Case Study Final

Aakar Kale

12/12/2019

reading the dataset

```
GC <- read.csv("/Users/aakarkale/Desktop/CSUEB/Data Mining/GermanCredit.csv") #dataset is read thorugh
missing(GC) #dataset is checked for any missing value
```

```
## [1] FALSE
```

```
#View(GC)
str(GC)
```

```
## 'data.frame':    1000 obs. of  32 variables:
##  $ OBS.            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ CHK_ACCT        : int  0 1 3 0 0 3 3 1 3 1 ...
##  $ DURATION        : int  6 48 12 42 24 36 24 36 12 30 ...
##  $ HISTORY         : int  4 2 4 2 3 2 2 2 2 4 ...
##  $ NEW_CAR         : int  0 0 0 0 1 0 0 0 0 1 ...
##  $ USED_CAR        : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ FURNITURE       : int  0 0 0 1 0 0 1 0 0 0 ...
##  $ RADIO.TV        : int  1 1 0 0 0 0 0 0 1 0 ...
##  $ EDUCATION       : int  0 0 1 0 0 1 0 0 0 0 ...
##  $ RETRAINING      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AMOUNT          : int  1169 5951 2096 7882 4870 9055 2835 6948 3059 5234 ...
##  $ SAV_ACCT        : int  4 0 0 0 0 4 2 0 3 0 ...
##  $ EMPLOYMENT      : int  4 2 3 3 2 2 4 2 3 0 ...
##  $ INSTALL_RATE    : int  4 2 2 2 3 2 3 2 2 4 ...
##  $ MALE_DIV        : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ MALE_SINGLE     : int  1 0 1 1 1 1 1 1 0 0 ...
##  $ MALE_MAR_or_WID : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ CO.APPLICANT    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ GUARANTOR       : int  0 0 0 1 0 0 0 0 0 0 ...
##  $ PRESENT_RESIDENT: int  4 2 3 4 4 4 4 2 4 2 ...
##  $ REAL_ESTATE     : int  1 1 1 0 0 0 0 0 1 0 ...
##  $ PROP_UNKN_NONE  : int  0 0 0 0 1 1 0 0 0 0 ...
##  $ AGE             : int  67 22 49 45 53 35 53 35 61 28 ...
##  $ OTHER_INSTALL   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RENT            : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ OWN_RES         : int  1 1 1 0 0 0 1 0 1 1 ...
##  $ NUM_CREDITS     : int  2 1 1 1 2 1 1 1 1 2 ...
##  $ JOB             : int  2 2 1 2 2 1 2 3 1 3 ...
##  $ NUM_DEPENDENTS  : int  1 1 2 2 2 2 1 1 1 1 ...
##  $ TELEPHONE       : int  1 0 0 0 0 1 0 1 0 0 ...
##  $ FOREIGN         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RESPONSE        : int  1 0 1 1 0 1 1 1 1 0 ...
```

Q1. Review the predictor variables and guess what their role in a credit decision might be. Are there any surprise in the data?

```
GC$PRESENT_RESIDENT <- GC$PRESENT_RESIDENT - 1
GC <- GC[,c(-1,-22)]

GC$ANOTHER_OBJECTIVE <- ifelse(GC$NEW_CAR+GC$USED_CAR+GC$FURNITURE+GC$RADIO.TV+GC$EDUCATION+GC$RETRAINII

GC$Female <- ifelse(GC$MALE_DIV+GC$MALE_MAR_or_WID+GC$MALE_SINGLE==0, 1, 0)

GC$PRESENT_RESIDENT <- factor(GC$PRESENT_RESIDENT, levels = c(0, 1, 2, 3), labels=c("<=1_year","1-2_yea

GC$EMPLOYMENT <- factor(GC$EMPLOYMENT, levels = c(0,1,2,3,4), labels = c("Unemployed", "1year","1-3year

GC$JOB <- factor(GC$JOB, levels = c(0, 1, 2, 3), labels=c("Uemployed","Unskilled-employee","Skilled emp

GC$CHK_ACCT <- factor(GC$CHK_ACCT, levels=c(0,1,2,3), labels = c("<0DM","0-200DM","200DM","No_checking_

GC$HISTORY <- factor(GC$HISTORY, levels = c(0,1,2,3,4), labels = c("No_credits","Paid","Existing_paid",

GC$SAV_ACCT <- factor(GC$SAV_ACCT, levels=c(0,1,2,3,4), labels = c("<
                                                        100DM","101-500DM","501-1000DM","1000DM","no_sa

NEW_GC <- GC
head(GC)
```

```
##              CHK_ACCT DURATION           HISTORY NEW_CAR USED_CAR
## 1                <0DM        6 important_account       0        0
## 2             0-200DM       48     Existing_paid       0        0
## 3 No_checking_account       12 important_account       0        0
## 4                <0DM       42     Existing_paid       0        0
## 5                <0DM       24            Unpaid       1        0
## 6 No_checking_account       36     Existing_paid       0        0
##   FURNITURE RADIO.TV EDUCATION RETRAINING AMOUNT
## 1         0        1         0          0   1169
## 2         0        1         0          0   5951
## 3         0        0         1          0   2096
## 4         1        0         0          0   7882
## 5         0        0         0          0   4870
## 6         0        0         1          0   9055
##                                                      SAV_ACCT
## 1                                           no_saving_account
## 2 <\n                                                   100DM
## 3 <\n                                                   100DM
## 4 <\n                                                   100DM
## 5 <\n                                                   100DM
## 6                                           no_saving_account
##   EMPLOYMENT INSTALL_RATE MALE_DIV MALE_SINGLE MALE_MAR_or_WID
## 1   >=7years            4        0           1               0
## 2    1-3year            2        0           0               0
## 3    4-6year            2        0           1               0
## 4    4-6year            2        0           1               0
```

```
## 5    1-3year            3        0           1               0
## 6    1-3year            2        0           1               0
##   CO.APPLICANT GUARANTOR PRESENT_RESIDENT REAL_ESTATE AGE OTHER_INSTALL
## 1            0         0        >=3_years           1  67             0
## 2            0         0        1-2_years           1  22             0
## 3            0         0         2-3_year           1  49             0
## 4            0         1        >=3_years           0  45             0
## 5            0         0        >=3_years           0  53             0
## 6            0         0        >=3_years           0  35             0
##   RENT OWN_RES NUM_CREDITS                JOB NUM_DEPENDENTS TELEPHONE
## 1    0       1           2   Skilled employee              1         1
## 2    0       1           1   Skilled employee              1         0
## 3    0       1           1 Unskilled-employee              2         0
## 4    0       0           1   Skilled employee              2         0
## 5    0       0           2   Skilled employee              2         0
## 6    0       0           1 Unskilled-employee              2         1
##   FOREIGN RESPONSE ANOTHER_OBJECTIVE Female
## 1       0        1                 0      0
## 2       0        0                 0      1
## 3       0        1                 0      0
## 4       0        1                 0      0
## 5       0        0                 0      0
## 6       0        1                 0      0
```

```r
head(NEW_GC)
```

```
##             CHK_ACCT DURATION           HISTORY NEW_CAR USED_CAR
## 1               <0DM        6 important_account       0        0
## 2            0-200DM       48     Existing_paid       0        0
## 3 No_checking_account       12 important_account       0        0
## 4               <0DM       42     Existing_paid       0        0
## 5               <0DM       24            Unpaid       1        0
## 6 No_checking_account       36     Existing_paid       0        0
##   FURNITURE RADIO.TV EDUCATION RETRAINING AMOUNT
## 1         0        1         0          0   1169
## 2         0        1         0          0   5951
## 3         0        0         1          0   2096
## 4         1        0         0          0   7882
## 5         0        0         0          0   4870
## 6         0        0         1          0   9055
##                                                             SAV_ACCT
## 1                                                  no_saving_account
## 2 <\n                                                          100DM
## 3 <\n                                                          100DM
## 4 <\n                                                          100DM
## 5 <\n                                                          100DM
## 6                                                  no_saving_account
##   EMPLOYMENT INSTALL_RATE MALE_DIV MALE_SINGLE MALE_MAR_or_WID
## 1   >=7years            4        0           1               0
## 2    1-3year            2        0           0               0
## 3    4-6year            2        0           1               0
## 4    4-6year            2        0           1               0
## 5    1-3year            3        0           1               0
## 6    1-3year            2        0           1               0
```

```
##   CO.APPLICANT GUARANTOR PRESENT_RESIDENT REAL_ESTATE AGE OTHER_INSTALL
## 1            0         0        >=3_years           1  67             0
## 2            0         0        1-2_years           1  22             0
## 3            0         0         2-3_year           1  49             0
## 4            0         1        >=3_years           0  45             0
## 5            0         0        >=3_years           0  53             0
## 6            0         0        >=3_years           0  35             0
##   RENT OWN_RES NUM_CREDITS                JOB NUM_DEPENDENTS TELEPHONE
## 1    0       1           2   Skilled employee              1         1
## 2    0       1           1   Skilled employee              1         0
## 3    0       1           1 Unskilled-employee              2         0
## 4    0       0           1   Skilled employee              2         0
## 5    0       0           2   Skilled employee              2         0
## 6    0       0           1 Unskilled-employee              2         1
##   FOREIGN RESPONSE ANOTHER_OBJECTIVE Female
## 1       0        1                 0      0
## 2       0        0                 0      1
## 3       0        1                 0      0
## 4       0        1                 0      0
## 5       0        0                 0      0
## 6       0        1                 0      0
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
AMOUNT.mean = GC %>% dplyr::select(AMOUNT,RESPONSE) %>% group_by(RESPONSE) %>% summarise(m =mean(AMOUNT)
AMOUNT.mean
```

```
## # A tibble: 2 x 2
##   RESPONSE     m
##      <int> <dbl>
## 1        0 3938.
## 2        1 2985.
```

```
DURATION.mean = GC %>% dplyr::select(DURATION,RESPONSE) %>%group_by(RESPONSE) %>% summarise( m =mean(DUR
DURATION.mean
```

```
## # A tibble: 2 x 2
##   RESPONSE     m
##      <int> <dbl>
## 1        0  24.9
## 2        1  19.2
```

```r
INSTALL_RATE.median = GC %>% dplyr::select(INSTALL_RATE,RESPONSE) %>%group_by(RESPONSE) %>% summarise( m
INSTALL_RATE.median
```

```
## # A tibble: 2 x 2
##   RESPONSE     m
##      <int> <dbl>
## 1        0     4
## 2        1     3
```

```r
AGE.median = GC %>% dplyr::select(AGE,RESPONSE) %>%group_by(RESPONSE) %>% summarise( m =median(AGE))
AGE.median
```

```
## # A tibble: 2 x 2
##   RESPONSE     m
##      <int> <dbl>
## 1        0    31
## 2        1    34
```

In this dataset there were 4 categories in Present_Resident so one has to be substracted in order to have 0 to 3 levels. Real_estate and Prop_Unkn_none- either of them can be 0 but cannot be 0 at the same time. the Another-objective option is need and should be added to the data set. So the Female option has been added.

At the end of this chunk, median values for bad records is lesser than that of good records in age variable, it might be premature to say young people tend to have bad credit records, but we can safely assume it tends to be riskier. In case of installment_rate variable great difference between the good and bad records, we see that bad records have more median value than good ones.

For the amount variable, we observe that the amount for bad records is larger in general as compared to good ones.

#Q2. Divide the data into trainning and validatin partitions, and develop classification models using following data mining techniques in R: logistic regression, classification trees, and neural networks.

#Q.3.Choose one modelfrom each technique and report the confusion matrix and the cost/gain matrix for the validation data. Which technique has the highest net profit?

```r
#install.packages("e1071")
library(e1071)

#creating model for logistic regression
set.seed(2)
dim(GC)
```

```
## [1] 1000   32
```

```r
training_rows <- sample(c(1:1000), 800) #sample is taken for first 1000 rows
train_data <- GC[training_rows,]#training data was made
valid_data <- GC[-training_rows,]#test data was made

#Model
glm <- glm(RESPONSE~., data = train_data, family="binomial") #logistic model was created
options(scipen = 999)
summary(glm) #summary of the model was shown
```

```
##
## Call:
## glm(formula = RESPONSE ~ ., family = "binomial", data = train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8202  -0.5702   0.3339   0.6433   2.6268
##
## Coefficients: (2 not defined because of singularities)
##                                               Estimate  Std. Error z value
## (Intercept)                                 2.11214477  1.20608385   1.751
## CHK_ACCT0-200DM                             0.29107351  0.26019593   1.119
## CHK_ACCT200DM                               0.83669433  0.40974880   2.042
## CHK_ACCTNo_checking_account                 1.74864373  0.27114315   6.449
## DURATION                                   -0.04129844  0.01096185  -3.767
## HISTORYPaid                                -0.53216904  0.64625498  -0.823
## HISTORYExisting_paid                        0.45262152  0.52563200   0.861
## HISTORYUnpaid                               0.93837640  0.57996762   1.618
## HISTORYimportant_account                    1.74311947  0.55201223   3.158
## NEW_CAR                                    -1.21943853  0.44985977  -2.711
## USED_CAR                                    0.28640087  0.55961942   0.512
## FURNITURE                                  -0.50103443  0.46516888  -1.077
## RADIO.TV                                   -0.26949220  0.45286263  -0.595
## EDUCATION                                  -1.71204797  0.58389101  -2.932
## RETRAINING                                 -0.62627216  0.50930540  -1.230
## AMOUNT                                     -0.00009651  0.00005174  -1.865
## SAV_ACCT101-500DM                           0.65628338  0.34873255   1.882
## SAV_ACCT501-1000DM                          0.08790008  0.41729739   0.211
## SAV_ACCT1000DM                              1.52075531  0.59928769   2.538
## SAV_ACCTno_saving_account                   1.27755014  0.31641209   4.038
## EMPLOYMENT1year                             0.55675493  0.51244797   1.086
## EMPLOYMENT1-3year                           0.91584377  0.49825398   1.838
## EMPLOYMENT4-6year                           1.48245532  0.53767085   2.757
## EMPLOYMENT>=7years                          0.95716512  0.49695760   1.926
## INSTALL_RATE                               -0.32647427  0.10327849  -3.161
## MALE_DIV                                   -0.49739632  0.46739684  -1.064
## MALE_SINGLE                                 0.47954953  0.24332160   1.971
## MALE_MAR_or_WID                             0.29487588  0.37460866   0.787
## CO.APPLICANT                               -0.52781097  0.47492591  -1.111
## GUARANTOR                                   1.60909207  0.53569671   3.004
## PRESENT_RESIDENT1-2_years                  -0.89870048  0.34755735  -2.586
## PRESENT_RESIDENT2-3_year                   -0.60297426  0.39077102  -1.543
## PRESENT_RESIDENT>=3_years                  -0.36738986  0.35719395  -1.029
## REAL_ESTATE                                 0.24298658  0.24930449   0.975
## AGE                                         0.01465486  0.01081263   1.355
## OTHER_INSTALL                              -0.50768698  0.24671306  -2.058
## RENT                                        0.06479400  0.41056146   0.158
## OWN_RES                                     0.39652354  0.35390425   1.120
## NUM_CREDITS                                -0.57066899  0.22981372  -2.483
## JOBUnskilled-employee                      -0.91072238  0.71230295  -1.279
## JOBSkilled employee                        -0.77042787  0.68233358  -1.129
## JOBhighly qualified employee/self employed -0.71717781  0.69277231  -1.035
## NUM_DEPENDENTS                             -0.19850678  0.28670864  -0.692
## TELEPHONE                                   0.49001752  0.23544794   2.081
```

```
## FOREIGN                                   1.15430233  0.64120810   1.800
## ANOTHER_OBJECTIVE                                 NA          NA      NA
## Female                                            NA          NA      NA
##                                             Pr(>|z|)
## (Intercept)                                 0.079904 .
## CHK_ACCT0-200DM                             0.263281
## CHK_ACCT200DM                               0.041155 *
## CHK_ACCTNo_checking_account           0.000000000112 ***
## DURATION                                    0.000165 ***
## HISTORYPaid                                 0.410243
## HISTORYExisting_paid                        0.389183
## HISTORYUnpaid                               0.105667
## HISTORYimportant_account                    0.001590 **
## NEW_CAR                                     0.006714 **
## USED_CAR                                    0.608806
## FURNITURE                                   0.281435
## RADIO.TV                                    0.551786
## EDUCATION                                   0.003366 **
## RETRAINING                                  0.218825
## AMOUNT                                      0.062125 .
## SAV_ACCT101-500DM                           0.059848 .
## SAV_ACCT501-1000DM                          0.833167
## SAV_ACCT1000DM                              0.011161 *
## SAV_ACCTno_saving_account             0.000053997425 ***
## EMPLOYMENT1year                             0.277275
## EMPLOYMENT1-3year                           0.066047 .
## EMPLOYMENT4-6year                           0.005830 **
## EMPLOYMENT>=7years                          0.054098 .
## INSTALL_RATE                                0.001572 **
## MALE_DIV                                    0.287245
## MALE_SINGLE                                 0.048741 *
## MALE_MAR_or_WID                             0.431190
## CO.APPLICANT                                0.266416
## GUARANTOR                                   0.002667 **
## PRESENT_RESIDENT1-2_years                   0.009716 **
## PRESENT_RESIDENT2-3_year                    0.122822
## PRESENT_RESIDENT>=3_years                   0.303694
## REAL_ESTATE                                 0.329730
## AGE                                         0.175307
## OTHER_INSTALL                               0.039609 *
## RENT                                        0.874600
## OWN_RES                                     0.262532
## NUM_CREDITS                                 0.013022 *
## JOBUnskilled-employee                       0.201052
## JOBSkilled employee                         0.258853
## JOBhighly qualified employee/self employed      0.300562
## NUM_DEPENDENTS                              0.488709
## TELEPHONE                                   0.037414 *
## FOREIGN                                     0.071829 .
## ANOTHER_OBJECTIVE                                 NA
## Female                                            NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 965.23  on 799  degrees of freedom
## Residual deviance: 671.65  on 755  degrees of freedom
## AIC: 761.65
##
## Number of Fisher Scoring iterations: 5
```

```r
pred_v <- predict(glm, valid_data[,-30], type = "response")
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type
## == : prediction from a rank-deficient fit may be misleading
```

```r
#prediction of the model was done
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```r
library(ggplot2)
confusionMatrix(as.factor(ifelse(pred_v>0.5, 1, 0)), as.factor(valid_data$RESPONSE))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  33  26
##          1  34 107
##
##                Accuracy : 0.7
##                  95% CI : (0.6314, 0.7626)
##     No Information Rate : 0.665
##     P-Value [Acc > NIR] : 0.1652
##
##                   Kappa : 0.3061
##
##  Mcnemar's Test P-Value : 0.3662
##
##             Sensitivity : 0.4925
##             Specificity : 0.8045
##          Pos Pred Value : 0.5593
##          Neg Pred Value : 0.7589
##              Prevalence : 0.3350
##          Detection Rate : 0.1650
##    Detection Prevalence : 0.2950
##       Balanced Accuracy : 0.6485
##
##        'Positive' Class : 0
##
```

```
#confusion matrix created
```

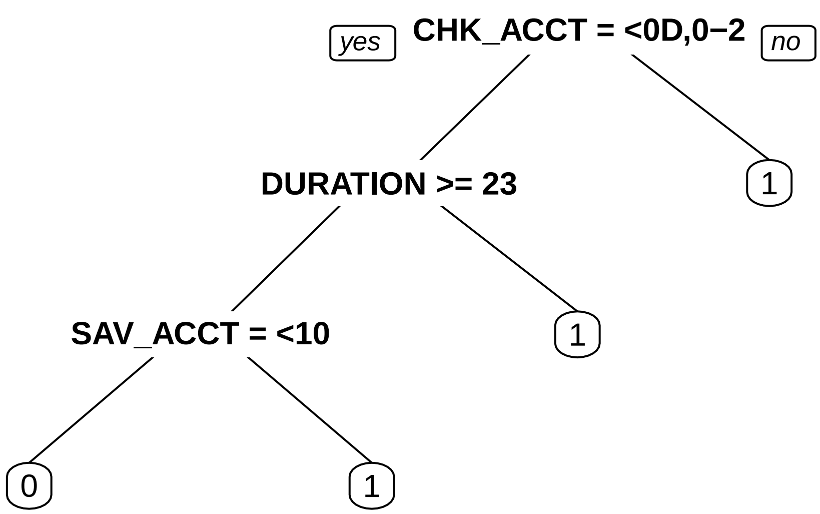
Logistic Regression Model Cost Metrix: Reference Bad Good Predited Bad 0 100*26=2600*
*Good 34*500=17000 0 Gain Matrix: Reference Bad Good Predicted
Bad 0 0 Good -500*34=-17000 100*107=10700 Logistic Regression Model, net profit is -6300.

Classification Tree

```
library(rpart)
library(rpart.plot)
set.seed(1)
training_rows <- sample(c(1:1000), 800)
train_data_tree <- NEW_GC[training_rows,]
valid_data_tree <- NEW_GC[-training_rows,]

#classification tree model
train_tree <- rpart(RESPONSE ~ ., data = train_data_tree, minbucket = 50, maxdepth = 10, model=TRUE, me
train_tree$cptable[which.min(train_tree$cptable[,"xerror"]),"CP"]
```

```
## [1] 0.01
```

```
pfit_tree <- prune(train_tree, cp = train_tree$cptable[which.min(train_tree$cptable[,"xerror"]),"CP"])
prp(train_tree)
```



```r
# predictions on validation set
pred_valid <- predict(train_tree, valid_data[,-30])
confusionMatrix(as.factor(1*(pred_valid[,2]>0.5)), as.factor(valid_data$RESPONSE), positive = "1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  19  12
##          1  48 121
##
##                Accuracy : 0.7
##                  95% CI : (0.6314, 0.7626)
##     No Information Rate : 0.665
##     P-Value [Acc > NIR] : 0.1652
##
##                   Kappa : 0.2231
##
##  Mcnemar's Test P-Value : 0.000006228
##
##             Sensitivity : 0.9098
##             Specificity : 0.2836
##          Pos Pred Value : 0.7160
##          Neg Pred Value : 0.6129
##              Prevalence : 0.6650
##          Detection Rate : 0.6050
##    Detection Prevalence : 0.8450
##       Balanced Accuracy : 0.5967
##
##        'Positive' Class : 1
##
```

Classification tree model, Cost Metrix: Reference Bad Good Predited Bad 0 100*12=1200 Good 48*500=31500 0 Gain Matrix: Reference Bad Good Predicted Bad 0 0 Good -500*48=-31500 100*121=19200 Classification Tree Model, net profit is -12300.

NeuralNet Model

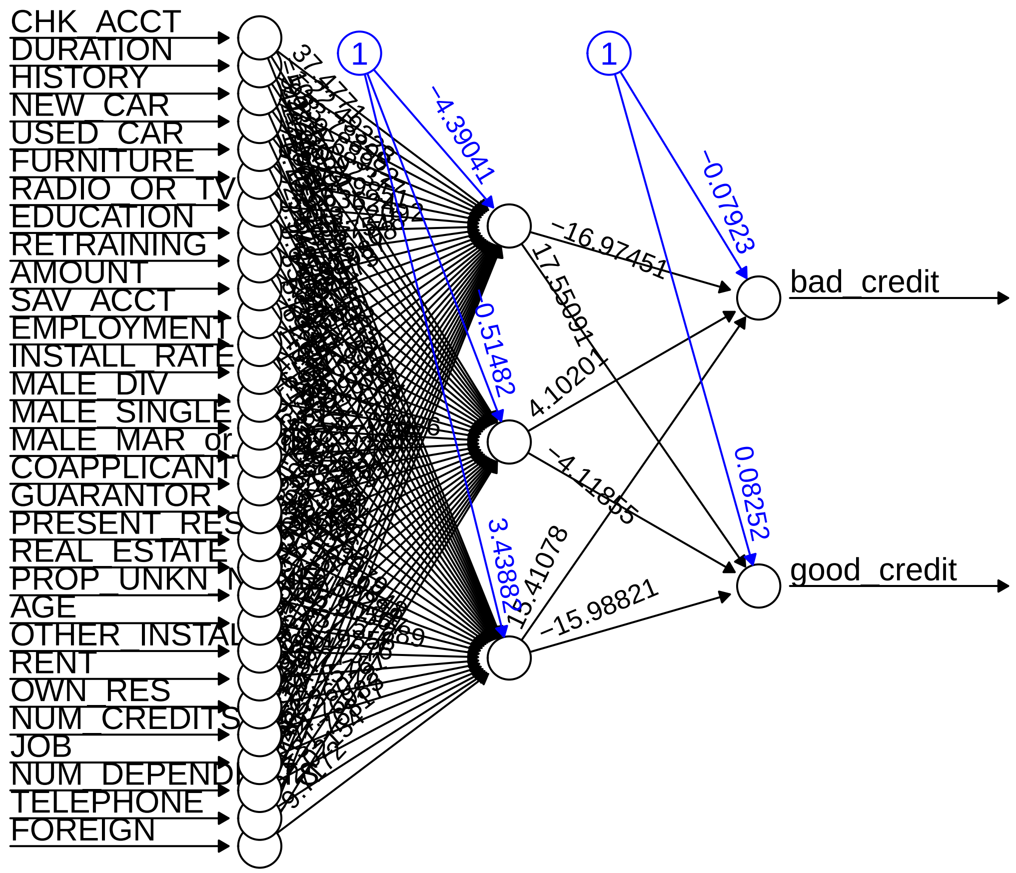
```r
library("neuralnet")
```

```
##
## Attaching package: 'neuralnet'

## The following object is masked from 'package:dplyr':
##
##     compute
```

```
NN_GC <- read.csv("/Users/aakarkale/Desktop/CSUEB/Data Mining/GermanCredit.csv")
scale <- preProcess(NN_GC, method = c("range"))
GC_scale <- predict(scale, NN_GC)
GC_scale$good_credit <- GC_scale$RESPONSE == 1
GC_scale$bad_credit <- GC_scale$RESPONSE == 0

set.seed(1)
training_rows <- sample(c(1:1000), 800)
train_data_nn <- GC_scale[training_rows,]
valid_data_nn <- GC_scale[-training_rows,]

colnames(train_data_nn)[8] <- "RADIO_OR_TV"
colnames(train_data_nn)[18] <- "COAPPLICANT"
colnames(train_data_nn)
```

```
##  [1] "OBS."             "CHK_ACCT"         "DURATION"
##  [4] "HISTORY"          "NEW_CAR"          "USED_CAR"
##  [7] "FURNITURE"        "RADIO_OR_TV"      "EDUCATION"
## [10] "RETRAINING"       "AMOUNT"           "SAV_ACCT"
## [13] "EMPLOYMENT"       "INSTALL_RATE"     "MALE_DIV"
## [16] "MALE_SINGLE"      "MALE_MAR_or_WID"  "COAPPLICANT"
## [19] "GUARANTOR"        "PRESENT_RESIDENT" "REAL_ESTATE"
## [22] "PROP_UNKN_NONE"   "AGE"              "OTHER_INSTALL"
## [25] "RENT"             "OWN_RES"          "NUM_CREDITS"
## [28] "JOB"              "NUM_DEPENDENTS"   "TELEPHONE"
## [31] "FOREIGN"          "RESPONSE"         "good_credit"
## [34] "bad_credit"
```

```
nn <- neuralnet(bad_credit+good_credit~CHK_ACCT+DURATION+HISTORY+NEW_CAR+USED_CAR+FURNITURE+RADIO_OR_TV

plot(nn, rep="best")
```




```
predict <- neuralnet::compute(nn, valid_data_nn[,2:31])

predicted.class <- apply(predict$net.result,1,which.max)-1
confusionMatrix(as.factor(predicted.class), as.factor(valid_data_nn$RESPONSE))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  26  19
##          1  41 114
##
##                Accuracy : 0.7
##                  95% CI : (0.6314, 0.7626)
##     No Information Rate : 0.665
##     P-Value [Acc > NIR] : 0.165172
##
##                   Kappa : 0.267
##
##  Mcnemar's Test P-Value : 0.006706
##
##             Sensitivity : 0.3881
##             Specificity : 0.8571
##          Pos Pred Value : 0.5778
##          Neg Pred Value : 0.7355
```

```
##             Prevalence : 0.3350
##         Detection Rate : 0.1300
##   Detection Prevalence : 0.2250
##      Balanced Accuracy : 0.6226
##
##       'Positive' Class : 0
##
```

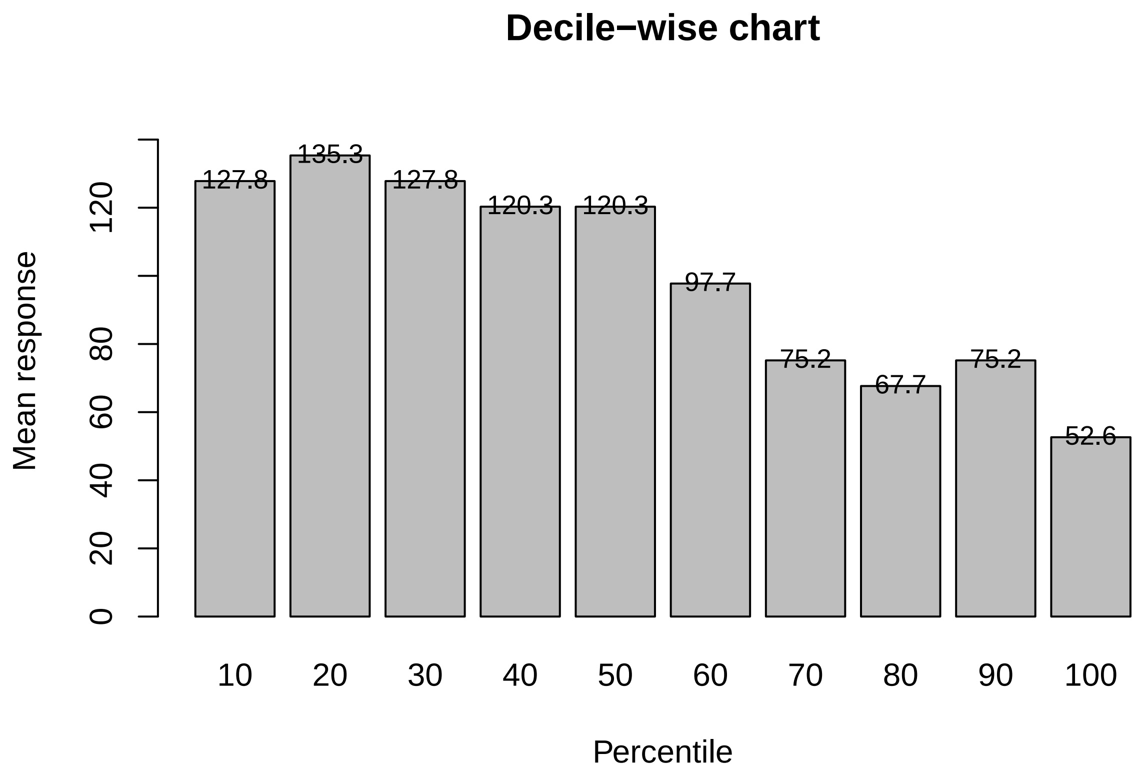
Neural network model, Cost Metrix: Reference Bad Good Predited Bad 0 100*19=1900 Good 41*500=20500 0 Gain Matrix: Reference Bad Good Predicted Bad 0 0 Good -500*41=-20500 100*114=11400 Neuralnet Model, net profit is -9100.

So by looking over all the models,the logistic regression model provides the best net profit.

4.Let's try and improve our performance. Rather than accept the default classification of all applicants' credit status, use the estimated probabilities (propensities) from the logistic regression (where success means 1) as a basis for selecting the best credit risks first, followed by poorer-risk applicants. Create a vector containing the net profit for each record in the validation set. Use this vestor to create a decile-wise lift chart for the validation set that incorporates the net profit.

Problem (a): How far into the validation data should you go to get maximum net profit? (often, this is specified as a percentile or rounded to deciles.)

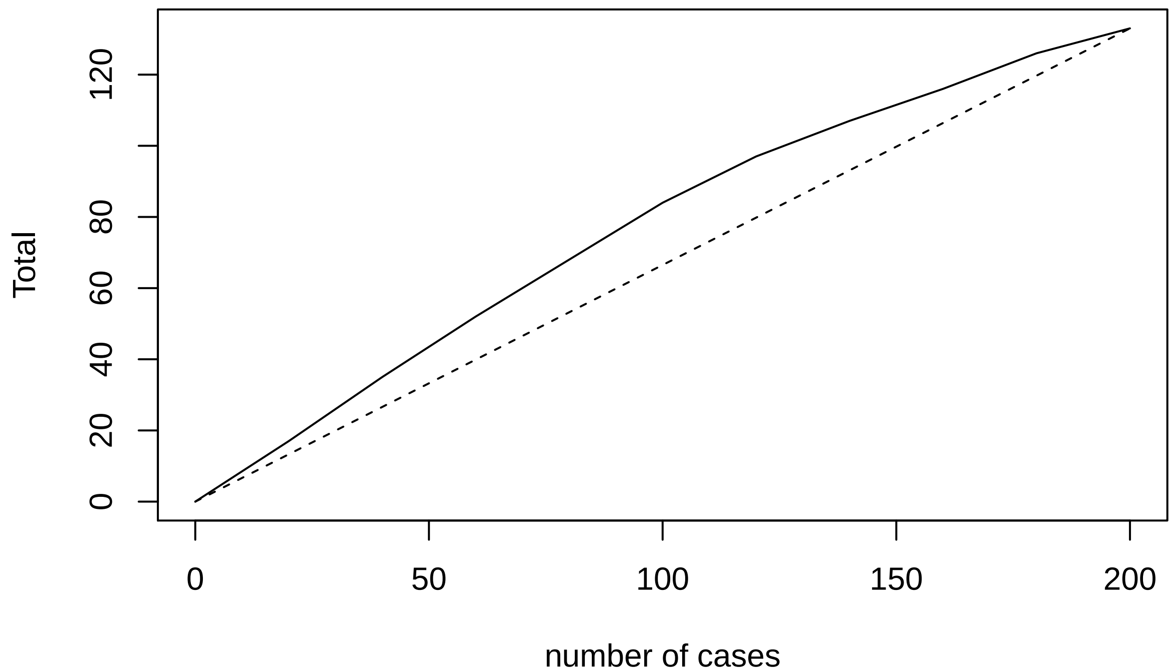
```
netprofit <- data.frame(Predicted = pred_v, Actual = valid_data$RESPONSE)
netprofit <- netprofit[order(-netprofit$Predicted),]
netprofit$net_profit <- netprofit$Actual*100

net_profit <- as.vector(netprofit$net_profit)
library(gains)
gain <- gains(net_profit, netprofit$Predicted, groups=10)
heights <- gain$mean.resp/mean(netprofit$Actual)
midpoints <- barplot(heights, names.arg = gain$depth, ylim = c(0,150),
xlab = "Percentile", ylab = "Mean response", main = "Decile-wise chart")
text(midpoints, heights+0.5, labels=round(heights, 1), cex = 0.8)
```



From this chart, we can easily see that we can use model to select the top 50% data with the highest propensities to get maximum net profit.

Problem (b):if this logistic regression model is used to score to future applicants, what "probability of success" cutoff should be used in extending credit?

```
# plot lift chart
plot(c(0,gain$cume.pct.of.total*sum(netprofit$Actual))~c(0,gain$cume.obs),
xlab="number of cases", ylab="Total", main="", type="l")
lines(c(0,sum(netprofit$Actual))~c(0, dim(netprofit)[1]), lty=2)
```




```
# plot a ROC curve
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

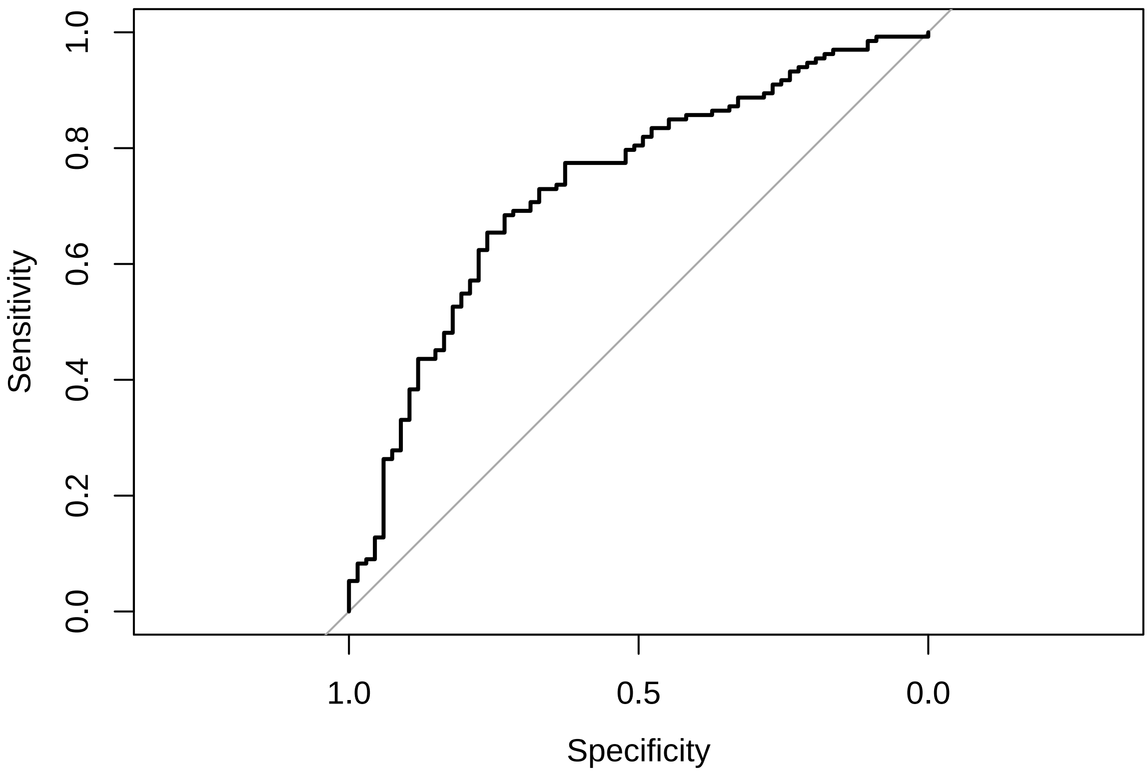
```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
r <- roc(netprofit$Actual, netprofit$Predicted)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
plot.roc(r)
```




```r
auc(r)
```

```
## Area under the curve: 0.7356
```

```r
cut_off <- netprofit$Predicted[round(length(netprofit$Predicted)*0.5)]
cut_off
```

```
## [1] 0.7562256
```

So, 0.756 cutoff value should be used in extending credit.

In this case study, I can conclude that logistic regression model is the best model. However, the bank cannot be guaranteed to bave benefit using the highest accruraccy model. The top 50% of the data provides the best profit. The best decision should be made by using the cutoff value or the top 30% of the validation data.

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this: