IIMT2641 Assignment 4

Sibo Ding

Spring 2023

Load the Data

```
loans <- read.csv("Loans.csv")
head(loans) # First 6 rows
```

```
##   CreditPolicy Purpose.CC Purpose.DC Purpose.Edu Purpose.MP
Purpose.SB IntRate
## 1            1          0          1           0          0
0  0.1189
## 2            1          1          0           0          0
0  0.1071
## 3            1          0          1           0          0
0  0.1357
## 4            1          0          1           0          0
0  0.1008
## 5            1          1          0           0          0
0  0.1426
## 6            1          1          0           0          0
0  0.0788
##   Installment LogAnnualInc   Dti Fico DaysWithCrLine RevolBal
RevolUtil
## 1      829.10     11.35041 19.48  737       5639.958    28854
52.1
## 2      228.22     11.08214 14.29  707       2760.000    33623
76.7
## 3      366.86     10.37349 11.63  682       4710.000     3511
25.6
## 4      162.34     11.35041  8.10  712       2699.958    33667
73.2
## 5      102.92     11.29973 14.97  667       4066.000     4740
39.5
## 6      125.13     11.90497 16.98  727       6120.042    50807
51.0
##   InqLast6mths Delinq2yrs PubRec NotFullyPaid
## 1            0          0      0            0
## 2            0          0      0            0
## 3            1          0      0            0
## 4            1          0      0            0
## 5            0          1      0            0
## 6            0          0      0            0
```

```
dim(loans) # Number of observations and variables
```

```
## [1] 9578   18

names(loans) # Names of variables

##  [1] "CreditPolicy"   "Purpose.CC"     "Purpose.DC"
"Purpose.Edu"
##  [5] "Purpose.MP"     "Purpose.SB"     "IntRate"
"Installment"
##  [9] "LogAnnualInc"   "Dti"            "Fico"
"DaysWithCrLine"
## [13] "RevolBal"       "RevolUtil"      "InqLast6mths"   "Delinq2yrs"
## [17] "PubRec"         "NotFullyPaid"

# Change to the categorical/factor variable
loans$NotFullyPaid <- as.factor(loans$NotFullyPaid)
```

Train-test Split

```
library(caTools)
set.seed(12)
# Randomly split the dataset with 70% in the training set
spl <- sample.split(loans$NotFullyPaid, SplitRatio = 0.7)
table(spl) # Number of TRUE/FALSE data

## spl
## FALSE  TRUE
##  2873  6705

train <- loans |> subset(spl == TRUE) # Training set
test <- loans |> subset(spl == FALSE) # Test set
```

Baseline Model Accuracy

```
table(test$NotFullyPaid)["0"] / length(test$NotFullyPaid)

##         0
## 0.8398886
```

Logistic Regression

```
model1 <- glm(NotFullyPaid ~ ., data = train, family = binomial)
summary(model1)

##
## Call:
## glm(formula = NotFullyPaid ~ ., family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2864  -0.6201  -0.4964  -0.3652   2.5926
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     7.463e+00  1.544e+00   4.835 1.33e-06 ***
```

```
## CreditPolicy    -3.757e-01  1.012e-01  -3.713 0.000205 ***
## Purpose.CC      -5.588e-01  1.279e-01  -4.369 1.25e-05 ***
## Purpose.DC      -3.289e-01  8.631e-02  -3.811 0.000138 ***
## Purpose.Edu      1.789e-02  1.819e-01   0.098 0.921649
## Purpose.MP      -2.608e-01  1.930e-01  -1.351 0.176656
## Purpose.SB       4.482e-01  1.358e-01   3.299 0.000969 ***
## IntRate          4.133e+00  2.076e+00   1.991 0.046448 *
## Installment      1.262e-03  2.077e-04   6.076 1.24e-09 ***
## LogAnnualInc    -3.866e-01  7.080e-02  -5.460 4.75e-08 ***
## Dti             -1.997e-03  5.423e-03  -0.368 0.712645
## Fico            -7.964e-03  1.694e-03  -4.701 2.59e-06 ***
## DaysWithCrLine   5.465e-06  1.608e-05   0.340 0.733912
## RevolBal         2.928e-06  1.141e-06   2.566 0.010299 *
## RevolUtil        1.423e-03  1.530e-03   0.930 0.352410
## InqLast6mths     7.029e-02  1.682e-02   4.178 2.94e-05 ***
## Delinq2yrs      -1.178e-01  6.888e-02  -1.710 0.087350 .
## PubRec           2.059e-01  1.191e-01   1.728 0.083903 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5896.6  on 6704  degrees of freedom
## Residual deviance: 5507.4  on 6687  degrees of freedom
## AIC: 5543.4
##
## Number of Fisher Scoring iterations: 5
```

```
# Significant independent variables (with p < 0.05)
# Note: `Intercept` is not an independent variable
which(summary(model1)$coefficients[, 4] < 0.05)
```

```
##  (Intercept) CreditPolicy   Purpose.CC   Purpose.DC   Purpose.SB
IntRate
##            1            2            3            4            7
8
##  Installment LogAnnualInc         Fico     RevolBal InqLast6mths
##            9           10           12           14           16
```

Differences between Two Logits

```
coef_fico <- summary(model1)$coefficients["Fico", 1]
coef_fico * (700 - 710)
```

```
## [1] 0.07963508
```

Predict the Test Set

```
predict_test1 <- predict(model1, type = "response", newdata = test)

# Confusion matrix for out-of-sample prediction at threshold value 0.5
```

```
confusion_matrix <- table(test$NotFullyPaid, predict_test1 > 0.5)
confusion_matrix
```

```
##
##     FALSE TRUE
##   0  2400   13
##   1   445   15
```

```
# Accuracy
(confusion_matrix[1, 1] + confusion_matrix[2, 2]) /
length(test$NotFullyPaid)
```

```
## [1] 0.8405848
```

```
# Baseline model accuracy
table(test$NotFullyPaid)["0"] / length(test$NotFullyPaid)
```

```
##         0
## 0.8398886
```

The logistic regression model is slightly more accurate than the baseline model.

Logistic Regression Using `IntRate`

```
model2 <- glm(NotFullyPaid ~ IntRate, data = train, family = binomial)
summary(model2)
```

```
##
## Call:
## glm(formula = NotFullyPaid ~ IntRate, family = binomial, data =
train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1039  -0.6296  -0.5366  -0.4160   2.3192
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.8590     0.1697  -22.75   <2e-16 ***
## IntRate      17.3673     1.2717   13.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5896.6  on 6704  degrees of freedom
## Residual deviance: 5702.9  on 6703  degrees of freedom
## AIC: 5706.9
##
## Number of Fisher Scoring iterations: 4
```

`IntRate` is significant in this model at 0.1%. It is also significant in the first model at 5%.
First, this difference in significance level is acceptable. Second, this difference may be because some information in the second model is explained by other independent variables in the first model.

Predict the Test Set

```
predict_test2 <- predict(model2, type = "response", newdata = test)

# Highest predicted probability
max(predict_test2)

## [1] 0.4562598

# No. of loans would not be paid back in full
table(predict_test2 > 0.5)["TRUE"]

## <NA>
##   NA
```