Universal Bank Analysis

Homer Kay

4/5/2018

```
## EDA
UB2<- read.csv("~/Desktop/Data Mining/UniversalBank.csv")
## Removing ZipCode and ID
UB <- UB2[,c(-1,-5)]
str(UB)
```

```
## 'data.frame':    5000 obs. of  12 variables:
##  $ Age               : int  25 45 39 35 35 37 53 50 35 34 ...
##  $ Experience        : int  1 19 15 9 8 13 27 24 10 9 ...
##  $ Income            : int  49 34 11 100 45 29 72 22 81 180 ...
##  $ Family            : int  4 3 1 1 4 4 2 1 3 1 ...
##  $ CCAvg             : num  1.6 1.5 1 2.7 1 0.4 1.5 0.3 0.6 8.9 ...
##  $ Education         : int  1 1 1 2 2 2 2 3 2 3 ...
##  $ Mortgage          : int  0 0 0 0 0 155 0 0 104 0 ...
##  $ Personal.Loan     : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ Securities.Account: int  1 1 0 0 0 0 0 0 0 0 ...
##  $ CD.Account        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Online            : int  0 0 0 0 0 1 1 0 1 0 ...
##  $ CreditCard        : int  0 0 0 0 1 0 0 1 0 0 ...
```

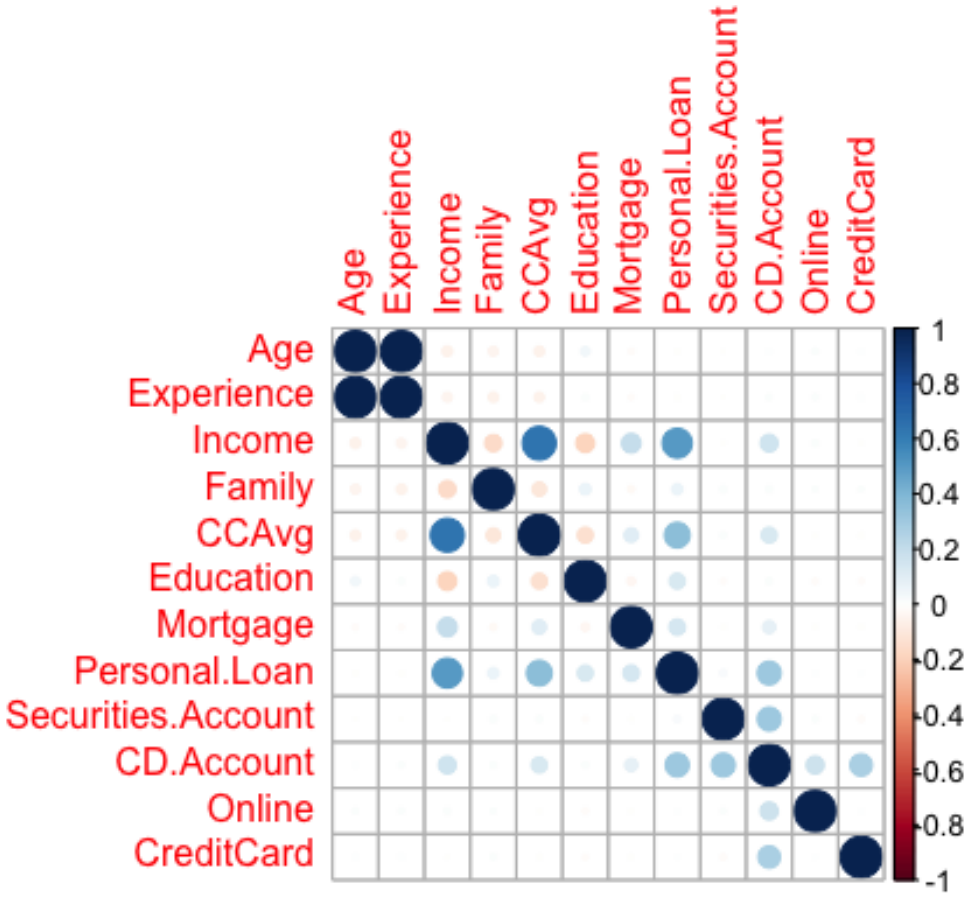
```
## STR shows which varaibles are numeric, integer, factor, etc.
library(corrplot)
```

```
## corrplot 0.84 loaded
```

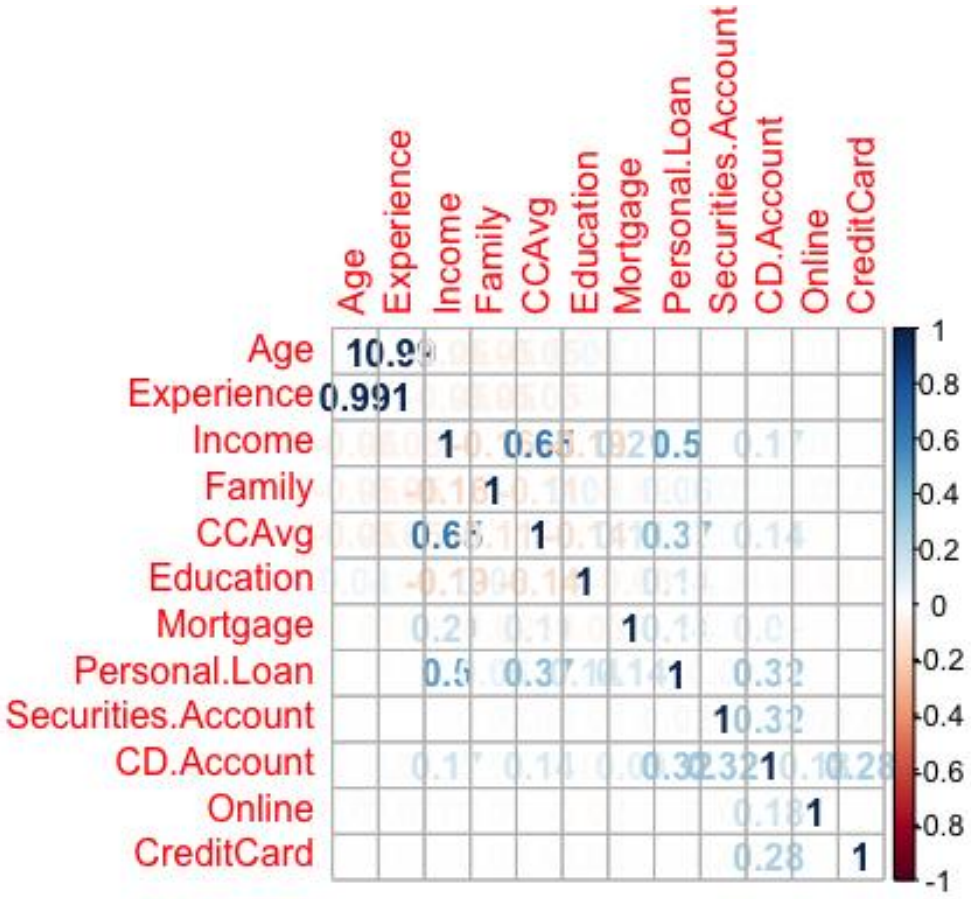
```
## Make correlation matrix with numeric predictors
CORR_MATRIX <- cor(UB[,])
CORR_MATRIX
```

```
##                              Age   Experience       Income       Family
## Age                 1.0000000000  0.994214857 -0.055268618 -0.04641766
## Experience          0.9942148570  1.000000000 -0.046574178 -0.05256315
## Income             -0.0552686182 -0.046574178  1.000000000 -0.15750079
## Family             -0.0464176636 -0.052563147 -0.157500785  1.00000000
## CCAvg              -0.0520121791 -0.050076511  0.645983670 -0.10927451
## Education           0.0413343834  0.013151813 -0.187524257  0.06492891
## Mortgage           -0.0125385869 -0.010581552  0.206806228 -0.02044493
## Personal.Loan      -0.0077256172 -0.007413098  0.502462292  0.06136704
## Securities.Account -0.0004362422 -0.001232134 -0.002616497  0.01999408
## CD.Account          0.0080425521  0.010353331  0.169738080  0.01411036
## Online              0.0137024021  0.013897900  0.014205919  0.01035404
## CreditCard          0.0076810368  0.008967447 -0.002385008  0.01158807
##                           CCAvg   Education     Mortgage Personal.Loan
```

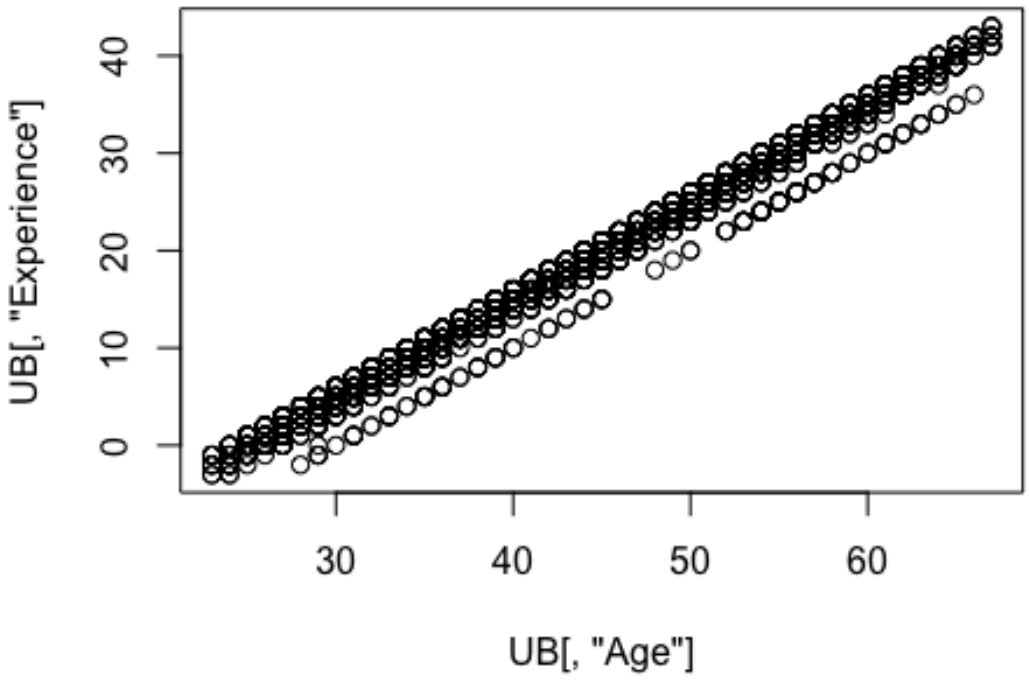
```
## Age                -0.052012179  0.04133438 -0.012538587  -0.007725617
## Experience         -0.050076511  0.01315181 -0.010581552  -0.007413098
## Income              0.645983670 -0.18752426  0.206806228   0.502462292
## Family             -0.109274506  0.06492891 -0.020444931   0.061367044
## CCAvg               1.000000000 -0.13612392  0.109904723   0.366888736
## Education          -0.136123922  1.00000000 -0.033327125   0.136721550
## Mortgage            0.109904723 -0.03332712  1.000000000   0.142095236
## Personal.Loan       0.366888736  0.13672155  0.142095236   1.000000000
## Securities.Account  0.015086311 -0.01081201 -0.005410970   0.021953882
## CD.Account          0.136533655  0.01393389  0.089311058   0.316354829
## Online             -0.003611009 -0.01500382 -0.005994898   0.006277815
## CreditCard         -0.006689494 -0.01101413 -0.007230919   0.002801509
##                    Securities.Account  CD.Account       Online
## Age                     -0.0004362422 0.008042552  0.013702402
## Experience              -0.0012321344 0.010353331  0.013897900
## Income                  -0.0026164967 0.169738080  0.014205919
## Family                   0.0199940798 0.014110365  0.010354036
## CCAvg                    0.0150863114 0.136533655 -0.003611009
## Education               -0.0108120136 0.013933888 -0.015003821
## Mortgage                -0.0054109700 0.089311058 -0.005994898
## Personal.Loan            0.0219538822 0.316354829  0.006277815
## Securities.Account       1.0000000000 0.317034416  0.012627470
## CD.Account               0.3170344157 1.000000000  0.175880016
## Online                   0.0126274704 0.175880016  1.000000000
## CreditCard              -0.0150283189 0.278644365  0.004209656
##                     CreditCard
## Age                 0.007681037
## Experience          0.008967447
## Income             -0.002385008
## Family              0.011588066
## CCAvg              -0.006689494
## Education          -0.011014134
## Mortgage           -0.007230919
## Personal.Loan       0.002801509
## Securities.Account -0.015028319
## CD.Account          0.278644365
## Online              0.004209656
## CreditCard          1.000000000

corrplot(CORR_MATRIX)
```




```
## Output plot with numeric values.
corrplot(CORR_MATRIX, method ="number")
```




```
plot_DISTANCE<-plot(UB[,"Age"],UB[,"Experience"])
```




```
summary(UB$Personal.Loan)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   0.096   0.000   1.000

summary(UB)

##       Age           Experience       Income           Family
##  Min.   :23.00   Min.   :-3.0   Min.   :  8.00   Min.   :1.000
##  1st Qu.:35.00   1st Qu.:10.0   1st Qu.: 39.00   1st Qu.:1.000
##  Median :45.00   Median :20.0   Median : 64.00   Median :2.000
##  Mean   :45.34   Mean   :20.1   Mean   : 73.77   Mean   :2.396
##  3rd Qu.:55.00   3rd Qu.:30.0   3rd Qu.: 98.00   3rd Qu.:3.000
##  Max.   :67.00   Max.   :43.0   Max.   :224.00   Max.   :4.000
##      CCAvg          Education        Mortgage     Personal.Loan
##  Min.   : 0.000   Min.   :1.000   Min.   :  0.0   Min.   :0.000
##  1st Qu.: 0.700   1st Qu.:1.000   1st Qu.:  0.0   1st Qu.:0.000
##  Median : 1.500   Median :2.000   Median :  0.0   Median :0.000
##  Mean   : 1.938   Mean   :1.881   Mean   : 56.5   Mean   :0.096
##  3rd Qu.: 2.500   3rd Qu.:3.000   3rd Qu.:101.0   3rd Qu.:0.000
##  Max.   :10.000   Max.   :3.000   Max.   :635.0   Max.   :1.000
##  Securities.Account   CD.Account         Online         CreditCard
##  Min.   :0.0000     Min.   :0.0000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.0000     1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000
```

```
##  Median :0.0000     Median :0.0000   Median :1.0000   Median :0.000
##  Mean   :0.1044     Mean   :0.0604   Mean   :0.5968   Mean   :0.294
##  3rd Qu.:0.0000     3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.000
##  Max.   :1.0000     Max.   :1.0000   Max.   :1.0000   Max.   :1.000
```

```
## Partitioning
# Dummy Variables already created.

# Partition
set.seed(123)
# Get rid of extra variables
UB.index <- UB[order(runif(5000)), ]#randomized the observations
train <- UB.index[1:3500, ] #create training set
valid <- UB.index[3501:5000, ] #create validation set
dim(train)
```

```
## [1] 3500   12
```

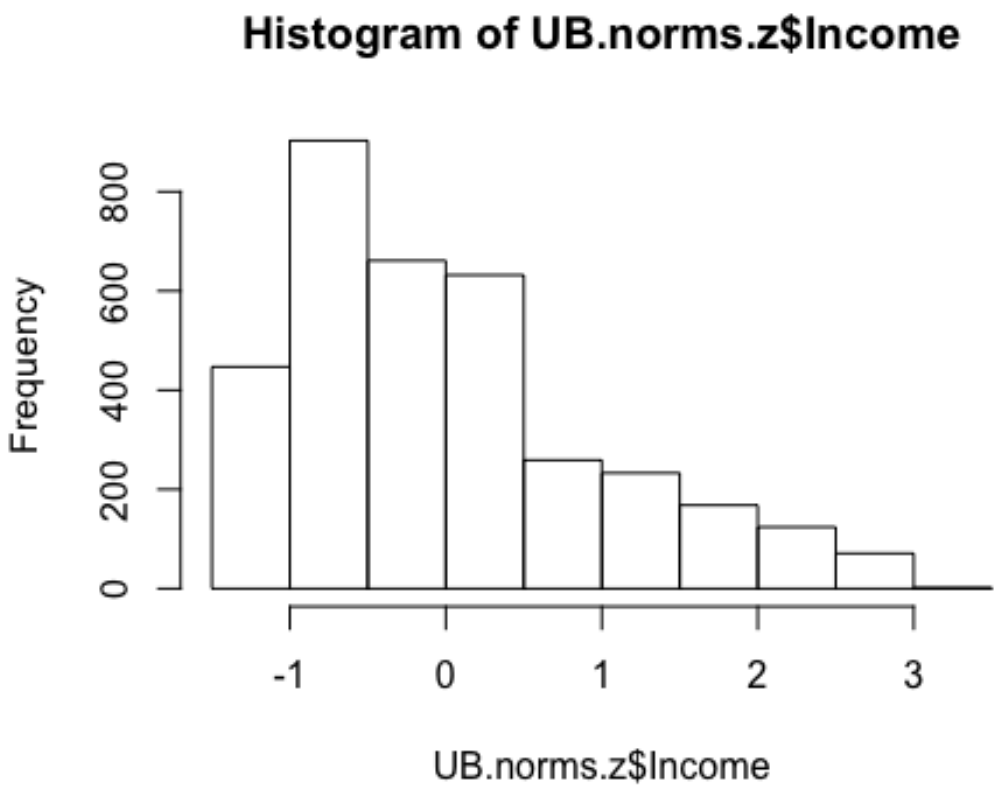
```
dim(valid)
```

```
## [1] 1500   12
```

```
##Creating copies for multiple models later on.
train2 <- UB.index[1:3500, ] #create training set
valid2 <- UB.index[3501:5000, ] #create validation set

#KNN
## Scaling with Z-Standardization
UB.norm <- train[,1:7]
UB.norms.z <- as.data.frame(scale(UB.norm))
range(UB.norms.z$Income)
```

```
## [1] -1.433659  3.307680
```

```
hist(UB.norms.z$Income)
```




```
train.knn <- cbind(UB.norms.z, train$Personal.Loan)
colnames(train.knn)[8] <- "Personal.Loan" #rename
names(train.knn)
```

```
## [1] "Age"           "Experience"    "Income"        "Family"
## [5] "CCAvg"         "Education"     "Mortgage"      "Personal.Loan"
```

```
## [1] "
summary(train.knn)
```

```
##       Age               Experience          Income             Family
##  Min.   :-1.95959   Min.   :-2.02529   Min.   :-1.4337   Min.   :-1.2111
##  1st Qu.:-0.82698   1st Qu.:-0.82846   1st Qu.:-0.7532   1st Qu.:-1.2111
##  Median :-0.04286   Median :-0.02333   Median :-0.2044   Median :-0.3412
##  Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.82837   3rd Qu.: 0.84709   3rd Qu.: 0.4541   3rd Qu.: 0.5286
##  Max.   : 1.87386   Max.   : 1.97863   Max.   : 3.3077   Max.   : 1.3985
##      CCAvg            Education         Mortgage       Personal.Loan
##  Min.   :-1.1049   Min.   :-1.054   Min.   :-0.5612   Min.   :0.00000
##  1st Qu.:-0.7021   1st Qu.:-1.054   1st Qu.:-0.5612   1st Qu.:0.00000
##  Median :-0.2416   Median : 0.139   Median :-0.5612   Median :0.00000
##  Mean   : 0.0000   Mean   : 0.000   Mean   : 0.0000   Mean   :0.08971
##  3rd Qu.: 0.3339   3rd Qu.: 1.332   3rd Qu.: 0.4368   3rd Qu.:0.00000
##  Max.   : 4.6502   Max.   : 1.332   Max.   : 5.4712   Max.   :1.00000
```

```r
UB.binary <- train[,9:12]

train.knn.total <- cbind(train.knn, UB.binary)

library(class)
train.knn.predictors <- train.knn.total[,-8]

train.knn.target <- train.knn[,8]

valid.norms.z <- as.data.frame(scale(valid[,1:7]))
valid.knn.predictors <- cbind(valid.norms.z, valid[,9:12])

valid.knn.target <- valid[,8]

set.seed(123)
preds <- knn(train=train.knn.predictors, test = valid.knn.predictors,
             cl=train.knn.target, k=1, prob=TRUE)

CONF_MATRIX<-table(preds,valid.knn.target)
CONF_MATRIX

##      valid.knn.target
## preds    0    1
##     0 1318   46
##     1   16  120

TruPo.1 <- CONF_MATRIX[2,2]
TruNeg.1 <-CONF_MATRIX[1,1]
FalPo.1 <- CONF_MATRIX[2,1]
FalNeg.1<- CONF_MATRIX[1,2]

### Trying K=SQRT(3500) = 59
set.seed(123)
preds <- knn(train=train.knn.predictors, test = valid.knn.predictors,
             cl=train.knn.target, k=59, prob=TRUE)
CONF_MATRIX<-table(preds,valid.knn.target)
CONF_MATRIX

##      valid.knn.target
## preds    0    1
##     0 1334  119
##     1    0   47

TruPo.1 <- CONF_MATRIX[2,2]
TruNeg.1 <-CONF_MATRIX[1,1]
FalPo.1 <- CONF_MATRIX[2,1]
FalNeg.1<- CONF_MATRIX[1,2]

## Trying K= 30
set.seed(123)
```

```r
preds <- knn(train=train.knn.predictors, test = valid.knn.predictors,
             cl=train.knn.target, k=30, prob=TRUE)
CONF_MATRIX<-table(preds,valid.knn.target)
CONF_MATRIX
```

```
##      valid.knn.target
## preds    0    1
##     0 1334  102
##     1    0   64
```

```r
## Trying K=5
set.seed(123)
preds <- knn(train=train.knn.predictors, test = valid.knn.predictors,
             cl=train.knn.target, k=5, prob=TRUE)
CONF_MATRIX<-table(preds,valid.knn.target)
CONF_MATRIX
```

```
##      valid.knn.target
## preds    0    1
##     0 1331   61
##     1    3  105
```

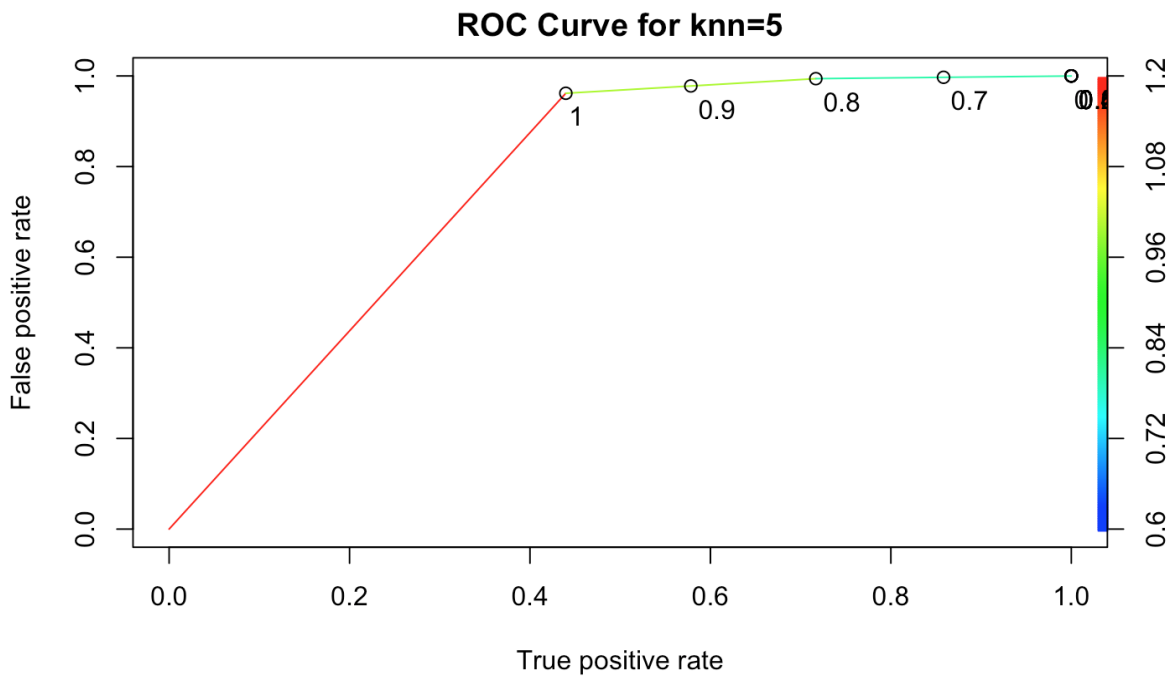
```r
prob <- attr(preds, "prob") #take out the raw probabilities from model

library(ROCR)
```

```
## Loading required package: gplots
```

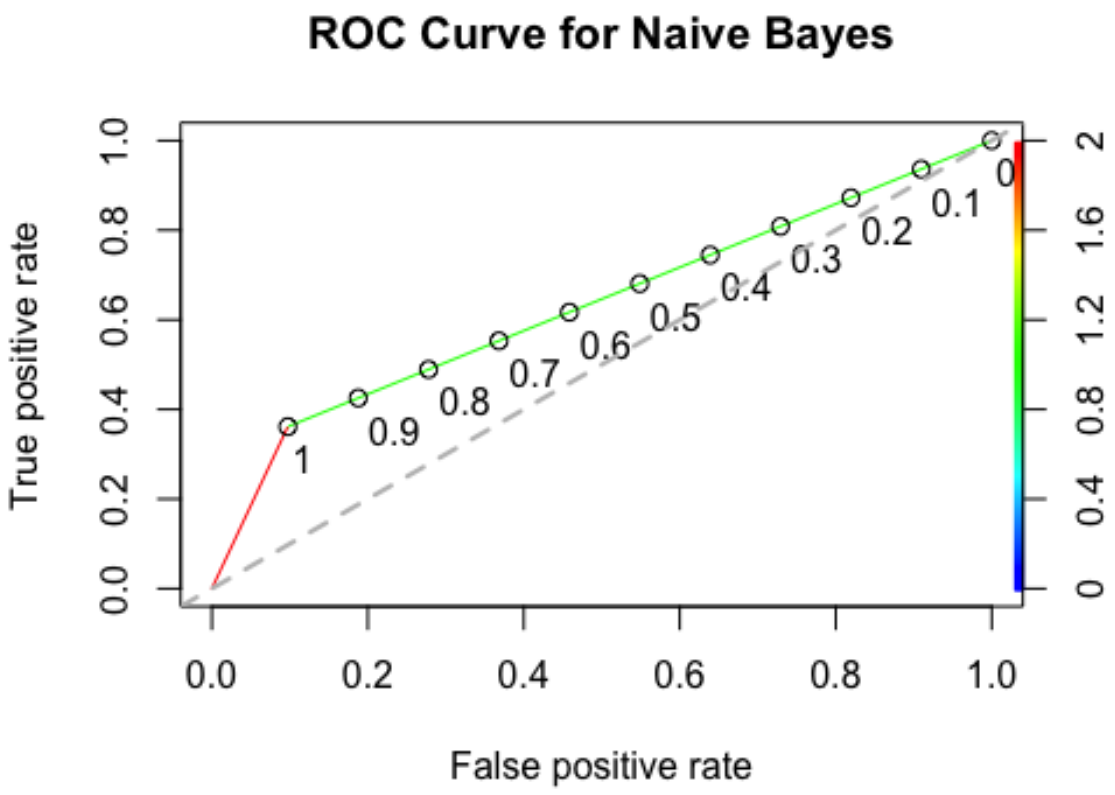
```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```r
library(gplots)

pred_knn <- prediction(prob, valid.knn.target)
perf_knn <- performance(pred_knn, "tpr", "fpr")
plot(perf_knn, colorize=TRUE, print.cutoffs.at=seq(0,1,by=0.1), text.adj=c(-0
.2,1.7),
     main = "ROC Curve for knn=59")
abline(a=0,b=1,lwd=2,lty=2,col="gray")
```




```
##Naive Bayes
library(e1071)
##Output must be factor
train$Personal.Loan <- as.factor(train$Personal.Loan)
train$Education <- as.factor(train$Education)
train$Family <- as.factor(train$Family)

UB_classifier <- naiveBayes(Personal.Loan~., data = train)
UB.pred <- predict(UB_classifier, valid[,c(-8)])

CONF_MATRIX.nd <- table(UB.pred,valid$Personal.Loan)

PROB.nb <- ifelse(UB.pred == 1, 1, 0)
pred.nb <- prediction(PROB.nb, valid.knn.target)
perf.nb <- performance(pred.nb, "tpr", "fpr")


plot(perf.nb, colorize=TRUE, print.cutoffs.at=seq(0,1,by=0.1), text.adj=c(-0.
2,1.7),
     main = "ROC Curve for Naive Bayes")
abline(a=0,b=1,lwd=2,lty=2,col="gray")
```




```
##Logistic Regression
# Logit Model, all variables
LOGIT_MODEL <- glm(Personal.Loan~., family=binomial(), data=train2)
summary(LOGIT_MODEL)

##
## Call:
## glm(formula = Personal.Loan ~ ., family = binomial(), data = train2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1233  -0.2112  -0.0914  -0.0384   3.8369
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -8.8292555  2.1385509  -4.129 3.65e-05 ***
## Age                -0.1548799  0.0823129  -1.882 0.059890 .
## Experience          0.1631397  0.0818170   1.994 0.046156 *
## Income              0.0488327  0.0029569  16.515  < 2e-16 ***
## Family              0.7134246  0.0880726   8.100 5.48e-16 ***
## CCAvg               0.1524097  0.0479231   3.180 0.001471 **
## Education           1.6021682  0.1340869  11.949  < 2e-16 ***
## Mortgage            0.0009415  0.0006721   1.401 0.161258
```

```
## Securities.Account -1.4014656  0.3674692  -3.814 0.000137 ***
## CD.Account          4.0467669  0.3855576  10.496  < 2e-16 ***
## Online             -0.7307465  0.1874038  -3.899 9.65e-05 ***
## CreditCard         -1.0821699  0.2428038  -4.457 8.31e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2113.13  on 3499  degrees of freedom
## Residual deviance:  913.55  on 3488  degrees of freedom
## AIC: 937.55
##
## Number of Fisher Scoring iterations: 7

library(DiscriMiner)
logLik(LOGIT_MODEL)

## 'log Lik.' -456.7735 (df=12)

## Logit Predictions
pred.LR<-predict(LOGIT_MODEL,valid2)
LOGITS <- data.frame(pred.LR)
ODDS <- exp(LOGITS)
PROBABILITIES <- ODDS/(ODDS+1)
library(psych)
describe(PROBABILITIES)

##         vars    n mean   sd median trimmed  mad min max range skew
## pred.LR    1 1500  0.1 0.23   0.01    0.04 0.01   0   1     1 2.73
##         kurtosis   se
## pred.LR     6.48 0.01

describe(ODDS)

##         vars    n mean   sd median trimmed  mad min    max  range  skew
## pred.LR    1 1500 4.06 48.9   0.01    0.04 0.01   0 1535.3 1535.3 24.34
##         kurtosis   se
## pred.LR   693.72 1.26

SC_PROB <- data.frame(fitted(LOGIT_MODEL))

PREDICTIONS <- ifelse(PROBABILITIES>.50,1,0)
data.frame(PREDICTIONS)

##      pred.LR
## 4390       0
## 3813       0
## 2153       0
## 1478       0
## 2647       1
## 4302       1
```

```r
table(PREDICTIONS,valid2$Personal.Loan)
```

```
##
## PREDICTIONS    0    1
##           0 1315   60
##           1   19  106
```

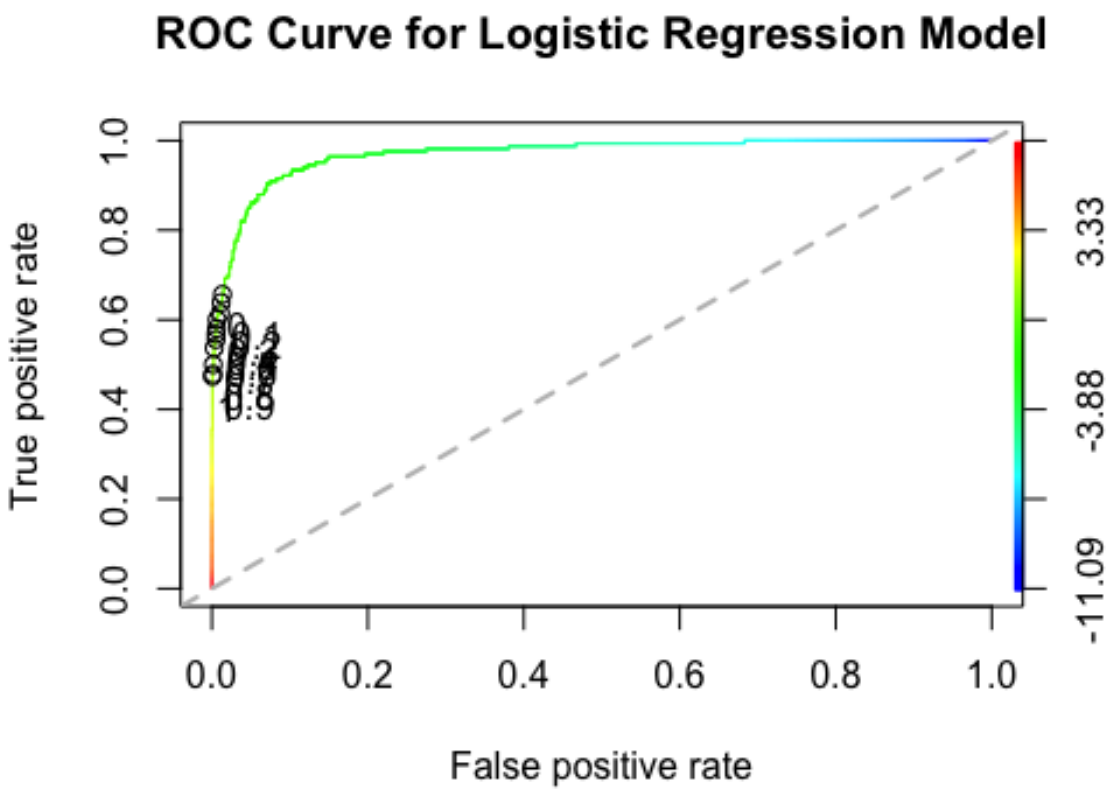
```r
##Reduced Logit Model ## Reduced Age and Mortgage
train_reduced <- train2[,c(-1,-7)]
LOGIT_MODEL_Red <- glm(Personal.Loan~., family=binomial(), data=train_reduced
)
summary(LOGIT_MODEL_Red)
```

```
##
## Call:
## glm(formula = Personal.Loan ~ ., family = binomial(), data = train_reduced
)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1257  -0.2170  -0.0910  -0.0385   3.8373
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -12.616891   0.675106 -18.689  < 2e-16 ***
## Experience           0.009550   0.007715   1.238 0.215794
## Income               0.049429   0.002935  16.840  < 2e-16 ***
## Family               0.712885   0.088302   8.073 6.84e-16 ***
## CCAvg                0.146442   0.047476   3.085 0.002039 **
## Education            1.547140   0.131210  11.791  < 2e-16 ***
## Securities.Account  -1.413216   0.367096  -3.850 0.000118 ***
## CD.Account           4.087821   0.383951  10.647  < 2e-16 ***
## Online              -0.733523   0.186710  -3.929 8.54e-05 ***
## CreditCard          -1.110802   0.242451  -4.582 4.62e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2113.13  on 3499  degrees of freedom
## Residual deviance:  919.04  on 3490  degrees of freedom
## AIC: 939.04
##
## Number of Fisher Scoring iterations: 7
```

```r
library(DiscriMiner)
logLik(LOGIT_MODEL_Red)
```

```
## 'log Lik.' -459.5184 (df=10)
```

```
## Logit Predictions
pred.LR2<-predict(LOGIT_MODEL_Red,valid2)
LOGITS.R <- data.frame(pred.LR2)
ODDS.R <- exp(LOGITS.R)
PROBABILITIES.R <- ODDS.R/(ODDS.R+1)
library(psych)
describe(PROBABILITIES.R)

##          vars    n mean   sd median trimmed  mad min max range skew
## pred.LR2    1 1500  0.1 0.23   0.01    0.04 0.01   0   1     1 2.73
##          kurtosis   se
## pred.LR2     6.47 0.01

describe(ODDS.R)

##          vars    n mean    sd median trimmed  mad min     max   range
## pred.LR2    1 1500 3.83 41.31   0.01    0.04 0.01   0 1012.66 1012.66
##           skew kurtosis   se
## pred.LR2 18.84   397.97 1.07

SC_PROB.R <- data.frame(fitted(LOGIT_MODEL_Red))

PREDICTIONS.R <- ifelse(PROBABILITIES.R>.50,1,0)
data.frame(PREDICTIONS.R)

##      pred.LR2
## 4390        0
## 3813        0
## 2647        1
## 4302        1

table(PREDICTIONS.R,valid2$Personal.Loan)

##
## PREDICTIONS.R    0    1
##             0 1316   57
##             1   18  109

##ROC Curve Reduced Model
library(ROCR)
pred_logit <- prediction(pred.LR2, valid2$Personal.Loan)
perf_logit <- performance(pred_logit, "tpr", "fpr")
plot(perf_logit, colorize=TRUE, print.cutoffs.at=seq(0,1,by=0.1), text.adj=c(
-0.2,1.7),
     main = "ROC Curve for Logistic Regression Model")
abline(a=0,b=1,lwd=2,lty=2,col="gray")
```




```
# Linear Discriminant Analysis
library(DiscriMiner)
LDA_MODEL <- linDA((train2)[,1:7,9:12],train2$Personal.Loan)
summary(LDA_MODEL)

##                Length Class  Mode
## functions          16   -none- numeric
## confusion           4   table  numeric
## scores           7000   -none- numeric
## classification 3500   factor numeric
## error_rate          1   -none- numeric
## specs               6   -none- list

LDA_MODEL$functions

##                          0             1
## constant   -2.260455e+02 -2.345341e+02
## Age         1.757973e+01  1.738439e+01
## Experience -1.728831e+01 -1.708560e+01
## Income      8.755441e-02  1.384491e-01
## Family      1.536688e+00  2.200747e+00
## CCAvg      -5.706179e-01 -2.431302e-01
## Education  -3.371475e+00 -1.992663e+00
## Mortgage    3.956897e-03  6.908163e-03
```

```r
LDA_MODEL$scores

##             0        1
## 3934 211.0253 205.4117
## 4088 213.3355 217.6438

LDA_MODEL$classification

##    [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## Levels: 0 1

#Predictions
library(MASS)
LDA_MODEL_MASS<-lda(Personal.Loan~.,train2)
LDA_MODEL_MASS

## Call:
## lda(Personal.Loan ~ ., data = train2)
##
## Prior probabilities of groups:
##          0          1
## 0.91028571 0.08971429
##
## Group means:
##        Age Experience    Income   Family    CCAvg Education  Mortgage
## 0 45.55461   20.32423  66.48274 2.365662 1.721067  1.851852  52.17483
## 1 44.85669   19.69745 142.61465 2.662420 3.937006  2.203822 105.22293
##   Securities.Account CD.Account    Online CreditCard
## 0          0.1029504 0.03546767 0.6045198  0.2944131
## 1          0.1242038 0.30573248 0.6146497  0.3025478
##
## Coefficients of linear discriminants:
##                              LD1
## Age                -0.0687048754
## Experience          0.0707985457
## Income              0.0192254494
## Family              0.2533718035
## CCAvg               0.1152940020
## Education           0.5287502253
## Mortgage            0.0006992763
## Securities.Account -0.5467550298
## CD.Account          2.4877072176
## Online             -0.2074410605
## CreditCard         -0.2982340762

LDA.Pred <- predict(LDA_MODEL_MASS,valid2)
LDA.Pred$class #class prediction

##    [1] 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
## [1497] 0 0 1 1
## Levels: 0 1
```

```r
table(LDA.Pred$class, valid2$Personal.Loan)
```

```
##
##        0    1
##   0 1304   67
##   1   30   99
```

```r
library(ROCR)
pred_lda <- prediction(LDA.Pred$posterior[,2], valid2$Personal.Loan)
perf_lda <- performance(pred_lda, "tpr", "fpr")
plot(perf_lda, colorize=TRUE, print.cutoffs.at=seq(0,1,by=0.1), text.adj=c(-0
.2,1.7),
     main = "ROC Curve for Linear Discriminant Model")
abline(a=0,b=1,lwd=2,lty=2,col="gray")
```