Load Data Prediction

2024-11-11

```r
# Install and load necessary libraries
if(!require(dplyr)) install.packages("dplyr", dependencies=TRUE)
```

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

```r
if(!require(ggplot2)) install.packages("ggplot2", dependencies=TRUE)
```

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

```r
if(!require(caret)) install.packages("caret", dependencies=TRUE)
```

```
## Loading required package: caret

## Warning: package 'caret' was built under R version 4.4.2

## Loading required package: lattice
```

```r
if(!require(reshape2)) install.packages("reshape2", dependencies=TRUE)
```

```
## Loading required package: reshape2

## Warning: package 'reshape2' was built under R version 4.4.2
```

```r
if(!require(GGally)) install.packages("GGally", dependencies=TRUE)
```

```
## Loading required package: GGally

## Warning: package 'GGally' was built under R version 4.4.2

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```r
library(GGally)
library(dplyr)
library(ggplot2)
library(caret)
library(reshape2)
```

```r
# Load the dataset
loan_data <- read.csv("Data/loan_data.csv")
```

```r
# Display summary and structure of the dataset
summary(loan_data)
```

```
##    person_age     person_gender      person_education   person_income    
##  Min.   : 20.00   Length:45000       Length:45000       Min.   :   8000  
##  1st Qu.: 24.00   Class :character   Class :character   1st Qu.:  47204  
##  Median : 26.00   Mode  :character   Mode  :character   Median :  67048  
##  Mean   : 27.76                                         Mean   :  80319  
##  3rd Qu.: 30.00                                         3rd Qu.:  95789  
##  Max.   :144.00                                         Max.   :7200766  
##  person_emp_exp   person_home_ownership   loan_amnt     loan_intent       
##  Min.   :  0.00   Length:45000          Min.   :  500   Length:45000      
##  1st Qu.:  1.00   Class :character      1st Qu.: 5000   Class :character  
##  Median :  4.00   Mode  :character      Median : 8000   Mode  :character  
##  Mean   :  5.41                         Mean   : 9583                     
##  3rd Qu.:  8.00                         3rd Qu.:12237                     
##  Max.   :125.00                         Max.   :35000                     
##  loan_int_rate   loan_percent_income cb_person_cred_hist_length  credit_score  
##  Min.   : 5.42   Min.   :0.0000      Min.   : 2.000             Min.   :390.0  
##  1st Qu.: 8.59   1st Qu.:0.0700      1st Qu.: 3.000             1st Qu.:601.0  
##  Median :11.01   Median :0.1200      Median : 4.000             Median :640.0  
##  Mean   :11.01   Mean   :0.1397      Mean   : 5.867             Mean   :632.6  
##  3rd Qu.:12.99   3rd Qu.:0.1900      3rd Qu.: 8.000             3rd Qu.:670.0  
##  Max.   :20.00   Max.   :0.6600      Max.   :30.000             Max.   :850.0  
##  previous_loan_defaults_on_file  loan_status    
##  Length:45000                   Min.   :0.0000  
##  Class :character               1st Qu.:0.0000  
##  Mode  :character               Median :0.0000  
##                                 Mean   :0.2222  
##                                 3rd Qu.:0.0000  
##                                 Max.   :1.0000
```

```r
str(loan_data)
```

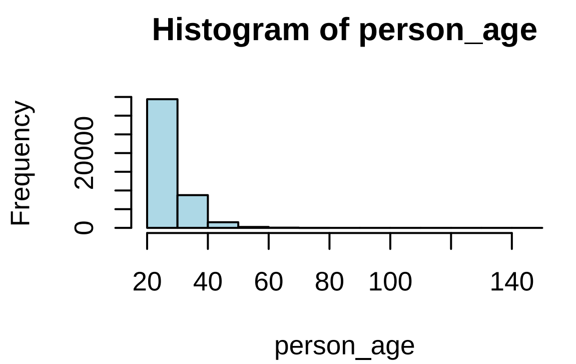
```
## 'data.frame':    45000 obs. of  14 variables:
##  $ person_age                    : num  22 21 25 23 24 21 26 24 24 21 ...
##  $ person_gender                 : chr  "female" "female" "female" "female" ...
##  $ person_education              : chr  "Master" "High School" "High School" "Bachelor" ...
##  $ person_income                 : num  71948 12282 12438 79753 66135 ...
##  $ person_emp_exp                : int  0 0 3 0 1 0 1 5 3 0 ...
##  $ person_home_ownership         : chr  "RENT" "OWN" "MORTGAGE" "RENT" ...
##  $ loan_amnt                     : num  35000 1000 5500 35000 35000 2500 35000 35000 35000 1600 ...
##  $ loan_intent                   : chr  "PERSONAL" "EDUCATION" "MEDICAL" "MEDICAL" ...
##  $ loan_int_rate                 : num  16 11.1 12.9 15.2 14.3 ...
##  $ loan_percent_income           : num  0.49 0.08 0.44 0.44 0.53 0.19 0.37 0.37 0.35 0.13 ...
##  $ cb_person_cred_hist_length    : num  3 2 3 2 4 2 3 4 2 3 ...
##  $ credit_score                  : int  561 504 635 675 586 532 701 585 544 640 ...
##  $ previous_loan_defaults_on_file: chr  "No" "Yes" "No" "No" ...
##  $ loan_status                   : int  1 0 1 1 1 1 1 1 1 1 ...
```

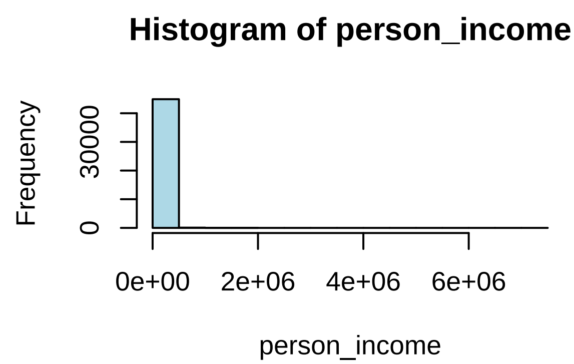
```r
# Check for missing values
colSums(is.na(loan_data))
```

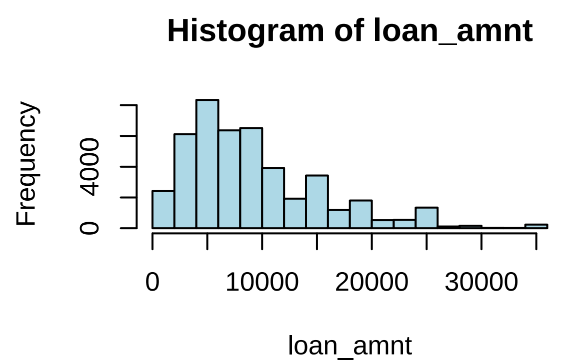
```
##                     person_age                  person_gender 
##                              0                              0
```

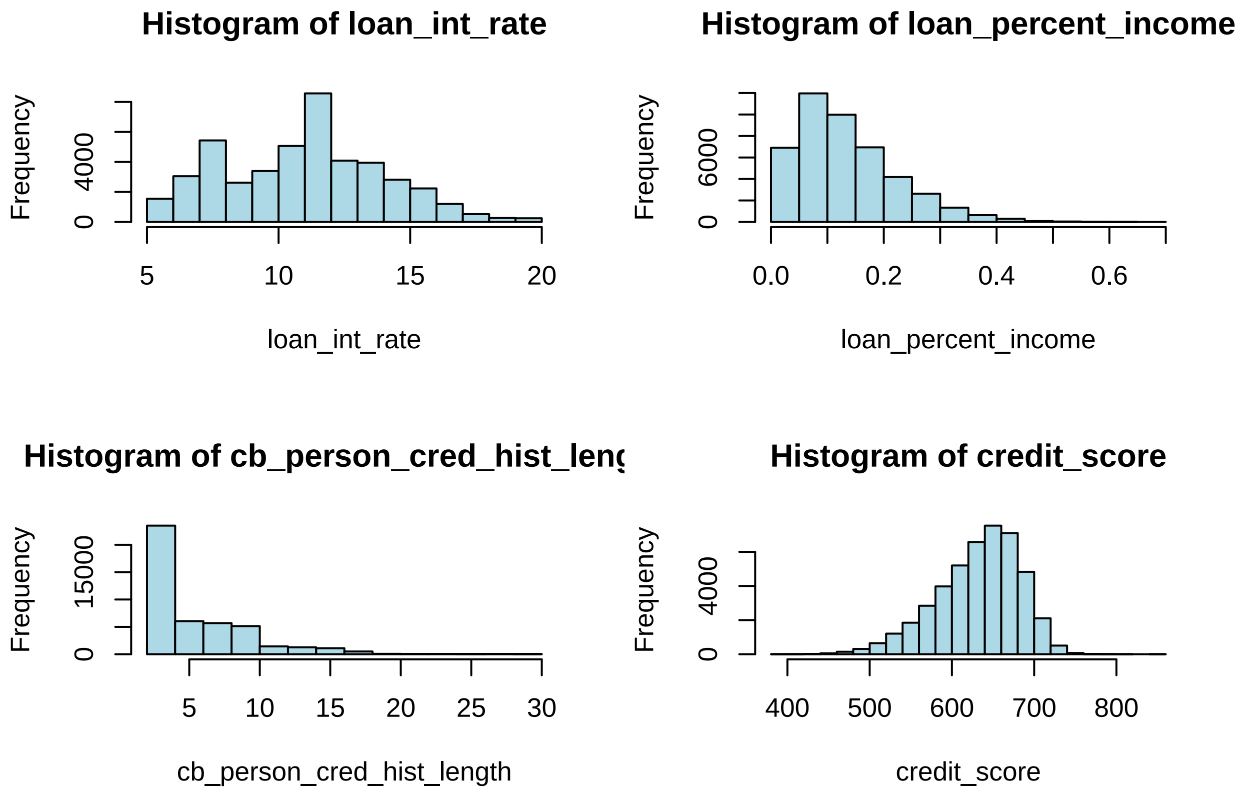
```
##                 person_education                    person_income
##                                0                                0
##                   person_emp_exp            person_home_ownership
##                                0                                0
##                       loan_amnt                      loan_intent
##                                0                                0
##                    loan_int_rate              loan_percent_income
##                                0                                0
##       cb_person_cred_hist_length                     credit_score
##                                0                                0
## previous_loan_defaults_on_file                      loan_status
##                                0                                0
```

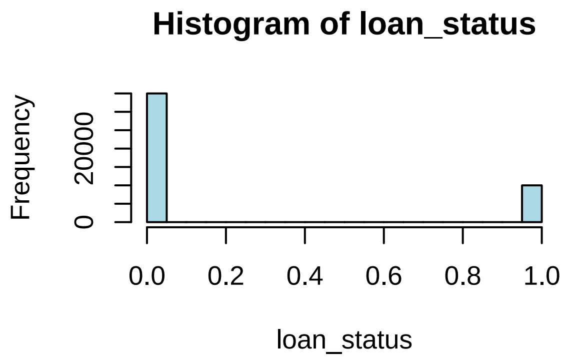
```r
# Plot Histograms for Numeric Variables
numeric_vars <- sapply(loan_data, is.numeric)
par(mfrow=c(2,2))
for (col in names(loan_data)[numeric_vars]) {
  hist(loan_data[[col]], main=paste("Histogram of", col), xlab=col, col="lightblue", border="black")
}
```





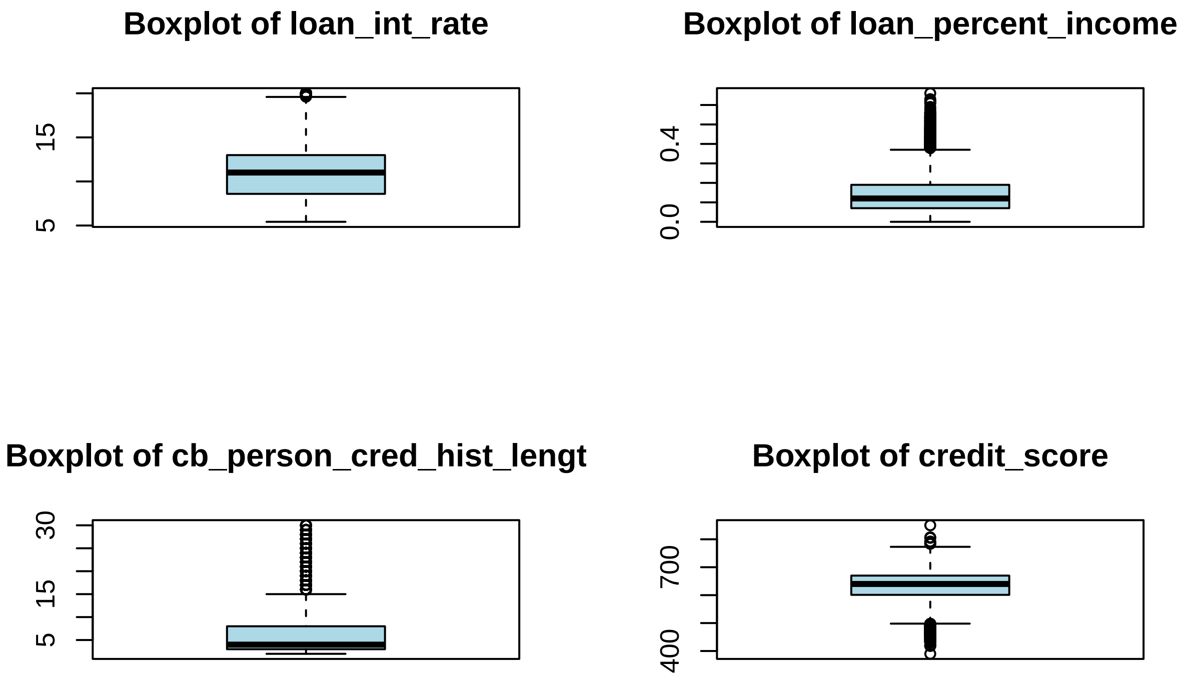


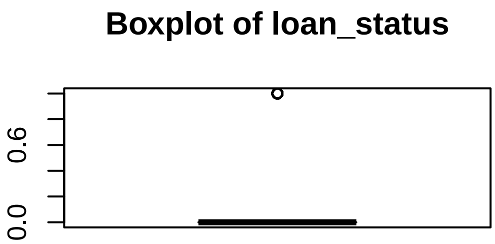

Histogram of loan_int_rate
Frequency
0
4000
5
10
15
20
loan_int_rate
Histogram of loan_percent_income
Frequency
0
6000
0.0
0.2
0.4
0.6
loan_percent_income
Histogram of cb_person_cred_hist_leng
Frequency
0
15000
5
10
15
20
25
30
cb_person_cred_hist_length
Histogram of credit_score
Frequency
0
4000
400
500
600
700
800
credit_score




```r
# Boxplots for Numeric Variables
par(mfrow=c(2,2))
for (col in names(loan_data)[numeric_vars]) {
  boxplot(loan_data[[col]], main=paste("Boxplot of", col), col="lightblue")
}
```


Boxplot of person_age
20
80
140
Boxplot of person_income
0e+00
6e+06
Boxplot of person_emp_exp
0
60
120
Boxplot of loan_amnt
0
20000


Boxplot of loan_int_rate
5
15
Boxplot of loan_percent_income
0.0
0.4
Boxplot of cb_person_cred_hist_lengt
5
15
30
Boxplot of credit_score
400
700




```r
# Bar plots for Categorical Variables
cat_vars <- sapply(loan_data, is.factor)
for (col in names(loan_data)[cat_vars]) {
  print(table(loan_data[[col]]))
  barplot(table(loan_data[[col]]), main=paste("Barplot of", col), col="lightgreen")
}
```

```r
# Cap 'person_age' to a maximum of 100
loan_data$person_age <- ifelse(loan_data$person_age > 100, 100, loan_data$person_age)

# Cap 'person_income' to the 99th percentile to reduce the impact of extreme values
income_cap <- quantile(loan_data$person_income, 0.99)
loan_data$person_income <- ifelse(loan_data$person_income > income_cap, income_cap, loan_data$person_in
```

```r
# Convert 'previous_loan_defaults_on_file' from Yes/No to 1/0
loan_data$previous_loan_defaults_on_file <- ifelse(loan_data$previous_loan_defaults_on_file == "Yes", 1

# Convert binary categorical features (e.g., gender) to 0 and 1
loan_data$person_gender <- ifelse(loan_data$person_gender == "female", 0, 1)

# One-Hot Encoding for other categorical variables (education, home_ownership, loan_intent)
loan_data <- cbind(loan_data, model.matrix(~person_education + person_home_ownership + loan_intent - 1,

loan_data <- loan_data[, !(names(loan_data) %in% c("person_education", "person_home_ownership", "loan_i
```

```r
# Scaling numerical features
scaling_vars <- c("person_income", "loan_amnt", "loan_int_rate", "loan_percent_income", "credit_score")
```

```r
scaler <- preProcess(loan_data[, scaling_vars], method = c("center", "scale"))
loan_data[, scaling_vars] <- predict(scaler, loan_data[, scaling_vars])
```

```r
# Correlation matrix of numeric variables
cor_matrix <- cor(loan_data[, numeric_vars])
print(cor_matrix)
```

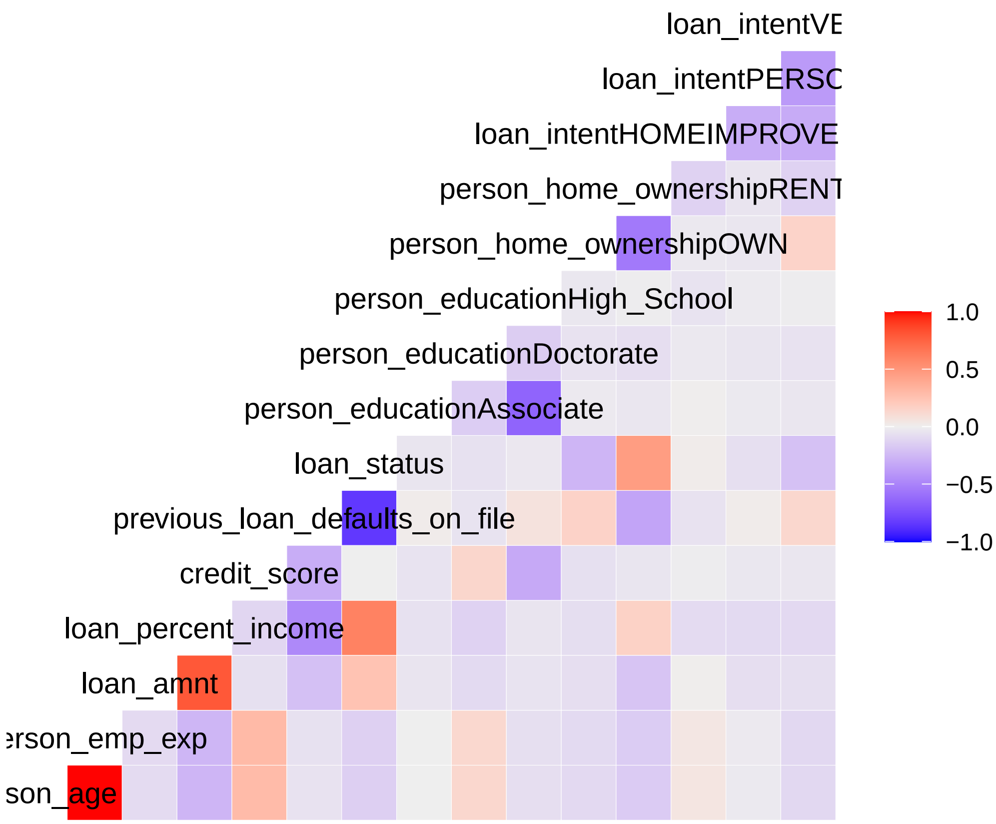
```
##                                  person_age person_emp_exp    loan_amnt
## person_age                      1.000000000   0.9516695320  0.051442959
## person_emp_exp                  0.951669532   1.0000000000  0.044589394
## loan_amnt                       0.051442959   0.0445893936  1.000000000
## loan_percent_income            -0.043065469  -0.0398615277  0.593011449
## credit_score                    0.178045324   0.1861961342  0.009074282
## previous_loan_defaults_on_file -0.025784767  -0.0292308430 -0.059008529
## loan_status                    -0.021315263  -0.0204812589  0.107714467
## person_educationAssociate       0.038875662   0.0366664958  0.004287908
## person_educationDoctorate       0.114698118   0.1066929377  0.006514700
## person_educationHigh School     0.005008056   0.0083725237 -0.003788349
## person_home_ownershipOWN       -0.003691109   0.0004901486 -0.025289845
## person_home_ownershipRENT      -0.036262929  -0.0344986463 -0.136521142
## loan_intentHOMEIMPROVEMENT      0.069110173   0.0581641168  0.045656894
## loan_intentPERSONAL             0.027038520   0.0268628239  0.001476392
## loan_intentVENTURE             -0.007327688  -0.0061556040  0.005500205
##                                loan_percent_income credit_score
## person_age                           -0.0430654686  0.178045324
## person_emp_exp                       -0.0398615277  0.186196134
## loan_amnt                             0.5930114493  0.009074282
## loan_percent_income                   1.0000000000 -0.011483096
## credit_score                         -0.0114830959  1.000000000
## previous_loan_defaults_on_file       -0.2032518569 -0.183005161
## loan_status                           0.3848803800 -0.007647176
## person_educationAssociate             0.0040587046 -0.038673191
## person_educationDoctorate             0.0003949995  0.082867927
## person_educationHigh School           0.0001001216 -0.164694833
## person_home_ownershipOWN              0.0529003909 -0.002891385
## person_home_ownershipRENT             0.1252820957 -0.005051217
## loan_intentHOMEIMPROVEMENT           -0.0156041197  0.010227720
## loan_intentPERSONAL                  -0.0077132071  0.003794876
## loan_intentVENTURE                    0.0016012805  0.009705433
##                                previous_loan_defaults_on_file  loan_status
## person_age                                       -0.025784767 -0.021315263
## person_emp_exp                                   -0.029230843 -0.020481259
## loan_amnt                                        -0.059008529  0.107714467
## loan_percent_income                              -0.203251857  0.384880380
## credit_score                                     -0.183005161 -0.007647176
## previous_loan_defaults_on_file                    1.000000000 -0.543096081
## loan_status                                      -0.543096081  1.000000000
## person_educationAssociate                         0.010979380 -0.002764610
## person_educationDoctorate                        -0.019599941  0.001832753
## person_educationHigh School                       0.029649902  0.001276836
## person_home_ownershipOWN                          0.053155501 -0.093666297
## person_home_ownershipRENT                        -0.138272502  0.255239005
## loan_intentHOMEIMPROVEMENT                       -0.021712749  0.033838061
## loan_intentPERSONAL                               0.004153455 -0.022487808
```

```
## loan_intentVENTURE                                0.052121963 -0.085991524
##                                person_educationAssociate
## person_age                                  0.0388756616
## person_emp_exp                              0.0366664958
## loan_amnt                                   0.0042879084
## loan_percent_income                         0.0040587046
## credit_score                               -0.0386731913
## previous_loan_defaults_on_file              0.0109793799
## loan_status                                -0.0027646098
## person_educationAssociate                   1.0000000000
## person_educationDoctorate                  -0.0714465246
## person_educationHigh School                -0.3636355046
## person_home_ownershipOWN                    0.0030895453
## person_home_ownershipRENT                  -0.0046291826
## loan_intentHOMEIMPROVEMENT                  0.0158945288
## loan_intentPERSONAL                         0.0007302009
## loan_intentVENTURE                         -0.0050270213
##                                person_educationDoctorate
## person_age                                  0.1146981183
## person_emp_exp                              0.1066929377
## loan_amnt                                   0.0065147001
## loan_percent_income                         0.0003949995
## credit_score                                0.0828679274
## previous_loan_defaults_on_file             -0.0195999407
## loan_status                                 0.0018327529
## person_educationAssociate                  -0.0714465246
## person_educationDoctorate                   1.0000000000
## person_educationHigh School                -0.0712195557
## person_home_ownershipOWN                    0.0025210759
## person_home_ownershipRENT                  -0.0104911106
## loan_intentHOMEIMPROVEMENT                  0.0105034961
## loan_intentPERSONAL                         0.0024376904
## loan_intentVENTURE                         -0.0004535703
##                                person_educationHigh School
## person_age                                    0.0050080561
## person_emp_exp                                0.0083725237
## loan_amnt                                    -0.0037883490
## loan_percent_income                           0.0001001216
## credit_score                                 -0.1646948327
## previous_loan_defaults_on_file                0.0296499025
## loan_status                                   0.0012768359
## person_educationAssociate                    -0.3636355046
## person_educationDoctorate                    -0.0712195557
## person_educationHigh School                   1.0000000000
## person_home_ownershipOWN                     -0.0059112063
## person_home_ownershipRENT                     0.0017234285
## loan_intentHOMEIMPROVEMENT                   -0.0124807573
## loan_intentPERSONAL                          -0.0008299693
## loan_intentVENTURE                            0.0015659514
##                                person_home_ownershipOWN
## person_age                                -0.0036911086
## person_emp_exp                             0.0004901486
## loan_amnt                                 -0.0252898449
## loan_percent_income                        0.0529003909
```

```
## credit_score                                -0.0028913852
## previous_loan_defaults_on_file               0.0531555014
## loan_status                                 -0.0936662968
## person_educationAssociate                    0.0030895453
## person_educationDoctorate                    0.0025210759
## person_educationHigh School                 -0.0059112063
## person_home_ownershipOWN                     1.0000000000
## person_home_ownershipRENT                   -0.2762607597
## loan_intentHOMEIMPROVEMENT                   0.0102938987
## loan_intentPERSONAL                          0.0049861053
## loan_intentVENTURE                           0.0910384428
##                                 person_home_ownershipRENT
## person_age                                   -0.036262929
## person_emp_exp                               -0.034498646
## loan_amnt                                    -0.136521142
## loan_percent_income                           0.125282096
## credit_score                                 -0.005051217
## previous_loan_defaults_on_file               -0.138272502
## loan_status                                   0.255239005
## person_educationAssociate                    -0.004629183
## person_educationDoctorate                    -0.010491111
## person_educationHigh School                   0.001723429
## person_home_ownershipOWN                     -0.276260760
## person_home_ownershipRENT                     1.000000000
## loan_intentHOMEIMPROVEMENT                   -0.054950837
## loan_intentPERSONAL                          -0.014433477
## loan_intentVENTURE                           -0.037609916
##                                 loan_intentHOMEIMPROVEMENT loan_intentPERSONAL
## person_age                                      0.06911017        0.0270385202
## person_emp_exp                                  0.05816412        0.0268628239
## loan_amnt                                       0.04565689        0.0014763918
## loan_percent_income                            -0.01560412       -0.0077132071
## credit_score                                    0.01022772        0.0037948760
## previous_loan_defaults_on_file                 -0.02171275        0.0041534548
## loan_status                                     0.03383806       -0.0224878076
## person_educationAssociate                       0.01589453        0.0007302009
## person_educationDoctorate                       0.01050350        0.0024376904
## person_educationHigh School                    -0.01248076       -0.0008299693
## person_home_ownershipOWN                        0.01029390        0.0049861053
## person_home_ownershipRENT                      -0.05495084       -0.0144334770
## loan_intentHOMEIMPROVEMENT                      1.00000000       -0.1548681218
## loan_intentPERSONAL                            -0.15486812        1.0000000000
## loan_intentVENTURE                             -0.15814681       -0.2059357521
##                                 loan_intentVENTURE
## person_age                           -0.0073276876
## person_emp_exp                       -0.0061556040
## loan_amnt                             0.0055002055
## loan_percent_income                   0.0016012805
## credit_score                          0.0097054325
## previous_loan_defaults_on_file        0.0521219628
## loan_status                          -0.0859915240
## person_educationAssociate            -0.0050270213
## person_educationDoctorate            -0.0004535703
## person_educationHigh School           0.0015659514
```

```
## person_home_ownershipOWN            0.0910384428
## person_home_ownershipRENT          -0.0376099162
## loan_intentHOMEIMPROVEMENT         -0.1581468072
## loan_intentPERSONAL                -0.2059357521
## loan_intentVENTURE                  1.0000000000
```

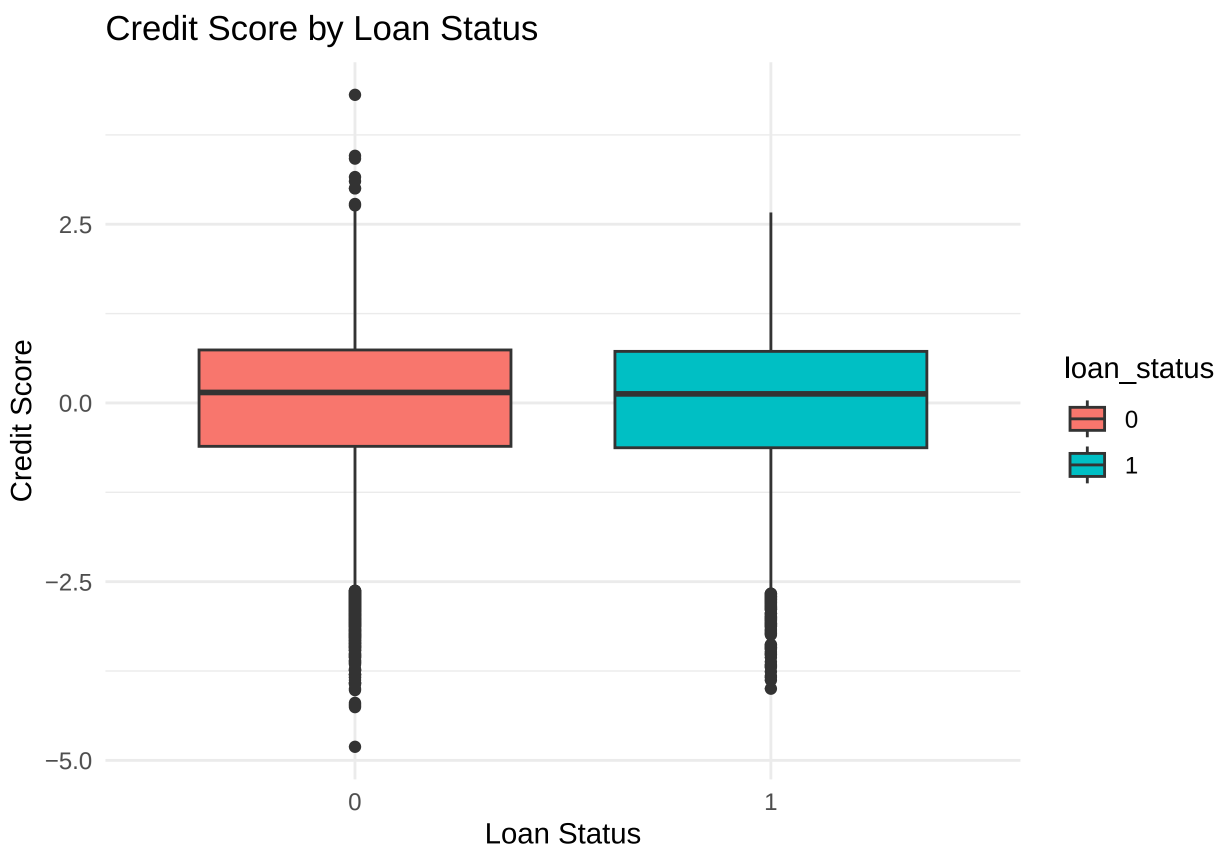
```r
# Visualize the correlation matrix
ggcorr(cor_matrix, label = FALSE, label_round = 2, low = "blue", high = "red")
```




```r
# Convert 'loan_status' to a factor for visualization
loan_data$loan_status <- as.factor(loan_data$loan_status)

# Visualize the distribution of loan amount by loan status
ggplot(loan_data, aes(x=loan_status, y=loan_amnt, fill=loan_status)) +
  geom_boxplot() +
  labs(title="Loan Amount by Loan Status", x="Loan Status", y="Loan Amount") +
  theme_minimal()
```


```r
# Visualize the distribution of credit score by loan status
ggplot(loan_data, aes(x=loan_status, y=credit_score, fill=loan_status)) +
  geom_boxplot() +
  labs(title="Credit Score by Loan Status", x="Loan Status", y="Credit Score") +
  theme_minimal()
```




```r
# Save the preprocessed data to a CSV file
write.csv(loan_data, "cleaned_loan_data.csv", row.names = FALSE)
```

```r
# Display summary and structure of the cleaned data
summary(loan_data)
```

```
##    person_age     person_gender   person_income     person_emp_exp
##  Min.   : 20.00   Min.   :0.000   Min.   :-1.5406   Min.   :  0.00
##  1st Qu.: 24.00   1st Qu.:0.000   1st Qu.:-0.6783   1st Qu.:  1.00
##  Median : 26.00   Median :1.000   Median :-0.2419   Median :  4.00
##  Mean   : 27.76   Mean   :0.552   Mean   : 0.0000   Mean   :  5.41
##  3rd Qu.: 30.00   3rd Qu.:1.000   3rd Qu.: 0.3902   3rd Qu.:  8.00
##  Max.   :100.00   Max.   :1.000   Max.   : 4.2537   Max.   :125.00
##    loan_amnt        loan_int_rate       loan_percent_income
##  Min.   :-1.4384   Min.   :-1.87545   Min.   :-1.6021
##  1st Qu.:-0.7258   1st Qu.:-0.81127   1st Qu.:-0.7995
##  Median :-0.2507   Median : 0.00114   Median :-0.2262
##  Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000
##  3rd Qu.: 0.4203   3rd Qu.: 0.66583   3rd Qu.: 0.5765
##  Max.   : 4.0249   Max.   : 3.01912   Max.   : 5.9656
##  cb_person_cred_hist_length  credit_score     previous_loan_defaults_on_file
##  Min.   : 2.000             Min.   :-4.8102   Min.   :0.000
##  1st Qu.: 3.000             1st Qu.:-0.6267   1st Qu.:0.000
##  Median : 4.000             Median : 0.1465   Median :1.000
##  Mean   : 5.867             Mean   : 0.0000   Mean   :0.508
##  3rd Qu.: 8.000             3rd Qu.: 0.7414   3rd Qu.:1.000
##  Max.   :30.000             Max.   : 4.3103   Max.   :1.000
```

```
##  loan_status person_educationAssociate person_educationBachelor
##  0:35000     Min.   :0.0000            Min.   :0.0000
##  1:10000     1st Qu.:0.0000            1st Qu.:0.0000
##              Median :0.0000            Median :0.0000
##              Mean   :0.2673            Mean   :0.2978
##              3rd Qu.:1.0000            3rd Qu.:1.0000
##              Max.   :1.0000            Max.   :1.0000
##  person_educationDoctorate person_educationHigh School person_educationMaster
##  Min.   :0.0000            Min.   :0.000              Min.   :0.0000
##  1st Qu.:0.0000            1st Qu.:0.000              1st Qu.:0.0000
##  Median :0.0000            Median :0.000              Median :0.0000
##  Mean   :0.0138            Mean   :0.266              Mean   :0.1551
##  3rd Qu.:0.0000            3rd Qu.:1.000              3rd Qu.:0.0000
##  Max.   :1.0000            Max.   :1.000              Max.   :1.0000
##  person_home_ownershipOTHER person_home_ownershipOWN person_home_ownershipRENT
##  Min.   :0.0000             Min.   :0.00000          Min.   :0.000
##  1st Qu.:0.0000             1st Qu.:0.00000          1st Qu.:0.000
##  Median :0.0000             Median :0.00000          Median :1.000
##  Mean   :0.0026             Mean   :0.06558          Mean   :0.521
##  3rd Qu.:0.0000             3rd Qu.:0.00000          3rd Qu.:1.000
##  Max.   :1.0000             Max.   :1.00000          Max.   :1.000
##  loan_intentEDUCATION loan_intentHOMEIMPROVEMENT loan_intentMEDICAL
##  Min.   :0.0000       Min.   :0.0000             Min.   :0.00
##  1st Qu.:0.0000       1st Qu.:0.0000             1st Qu.:0.00
##  Median :0.0000       Median :0.0000             Median :0.00
##  Mean   :0.2034       Mean   :0.1063             Mean   :0.19
##  3rd Qu.:0.0000       3rd Qu.:0.0000             3rd Qu.:0.00
##  Max.   :1.0000       Max.   :1.0000             Max.   :1.00
##  loan_intentPERSONAL loan_intentVENTURE
##  Min.   :0.0000      Min.   :0.0000
##  1st Qu.:0.0000      1st Qu.:0.0000
##  Median :0.0000      Median :0.0000
##  Mean   :0.1678      Mean   :0.1738
##  3rd Qu.:0.0000      3rd Qu.:0.0000
##  Max.   :1.0000      Max.   :1.0000
```

```r
str(loan_data)
```

```
## 'data.frame':    45000 obs. of  24 variables:
##  $ person_age                    : num  22 21 25 23 24 21 26 24 24 21 ...
##  $ person_gender                 : num  0 0 0 0 1 0 0 0 0 0 ...
##  $ person_income                 : num  -0.1341 -1.4464 -1.4429 0.0376 -0.262 ...
##  $ person_emp_exp                : int  0 0 3 0 1 0 1 5 3 0 ...
##  $ loan_amnt                     : num  4.025 -1.359 -0.647 4.025 4.025 ...
##  $ loan_int_rate                 : num  1.683 0.0448 0.6256 1.4178 1.0955 ...
##  $ loan_percent_income           : num  4.016 -0.685 3.443 3.443 4.475 ...
##  $ cb_person_cred_hist_length    : num  3 2 3 2 4 2 3 4 2 3 ...
##  $ credit_score                  : num  -1.4198 -2.5499 0.0474 0.8405 -0.9241 ...
##  $ previous_loan_defaults_on_file: num  0 1 0 0 0 0 0 0 0 0 ...
##  $ loan_status                   : Factor w/ 2 levels "0","1": 2 1 2 2 2 2 2 2 2 2 ...
##  $ person_educationAssociate     : num  0 0 0 0 0 0 0 0 1 0 ...
##  $ person_educationBachelor      : num  0 0 0 1 0 0 1 0 0 0 ...
##  $ person_educationDoctorate     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ person_educationHigh School   : num  0 1 1 0 0 1 0 1 0 1 ...
##  $ person_educationMaster        : num  1 0 0 0 1 0 0 0 0 0 ...
```

```
##  $ person_home_ownershipOTHER    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ person_home_ownershipOWN      : num  0 1 0 0 0 1 0 0 0 1 ...
##  $ person_home_ownershipRENT     : num  1 0 0 1 1 0 1 1 1 0 ...
##  $ loan_intentEDUCATION          : num  0 1 0 0 0 0 1 0 0 0 ...
##  $ loan_intentHOMEIMPROVEMENT    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ loan_intentMEDICAL            : num  0 0 1 1 1 0 0 1 0 0 ...
##  $ loan_intentPERSONAL           : num  1 0 0 0 0 0 0 0 1 0 ...
##  $ loan_intentVENTURE            : num  0 0 0 0 0 1 0 0 0 1 ...
```

Load libraries necessary for models

```r
if(!require(MASS)) install.packages("MASS", dependencies=TRUE)
```

```
## Loading required package: MASS

##
## Attaching package: 'MASS'

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

```r
if(!require(e1071)) install.packages("e1071", dependencies=TRUE)
```

```
## Loading required package: e1071

## Warning: package 'e1071' was built under R version 4.4.2
```

```r
library(MASS)
library(e1071)  # Needed for confusion matrix calculations
```

Define control parameters for different resampling techniques

```r
set.seed(123)
control_loocv = trainControl(method = "LOOCV")
control_cv10 = trainControl(method = "cv", number = 10)
control_cv5 = trainControl(method = "cv", number = 5)
```

Logistic Regression

```r
# Perform forward selection to use for the logistic regression models
null_model = glm(loan_status ~ 1, data = loan_data, family = "binomial")
full_model = glm(loan_status ~ ., data = loan_data, family = "binomial")
forward_model = step(null_model, scope = list(lower = null_model, upper = full_model), direction = "for
```

```
## Start:  AIC=47675.56
## loan_status ~ 1
##
##                                  Df Deviance   AIC
## + previous_loan_defaults_on_file  1    30488 30492
## + loan_percent_income             1    41352 41356
## + loan_int_rate                   1    42561 42565
## + person_home_ownershipRENT       1    44610 44614
## + person_income                   1    44741 44745
## + loan_amnt                       1    47175 47179
## + person_home_ownershipOWN        1    47184 47188
## + loan_intentVENTURE              1    47313 47317
## + loan_intentEDUCATION            1    47480 47484
## + loan_intentMEDICAL              1    47490 47494
## + loan_intentHOMEIMPROVEMENT      1    47624 47628
```

```
## + loan_intentPERSONAL            1    47650 47654
## + person_age                     1    47653 47657
## + person_emp_exp                 1    47654 47658
## + cb_person_cred_hist_length     1    47664 47668
## + person_home_ownershipOTHER     1    47666 47670
## + credit_score                   1    47671 47675
## <none>                                47674 47676
## + person_educationMaster         1    47673 47677
## + person_educationBachelor       1    47673 47677
## + person_educationAssociate      1    47673 47677
## + person_educationDoctorate      1    47673 47677
## + `person_educationHigh School`  1    47673 47677
## + person_gender                  1    47674 47678
##
## Step:  AIC=30491.79
## loan_status ~ previous_loan_defaults_on_file
##
##                                Df Deviance   AIC
## + loan_percent_income           1    26333 26339
## + loan_int_rate                 1    27139 27145
## + person_home_ownershipRENT     1    28288 28294
## + person_income                 1    28451 28457
## + credit_score                  1    29624 29630
## + person_home_ownershipOWN      1    30126 30132
## + loan_amnt                     1    30153 30159
## + loan_intentVENTURE            1    30248 30254
## + loan_intentMEDICAL            1    30358 30364
## + loan_intentEDUCATION          1    30361 30367
## + person_emp_exp                1    30410 30416
## + person_age                    1    30414 30420
## + cb_person_cred_hist_length    1    30443 30449
## + loan_intentHOMEIMPROVEMENT    1    30459 30465
## + loan_intentPERSONAL           1    30461 30467
## + `person_educationHigh School` 1    30468 30474
## + person_educationMaster        1    30472 30478
## + person_educationDoctorate     1    30484 30490
## + person_home_ownershipOTHER    1    30485 30491
## <none>                               30488 30492
## + person_educationBachelor      1    30486 30492
## + person_educationAssociate     1    30487 30493
## + person_gender                 1    30488 30494
##
## Step:  AIC=26338.79
## loan_status ~ previous_loan_defaults_on_file + loan_percent_income
##
##                                Df Deviance   AIC
## + loan_int_rate                 1    23109 23117
## + person_home_ownershipRENT     1    24979 24987
## + credit_score                  1    25613 25621
## + loan_amnt                     1    25698 25706
## + person_income                 1    25742 25750
## + person_home_ownershipOWN      1    25835 25843
## + loan_intentVENTURE            1    26027 26035
## + loan_intentEDUCATION          1    26179 26187
```

```
## + loan_intentMEDICAL            1    26226 26234
## + loan_intentHOMEIMPROVEMENT     1    26262 26270
## + person_emp_exp                 1    26299 26307
## + loan_intentPERSONAL            1    26305 26313
## + person_age                     1    26305 26313
## + cb_person_cred_hist_length     1    26314 26322
## + `person_educationHigh School`  1    26316 26324
## + person_educationMaster         1    26324 26332
## + person_educationDoctorate      1    26330 26338
## + person_educationBachelor       1    26330 26338
## <none>                                26333 26339
## + person_home_ownershipOTHER     1    26332 26340
## + person_educationAssociate      1    26332 26340
## + person_gender                  1    26333 26341
##
## Step:  AIC=23117.41
## loan_status ~ previous_loan_defaults_on_file + loan_percent_income +
##     loan_int_rate
##
##                                 Df Deviance   AIC
## + person_home_ownershipRENT      1    22013 22023
## + loan_amnt                      1    22015 22025
## + person_income                  1    22294 22304
## + credit_score                   1    22440 22450
## + person_home_ownershipOWN       1    22663 22673
## + loan_intentVENTURE             1    22814 22824
## + loan_intentEDUCATION           1    22996 23006
## + loan_intentMEDICAL             1    23018 23028
## + loan_intentHOMEIMPROVEMENT     1    23049 23059
## + person_emp_exp                 1    23062 23072
## + person_age                     1    23071 23081
## + cb_person_cred_hist_length     1    23078 23088
## + loan_intentPERSONAL            1    23085 23095
## + `person_educationHigh School`  1    23092 23102
## + person_educationMaster         1    23101 23111
## + person_educationDoctorate      1    23104 23114
## + person_educationBachelor       1    23106 23116
## <none>                                23109 23117
## + person_educationAssociate      1    23109 23119
## + person_home_ownershipOTHER     1    23109 23119
## + person_gender                  1    23109 23119
##
## Step:  AIC=22023.34
## loan_status ~ previous_loan_defaults_on_file + loan_percent_income +
##     loan_int_rate + person_home_ownershipRENT
##
##                                 Df Deviance   AIC
## + credit_score                   1    21373 21385
## + loan_amnt                      1    21397 21409
## + person_income                  1    21639 21651
## + loan_intentVENTURE             1    21727 21739
## + person_home_ownershipOWN       1    21877 21889
## + loan_intentEDUCATION           1    21896 21908
## + loan_intentHOMEIMPROVEMENT     1    21925 21937
```

```
## + loan_intentMEDICAL            1    21942 21954
## + person_emp_exp                 1    21979 21991
## + loan_intentPERSONAL            1    21987 21999
## + person_age                     1    21988 22000
## + cb_person_cred_hist_length     1    21989 22001
## + `person_educationHigh School`  1    21995 22007
## + person_educationMaster         1    22006 22018
## + person_home_ownershipOTHER     1    22007 22019
## + person_educationBachelor       1    22008 22020
## + person_educationDoctorate      1    22010 22022
## <none>                                22013 22023
## + person_educationAssociate      1    22012 22024
## + person_gender                  1    22013 22025
##
## Step:  AIC=21385.04
## loan_status ~ previous_loan_defaults_on_file + loan_percent_income +
##     loan_int_rate + person_home_ownershipRENT + credit_score
##
##                                 Df Deviance   AIC
## + loan_amnt                      1    20805 20819
## + person_income                  1    21036 21050
## + loan_intentVENTURE             1    21109 21123
## + person_home_ownershipOWN       1    21236 21250
## + loan_intentEDUCATION           1    21258 21272
## + loan_intentHOMEIMPROVEMENT     1    21287 21301
## + loan_intentMEDICAL             1    21302 21316
## + loan_intentPERSONAL            1    21346 21360
## + person_home_ownershipOTHER     1    21369 21383
## <none>                                21373 21385
## + person_emp_exp                 1    21373 21387
## + cb_person_cred_hist_length     1    21373 21387
## + person_educationDoctorate      1    21373 21387
## + person_educationMaster         1    21373 21387
## + person_gender                  1    21373 21387
## + person_educationBachelor       1    21373 21387
## + `person_educationHigh School`  1    21373 21387
## + person_educationAssociate      1    21373 21387
## + person_age                     1    21373 21387
##
## Step:  AIC=20819.21
## loan_status ~ previous_loan_defaults_on_file + loan_percent_income +
##     loan_int_rate + person_home_ownershipRENT + credit_score +
##     loan_amnt
##
##                                 Df Deviance   AIC
## + person_home_ownershipOWN       1    20516 20532
## + loan_intentVENTURE             1    20531 20547
## + loan_intentEDUCATION           1    20693 20709
## + loan_intentHOMEIMPROVEMENT     1    20699 20715
## + loan_intentMEDICAL             1    20753 20769
## + loan_intentPERSONAL            1    20778 20794
## + person_income                  1    20798 20814
## + person_age                     1    20802 20818
## + person_home_ownershipOTHER     1    20803 20819
```

```
## <none>                              20805 20819
## + person_emp_exp                 1    20804 20820
## + cb_person_cred_hist_length     1    20804 20820
## + person_gender                  1    20805 20821
## + person_educationMaster         1    20805 20821
## + `person_educationHigh School`  1    20805 20821
## + person_educationBachelor       1    20805 20821
## + person_educationDoctorate      1    20805 20821
## + person_educationAssociate      1    20805 20821
##
## Step:  AIC=20531.49
## loan_status ~ previous_loan_defaults_on_file + loan_percent_income +
##     loan_int_rate + person_home_ownershipRENT + credit_score +
##     loan_amnt + person_home_ownershipOWN
##
##                                 Df Deviance   AIC
## + loan_intentVENTURE             1    20269 20287
## + loan_intentHOMEIMPROVEMENT     1    20404 20422
## + loan_intentEDUCATION           1    20406 20424
## + loan_intentMEDICAL             1    20469 20487
## + loan_intentPERSONAL            1    20490 20508
## + person_income                  1    20509 20527
## + person_age                     1    20511 20529
## <none>                              20516 20532
## + person_emp_exp                 1    20514 20532
## + cb_person_cred_hist_length     1    20514 20532
## + person_home_ownershipOTHER     1    20515 20533
## + person_gender                  1    20515 20533
## + person_educationMaster         1    20515 20533
## + person_educationBachelor       1    20515 20533
## + `person_educationHigh School`  1    20515 20533
## + person_educationDoctorate      1    20515 20533
## + person_educationAssociate      1    20516 20534
##
## Step:  AIC=20287.09
## loan_status ~ previous_loan_defaults_on_file + loan_percent_income +
##     loan_int_rate + person_home_ownershipRENT + credit_score +
##     loan_amnt + person_home_ownershipOWN + loan_intentVENTURE
##
##                                 Df Deviance   AIC
## + loan_intentEDUCATION           1    20076 20096
## + loan_intentHOMEIMPROVEMENT     1    20199 20219
## + loan_intentPERSONAL            1    20203 20223
## + loan_intentMEDICAL             1    20256 20276
## + person_income                  1    20260 20280
## + person_age                     1    20266 20286
## <none>                              20269 20287
## + person_gender                  1    20268 20288
## + person_emp_exp                 1    20268 20288
## + person_home_ownershipOTHER     1    20268 20288
## + cb_person_cred_hist_length     1    20268 20288
## + person_educationMaster         1    20269 20289
## + person_educationBachelor       1    20269 20289
## + person_educationDoctorate      1    20269 20289
```

```
## + `person_educationHigh School`  1    20269 20289
## + person_educationAssociate      1    20269 20289
##
## Step:  AIC=20096.42
## loan_status ~ previous_loan_defaults_on_file + loan_percent_income +
##     loan_int_rate + person_home_ownershipRENT + credit_score +
##     loan_amnt + person_home_ownershipOWN + loan_intentVENTURE +
##     loan_intentEDUCATION
##
##                                 Df Deviance   AIC
## + loan_intentPERSONAL            1    19928 19950
## + loan_intentHOMEIMPROVEMENT     1    20045 20067
## + person_income                  1    20067 20089
## <none>                                20076 20096
## + person_gender                  1    20075 20097
## + person_home_ownershipOTHER     1    20076 20098
## + person_age                     1    20076 20098
## + loan_intentMEDICAL             1    20076 20098
## + person_educationMaster         1    20076 20098
## + person_educationBachelor       1    20076 20098
## + person_educationAssociate      1    20076 20098
## + cb_person_cred_hist_length     1    20076 20098
## + `person_educationHigh School`  1    20076 20098
## + person_emp_exp                 1    20076 20098
## + person_educationDoctorate      1    20076 20098
##
## Step:  AIC=19949.65
## loan_status ~ previous_loan_defaults_on_file + loan_percent_income +
##     loan_int_rate + person_home_ownershipRENT + credit_score +
##     loan_amnt + person_home_ownershipOWN + loan_intentVENTURE +
##     loan_intentEDUCATION + loan_intentPERSONAL
##
##                                 Df Deviance   AIC
## + loan_intentMEDICAL             1    19894 19918
## + person_income                  1    19918 19942
## + loan_intentHOMEIMPROVEMENT     1    19921 19945
## <none>                                19928 19950
## + person_gender                  1    19926 19950
## + person_home_ownershipOTHER     1    19927 19951
## + person_age                     1    19927 19951
## + person_educationMaster         1    19927 19951
## + person_educationBachelor       1    19928 19952
## + person_educationAssociate      1    19928 19952
## + cb_person_cred_hist_length     1    19928 19952
## + person_educationDoctorate      1    19928 19952
## + person_emp_exp                 1    19928 19952
## + `person_educationHigh School`  1    19928 19952
##
## Step:  AIC=19918.3
## loan_status ~ previous_loan_defaults_on_file + loan_percent_income +
##     loan_int_rate + person_home_ownershipRENT + credit_score +
##     loan_amnt + person_home_ownershipOWN + loan_intentVENTURE +
##     loan_intentEDUCATION + loan_intentPERSONAL + loan_intentMEDICAL
##
```

```
##                                  Df Deviance   AIC
## + person_income                   1    19886 19912
## <none>                                 19894 19918
## + person_gender                   1    19893 19919
## + person_home_ownershipOTHER      1    19893 19919
## + person_age                      1    19894 19920
## + person_educationMaster          1    19894 19920
## + person_educationBachelor        1    19894 19920
## + person_educationAssociate       1    19894 19920
## + cb_person_cred_hist_length      1    19894 19920
## + `person_educationHigh School`   1    19894 19920
## + loan_intentHOMEIMPROVEMENT      1    19894 19920
## + person_educationDoctorate       1    19894 19920
## + person_emp_exp                  1    19894 19920
##
## Step:  AIC=19912.31
## loan_status ~ previous_loan_defaults_on_file + loan_percent_income +
##     loan_int_rate + person_home_ownershipRENT + credit_score +
##     loan_amnt + person_home_ownershipOWN + loan_intentVENTURE +
##     loan_intentEDUCATION + loan_intentPERSONAL + loan_intentMEDICAL +
##     person_income
##
##                                  Df Deviance   AIC
## <none>                                 19886 19912
## + person_gender                   1    19885 19913
## + person_home_ownershipOTHER      1    19885 19913
## + person_age                      1    19886 19914
## + person_educationMaster          1    19886 19914
## + person_educationBachelor        1    19886 19914
## + person_educationAssociate       1    19886 19914
## + `person_educationHigh School`   1    19886 19914
## + loan_intentHOMEIMPROVEMENT      1    19886 19914
## + person_emp_exp                  1    19886 19914
## + person_educationDoctorate       1    19886 19914
## + cb_person_cred_hist_length      1    19886 19914
```

```r
# Display the summary of the selected model from forward selection
summary(forward_model)
```

```
##
## Call:
## glm(formula = loan_status ~ previous_loan_defaults_on_file +
##     loan_percent_income + loan_int_rate + person_home_ownershipRENT +
##     credit_score + loan_amnt + person_home_ownershipOWN + loan_intentVENTURE +
##     loan_intentEDUCATION + loan_intentPERSONAL + loan_intentMEDICAL +
##     person_income, family = "binomial", data = loan_data)
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -0.44948    0.04051 -11.095  < 2e-16 ***
## previous_loan_defaults_on_file -20.37853  102.79084  -0.198  0.84285
## loan_percent_income               1.43871    0.03979  36.161  < 2e-16 ***
## loan_int_rate                     0.99572    0.01958  50.859  < 2e-16 ***
## person_home_ownershipRENT         0.72750    0.04021  18.093  < 2e-16 ***
## credit_score                     -0.44892    0.01970 -22.786  < 2e-16 ***
```

```
## loan_amnt                        -0.70284    0.03973 -17.692  < 2e-16 ***
## person_home_ownershipOWN         -1.46824    0.10192 -14.406  < 2e-16 ***
## loan_intentVENTURE               -1.20613    0.05819 -20.729  < 2e-16 ***
## loan_intentEDUCATION             -0.90601    0.05252 -17.250  < 2e-16 ***
## loan_intentPERSONAL              -0.72165    0.05416 -13.325  < 2e-16 ***
## loan_intentMEDICAL               -0.28378    0.05027  -5.646 1.65e-08 ***
## person_income                     0.10714    0.03754   2.854  0.00432 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 47674  on 44999  degrees of freedom
## Residual deviance: 19886  on 44987  degrees of freedom
## AIC: 19912
##
## Number of Fisher Scoring iterations: 19
```

```r
# Extract the formula of the selected model for further evaluation
forward_formula = formula(forward_model)

# Fit the logistic model using the selected predictors from forward selection without resampling
logistic_model = glm(forward_formula, data = loan_data, family = "binomial")
print(logistic_model)
```

```
##
## Call:  glm(formula = forward_formula, family = "binomial", data = loan_data)
##
## Coefficients:
##                  (Intercept)  previous_loan_defaults_on_file
##                      -0.4495                        -20.3785
##          loan_percent_income                   loan_int_rate
##                       1.4387                          0.9957
##     person_home_ownershipRENT                   credit_score
##                       0.7275                         -0.4489
##                    loan_amnt        person_home_ownershipOWN
##                      -0.7028                         -1.4682
##           loan_intentVENTURE            loan_intentEDUCATION
##                      -1.2061                         -0.9060
##          loan_intentPERSONAL              loan_intentMEDICAL
##                      -0.7217                         -0.2838
##                person_income
##                       0.1071
##
## Degrees of Freedom: 44999 Total (i.e. Null);  44987 Residual
## Null Deviance:       47670
## Residual Deviance: 19890     AIC: 19910
```

```r
summary(logistic_model)
```

```
##
## Call:
## glm(formula = forward_formula, family = "binomial", data = loan_data)
##
## Coefficients:
```

```
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -0.44948    0.04051 -11.095  < 2e-16 ***
## previous_loan_defaults_on_file -20.37853  102.79084  -0.198  0.84285
## loan_percent_income              1.43871    0.03979  36.161  < 2e-16 ***
## loan_int_rate                    0.99572    0.01958  50.859  < 2e-16 ***
## person_home_ownershipRENT        0.72750    0.04021  18.093  < 2e-16 ***
## credit_score                    -0.44892    0.01970 -22.786  < 2e-16 ***
## loan_amnt                       -0.70284    0.03973 -17.692  < 2e-16 ***
## person_home_ownershipOWN        -1.46824    0.10192 -14.406  < 2e-16 ***
## loan_intentVENTURE              -1.20613    0.05819 -20.729  < 2e-16 ***
## loan_intentEDUCATION            -0.90601    0.05252 -17.250  < 2e-16 ***
## loan_intentPERSONAL             -0.72165    0.05416 -13.325  < 2e-16 ***
## loan_intentMEDICAL              -0.28378    0.05027  -5.646 1.65e-08 ***
## person_income                    0.10714    0.03754   2.854  0.00432 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 47674  on 44999  degrees of freedom
## Residual deviance: 19886  on 44987  degrees of freedom
## AIC: 19912
##
## Number of Fisher Scoring iterations: 19
```

```r
# Calculate the confusion matrix for the logistic model without resampling
logistic_preds = predict(logistic_model, loan_data, type = "response")
logistic_class = ifelse(logistic_preds > 0.5, 1, 0)
logistic_cm = confusionMatrix(as.factor(logistic_class), as.factor(loan_data$loan_status))

# Extract accuracy
accuracy = logistic_cm$overall['Accuracy']
print(accuracy)
```

```
##  Accuracy
## 0.8966667
```

```r
# Define resampling control parameters
set.seed(123)
control_cv5 = trainControl(method = "cv", number = 5)
control_cv10 = trainControl(method = "cv", number = 10)

# Apply resampling methods with the selected model from forward selection
# 5-fold Cross-Validation
logistic_cv5 = train(forward_formula, data = loan_data, method = "glm", family = "binomial", trControl =
print(logistic_cv5)
```

```
## Generalized Linear Model
##
## 45000 samples
##    12 predictor
##     2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 36000, 36000, 36000, 36000, 36000
```

```
## Resampling results:
##
##   Accuracy   Kappa
##   0.8961778  0.6956566
```

```
# 10-fold Cross-Validation
logistic_cv10 = train(forward_formula, data = loan_data, method = "glm", family = "binomial", trControl
print(logistic_cv10)
```

```
## Generalized Linear Model
##
## 45000 samples
##    12 predictor
##     2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 40500, 40500, 40500, 40500, 40500, 40500, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.8964444  0.6964036
```

1- **Null Hypothesis Testing**

Forward selection was used to iteratively add predictors that improved the model's fit, based on minimizing the Akaike Information Criterion (AIC). This process identified predictors that contribute meaningfully to explaining the variance in loan_status.

The results show the following significance levels:

- Significant Predictors (p-value < 0.05): previous_loan_defaults_on_file, loan_percent_income, loan_int_rate, person_home_ownershipRENT, credit_score, loan_amnt, person_home_ownershipOWN, loan_intentVENTURE, loan_intentEDUCATION, loan_intentPERSONAL, loan_intentMEDICAL, and person_income.
- Non-Significant Predictors (p-value > 0.05): Although included in the model through forward selection, previous_loan_defaults_on_file has a very high standard error and non-significant p-value, suggesting it may not strongly influence loan_status.

Forward selection allowed us to isolate these significant predictors by iteratively adding only the most relevant variables, resulting in a more efficient and interpretable model.

2- **Results**

Key results from the logistic regression model with forward selection include:

- Model Fit Metrics:
  - Null Deviance: 47674
  - Residual Deviance: 19886
  - AIC: 19912
  - Accuracy (without resampling): 0.8967
- Significant Predictors: Forward selection identified important predictors, including loan_amnt, loan_int_rate, loan_percent_income, credit_score, and specific categories of loan_intent and person_home_ownership. These predictors showed statistical significance and add substantial predictive power for loan_status.

- Cross-Validation Accuracy:
  - 5-fold Cross-Validation: Accuracy = 0.8962
  - 10-fold Cross-Validation: Accuracy = 0.8964

3- **Comparison of Results**

The model's accuracy without resampling is 0.8967, while the accuracy under 5-fold and 10-fold cross-validation is slightly lower at 0.8962 and 0.8964, respectively. This small difference in accuracy demonstrates that the model is stable and generalizes well to new data. The consistent cross-validation results further validate the predictors selected through forward selection, indicating that these features contribute to robust performance across different subsets of the data.

4- **Interpretations**

- Prediction: The forward selection process prioritized financial features, such as loan_amnt, loan_int_rate, loan_percent_income, and credit_score, which emerged as strong predictors. This suggests that these financial metrics are critical for predicting loan approval likelihood.
- Demographic Predictors: The forward selection process excluded demographic variables like person_gender and most person_education levels due to their low correlation to the response. This highlights that demographic characteristics may not be as influential as financial features in this context.
- Model Stability and Interpretability: Forward selection enabled us to build a good model with a select group of highly predictive variables. The similar accuracy across different cross-validation folds (5- and 10-fold) indicates that the model generalizes well and is less likely to overfit. By selecting only the most influential predictors, forward selection produced a model that is both effective.

```r
# Step 0: Check for high correlation among features
# Compute the correlation matrix for numerical features
cor_matrix = cor(loan_data[sapply(loan_data, is.numeric)])

# Set a correlation threshold (e.g., 0.9)
high_cor_threshold = 0.9

# Identify pairs of highly correlated features
high_cor_pairs = which(abs(cor_matrix) > high_cor_threshold, arr.ind = TRUE)
high_cor_pairs = high_cor_pairs[high_cor_pairs[,1] != high_cor_pairs[,2], ]  # Remove self-correlations

# Display highly correlated feature pairs
if (nrow(high_cor_pairs) > 0) {
  print("Highly correlated feature pairs:")
  for (i in seq_len(nrow(high_cor_pairs))) {
    row = high_cor_pairs[i, ]
    feature1 = rownames(cor_matrix)[row[1]]
    feature2 = colnames(cor_matrix)[row[2]]
    correlation_value = cor_matrix[row[1], row[2]]
    cat(feature1, "and", feature2, "with correlation:", correlation_value, "\n")
  }
} else {
  print("No highly correlated feature pairs found.")
}
```

```
## [1] "Highly correlated feature pairs:"
## person_emp_exp and person_age with correlation: 0.9516695
## person_age and person_emp_exp with correlation: 0.9516695
```

The high correlation (0.95) between person_emp_exp and person_age indicates redundancy. In QDA, such correlation can lead to instability in estimating class-specific covariance matrices, causing errors. Removing one of these features or using PCA can resolve this, ensuring QDA runs smoothly.

```r
# Identify near-zero variance predictors and highly correlated predictors
nzv = nearZeroVar(loan_data)
loan_data_filtered = loan_data[, -nzv]

# Check for highly correlated predictors
cor_matrix = cor(loan_data_filtered[sapply(loan_data_filtered, is.numeric)])
high_cor = findCorrelation(cor_matrix, cutoff = 0.9)
loan_data_filtered = loan_data_filtered[, -high_cor]

# Apply PCA to create uncorrelated components for QDA and LDA
# Exclude the target variable for PCA transformation
predictor_data = loan_data_filtered[, colnames(loan_data_filtered) != "loan_status"]
pca_model = preProcess(predictor_data, method = "pca", pcaComp = 10) # Adjust pcaComp as needed
pca_data = predict(pca_model, predictor_data)

# Combine PCA components with the target variable
pca_data$loan_status = loan_data_filtered$loan_status

# Step 4: Perform backward selection with logistic regression on PCA components
full_model = glm(loan_status ~ ., data = pca_data, family = "binomial")
backward_model = step(full_model, direction = "backward")
```

```
## Start:  AIC=27675.96
## loan_status ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 + PC8 +
##     PC9 + PC10
##
##        Df Deviance   AIC
## <none>       27654 27676
## - PC5   1    27669 27689
## - PC9   1    27672 27692
## - PC8   1    27741 27761
## - PC7   1    27752 27772
## - PC6   1    27755 27775
## - PC4   1    27926 27946
## - PC1   1    28270 28290
## - PC10  1    28739 28759
## - PC3   1    29273 29293
## - PC2   1    46039 46059
```

```r
# Display the summary of the selected model from backward selection
summary(backward_model)
```

```
##
## Call:
## glm(formula = loan_status ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 +
##     PC7 + PC8 + PC9 + PC10, family = "binomial", data = pca_data)
##
## Coefficients:
##             Estimate Std. Error  z value Pr(>|z|)
## (Intercept) -2.32496    0.02298 -101.173  < 2e-16 ***
## PC1         -0.27365    0.01140  -24.013  < 2e-16 ***
## PC2          1.65986    0.01740   95.403  < 2e-16 ***
## PC3          0.48526    0.01275   38.074  < 2e-16 ***
```

```
## PC4         -0.20905    0.01277  -16.367  < 2e-16 ***
## PC5          0.05181    0.01320    3.926 8.63e-05 ***
## PC6          0.15513    0.01563    9.924  < 2e-16 ***
## PC7          0.14740    0.01495    9.860  < 2e-16 ***
## PC8         -0.13594    0.01461   -9.307  < 2e-16 ***
## PC9         -0.05824    0.01364   -4.271 1.94e-05 ***
## PC10         0.51264    0.01612   31.811  < 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: 47674  on 44999  degrees of freedom
## Residual deviance: 27654  on 44989  degrees of freedom
## AIC: 27676
##
## Number of Fisher Scoring iterations: 6
```

```r
# Extract the formula of the selected model for further evaluation
backward_formula = formula(backward_model)

# Define resampling control parameters
set.seed(123)
control_cv5 = trainControl(method = "cv", number = 5)
control_cv10 = trainControl(method = "cv", number = 10)

# Plain QDA on selected PCA components (without resampling)
qda_model_plain = qda(backward_formula, data = pca_data)
qda_preds_plain = predict(qda_model_plain)$class
qda_cm_plain = confusionMatrix(qda_preds_plain, as.factor(pca_data$loan_status))
print("Plain QDA Confusion Matrix:")
```

```
## [1] "Plain QDA Confusion Matrix:"
```

```r
print(qda_cm_plain)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 32747  3680
##          1  2253  6320
##
##                Accuracy : 0.8682
##                  95% CI : (0.865, 0.8713)
##     No Information Rate : 0.7778
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5981
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9356
##             Specificity : 0.6320
##          Pos Pred Value : 0.8990
```

```
##          Neg Pred Value : 0.7372
##              Prevalence : 0.7778
##          Detection Rate : 0.7277
##    Detection Prevalence : 0.8095
##       Balanced Accuracy : 0.7838
##
##        'Positive' Class : 0
##
```

```r
# Plain LDA on selected PCA components (without resampling)
lda_model_plain = lda(backward_formula, data = pca_data)
lda_preds_plain = predict(lda_model_plain)$class
lda_cm_plain = confusionMatrix(lda_preds_plain, as.factor(pca_data$loan_status))
print("Plain LDA Confusion Matrix:")
```

```
## [1] "Plain LDA Confusion Matrix:"
```

```r
print(lda_cm_plain)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 33188  3876
##          1  1812  6124
##
##                Accuracy : 0.8736
##                  95% CI : (0.8705, 0.8767)
##     No Information Rate : 0.7778
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6052
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9482
##             Specificity : 0.6124
##          Pos Pred Value : 0.8954
##          Neg Pred Value : 0.7717
##              Prevalence : 0.7778
##          Detection Rate : 0.7375
##    Detection Prevalence : 0.8236
##       Balanced Accuracy : 0.7803
##
##        'Positive' Class : 0
##
```

```r
# QDA with 5-fold Cross-Validation on selected PCA components
qda_cv5 = train(backward_formula, data = pca_data, method = "qda", trControl = control_cv5)
print("5-Fold Cross-Validated QDA Results:")
```

```
## [1] "5-Fold Cross-Validated QDA Results:"
```

```r
print(qda_cv5)
```

```
## Quadratic Discriminant Analysis
##
```

```
## 45000 samples
##    10 predictor
##     2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 36000, 36000, 36000, 36000, 36000
## Resampling results:
##
##   Accuracy   Kappa
##   0.8670222  0.5946046
```

```r
# QDA with 10-fold Cross-Validation on selected PCA components
qda_cv10 = train(backward_formula, data = pca_data, method = "qda", trControl = control_cv10)
print("10-Fold Cross-Validated QDA Results:")
```

```
## [1] "10-Fold Cross-Validated QDA Results:"
```

```r
print(qda_cv10)
```

```
## Quadratic Discriminant Analysis
##
## 45000 samples
##    10 predictor
##     2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 40500, 40500, 40500, 40500, 40500, 40500, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.8676889  0.5966949
```

```r
# LDA with 5-fold Cross-Validation on selected PCA components
lda_cv5 = train(backward_formula, data = pca_data, method = "lda", trControl = control_cv5)
print("5-Fold Cross-Validated LDA Results:")
```

```
## [1] "5-Fold Cross-Validated LDA Results:"
```

```r
print(lda_cv5)
```

```
## Linear Discriminant Analysis
##
## 45000 samples
##    10 predictor
##     2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 36000, 36000, 36000, 36000, 36000
## Resampling results:
##
##   Accuracy   Kappa
##   0.8736222  0.6053371
```

```r
# LDA with 10-fold Cross-Validation on selected PCA components
lda_cv10 = train(backward_formula, data = pca_data, method = "lda", trControl = control_cv10)
print("10-Fold Cross-Validated LDA Results:")
```

```
## [1] "10-Fold Cross-Validated LDA Results:"
```

```r
print(lda_cv10)
```

```
## Linear Discriminant Analysis
##
## 45000 samples
##    10 predictor
##     2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 40500, 40500, 40500, 40500, 40500, 40500, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.8733778  0.6044724
```

1. Results

- Plain QDA:
  - Accuracy = 86.82%
  - Sensitivity = 93.56% (indicating it's effective at correctly identifying positive cases, e.g., likely loan approvals)
  - Specificity = 63.20% (indicating a moderate ability to identify negative cases)
- Plain LDA:
  - Accuracy = 87.36% (slightly higher than QDA)
  - Sensitivity = 94.82% (high detection of positive cases)
  - Specificity = 61.24% (slightly lower than QDA)
- 5-Fold Cross-Validation:
  - QDA: Accuracy = 86.70%
  - LDA: Accuracy = 87.36%
- 10-Fold Cross-Validation:
  - QDA: Accuracy = 86.77%
  - LDA: Accuracy = 87.34%

2. Comparison of Results The LDA model consistently outperformed the QDA model across all metrics, achieving slightly higher accuracy scores. This consistent difference suggests that LDA may offer more reliable predictions with this dataset. Both models displayed very similar accuracy between plain (non-resampled) and cross-validated results, with LDA having higher accuracies. These close results indicate good model stability and generalization to unseen data.

3. Interpretation The higher sensitivity of the LDA model reflects its stronger performance in identifying positive cases (e.g., likely loan approvals), while QDA showed slightly better specificity, meaning it may be better at identifying negative cases. Both models have moderate Kappa scores, indicating a reasonable level of agreement beyond random chance.

The use of PCA for dimensionality reduction, followed by backward selection, proved effective in reducing multicollinearity, particularly important for QDA's stability. The slight performance advantage of LDA suggests that the linear boundaries it assumes may better fit this dataset compared to QDA's quadratic boundaries, making LDA a more robust choice for this loan approval prediction task.

```r
# Start with the full model containing all PCA components
full_model = glm(loan_status ~ ., data = pca_data, family = "binomial")

# Step 2: Perform mixed selection (both directions)
mixed_model = step(full_model, direction = "both")
```

```
## Start:  AIC=27675.96
## loan_status ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 + PC8 +
##     PC9 + PC10
##
##        Df Deviance   AIC
## <none>       27654 27676
## - PC5   1    27669 27689
## - PC9   1    27672 27692
## - PC8   1    27741 27761
## - PC7   1    27752 27772
## - PC6   1    27755 27775
## - PC4   1    27926 27946
## - PC1   1    28270 28290
## - PC10  1    28739 28759
## - PC3   1    29273 29293
## - PC2   1    46039 46059
```

```r
# Step 3: Display the summary of the selected model
summary(mixed_model)
```

```
##
## Call:
## glm(formula = loan_status ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 +
##     PC7 + PC8 + PC9 + PC10, family = "binomial", data = pca_data)
##
## Coefficients:
##             Estimate Std. Error  z value Pr(>|z|)
## (Intercept) -2.32496    0.02298 -101.173  < 2e-16 ***
## PC1         -0.27365    0.01140  -24.013  < 2e-16 ***
## PC2          1.65986    0.01740   95.403  < 2e-16 ***
## PC3          0.48526    0.01275   38.074  < 2e-16 ***
## PC4         -0.20905    0.01277  -16.367  < 2e-16 ***
## PC5          0.05181    0.01320    3.926 8.63e-05 ***
## PC6          0.15513    0.01563    9.924  < 2e-16 ***
## PC7          0.14740    0.01495    9.860  < 2e-16 ***
## PC8         -0.13594    0.01461   -9.307  < 2e-16 ***
## PC9         -0.05824    0.01364   -4.271 1.94e-05 ***
## PC10         0.51264    0.01612   31.811  < 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: 47674  on 44999  degrees of freedom
## Residual deviance: 27654  on 44989  degrees of freedom
## AIC: 27676
##
## Number of Fisher Scoring iterations: 6
```

```r
# Extract the formula of the selected model for further evaluation
mixed_formula = formula(mixed_model)

# Step 4: Fit the final model using the selected predictors from mixed selection
# Plain LDA
lda_model_mixed = lda(mixed_formula, data = pca_data)
lda_preds_mixed = predict(lda_model_mixed)$class
lda_cm_mixed = confusionMatrix(lda_preds_mixed, as.factor(pca_data$loan_status))
print("Plain LDA Confusion Matrix (Mixed Selection):")
```

```
## [1] "Plain LDA Confusion Matrix (Mixed Selection):"
```

```r
print(lda_cm_mixed)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 33188  3876
##          1  1812  6124
##
##                Accuracy : 0.8736
##                  95% CI : (0.8705, 0.8767)
##     No Information Rate : 0.7778
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6052
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9482
##             Specificity : 0.6124
##          Pos Pred Value : 0.8954
##          Neg Pred Value : 0.7717
##              Prevalence : 0.7778
##          Detection Rate : 0.7375
##    Detection Prevalence : 0.8236
##       Balanced Accuracy : 0.7803
##
##        'Positive' Class : 0
##
```

```r
# Plain QDA
qda_model_mixed = qda(mixed_formula, data = pca_data)
qda_preds_mixed = predict(qda_model_mixed)$class
qda_cm_mixed = confusionMatrix(qda_preds_mixed, as.factor(pca_data$loan_status))
print("Plain QDA Confusion Matrix (Mixed Selection):")
```

```
## [1] "Plain QDA Confusion Matrix (Mixed Selection):"
```

```r
print(qda_cm_mixed)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
```

```
##           0 32747  3680
##           1  2253  6320
##
##                Accuracy : 0.8682
##                  95% CI : (0.865, 0.8713)
##     No Information Rate : 0.7778
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5981
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9356
##             Specificity : 0.6320
##          Pos Pred Value : 0.8990
##          Neg Pred Value : 0.7372
##              Prevalence : 0.7778
##          Detection Rate : 0.7277
##    Detection Prevalence : 0.8095
##       Balanced Accuracy : 0.7838
##
##        'Positive' Class : 0
##
```

```r
# Define resampling control parameters for cross-validation
set.seed(123)
control_cv5 = trainControl(method = "cv", number = 5)
control_cv10 = trainControl(method = "cv", number = 10)

# LDA and QDA with 5-fold Cross-Validation using selected predictors from mixed selection
lda_cv5_mixed = train(mixed_formula, data = pca_data, method = "lda", trControl = control_cv5)
print("5-Fold Cross-Validated LDA Results (Mixed Selection):")
```

```
## [1] "5-Fold Cross-Validated LDA Results (Mixed Selection):"
```

```r
print(lda_cv5_mixed)
```

```
## Linear Discriminant Analysis
##
## 45000 samples
##    10 predictor
##     2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 36000, 36000, 36000, 36000, 36000
## Resampling results:
##
##   Accuracy   Kappa
##   0.8731778  0.6041107
```

```r
qda_cv5_mixed = train(mixed_formula, data = pca_data, method = "qda", trControl = control_cv5)
print("5-Fold Cross-Validated QDA Results (Mixed Selection):")
```

```
## [1] "5-Fold Cross-Validated QDA Results (Mixed Selection):"
```

```r
print(qda_cv5_mixed)
```

```
## Quadratic Discriminant Analysis
##
## 45000 samples
##    10 predictor
##     2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 36000, 36000, 36000, 36000, 36000
## Resampling results:
##
##   Accuracy  Kappa
##   0.8678    0.5968822
```

```r
# LDA and QDA with 10-fold Cross-Validation using selected predictors from mixed selection
lda_cv10_mixed = train(mixed_formula, data = pca_data, method = "lda", trControl = control_cv10)
print("10-Fold Cross-Validated LDA Results (Mixed Selection):")
```

```
## [1] "10-Fold Cross-Validated LDA Results (Mixed Selection):"
```

```r
print(lda_cv10_mixed)
```

```
## Linear Discriminant Analysis
##
## 45000 samples
##    10 predictor
##     2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 40500, 40500, 40500, 40500, 40500, 40500, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.8739556  0.6062419
```

```r
qda_cv10_mixed = train(mixed_formula, data = pca_data, method = "qda", trControl = control_cv10)
print("10-Fold Cross-Validated QDA Results (Mixed Selection):")
```

```
## [1] "10-Fold Cross-Validated QDA Results (Mixed Selection):"
```

```r
print(qda_cv10_mixed)
```

```
## Quadratic Discriminant Analysis
##
## 45000 samples
##    10 predictor
##     2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 40500, 40500, 40500, 40500, 40500, 40500, ...
## Resampling results:
##
##   Accuracy   Kappa
```

```
## 0.8674667 0.5959751
```

1. Results

- Plain LDA:
  - Accuracy: 87.36%
  - Sensitivity: 94.82% (high ability to correctly identify positives)
  - Specificity: 61.24% (moderate ability to identify negatives)
- Plain QDA:
  - Accuracy: 86.82%
  - Sensitivity: 93.56%
  - Specificity: 63.20% (slightly higher than LDA)
- 5-Fold Cross-Validated Results:
  - LDA: Accuracy = 87.32%, Kappa = 0.6041
  - QDA: Accuracy = 86.78%, Kappa = 0.5969
- 10-Fold Cross-Validated Results:
  - LDA: Accuracy = 87.40%, Kappa = 0.6062
  - QDA: Accuracy = 86.75%, Kappa = 0.5960

2. Comparison of Results

The LDA model consistently outperformed the QDA model across plain and cross-validated settings, achieving higher accuracy and Kappa values. The difference between 5-fold and 10-fold cross-validated accuracies is minor, showing both models' robustness and generalization capability. Notably, LDA had slightly higher sensitivity, which indicates it was more effective in identifying true positives (loan approvals), while QDA had slightly better specificity, suggesting it performed marginally better at identifying true negatives (loan denials).

3. Interpretation

1. **Model Choice**: LDA appears to be slightly better suited to this data, given its consistently higher accuracy and sensitivity, suggesting that a linear decision boundary may better fit the distribution of loan approval data than a quadratic one.
2. **Stability Across Folds**: The small differences between 5-fold and 10-fold cross-validated results for both LDA and QDA indicate that the models are stable and unlikely to be overfitting, which supports the reliability of these models on new data.
3. **Sensitivity vs. Specificity**: While LDA shows higher sensitivity, QDA has slightly higher specificity. This trade-off suggests that LDA might be more effective when the goal is to maximize true positives (approvals), while QDA might be preferable when avoiding false positives (incorrectly approving loans) is more critical.
4. **Effectiveness of Mixed Selection**: The mixed selection process using PCA components allowed the model to retain the most predictive components while achieving reasonable dimensionality reduction. This approach minimized multicollinearity issues, which is especially beneficial for QDA's stability, and allowed both models to perform well on the filtered predictors.

CONCLUSION

Among the models tested, the Linear Discriminant Analysis (LDA) model with mixed selection is the best choice for predicting loan approval. LDA consistently achieved the highest accuracy across both plain and resampled (5-fold and 10-fold cross-validation) scenarios, suggesting that its linear boundaries align well with the data distribution. The subset selection process using mixed selection on PCA components allowed LDA to retain only the most predictive and independent variables, which improved model stability and interpretability by reducing multicollinearity. This approach not only helped avoid overfitting but also demonstrated that a smaller subset of well-chosen features can capture the key patterns in loan approval data effectively. Additionally, the high sensitivity of LDA indicates that it performs well in identifying likely loan approvals, which could be valuable in contexts prioritizing true positives. The small difference between plain and resampled accuracies highlights LDA's robustness, making it a reliable model for generalization to unseen data.