Class 6 - Classification Analysis

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
data = read.csv("G:/My Drive/Master-Data-Science/Semester_1/Business_Analytics/Data/index.csv", header=T
str(data)
```

```
## ’data.frame’:    1000 obs. of  21 variables:
##  $ Creditability                    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Account.Balance                  : int  1 1 2 1 1 1 1 1 4 2 ...
##  $ Duration.of.Credit..month.       : int  18 9 12 12 12 10 8 6 18 24 ...
##  $ Payment.Status.of.Previous.Credit: int  4 4 2 4 4 4 4 4 4 2 ...
##  $ Purpose                          : int  2 0 9 0 0 0 0 0 3 3 ...
##  $ Credit.Amount                    : int  1049 2799 841 2122 2171 2241 3398 1361 1098 3758 ...
##  $ Value.Savings.Stocks             : int  1 1 2 1 1 1 1 1 1 3 ...
##  $ Length.of.current.employment     : int  2 3 4 3 3 2 4 2 1 1 ...
##  $ Instalment.per.cent              : int  4 2 2 3 4 1 1 2 4 1 ...
##  $ Sex...Marital.Status             : int  2 3 2 3 3 3 3 3 2 2 ...
##  $ Guarantors                       : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Duration.in.Current.address      : int  4 2 4 2 4 3 4 4 4 4 ...
##  $ Most.valuable.available.asset    : int  2 1 1 1 2 1 1 1 3 4 ...
##  $ Age..years.                      : int  21 36 23 39 38 48 39 40 65 23 ...
##  $ Concurrent.Credits               : int  3 3 3 3 1 3 3 3 3 3 ...
##  $ Type.of.apartment                : int  1 1 1 1 2 1 2 2 2 1 ...
##  $ No.of.Credits.at.this.Bank       : int  1 2 1 2 2 2 2 1 2 1 ...
##  $ Occupation                       : int  3 3 2 2 2 2 2 2 1 1 ...
##  $ No.of.dependents                 : int  1 2 1 2 1 2 1 2 1 1 ...
##  $ Telephone                        : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Foreign.Worker                   : int  1 1 1 2 2 2 2 2 1 1 ...
```

```
table(data$Purpose)/1000*100
```

```
##
##    0    1    2    3    4    5    6    8    9   10
## 23.4 10.3 18.1 28.0  1.2  2.2  5.0  0.9  9.7  1.2
```

```r
data$Account.Balance <- replace(data$Account.Balance, data$Account.Balance==4, 3)
data$Account.Balance = factor(data$Account.Balance, levels = seq(1,3), labels = c('No Account', 'No bala

data$Payment.Status.of.Previous.Credit[data$Payment.Status.of.Previous.Credit <=1] =1
data$Payment.Status.of.Previous.Credit[data$Payment.Status.of.Previous.Credit ==2] = 2
data$Payment.Status.of.Previous.Credit[data$Payment.Status.of.Previous.Credit >=3] = 3
data$Payment.Status.of.Previous.Credit = factor(data$Payment.Status.of.Previous.Credit, levels = seq(1,

data$Value.Savings.Stocks[data$Value.Savings.Stocks == 4] = 3
data$Value.Savings.Stocks[data$Value.Savings.Stocks == 5] = 4
data$Value.Savings.Stocks = factor(data$Value.Savings.Stocks, levels = seq(1,4), labels = c('None','Bel

data$Length.of.current.employment[data$Length.of.current.employment == 2] = 1
data$Length.of.current.employment[data$Length.of.current.employment == 3] = 2
data$Length.of.current.employment[data$Length.of.current.employment == 4] = 3
data$Length.of.current.employment[data$Length.of.current.employment == 5] = 4
data$Length.of.current.employment = factor(data$Length.of.current.employment, levels = seq(1,4), labels

data$Sex...Marital.Status[data$Sex...Marital.Status <=2] = 1
data$Sex...Marital.Status[data$Sex...Marital.Status ==3] = 2
data$Sex...Marital.Status[data$Sex...Marital.Status ==4] = 3
data$Sex...Marital.Status = factor(data$Sex...Marital.Status, levels = seq(1,3), labels = c('Male Divor

data$No.of.Credits.at.this.Bank[data$No.of.Credits.at.this.Bank == 3] = 2
data$No.of.Credits.at.this.Bank = factor(data$No.of.Credits.at.this.Bank, levels = seq(1,2), labels = c

data$Guarantors[data$Guarantors >= 2] = 2
data$Guarantors = factor(data$Guarantors, levels = seq(1,2), labels = c('None','Yes'))

data$Concurrent.Credits[data$Concurrent.Credits <=2] = 1
data$Concurrent.Credits[data$Concurrent.Credits ==3] = 2
data$Concurrent.Credits = factor(data$Concurrent.Credits, levels = seq(1,2), labels = c('Other Banks or

data = data[-21]

data$Purpose[data$Purpose ==1] = 1
data$Purpose[data$Purpose ==2] = 2
data$Purpose[data$Purpose %in% c(3,4,5,6)] = 3
data$Purpose[data$Purpose %in% c(8,9,10,0)] = 4
data$Purpose = factor(data$Purpose, levels = seq(1,4), labels = c('New Car','Used Car','Home Related','
```

```r
str(data)
```

```
## 'data.frame':    1000 obs. of  20 variables:
##  $ Creditability                    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Account.Balance                  : Factor w/ 3 levels "No Account","No balance",..: 1 1 2 1 1 1 1
##  $ Duration.of.Credit..month.       : int  18 9 12 12 12 10 8 6 18 24 ...
##  $ Payment.Status.of.Previous.Credit: Factor w/ 3 levels "Some Problems",..: 3 3 2 3 3 3 3 3 3 2 ...
##  $ Purpose                          : Factor w/ 4 levels "New Car","Used Car",..: 2 4 4 4 4 4 4 4 3 3
##  $ Credit.Amount                    : int  1049 2799 841 2122 2171 2241 3398 1361 1098 3758 ...
##  $ Value.Savings.Stocks             : Factor w/ 4 levels "None","Below 100 DM",..: 1 1 2 1 1 1 1 1 1
##  $ Length.of.current.employment     : Factor w/ 4 levels "Below 1 year (including unemployed)",..: 1
##  $ Instalment.per.cent              : int  4 2 2 3 4 1 1 2 4 1 ...
##  $ Sex...Marital.Status            : Factor w/ 3 levels "Male Divorces/Single",..: 1 2 1 2 2 2 2 2
##  $ Guarantors                      : Factor w/ 2 levels "None","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Duration.in.Current.address     : int  4 2 4 2 4 3 4 4 4 4 ...
##  $ Most.valuable.available.asset   : int  2 1 1 1 2 1 1 1 3 4 ...
##  $ Age..years.                     : int  21 36 23 39 38 48 39 40 65 23 ...
##  $ Concurrent.Credits              : Factor w/ 2 levels "Other Banks or Dept Stores",..: 2 2 2 2 1
##  $ Type.of.apartment               : int  1 1 1 1 2 1 2 2 2 1 ...
##  $ No.of.Credits.at.this.Bank      : Factor w/ 2 levels "1","More than 1": 1 2 1 2 2 2 2 2 1 2 1 ...
##  $ Occupation                      : int  3 3 2 2 2 2 2 2 1 1 ...
##  $ No.of.dependents                : int  1 2 1 2 1 2 1 2 1 1 ...
##  $ Telephone                       : int  1 1 1 1 1 1 1 1 1 1 ...
```

Statistical Testing

Chi-square for

```
Categorical.Table = data.frame(
    'Variable' = character(),
  'p-value' = numeric()
)

for (i in colnames(data[,-c(1,3,6,14)])){
  test = chisq.test(table(data$Creditability,data[,i]))
  test2 = data.frame(i,test$p.value)
  Categorical.Table = rbind(Categorical.Table, test2)
}
Categorical.Table
```

```
##                                    i test.p.value
## 1                    Account.Balance 5.742621e-27
## 2  Payment.Status.of.Previous.Credit 1.557328e-12
## 3                            Purpose 2.760708e-04
## 4                 Value.Savings.Stocks 8.335937e-08
## 5        Length.of.current.employment 4.220685e-04
## 6                  Instalment.per.cent 1.400333e-01
## 7                 Sex...Marital.Status 1.043498e-02
## 8                           Guarantors 1.000000e+00
## 9          Duration.in.Current.address 8.615521e-01
## 10      Most.valuable.available.asset 2.858442e-05
## 11                  Concurrent.Credits 4.763431e-04
## 12                   Type.of.apartment 8.810311e-05
## 13          No.of.Credits.at.this.Bank 1.614375e-01
## 14                          Occupation 5.965816e-01
## 15                    No.of.dependents 1.000000e+00
## 16                           Telephone 2.788762e-01
```

```
#Numerical.Table = data.frame(
#    Variable = character(),
#    'mean.credit.worthy' = numeric(),
#    'mean.credit.nonworthy' = numeric(),
#    'p.value' = numeric()
#)
```

```r
#for (i in colnames(data[,c(3,6,14)])){
#  test = t.test(data[,i] ~ data$Creditability)
#  Numerical.Table[Variable] = i
#  Numerical.Table[mean.credit.worthy] = test$estimate[1]
#  Numerical.Table[mean.credit.nonworthy] = test$estimate[2]
#  Numerical.Table[p.value] = test$p.value
#}
#Numerical.Table
```

Train test split

```r
indexes = sample(1:nrow(data), size = 0.5*nrow(data))
Train = data[indexes,]
Test = data[-indexes,]
```

Logistic Regression

generalized linear model = glm()

- when y is discrete/binary

$$H_0 \ : \ B_j = 0H_1 : B_j \neq 0$$

Create initial model

```r
logisticmodel50 = glm(Creditability~Account.Balance+Payment.Status.of.Previous.Credit+Purpose+Value.Sav
summary(logisticmodel50)
```

```
##
## Call:
## glm(formula = Creditability ~ Account.Balance + Payment.Status.of.Previous.Credit +
##     Purpose + Value.Savings.Stocks + Length.of.current.employment +
##     Sex...Marital.Status + Most.valuable.available.asset + Type.of.apartment +
##     Concurrent.Credits + Duration.in.Current.address + Credit.Amount +
##     Age..years., family = "binomial", data = Train)
##
## Coefficients:
##                                                        Estimate
## (Intercept)                                           3.404e-01
## Account.BalanceNo balance                             7.183e-01
## Account.BalanceSome balance                           1.812e+00
## Payment.Status.of.Previous.CreditPaid Up              6.542e-01
## Payment.Status.of.Previous.CreditNo Problems(in this bank)  1.477e+00
## PurposeUsed Car                                      -6.604e-01
## PurposeHome Related                                  -9.039e-01
## PurposeOther                                               -1.285e+00
## Value.Savings.StocksBelow 100 DM                            3.880e-02
## Value.Savings.Stocks[100, 1000)                             1.688e+00
## Value.Savings.StocksAbove 1000 DM                           8.151e-01
## Length.of.current.employment[1,4)                           7.377e-02
## Length.of.current.employment[4,7)                           5.803e-01
## Length.of.current.employmentAbove 7                         1.797e-01
## Sex...Marital.StatusMale Married/Widowed                    2.890e-01
## Sex...Marital.StatusFemale                                 -3.569e-02
## Most.valuable.available.asset                              -2.757e-01
## Type.of.apartment                                           2.593e-01
## Concurrent.CreditsNone                                     -2.756e-02
## Duration.in.Current.address                                -9.225e-02
## Credit.Amount                                              -9.737e-05
## Age..years.                                                -5.150e-04
##                                                            Std. Error z value
## (Intercept)                                                 9.073e-01   0.375
## Account.BalanceNo balance                                   2.810e-01   2.556
## Account.BalanceSome balance                                 2.994e-01   6.052
## Payment.Status.of.Previous.CreditPaid Up                    3.907e-01   1.674
## Payment.Status.of.Previous.CreditNo Problems(in this bank)  4.101e-01   3.603
## PurposeUsed Car                                             5.161e-01  -1.280
## PurposeHome Related                                         4.814e-01  -1.878
## PurposeOther                                                4.658e-01  -2.758
## Value.Savings.StocksBelow 100 DM                            3.831e-01   0.101
## Value.Savings.Stocks[100, 1000)                             5.736e-01   2.943
## Value.Savings.StocksAbove 1000 DM                           3.370e-01   2.419
## Length.of.current.employment[1,4)                           3.033e-01   0.243
## Length.of.current.employment[4,7)                           3.775e-01   1.537
## Length.of.current.employmentAbove 7                         3.593e-01   0.500
## Sex...Marital.StatusMale Married/Widowed                    2.667e-01   1.084
## Sex...Marital.StatusFemale                                  3.940e-01  -0.091
## Most.valuable.available.asset                               1.267e-01  -2.176
## Type.of.apartment                                           2.394e-01   1.083
## Concurrent.CreditsNone                                      3.032e-01  -0.091
## Duration.in.Current.address                                 1.125e-01  -0.820
## Credit.Amount                                               4.439e-05  -2.193
## Age..years.                                                 1.230e-02  -0.042
##                                                            Pr(>|z|)    
## (Intercept)                                                0.707542    
## Account.BalanceNo balance                                  0.010575 *  
## Account.BalanceSome balance                                1.43e-09 ***
## Payment.Status.of.Previous.CreditPaid Up                   0.094053 .  
## Payment.Status.of.Previous.CreditNo Problems(in this bank) 0.000315 ***
## PurposeUsed Car                                            0.200698    
## PurposeHome Related                                        0.060430 .  
## PurposeOther                                               0.005821 ** 
## Value.Savings.StocksBelow 100 DM                           0.919339    
## Value.Savings.Stocks[100, 1000)                            0.003246 ** 
## Value.Savings.StocksAbove 1000 DM                          0.015584 *  
## Length.of.current.employment[1,4)                          0.807843    
## Length.of.current.employment[4,7)                          0.124219    
## Length.of.current.employmentAbove 7                        0.617073    
## Sex...Marital.StatusMale Married/Widowed                   0.278528    
## Sex...Marital.StatusFemale                                 0.927822
## Most.valuable.available.asset                              0.029547 *
## Type.of.apartment                                          0.278651
## Concurrent.CreditsNone                                     0.927577
## Duration.in.Current.address                                0.412394
## Credit.Amount                                              0.028292 *
## Age..years.                                                0.966605
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 603.93  on 499  degrees of freedom
## Residual deviance: 473.36  on 478  degrees of freedom
## AIC: 517.36
##
## Number of Fisher Scoring iterations: 5
```

Optimize model

```
logisticmodel50final = glm(Creditability~Account.Balance + Payment.Status.of.Previous.Credit + Purpose
summary(logisticmodel50final)
```

```
##
## Call:
## glm(formula = Creditability ~ Account.Balance + Payment.Status.of.Previous.Credit +
##     Purpose + Length.of.current.employment + Sex...Marital.Status,
##     family = "binomial", data = Train)
##
## Coefficients:
##                                                            Estimate Std. Error
## (Intercept)                                                -0.57135    0.58100
## Account.BalanceNo balance                                   0.72974    0.26092
## Account.BalanceSome balance                                 1.96784    0.28839
## Payment.Status.of.Previous.CreditPaid Up                    0.82492    0.34603
## Payment.Status.of.Previous.CreditNo Problems(in this bank)  1.54087    0.37450
## PurposeUsed Car                                            -0.47779    0.48047
## PurposeHome Related                                        -0.52828    0.44307
## PurposeOther                                               -0.99611    0.43196
## Length.of.current.employment[1,4)                           0.10372    0.29172
## Length.of.current.employment[4,7)                           0.37425    0.36156
## Length.of.current.employmentAbove 7                         0.09844    0.31952
## Sex...Marital.StatusMale Married/Widowed                    0.20574    0.24533
## Sex...Marital.StatusFemale                                  0.10520    0.37502
##                                                            z value Pr(>|z|)
## (Intercept)                                                 -0.983  0.32541
## Account.BalanceNo balance                                    2.797  0.00516 **
## Account.BalanceSome balance                                  6.824 8.88e-12 ***
## Payment.Status.of.Previous.CreditPaid Up                     2.384  0.01713 *
## Payment.Status.of.Previous.CreditNo Problems(in this bank)   4.114 3.88e-05 ***
## PurposeUsed Car                                             -0.994  0.32002
## PurposeHome Related                                         -1.192  0.23313
## PurposeOther                                              -2.306  0.02111 *
## Length.of.current.employment[1,4)                          0.356  0.72218
## Length.of.current.employment[4,7)                          1.035  0.30062
## Length.of.current.employmentAbove 7                        0.308  0.75801
## Sex...Marital.StatusMale Married/Widowed                   0.839  0.40168
## Sex...Marital.StatusFemale                                 0.281  0.77908
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 603.93  on 499  degrees of freedom
## Residual deviance: 503.63  on 487  degrees of freedom
## AIC: 529.63
##
## Number of Fisher Scoring iterations: 4
```

Obtain fitted values

```
fit50 = fitted.values(logisticmodel50final)
head(fit50)
```

```
##       408       343       590       542       114       944
## 0.5488750 0.9586358 0.6237571 0.7378883 0.8638148 0.5736276
```

Change binary response

```
thres = rep(0,500)
for (i in 1:500) {
  if(fit50[i]>0.5) {
    thres[i] = 1
  }
  else {
    thres[i] = 0
  }
}
str(thres)
```

```
##  num [1:500] 1 1 1 1 1 1 1 1 1 1 ...
```

```
str(Train$Creditability)
```

```
##  int [1:500] 1 1 1 1 1 0 1 0 1 0 ...
```

Create cross table

```
conf.mat = table(Train$Creditability, thres)
conf.mat
```

```
##    thres
##       0   1
##   0  57  89
##   1  33 321
```

Compute accuracy

```
sum(diag(conf.mat))/500*100
```

```
## [1] 75.6
```