account, Online, CreditCard. It also selected Experience2 but that one was not significant as we had seen in the EDA.
- When the forward model was added to the full model it selected the same thing as stepwise but included CCAVG which was significant and Mortgage which was not significant .
- As for LASSO it selected all of the attributes by Stepwise and included CCAvg.

```
[1] 0.0009615423
```

- Using different threholds

Changing the threshold

```
# Comparing stepwise with different threashold
print("Threashhold | Accuracy| Sensitivity| Specificy")
```

```
[1] "Threashhold | Accuracy| Sensitivity| Specificy"
```

```
saving=threshold(simpleLG,FALSE)
names(saving) <- c("0.3", "0.5", "0.7")
print("All_Attributes")
```

```
[1] "All_Attributes"
```

```
saving
```

```
$`0.3`
[1] 0.3000000 0.9573333 0.9747212 0.8064516
```

```
$`0.5`
[1] 0.5000000 0.9620000 0.9947955 0.6774194

$`0.7`
[1] 0.7000000 0.9553333 0.9992565 0.5741935
```

```r
# Comparing simpleLG with different threashold
print("Threashhold | Accuracy| Sensitivity| Specificy")
```

```
[1] "Threashhold | Accuracy| Sensitivity| Specificy"
```

```r
saving=threshold(stepWiseAIC,FALSE)
names(saving) <- c("0.3", "0.5", "0.7")
print("StepWiseAIC")
```

```
[1] "StepWiseAIC"
```

```r
print(saving)
```

```
$`0.3`
[1] 0.3000000 0.9573333 0.9747212 0.8064516

$`0.5`
[1] 0.5000000 0.9620000 0.9947955 0.6774194

$`0.7`
[1] 0.7000000 0.9553333 0.9992565 0.5741935
```

```r
# Comparing ,ForwardModel with different threashold
print("Threashhold | Accuracy| Sensitivity| Specificy")
```

```
[1] "Threashhold | Accuracy| Sensitivity| Specificy"
```

```r
saving=threshold(ForwardModel,FALSE)
names(saving) <- c("0.3", "0.5", "0.7")
print("ForwardModel")
```

```
[1] "ForwardModel"
```

```r
print(saving)
```

```
$`0.3`
[1] 0.3000000 0.9573333 0.9747212 0.8064516

$`0.5`
[1] 0.5000000 0.9620000 0.9947955 0.6774194

$`0.7`
[1] 0.7000000 0.9553333 0.9992565 0.5741935
```

```r
# Comparing LASSOmodel with different threashold
print("Threashhold | Accuracy| Sensitivity| Specificy")
```

```
[1] "Threashhold | Accuracy| Sensitivity| Specificy"
```

```r
saving=threshold(LASSOmodel,TRUE)
names(saving) <- c("0.3", "0.5", "0.7")
print("LASSO")
```

```
[1] "LASSO"
```

```r
print(saving)
```

```
$`0.3`
[1] 0.300000 0.822000 2.106320 0.283871

$`0.5`
[1] 0.5000000 0.8414286 2.1635688 0.2258065

$`0.7`
[1] 0.7000000 0.8528571 2.2000000 0.1677419
```

```r
# Comparingintuition,eda with different threashold
print("Threashhold | Accuracy| Sensitivity| Specificy")
```

```
[1] "Threashhold | Accuracy| Sensitivity| Specificy"
```

```r
saving=threshold(intuition,FALSE)
names(saving) <- c("0.3", "0.5", "0.7")
print("Intuition")
```

```
[1] "Intuition"
```

```r
print(saving)
```

```
$`0.3`
[1] 0.3000000 0.9573333 0.9747212 0.8064516

$`0.5`
[1] 0.5000000 0.9620000 0.9947955 0.6774194

$`0.7`
[1] 0.7000000 0.9553333 0.9992565 0.5741935
```

```r
# Comparing eda with different threashold
print("Threashhold | Accuracy| Sensitivity| Specificy")
```

```
[1] "Threashhold | Accuracy| Sensitivity| Specificy"
```

```r
saving=threshold(eda,FALSE)
names(saving) <- c("0.3", "0.5", "0.7")
print("EDA")
```

```
[1] "EDA"
```

```r
print(saving)
```

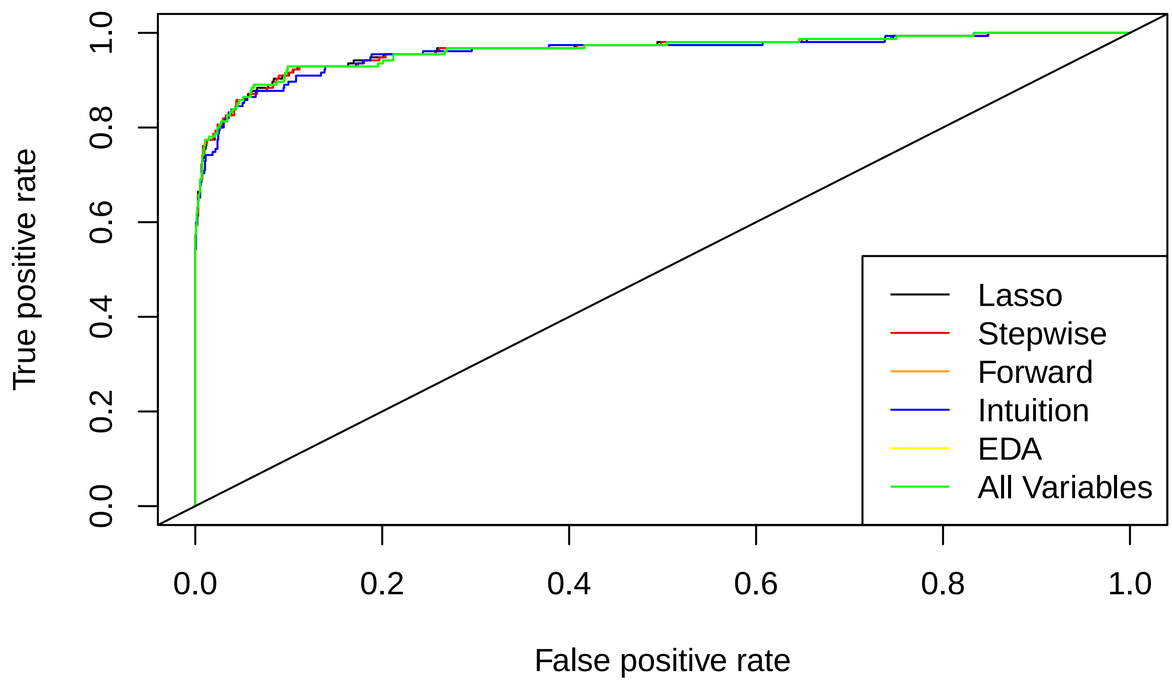
```
$`0.3`
[1] 0.3000000 0.9573333 0.9747212 0.8064516

$`0.5`
[1] 0.5000000 0.9620000 0.9947955 0.6774194

$`0.7`
[1] 0.7000000 0.9553333 0.9992565 0.5741935
```

Choosing 0.3 threshold based of the threashold results. Criterion Comparison of all models

```r
all_results
```

```
  StatisticResults  Full_Model   Step_Wise Forward_Model LASSO_model
1              AIC 820.3981611 818.0737507   820.3981611   0.0000000
2              BIC 906.6454165 891.9999697   906.6454165   0.0000000
3         Accuracy   0.9566667   0.9573333     0.9566667   0.9560000
4      Sensitivity   0.9747212   0.9747212     0.9747212   0.9747212
5      Specificity   0.8000000   0.8064516     0.8000000   0.7935484
    Intuition         EDA
1 836.8792021 844.4809292
2 892.3238663 899.9255934
3   0.9553333   0.9533333
4   0.9762082   0.9739777
5   0.7741935   0.7741935
```

The ROC of models

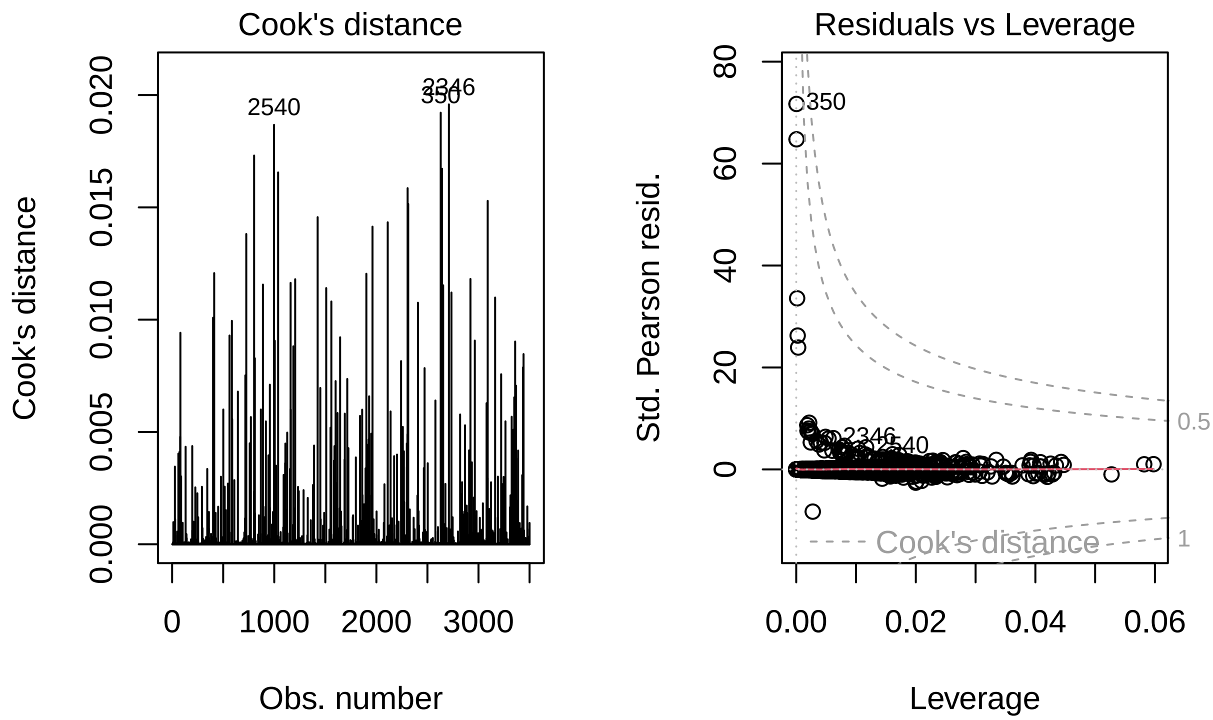


Verify Proportions in test and train manually + Distribution in train and test do represent that of the whole data

```
[1] "All data"

   No   Yes 
0.904 0.096 

[1] "Train"

        No        Yes 
0.90714286 0.09285714 

[1] "Test"

       No       Yes 
0.8966667 0.1033333 
```

- Assumptions via PLOTS of selected model, Stepwise and checking VIF
- Plots look normal and there seems to be multicolinarity among variables based on VIF

```
par(mfrow = c(1, 2))
#Cook's Distance Plot
plot(stepWiseAIC, 4)
#Standardized Residuals vs Leverage
plot(stepWiseAIC, 5)
```




```r
par(mfrow = c(1, 1))
# vifs
vif(stepWiseAIC)
```

	GVIF	Df	GVIF^(1/(2*Df))
Income	2.940809	1	1.714879
Family	1.529409	3	1.073381
CCAvg	1.516750	1	1.231564
Education	2.323075	2	1.234570
Securities.Account	1.291648	1	1.136507
CD.Account	1.936714	1	1.391659
Online	1.143566	1	1.069376
CreditCard	1.383602	1	1.176266

Conclusion from Part 1f

- The best model was step setting the threshold to 0.3 it gave an sensitivity of 94 and specificity of 72
- Due to the imbalance of amount of people with loan and without loan we do see we do see that the model favors no loan due to it but 72 compared to the 55 specificity was a great increase. This model is about trying to predict those who will say yes to Loan therefore Specificity is important.
- The attributes found useful were : Income, Family,CCavg, Education,Securites.Account, CD.Account, Online, CreditCard
- The threshold was set to 0.3 and it lead to a Sensitivity of 0.96 and specificy of 0.71
- These were variables seen in the EDA as related to the loan.

- Coefficients results: For every unit increase in income the odd of getting a loan are e^1.06 times higher For every unit increase in Family the odd of getting a loan are e^0.698209 times higher For every unit increase in CCAvg the odd of getting a loan are e^0.120635 times higher For every unit increase in Education the odd of getting a loan are e^1.713690 times higher For every unit increase in Securities.Account 1 the odd of getting a loan are e^-0.937183 times less likely For every unit increase in CD.Account1 the odd of getting a loan are e^3.840892 times higher For every unit increase in Online1 the odd of getting a loan are e^-0.673230 times less likely For every unit increase in CreditCard1 the odd of getting a loan are e^ -1.122701 times higher

##

EXTRA code that has been commented out

Re-run all models but removing age and Logging mortgage