Binary Classification with Logistic Regression

Serra Uzun, MSDS_410 FALL 2020

11/08/2020

Introduction

Logistic regression, unlike the linear regression, uses the logistic function that enables predictive modeling with datasets where response variable is binary. The probability of a particular event or class happening. The response (dependent) variable in the Universal Bank dataset that will be used in the analysis is binary. The binary response variable, Personal Loan, indicates whether the observation has obtained a Personal Loan from Universal Bank or not and comes with predictor variables that are related to an individual's personal information, accounts, and finances. The following report will conduct a detailed exploratory data analysis followed by multiple generalized linear models for extensive assessment and comparison to choose the best possible model for the most accurate true positive response rate.

1. The Dataset and Data Split

1.1 The Dataset

The Universal Bank dataset consists of 14 variables, that are all numeric, and 5000 observations. Out of 14 variables, there is one response variable, PersonalLoan, and 13 predictor variables related to the individual's personal information and finances. Below is the list of variables and their types:

- ID (continuous)
- Age (continuous)
- Experience (continuous)
- Income (continuous)
- ZIP Code (discrete)
- Family (discrete)
- CCAvg (continuous)
- Education (discrete)
- Mortgage (continuous)
- Personal Loan (binary)
- Securities Account (binary)
- CDAccount (binary)
- Online (binary)
- Credit Card (binary)

The variables ID and ZIP Code are removed from the dataset pre train and test data split as they are not predictors of Personal Loan.

1.2 The Train/Test Split

Before beginning our predictive model, we will perform a 70/30 split on our dataset consisting of 5,000 observations. This split aims to 'train' our model with 70 percent of the sample data and then 'test' it with the test dataset, which is 30 percent of the sample data. With the 70/30 split, we continue our model with a train dataset of 3,492 observations and a test dataset of 1,508 observations. While we will mainly use the train data for our models, we will be tapping back to the test data to cross-validate our models in the following sections.

2. Exploratory Data Analysis

Per our initial review of the Personal Loan variable, we find 3,148 observations without Personal Loan (0) and 344 observations with Personal Loan (1), %90, and %10 of the train dataset, respectively. We will start our EDA by looking into some of the continuous variables that are Income, Experience, Mortgage, and Age.

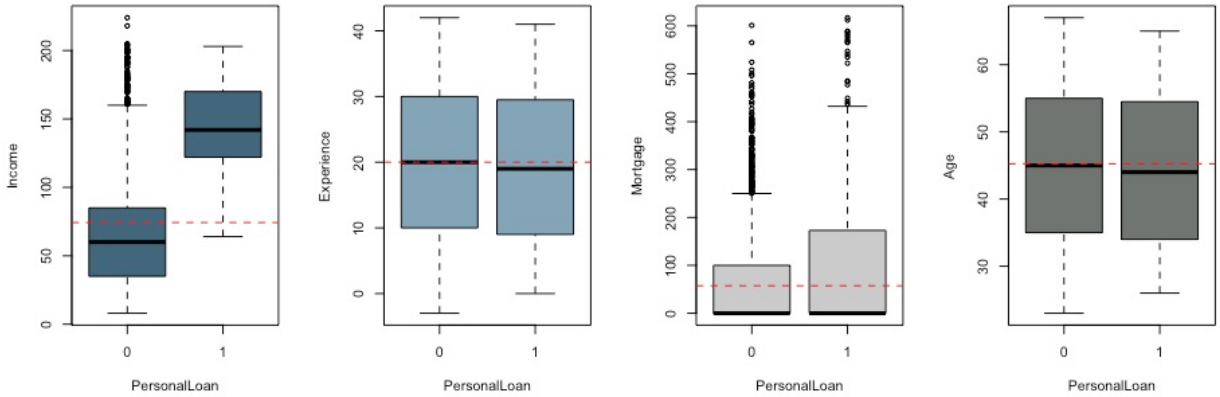

Figure 1: Boxplots for Income, Experience, Mortgage and Age

The boxplots for Income and Mortgage presented in Figure 1 show that a significant number of observations are above the maximum line while the mean is on the lower side of the plot in boxplots for Income where Personal Loan is 0, and for Mortgage. While these observations are passed the maximum line in the boxplot, they are not considered outliers, yet indicate that we are likely to have also a large number of low values within the variable, which is also suggested through the mean, presented in plots as the dashed red line. The boxplots for Experience and Age are shown to be very similar, which may indicate similarities in pattern and distribution and high correlation between the variables. The mean line falls around where the mean for each boxplot is, showing no significant skews towards either end of the spectrum. Next, we will explore the binary variables in our dataset against our response variable, which is also binary.

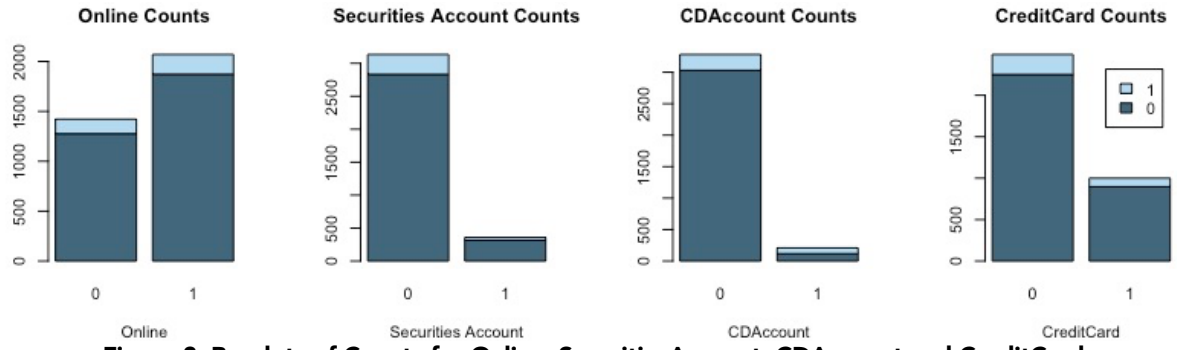

Figure 2: Barplots of Counts for Online, SecuritiesAccount, CDAccount and CreditCard

Figure 2 exhibits the counts for Online, SecuritiesAccount, CDAccount, and CreditCard in the context of Personal Loan, which is presented within the bar as different colors. Online Account holders count is the only one where the account holders are a larger group than non-account holders, whereas when we look at the remaining bar plots for Securities and CD account and Credit Cardholders, we see a relatively similar pattern throughout. The Personal Loan response variable is positive at a greater degree when the observation doesn't have the accounts listed above, except for Online Account. In the Online Account bar plot, we can see that Personal Loan of 1is presented in a comparatively equal proportion for both Online account holders and non-holders. So, we can conclude that Personal Loan response shows a similar pattern throughout all binary variables for accounts that have more non-account holders than the account holders, suggesting that potentially not having Securities, CD account or Credit Card are not significant indicators of the response variable, Personal Loan.

As the final step of our EDA, we generate plots for discrete variables but also gather variables such as CCAvg, Age, and Experience within 'bins' to see an indication of important patterns within the dataset per bin/group characteristic.

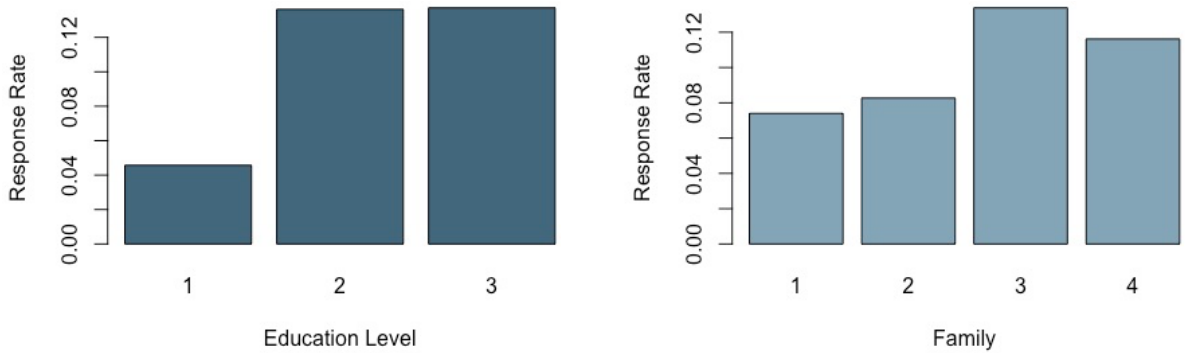

Figure 3: Barplots for Education Level and Family

Firstly, as we review Figure 3 with the bar plots for each unique Education Level and Family, we see that observations with an Education Level of 2 or higher are more likely to get a Personal Loan. This suggests a clear correlation between the likelihood of people with higher education levels getting a Personal Loan. On the other hand, when we look at the response rate per each discrete value under the Family variable, we see that Family type 3 has the highest response rate, followed by 4, 2, and finally 1. As the response rate drops between 3 and 4, we can't conclude as the Family increases, the response rate also increases, yet we can state that through our train dataset observations, Family type 3 or 4 is more likely to get a Personal Loan than 1 and 2.

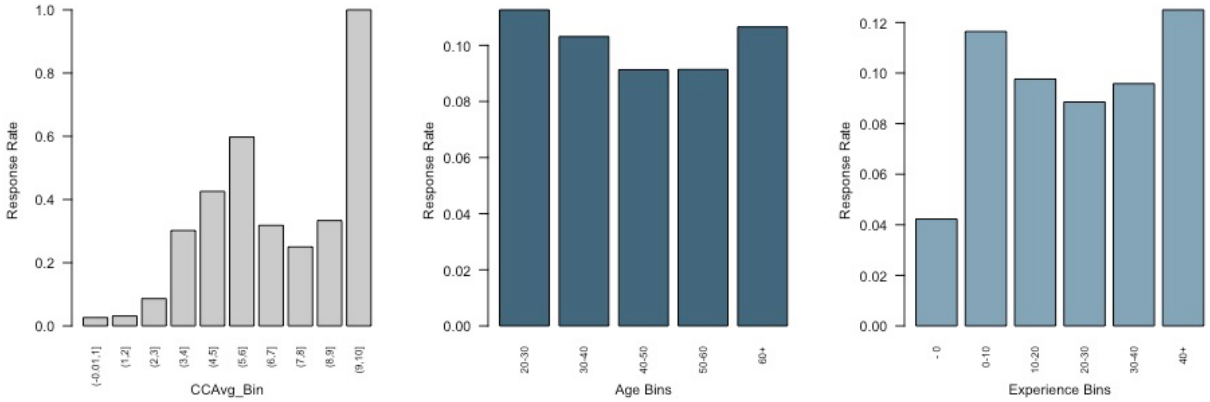

Figure 4: Barplots for Different Age, Experience and CCAvg Bins

Finally, when we review the bins that we created for Age and Experience, presented in Figure 4, we see a similar pattern that is a high response rate for bins with lower Age and Experience, then a decrease within the middle section of our plots and again an increase as the Age and Experience reaches to the bins with highest values. So, our plots for Age and Experience suggest that people who fall under the age groups 20-30, 30-40 and 60+, and Experience group 0-10 and 40+ have the highest response rate for Personal Loan. Lastly, as we divide CCAvg values to 10 bins, we see that the highest response rate by far is for CCAvg between 9 and 10, followed by between 4 and 6. Even though the bar plot for CCAvg Bins do not suggest a clear pattern, it shows us that observations with CCAvg between 9 and 10 are most likely to have a high response rate.

On a final note following the EDA presented above, when a correlation analysis is conducted on the dataset, we see that the most correlated variable with the response variable is Income. Some other noticeable correlations are seen between Age and Experience with 0.99, CCAvg and Income with 0.65, and between CCAvg and Personal Loan with 0.37.

3. Generalized Linear Models for Logistic Regression

At the start of our modeling process, we fit a naïve model as a baseline for future model comparisons. The baseline naïve model will run the binominal generalized linear model using Income, CCAvg, CDAccount, Education (as a factor), Family and SecuritiesAccount as predictor variables, and our response variable, PersonalLoan. Aside from the typical model summary, we will be producing the ROC Curve plot, which shows the trade-off between sensitivity and specificity, AUC (Area Under the ROC Curve) metric, which shows us the performance across all possible classification thresholds, and finally confusion matrix.

3.1 Baseline Naïve Model

The summary statistics of the naïve model we ran with the train dataset, with 3,492 observations, are presented in Figure 5 below. The results show a Log Likelihood of -439.54 and an AIC score of 895.07. While the AIC score is a crucial parameter, it does not add value to our analysis on its own yet will be used to compare models in the following steps. In addition, the naïve model has 8 Fisher Scoring iterations, which represents the optimal number of iterations for the maximum likelihood.

	Model.1		
	Dependent variable:		
	PersonalLoan		
Income	0.06***	Family	0.57***
	(0.003)		(0.09)
CCAvg	0.16***	SecuritiesAccount	-0.57*
	(0.05)		(0.34)
CDAccount	2.49***	Constant	-13.51***
	(0.33)		(0.66)
factor(Education)2	4.08***	Observations	3,492
	(0.31)	Log Likelihood	-439.54
factor(Education)3	4.10***	Akaike Inf. Crit.	895.07
	(0.31)	*Note:*	*p<0.1; **p<0.05; ***p<0.01

Figure 5: Model 1 (Naïve Model) Summary

The ROC Curve plot presented in Figure 6 shows the Model 1 curve to have an area of 0.9584 (95.84%) under the curve. Furthermore, if we review the confusion matrix obtained for Model 1 above Figure 6, we can see that our baseline model has a true negative rate of 94% and a true positive rate of 85%. This indicates that our model is better predicting the Personal Loan outcome of 0 than the outcome of 1.

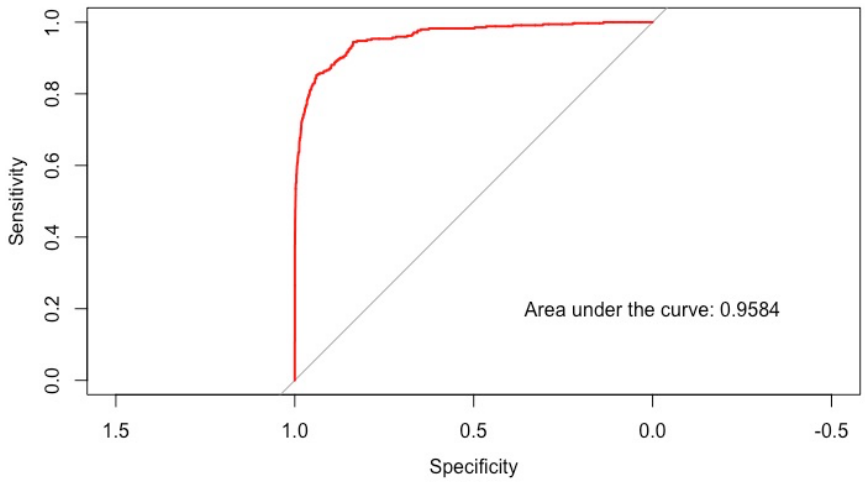

Figure 6: ROC Curve Plot for Model #1

Model #1 (Baseline)

	0	1
0	0.94	0.06
1	0.15	0.85

Table 1: Model #1 Confusion Matrix

3.2 Model #2 with Stepwise Variable Selection

For the next model, we will conduct using the same techniques used in section 3.1. We will first conduct a stepwise regression model to select variables for the Generalized Linear Model #2. The stepAIC() model we ran with the stepwise regression presented us with the results shown below:

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

		Df	Deviance	AIC
	<none>		946.21	966.21
+	Age	1	944.7	966.7
-	Experience	1	949.12	967.12
+	Mortgage	1	945.77	967.77
-	CCAvg	1	952.93	970.93
-	SecuritiesAccount	1	953.68	971.68
-	CreditCard	1	959.07	977.07
-	Online	1	962.81	980.81
-	Family	1	1012.09	1030.09
-	CDAccount	1	1037.74	1055.74
-	Education	1	1181.71	1199.71
-	Income	1	1507.74	1525.74

Table 2: Stepwise Regression Model Results

Through the Stepwise Variable Selection model, we determine that the variables that are optimal for modeling with the Universal Bank train dataset are Income, Education, CDAccount, Family, Online, CreditCard, SecuritiesAccount, CCAvg, and Experience. Predictor variables Mortgage and Age are not included in the list of selected variables. They present relatively low AIC scores that suggest minimal information loss than other predictor variables if they were not included. We will use these selected variables listed above for our Model #2.

Model.2	
	Dependent variable:
	PersonalLoan
Income	0.06***
	(0.004)
factor(Education)2	4.11***
	(0.32)
factor(Education)3	4.14***
	(0.32)
CDAccount	3.40***
	(0.41)
Family	0.57***
	(0.09)
Online	-0.81***
	(0.19)
CreditCard	-0.74***
	(0.24)
SecuritiesAccount	-0.78**
	(0.36)
CCAvg	0.17***
	(0.05)
Experience	0.01
	(0.01)
Constant	-13.28***
	(0.70)
Observations	3,492
Log Likelihood	-426.12
Akaike Inf. Crit.	874.24
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 7: Model #2 Summary

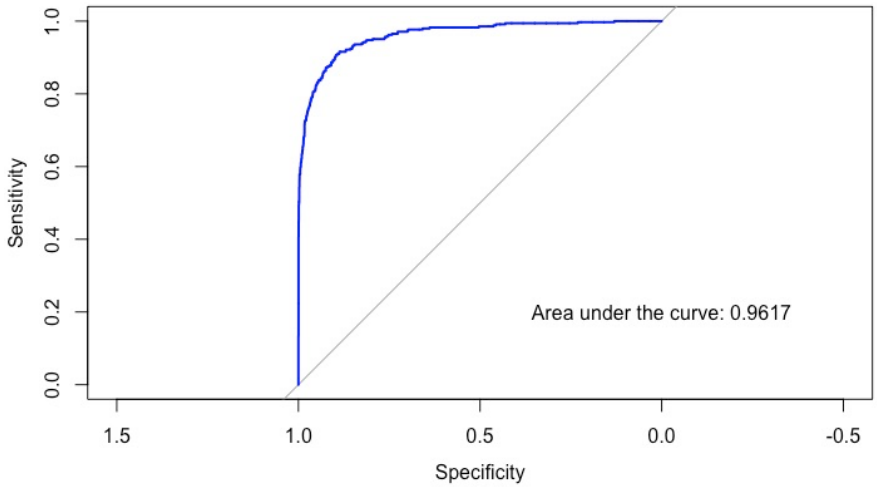

Figure 8: ROC Curve Plot for Model #2

Model #2 (Stepwise)

	0	1
0	0.90	0.10
1	0.09	0.91

Table 3: Model #2 Confusion Matrix

Model #2 results presented above in Table 3, Figure 7 and 8 show the model AIC score as 874.24, AUC as 0.9617 (96.17%). The AUC of 0.9617 is slightly higher than our baseline model. The confusion matrix for model #2 that we conducted with variables selected through the Stepwise Variable Selection Model shows that the true negative rate is 90%, and the true positive rate is 91%.

3.3 Model #3 with Random Selected Variables

After concluding Model #2 where we got the AIC score for each variable through our stepwise variable selection model and as we now have an idea of which variables are more significant than the others, we will be conducting a third model (Model #3) with a mix of variables that we found to be strong predictors of Personal Loan through our EDA. Through our EDA, we know that rising Education levels, whether they are Online Account holders or not, their Age group, Income, and CCAvg within certain groups (between 4-6 and 9-10) have shown an effect on the Personal Loan obtained by the individual. Another indicator of a good predictor variable we observed through our EDA was for Income. The boxplot for Income showed different boxplots for possible response variable outcome, that suggests a solid predictor. By this approach, we will be using Education Level (as Factor), Age, Income, Online, and CCAvg as the predictor variables in Model #3.

Model.3

	Dependent variable:
	PersonalLoan
Income	0.06***
	(0.003)
Online	-0.33*
	(0.17)
Age	0.01
	(0.01)
factor(Education)2	4.37***
	(0.29)
factor(Education)3	4.31***
	(0.29)
CCAvg	0.18***
	(0.05)
factor(Education)3	4.31***
	(0.29)
CCAvg	0.18***
	(0.05)
Constant	-12.13***
	(0.66)
Observations	3,492
Log Likelihood	-492.43
Akaike Inf. Crit.	998.85
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 9: Model #3 Summary

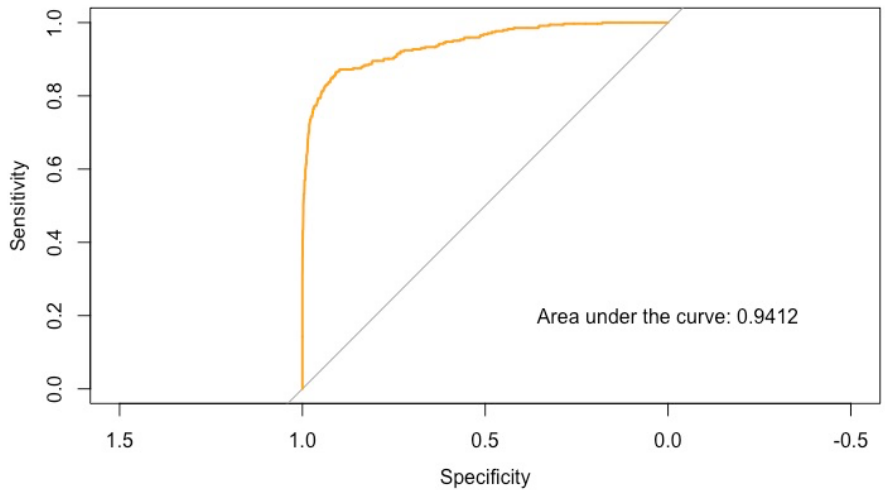

Figure 10: ROC Curve Plot for Model #3

Model #3 (Random)

	0	1
0	0.91	0.09
1	0.14	0.86

Table 4: Model #3 Confusion Matrix

Model #3 has an AIC score of 998.85 and an AUC of 0.9412. From Figure 9 where the Generalized Linear Model summary is presented, we can see that two variables have p-value above the alpha level, which are Age and Online. Finally, Table 4 showed the confusion matrix indicates a 91% true negative and 86% true positive accuracy.

4. Model Comparisons

So far, we conducted three models, Model #1 as the Baseline Naïve Model, Model #2 where we ran with the predictor variables we have determined as optimal through Stepwise Variable Selection Model and finally Model #3 where we used self-determined variables through our EDA. Firstly, we will start by comparing the AIC scores of each model.

	Model #1 (Baseline)	Model #2	Model #3
AIC Score	895.07	874.24	998.85

Table 5: Models AIC Score Comparison

Table 5 above presents the AIC scores for each of the models. Per these results, we can see that Model #3 is our worst performing model with an AIC of 998.85. On the other hand, we see that Model #2, where we used variables selected through Stepwise Regression, has the lowest AIC score, 874.24, out of all three models we ran with the train dataset. Hence, the best performing model when models are compared through the AIC scores. Following, we will look at the ROC Curve plots and AUC of each model for comparison.

	Model #1 (Baseline)	Model #2	Model #3
AUC (Area Under the Curve)	0.9584	0.9617	0.9412

Table 6: Models AUC Score Comparison

Both through Table 6 and Figure 11, we can see that the model with the highest AUC is Model #2 with 0.9617. The ROC Curve plot and AUCs' results present the same ranking of models, Model #2 being the best performing, Model #3 being the least successful, and Model #1 in the middle.

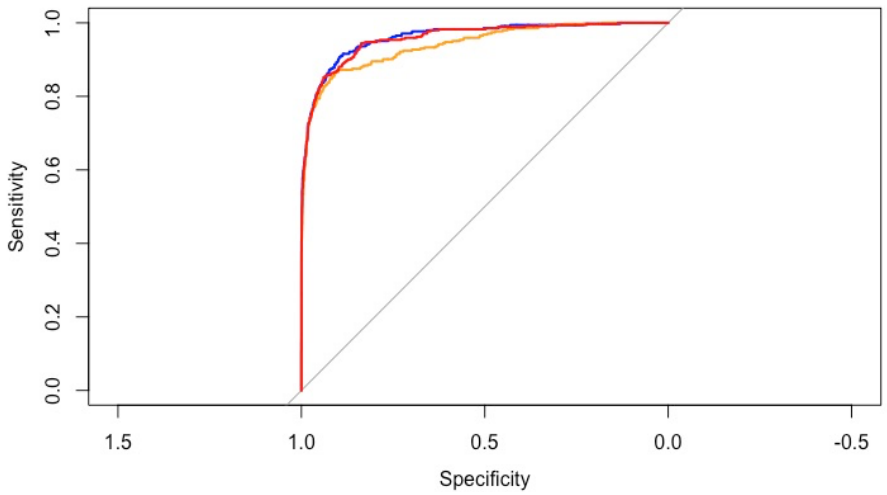

Figure 11: ROC Curve Plot for All Models

Finally, we will compare the confusion matrix of each model to determine which model best predicts the true negative and true positive.

	Model #1 (Baseline)		Model #2		Model #3	
	0	1	0	1	0	1
0	0.94	0.06	0.9	0.1	0.91	0.09
1	0.15	0.85	0.09	0.91	0.14	0.86

Table 7: Confusion Matrix of All Models

Per Table 7, the model with the highest true negative accuracy is Model #1, our baseline model. On the other hand, the model with the highest true positive accuracy is Model #2, which is the model with the variables selected through Stepwise Regression. Our analysis aims to get the model that gives us the most accurate true positive, the best model is Model #2. Out of all the Generalized Linear Models we have conducted using the train dataset, Model #2 is the best performing model throughout all our criteria from AIC to AUC and Confusion Matrix true positive accuracy.

That said, per Stepwise Regression and Model #2, we determine that Income is the most significant predictor in predicting Personal Loan response variables, followed by Education, CDAccount, and Family. Referring back to section 2, our EDA has suggested the possibility of this outcome through the Income boxplot and Education and Family bar plots.

Next, to assess these three models' predictive accuracy, we will be conducting the same models using the test (out-sample) dataset.

5. Predictive Accuracy

As the final step of our analysis, we will bring in the test dataset representing 30% of the Universal Bank dataset that we initially split in Section 1. The test data set consists of 1,508 observations. All three models are once again conducted using the test dataset and output the AUC results presented in Table 8 below:

	Model #1 (Baseline)	Model #2	Model #3
AUC (Area Under the Curve)	0.9591	0.9635	0.9311

Table 8: Models ran with Test Dataset AUC Score Comparison

Upon reviewing each model's AUC results, we see that the switch from the train dataset to the test dataset did not generate any model ranking changes. Model #2 with 0.9635 AUC is the highest performing model, while Model #3 with 0.9311 is the lowest-performing model. Model #1, the baseline model, performed much better than Model #3 with an AUC of 0.9591, relatively close to the best performing model, Model #2. Finally, we will look at the confusion matrix of each model, presented in Table 9.

	Model #1 (Baseline)		**Model #2**		**Model #3**	
	0	1	0	1	0	1
0	0.96	0.04	0.91	0.09	0.94	0.06
1	0.15	0.85	0.1	0.90	0.17	0.83

Table 9: Confusion Matrix with Test Dataset of All Models

Table 9 shows consistent results with previous model comparisons, and all results obtained both through train and test dataset. We see that with the test dataset, the model ranking remained the same, Model #2 being the best performing with the true positive accuracy of 90%. Like the in-sample model results, Model #1 is the model with the highest true negative rate. These results finalize our ranking of models as Model #2 being the best, Model #1 as the mediocre, and Model #3 as the worst-performing.

Conclusion

As a conclusion of our analysis, Model #2 has consistently given the best results and highest accuracies both with in-sample and out-sample datasets. The predictor variables Income, Education, CDAccount, Family, Online, CreditCard, SecuritiesAccount, CCAvg, and Experience are the optimal variables to predict the Personal Loan response variable.

```
# Assignment_8
# Serra Uzun
# 11.08.2020

install.packages("ggplot2")
library(ggplot2)
library(stargazer)
library(plyr)
library(psych)
library(lessR)
library(corrplot)
library(pROC)
out.path <- '/Users/serrauzun/Desktop/MSDS_410_Supervised/Assignment #8';

unibank <- UniversalBank

dim(unibank)
str(unibank)
head(unibank)

#remove ID and ZIPCODE from main dataset
colnames(unibank)
unibank <- unibank[,-c(1,5)]
head(unibank)

set.seed(12345)
unibank$u <- runif(n=dim(unibank)[1],min=0,max=1)

# Create train/test split;
train.df <- subset(unibank, u<0.70)
test.df <- subset(unibank, u>=0.70)
# Check your data split. The sum of the parts should equal the whole. # Do your totals add up?
dim(unibank)[1]
dim(train.df)[1]
dim(test.df)[1]
dim(train.df)[1] + dim(test.df)[1]

#####################################################################
# Response rates for discrete variables
#####################################################################

summary(train.df)
cor(train.df)
corrplot(cor(train.df))


PersonalLoan <- train.df$PersonalLoan
```

```
Age <- train.df$Age
Online <- train.df$Online
SecuritiesAccount <- train.df$SecuritiesAccount
CDAccount <- train.df$CDAccount
CreditCard <- train.df$CreditCard

par(mfrow=c(1,4))
online_table <- table(PersonalLoan, Online)
barplot(online_table, main="Online Counts",
      xlab="Online", col=c('#42687C','#B3DAF1'))
online_table
Securities_table <- table(PersonalLoan, SecuritiesAccount)
barplot(Securities_table, main="Securities Account Counts",
      xlab="Securities Account", col=c('#42687C','#B3DAF1'))


CDAccount_table <- table(PersonalLoan, CDAccount)
barplot(CDAccount_table, main="CDAccount Counts",
      xlab="CDAccount", col=c('#42687C','#B3DAF1'))


CreditCard_table <- table(PersonalLoan, CreditCard)
barplot(CreditCard_table, main="CreditCard Counts",
      xlab="CreditCard", col=c('#42687C','#B3DAF1'),
      legend = rownames(CreditCard_table))

table(train.df$Online)
multi.hist(train.df[,-c(13)])
colnames(train.df)

colnames(train.df)
par(mfrow=c(1,4))
boxplot(Income ~ PersonalLoan, data = train.df, col='#42687C',xlab='PersonalLoan', ylab='Income' )
abline(h=mean(train.df$Income), col='red',lwd=1 , lty = 2 )
boxplot(Experience ~ PersonalLoan, data = train.df, col='#84A5B8',xlab='PersonalLoan',
ylab='Experience' )
abline(h=mean(train.df$Experience), col='red',lwd=1 , lty = 2 )
boxplot(Mortgage ~ PersonalLoan, data = train.df, col='#CBCBCB',xlab='PersonalLoan',
ylab='Mortgage' )
abline(h=mean(train.df$Mortgage), col='red',lwd=1 , lty = 2 )
boxplot(Age ~ PersonalLoan, data = train.df, col='#707571',xlab='PersonalLoan', ylab='Age' )
abline(h=mean(train.df$Age), col='red',lwd=1 , lty = 2 )


par(mfrow=c(1,2))
response.Education <-
aggregate(train.df$PersonalLoan,by=list(Education=train.df$Education),FUN=mean)
```

```
barplot(height=response.Education$x,names.arg=response.Education$Education,xlab='Education
Level',ylab='Response Rate', col = '#42687C')

response.Family <- aggregate(train.df$PersonalLoan,by=list(Family=train.df$Family),FUN=mean)
barplot(height=response.Family$x,names.arg=response.Family$Family, xlab='Family',ylab='Response
Rate', col = '#84A5B8')

######################
# Bins
######################
par(mfrow=c(1,3))
train.df$CCAvg_Bins <- cut(train.df$CCAvg,breaks=10)
table(train.df$CCAvg_Bins)
response.CCAvg_Bins <-
aggregate(train.df$PersonalLoan,by=list(CCAvg_Bins=train.df$CCAvg_Bins),FUN=mean)
barplot(height=response.CCAvg_Bins$x,names.arg=response.CCAvg_Bins$CCAvg_Bins,xlab='CCAvg_
Bin',ylab='Response Rate',las=2,cex.names=0.75, col = '#CBCBCB')

train.df$Age_Bins <- ifelse(Age <= 30, "20-30",ifelse(Age > 30 & Age <= 40, "30-40", ifelse(Age > 40 &
Age <= 50, "40-50",ifelse(Age > 50 & Age <= 60, "50-60",ifelse(Age > 60,"60+","NA")))))
table(train.df$Age_Bins)
response.Age_Bins <-
aggregate(train.df$PersonalLoan,by=list(Age_Bins=train.df$Age_Bins),FUN=mean)
barplot(height=response.Age_Bins$x,names.arg=response.Age_Bins$Age_Bins,xlab='Age
Bins',ylab='Response Rate',las=2,cex.names=0.75 , col = '#42687C')

Experience <- train.df$Experience
train.df$Exp_Bins <- ifelse(Experience <= 0, "- 0",ifelse(Experience > 0 & Experience <= 10, "0-10",
ifelse(Experience > 10 & Experience <= 20, "10-20",ifelse(Experience > 20 & Experience <= 30, "20-
30",ifelse(Experience > 30 & Experience <= 40,"30-40",ifelse(Experience > 40,"40+","NA"))))))
table(train.df$Exp_Bins)
response.Exp_Bins <- aggregate(train.df$PersonalLoan,by=list(Exp_Bins=train.df$Exp_Bins),FUN=mean)
barplot(height=response.Exp_Bins$x,names.arg=response.Exp_Bins$Exp_Bins,xlab='Experience
Bins',ylab='Response Rate',las=2,cex.names=0.75, col = '#84A5B8')

#################################
#cleaning of dataset for modeling
colnames(train.df)
train.df <- train.df[,-c(13:16)]
colnames(train.df)
dim(train.df)

#################################
#Variable Selection
#################################

library(MASS)
```

```
#limits
upper.glm <- glm(PersonalLoan ~ ., data=train.df,family=c('binomial'))
lower.glm <- glm(PersonalLoan ~ 1, data=train.df,family=c('binomial'))
sqft.glm <- glm(PersonalLoan ~ Income, data=train.df, family=c('binomial'));

#forward
forward.glm <-
stepAIC(object=lower.glm,scope=list(upper=formula(upper.glm),lower=~1),direction=c('forward'))
summary(forward.glm)
stepAIC(forward.glm)

#backward
backward.glm <- stepAIC(object=upper.glm,direction=c('backward'))
summary(backward.glm)
stepAIC(backward.glm)

#stepwise
stepwise.glm <- stepAIC(object=sqft.glm,scope=list(upper=formula(upper.glm),lower=~1),
direction=c('both'));
summary(stepwise.glm)
stepwise.glm


##############################################
##############################################
# Model 1
model.1 <- glm(PersonalLoan ~ Income+CCAvg+CDAccount+factor(Education)+Family
            +SecuritiesAccount, data=train.df, family=c('binomial'))

summary(model.1)

file.name.1 <- 'Model.1.html';
stargazer(model.1, type=c('html'),out=paste(out.path,file.name.1,sep=''),
        title=c('Model.1'),
        align=TRUE, digits=2, digits.extra=2, initial.zero=TRUE, median=TRUE)

roc.1 <- roc(response=train.df$PersonalLoan, predictor=model.1$fitted.values)
print(roc.1)
auc.1 <- auc(roc.1)
auc.1
text1 = "Area under the curve: 0.9584"

par(mfrow=c(1,1))
plot(roc.1, col= 'red', lwd=2)
text(0.0,0.2,text1)
```

```
roc.specs <-
coords(roc=roc.1,x=c('best'),input=c('threshold','specificity','sensitivity'),ret=c('threshold','specificity','sen
sitivity'),as.list=TRUE)

train.df$ModelScores <- model.1$fitted.values;
train.df$classes <- ifelse(train.df$ModelScores>roc.specs$threshold,1,0);

# Rough confusion matrix using counts;
table(train.df$PersonalLoan, train.df$classes)
t1 <- table(train.df$PersonalLoan, train.df$classes);
r1 <- apply(t1,MARGIN=1,FUN=sum);
# Normalize confusion matrix to rates;
t1/r1

##############################################
##############################################
# Model 2
model.2 <- glm(PersonalLoan ~ Income + factor(Education) + CDAccount + Family + Online +
CreditCard + SecuritiesAccount + CCAvg + Experience, data=train.df, family=c('binomial'))

summary(model.2)

file.name.2 <- 'Model.2.html';
stargazer(model.2, type=c('html'),out=paste(out.path,file.name.2,sep=''),
        title=c('Model.2'),
        align=TRUE, digits=2, digits.extra=2, initial.zero=TRUE, median=TRUE)

roc.2 <- roc(response=train.df$PersonalLoan, predictor=model.2$fitted.values)
print(roc.2)
auc.2 <- auc(roc.2)
auc.2
text2 = "Area under the curve: 0.9617"

par(mfrow=c(1,1))
plot(roc.2, col= 'blue', lwd=2)
plot(roc.3, col= 'orange', lwd=2, add = TRUE)
plot(roc.1, col= 'red', lwd=2, add = TRUE)


roc.specs2 <-
coords(roc=roc.2,x=c('best'),input=c('threshold','specificity','sensitivity'),ret=c('threshold','specificity','sen
sitivity'),as.list=TRUE)

train.df$ModelScores2 <- model.2$fitted.values;
train.df$classes2 <- ifelse(train.df$ModelScores2>roc.specs2$threshold,1,0);

table(train.df$PersonalLoan, train.df$classes2)
```

```
t2 <- table(train.df$PersonalLoan, train.df$classes2);
r2 <- apply(t2,MARGIN=1,FUN=sum);
t2/r2


##############################################
##############################################
# Model 3
model.3 <- glm(PersonalLoan ~ Income+Online+Age+factor(Education)+CCAvg, data=train.df,
family=c('binomial'))

summary(model.3)

file.name.3 <- 'Model.3.html';
stargazer(model.3, type=c('html'),out=paste(out.path,file.name.3,sep=''),
        title=c('Model.3'),
        align=TRUE, digits=2, digits.extra=2, initial.zero=TRUE, median=TRUE)

roc.3 <- roc(response=train.df$PersonalLoan, predictor=model.3$fitted.values)
print(roc.3)
plot(roc.3)
auc.3 <- auc(roc.3);
auc.3

text3 = "Area under the curve: 0.9412"

par(mfrow=c(1,1))
plot(roc.3, col= 'orange', lwd=2)
text(0.0,0.2,text3)

roc.specs3 <-
coords(roc=roc.3,x=c('best'),input=c('threshold','specificity','sensitivity'),ret=c('threshold','specificity','sen
sitivity'),as.list=TRUE)

train.df$ModelScores3 <- model.3$fitted.values;
train.df$classes3 <- ifelse(train.df$ModelScores3>roc.specs3$threshold,1,0);

table(train.df$PersonalLoan, train.df$classes3)
t3 <- table(train.df$PersonalLoan, train.df$classes3);
r3 <- apply(t3,MARGIN=1,FUN=sum);
t3/r3

########
########
# Model 1 - TEST
model.1test <- glm(PersonalLoan ~ Income+CCAvg+CDAccount+factor(Education)+Family
            +SecuritiesAccount, data=test.df, family=c('binomial'))
```

```
summary(model.1test)

file.name.1test <- 'Model.1test.html';
stargazer(model.1test, type=c('html'),out=paste(out.path,file.name.1test,sep=''),
        title=c('Model.1test'),
        align=TRUE, digits=2, digits.extra=2, initial.zero=TRUE, median=TRUE)

roc.1test <- roc(response=test.df$PersonalLoan, predictor=model.1test$fitted.values)
print(roc.1test)
auc.1test <- auc(roc.1test)
auc.1test
text1test = "Area under the curve: 0.9591"

par(mfrow=c(1,1))
plot(roc.1test, col= 'red', lwd=2)
text(0.0,0.2,text1test)

roc.specs_test <-
coords(roc=roc.1test,x=c('best'),input=c('threshold','specificity','sensitivity'),ret=c('threshold','specificity','
sensitivity'),as.list=TRUE)

test.df$ModelScores <- model.1test$fitted.values;
test.df$classes <- ifelse(test.df$ModelScores>roc.specs_test$threshold,1,0)


table(test.df$PersonalLoan, test.df$classes)
t1test <- table(test.df$PersonalLoan, test.df$classes);
r1test <- apply(t1test,MARGIN=1,FUN=sum);

t1test/r1test

################################################
################################################
# Model 2 TEST
model.2test <- glm(PersonalLoan ~ Income + factor(Education) + CDAccount + Family + Online +
CreditCard + SecuritiesAccount + CCAvg + Experience, data=test.df, family=c('binomial'))

summary(model.2test)

file.name.2test <- 'Model.2test.html';
stargazer(model.2test, type=c('html'),out=paste(out.path,file.name.2test,sep=''),
        title=c('Model.2test'),
        align=TRUE, digits=2, digits.extra=2, initial.zero=TRUE, median=TRUE)

roc.2test <- roc(response=test.df$PersonalLoan, predictor=model.2test$fitted.values)
print(roc.2test)
```

```
auc.2test <- auc(roc.2test)
auc.2test
text2test = "Area under the curve: 0.9635"

par(mfrow=c(1,1))
plot(roc.2test, col= 'red', lwd=2)
text(0.0,0.2,text1test)

roc.specs2test <-
coords(roc=roc.2test,x=c('best'),input=c('threshold','specificity','sensitivity'),ret=c('threshold','specificity','
sensitivity'),as.list=TRUE)

test.df$ModelScores2 <- model.2test$fitted.values;
test.df$classes2 <- ifelse(test.df$ModelScores2>roc.specs2test$threshold,1,0);

table(test.df$PersonalLoan, test.df$classes2)
t2test <- table(test.df$PersonalLoan, test.df$classes2);
r2test <- apply(t2test,MARGIN=1,FUN=sum);
t2test/r2test

################################################
################################################
# Model 3 TEST
model.3test <- glm(PersonalLoan ~ Income+Online+Age+factor(Education)+CCAvg, data=test.df,
family=c('binomial'))

summary(model.3test)

file.name.3test <- 'Model.3test.html';
stargazer(model.3test, type=c('html'),out=paste(out.path,file.name.3test,sep=''),
        title=c('Model.3test'),
        align=TRUE, digits=2, digits.extra=2, initial.zero=TRUE, median=TRUE)

roc.3test <- roc(response=test.df$PersonalLoan, predictor=model.3test$fitted.values)
print(roc.3test)
plot(roc.3test)
auc.3test <- auc(roc.3test)
auc.3test
text3test = "Area under the curve: 0.9311"

par(mfrow=c(1,1))
plot(roc.3test, col= 'red', lwd=2)
text(0.0,0.2,text3test)
```

```
roc.specs3test <-
coords(roc=roc.3test,x=c('best'),input=c('threshold','specificity','sensitivity'),ret=c('threshold','specificity','
sensitivity'),as.list=TRUE)

test.df$ModelScores3 <- model.3test$fitted.values;
test.df$classes3 <- ifelse(test.df$ModelScores3>roc.specs3test$threshold,1,0);

table(test.df$PersonalLoan, test.df$classes3)
t3test <- table(test.df$PersonalLoan, test.df$classes3);
r3test <- apply(t3test,MARGIN=1,FUN=sum);
t3test/r3test
```