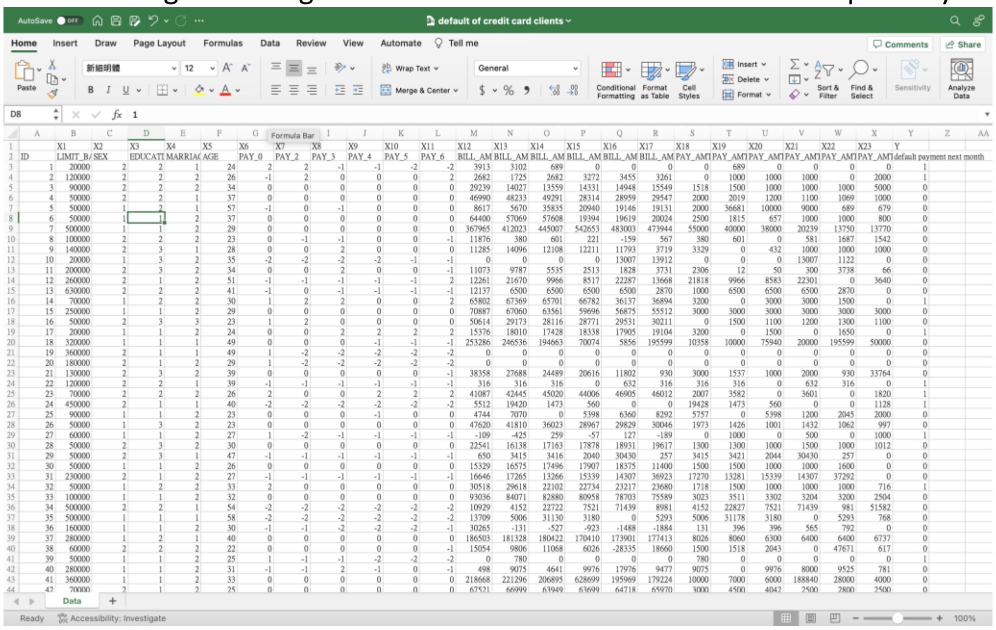
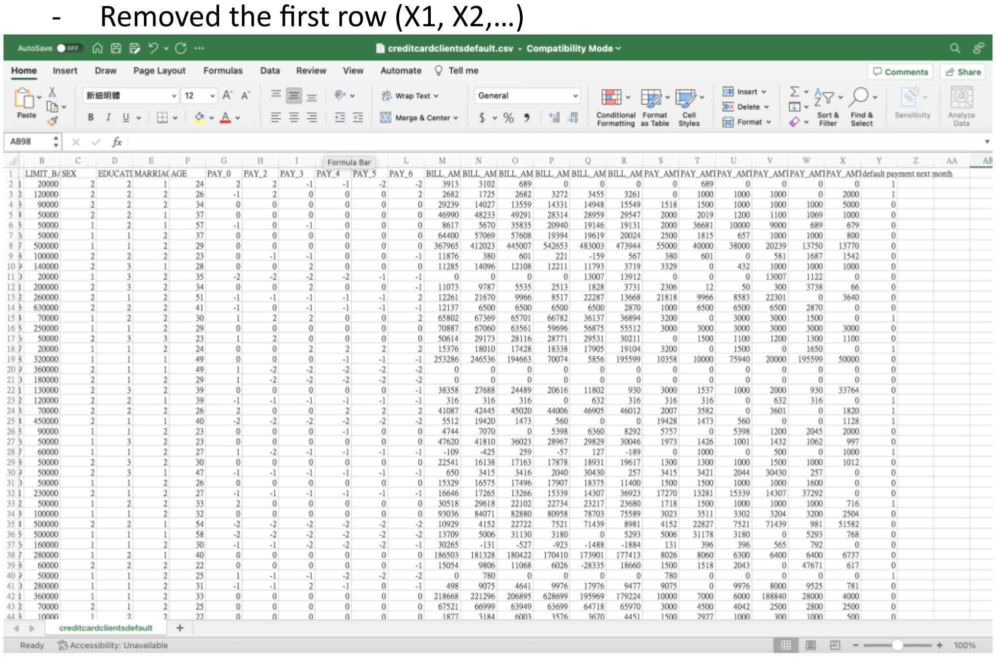
Riddhi Mistry
IMT 572 Mini Final Project - Credit Card Clients Default Analysis

The following is the original dataset downloaded from the UCI Repository:



Modified Dataset:

- Converted the file into .csv from .xlsx
- Removed the first row (X1, X2,…)



Summary Statistics Table:

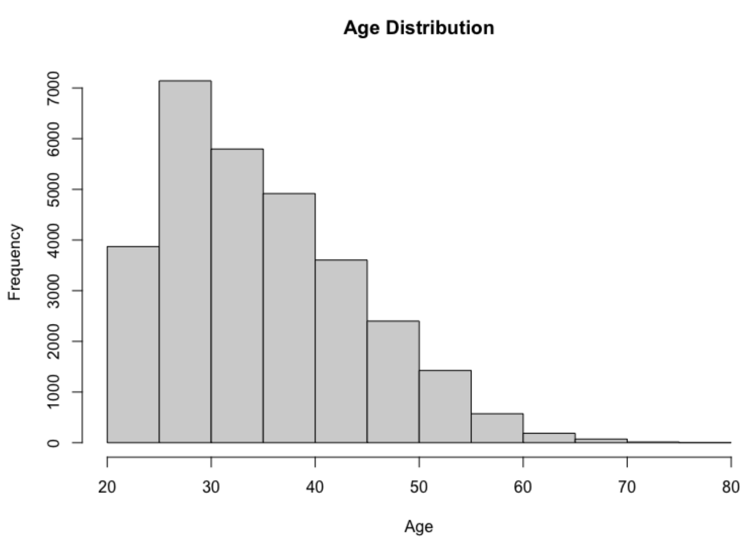
```
       ID           LIMIT_BAL            SEX          EDUCATION        MARRIAGE          AGE       
 Min.   :    1   Min.   :  10000   Min.   :1.000   Min.   :0.000   Min.   :0.000   Min.   :21.00  
 1st Qu.: 7501   1st Qu.:  50000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:28.00  
 Median :15000   Median : 140000   Median :2.000   Median :2.000   Median :2.000   Median :34.00  
 Mean   :15000   Mean   : 167484   Mean   :1.604   Mean   :1.853   Mean   :1.552   Mean   :35.49  
 3rd Qu.:22500   3rd Qu.: 240000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:41.00  
 Max.   :30000   Max.   :1000000   Max.   :2.000   Max.   :6.000   Max.   :3.000   Max.   :79.00  
     PAY_0             PAY_2             PAY_3             PAY_4             PAY_5             PAY_6        
 Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000  
 1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000  
 Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 0.0000  
 Mean   :-0.0167   Mean   :-0.1338   Mean   :-0.1662   Mean   :-0.2207   Mean   :-0.2662   Mean   :-0.2911  
 3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000  
 Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000  
   BILL_AMT1         BILL_AMT2        BILL_AMT3         BILL_AMT4         BILL_AMT5        BILL_AMT6      
 Min.   :-165580   Min.   :-69777   Min.   :-157264   Min.   :-170000   Min.   :-81334   Min.   :-339603  
 1st Qu.:   3559   1st Qu.:  2985   1st Qu.:   2666   1st Qu.:   2327   1st Qu.:  1763   1st Qu.:   1256  
 Median :  22382   Median : 21200   Median :  20088   Median :  19052   Median : 18104   Median :  17071  
 Mean   :  51223   Mean   : 49179   Mean   :  47013   Mean   :  43263   Mean   : 40311   Mean   :  38872  
 3rd Qu.:  67091   3rd Qu.: 64006   3rd Qu.:  60165   3rd Qu.:  54506   3rd Qu.: 50190   3rd Qu.:  49198  
 Max.   : 964511   Max.   :983931   Max.   :1664089   Max.   : 891586   Max.   :927171   Max.   : 961664  
    PAY_AMT1         PAY_AMT2          PAY_AMT3         PAY_AMT4         PAY_AMT5           PAY_AMT6       
 Min.   :     0   Min.   :      0   Min.   :     0   Min.   :     0   Min.   :     0.0   Min.   :     0.0  
 1st Qu.:  1000   1st Qu.:    833   1st Qu.:   390   1st Qu.:   296   1st Qu.:   252.5   1st Qu.:   117.8  
 Median :  2100   Median :   2009   Median :  1800   Median :  1500   Median :  1500.0   Median :  1500.0  
 Mean   :  5664   Mean   :   5921   Mean   :  5226   Mean   :  4826   Mean   :  4799.4   Mean   :  5215.5  
 3rd Qu.:  5006   3rd Qu.:   5000   3rd Qu.:  4505   3rd Qu.:  4013   3rd Qu.:  4031.5   3rd Qu.:  4000.0  
 Max.   :873552   Max.   :1684259   Max.   :896040   Max.   :621000   Max.   :426529.0   Max.   :528666.0  
 default.payment.next.month
 Min.   :0.0000            
 1st Qu.:0.0000            
 Median :0.0000            
 Mean   :0.2212            
 3rd Qu.:0.0000            
 Max.   :1.0000            
```

Exploratory Data Analysis

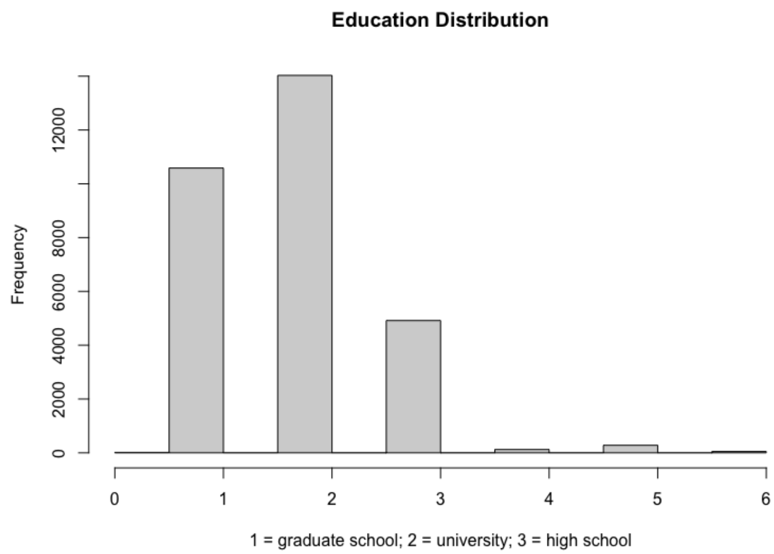
1. Since the variables for Sex, Education, and Marriage are numeric, there is no need for factorization and the variables can be used in the regression models directly.

2. We check for missing values using any_missing <- any(is.na(d)) and it outputs FALSE. Which means there are no missing values in the dataset.

3. Plotting a histogram to understand age distribution of credit card clients.



4. Plotting a graph to understand the distribution of the level of education.



Choosing predictor variables:

LIMIT_BAL: Credit card limit balance tells how likely people are to default, as clients with higher credit limits might be more likely.

EDUCATION: Education level can also be a crucial factor. Higher education level might correlate with lower default rates.

AGE: Age can also be an important factor since younger individuals might default more while older clients can have lower default rates with experience.

PAY_0 to PAY_6: These variables are repayment status numbered by month. They can be a strong predictor since clients with a greater number of delayed months or consistently delayed payments can be more likely to default.

BILL_AMT1 to BILL_AMT6: Higher bill amount can lead to more chances of default since it means higher financial stress. Hence it is a strong predictor too.

PAY_AMT1 to PAY_AMT6: Monthly payment amounts can also talk strongly about the possibility of default for a client. If the client has had consistently low payments or shows a trend of decreasing monthly payment amounts, they have higher chances of default.

Thus, the above variables will be used in the models.

Logit Regression

Summary:

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.253e+00  6.694e-02 -18.725  < 2e-16 ***
LIMIT_BAL   -7.163e-07  1.560e-07  -4.592 4.39e-06 ***
EDUCATION   -9.491e-02  2.080e-02  -4.564 5.03e-06 ***
AGE          1.128e-02  1.629e-03   6.924 4.38e-12 ***
PAY_0        5.779e-01  1.768e-02  32.681  < 2e-16 ***
PAY_2        8.548e-02  2.016e-02   4.239 2.25e-05 ***
PAY_3        7.247e-02  2.258e-02   3.209  0.00133 **
PAY_4        2.419e-02  2.499e-02   0.968  0.33305
PAY_5        3.500e-02  2.687e-02   1.303  0.19260
PAY_6        7.074e-03  2.212e-02   0.320  0.74912
BILL_AMT1   -5.475e-06  1.136e-06  -4.817 1.46e-06 ***
BILL_AMT2    2.388e-06  1.501e-06   1.590  0.11177
BILL_AMT3    1.324e-06  1.322e-06   1.002  0.31654
BILL_AMT4   -1.832e-07  1.350e-06  -0.136  0.89208
BILL_AMT5    6.161e-07  1.518e-06   0.406  0.68488
BILL_AMT6    3.715e-07  1.192e-06   0.312  0.75534
PAY_AMT1    -1.363e-05  2.306e-06  -5.912 3.38e-09 ***
PAY_AMT2    -9.612e-06  2.098e-06  -4.580 4.64e-06 ***
PAY_AMT3    -2.755e-06  1.726e-06  -1.596  0.11045
PAY_AMT4    -4.013e-06  1.790e-06  -2.243  0.02493 *
PAY_AMT5    -3.415e-06  1.777e-06  -1.921  0.05469 .
PAY_AMT6    -2.101e-06  1.300e-06  -1.616  0.10602
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 31705  on 29999  degrees of freedom
Residual deviance: 27911  on 29978  degrees of freedom
AIC: 27955

Number of Fisher Scoring iterations: 6
```

According to the logistic regression summary, coefficients like LIMIT_BAL, EDUCATION, AGE, payment history (PAY_0, PAY_2) bill amount for first month (BILL_AMT1) and payment amounts for first two months (PAY_AMT1, PAY_AMT2) are significant.
AIC Score is 27955, which is reasonable for the model.

Marginal Effects Analysis:

```
Marginal Effects:
                dF/dx   Std. Err.        z     P>|z|
LIMIT_BAL -1.0932e-07  2.3760e-08 -4.6010 4.204e-06 ***
EDUCATION -1.4485e-02  3.1721e-03 -4.5663 4.964e-06 ***
AGE        1.7208e-03  2.4856e-04  6.9232 4.415e-12 ***
PAY_0      8.8185e-02  2.6838e-03 32.8579 < 2.2e-16 ***
PAY_2      1.3044e-02  3.0812e-03  4.2334 2.301e-05 ***
PAY_3      1.1059e-02  3.4464e-03  3.2090  0.001332 **
PAY_4      3.6913e-03  3.8132e-03  0.9680  0.333035
PAY_5      5.3420e-03  4.1002e-03  1.3029  0.192623
PAY_6      1.0796e-03  3.3759e-03  0.3198  0.749119
BILL_AMT1 -8.3547e-07  1.7311e-07 -4.8261 1.392e-06 ***
BILL_AMT2  3.6436e-07  2.2900e-07  1.5911  0.111593
BILL_AMT3  2.0213e-07  2.0178e-07  1.0017  0.316475
BILL_AMT4 -2.7960e-08  2.0607e-07 -0.1357  0.892076
BILL_AMT5  9.4024e-08  2.3169e-07  0.4058  0.684871
BILL_AMT6  5.6695e-08  1.8194e-07  0.3116  0.755335
PAY_AMT1  -2.0807e-06  3.4951e-07 -5.9533 2.627e-09 ***
PAY_AMT2  -1.4668e-06  3.1852e-07 -4.6050 4.125e-06 ***
PAY_AMT3  -4.2040e-07  2.6325e-07 -1.5970  0.110275
PAY_AMT4  -6.1243e-07  2.7290e-07 -2.2442  0.024821 *
PAY_AMT5  -5.2115e-07  2.7111e-07 -1.9223  0.054571 .
PAY_AMT6  -3.2056e-07  1.9822e-07 -1.6172  0.105840
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

df/dx represents marginal effects or the change in probability for default.

According to the marginal effects table, the significant variables have the following interpretation:

1. LIMIT_BAL: Decrease in credit limit balance shows increase in default probability.
2. EDUCATION: Lower level of education indicates increase in default probability.
3. AGE: Higher clients have slightly higher probability of default
4. PAY_0 to PAY_3: Higher chances of default for payments that are more delayed.
5. BILL_AMT1: Higher billing amounts are associated with higher probability of default in next month.
6. PAY_AMT1, PAY_AMT2, PAY_AMT4, PAY_AMT5: Higher pay amounts can show a decrease in the probability of default whereas lower can indicate higher default probability.

Probit Regression

Summary:

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.649e-01  3.804e-02 -20.108  < 2e-16 ***
LIMIT_BAL   -3.907e-07  8.600e-08  -4.543 5.56e-06 ***
EDUCATION   -5.195e-02  1.166e-02  -4.457 8.29e-06 ***
AGE          6.706e-03  9.340e-04   7.180 6.99e-13 ***
PAY_0        3.120e-01  1.017e-02  30.675  < 2e-16 ***
PAY_2        5.183e-02  1.184e-02   4.378 1.20e-05 ***
PAY_3        3.891e-02  1.303e-02   2.986  0.00283 **
PAY_4        1.113e-02  1.445e-02   0.770  0.44109
PAY_5        2.026e-02  1.560e-02   1.298  0.19417
PAY_6        3.086e-03  1.281e-02   0.241  0.80970
BILL_AMT1   -2.612e-06  5.709e-07  -4.575 4.76e-06 ***
BILL_AMT2    9.157e-07  7.765e-07   1.179  0.23828
BILL_AMT3    5.731e-07  7.071e-07   0.811  0.41761
BILL_AMT4    3.304e-08  7.218e-07   0.046  0.96350
BILL_AMT5    2.744e-07  8.120e-07   0.338  0.73538
BILL_AMT6    6.350e-08  6.355e-07   0.100  0.92041
PAY_AMT1    -6.204e-06  1.093e-06  -5.676 1.38e-08 ***
PAY_AMT2    -4.207e-06  9.913e-07  -4.244 2.19e-05 ***
PAY_AMT3    -1.454e-06  8.825e-07  -1.647  0.09954 .
PAY_AMT4    -1.774e-06  9.029e-07  -1.965  0.04940 *
PAY_AMT5    -1.452e-06  9.057e-07  -1.603  0.10891
PAY_AMT6    -8.820e-07  6.595e-07  -1.337  0.18110
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 31705  on 29999  degrees of freedom
Residual deviance: 28113  on 29978  degrees of freedom
AIC: 28157

Number of Fisher Scoring iterations: 6
```

Coefficients like LIMIT_BAL, EDUCATION, AGE, payment history (PAY_0, PAY_2) bill amount for first month (BILL_AMT1) and payment amounts for first two months (PAY_AMT1, PAY_AMT2) are highly significant here too, as in logit.
AIC Score is 28157, which is greater than logit.

Marginal Effects:

```
Marginal Effects:
                dF/dx   Std. Err.       z     P>|z|
LIMIT_BAL -1.0854e-07  2.3876e-08 -4.5460 5.468e-06 ***
EDUCATION -1.4435e-02  3.2375e-03 -4.4586 8.251e-06 ***
AGE        1.8631e-03  2.5954e-04  7.1786 7.045e-13 ***
PAY_0      8.6682e-02  2.8238e-03 30.6968 < 2.2e-16 ***
PAY_2      1.4401e-02  3.2913e-03  4.3754 1.212e-05 ***
PAY_3      1.0811e-02  3.6209e-03  2.9858  0.002829 **
PAY_4      3.0934e-03  4.0156e-03  0.7704  0.441088
PAY_5      5.6275e-03  4.3344e-03  1.2983  0.194174
PAY_6      8.5732e-04  3.5601e-03  0.2408  0.809701
BILL_AMT1 -7.2563e-07  1.5849e-07 -4.5785 4.684e-06 ***
BILL_AMT2  2.5442e-07  2.1571e-07  1.1795  0.238215
BILL_AMT3  1.5924e-07  1.9644e-07  0.8106  0.417591
BILL_AMT4  9.1784e-09  2.0054e-07  0.0458  0.963495
BILL_AMT5  7.6244e-08  2.2559e-07  0.3380  0.735377
BILL_AMT6  1.7643e-08  1.7657e-07  0.0999  0.920405
PAY_AMT1  -1.7237e-06  3.0274e-07 -5.6936 1.244e-08 ***
PAY_AMT2  -1.1689e-06  2.7479e-07 -4.2538 2.102e-05 ***
PAY_AMT3  -4.0385e-07  2.4513e-07 -1.6475  0.099455 .
PAY_AMT4  -4.9298e-07  2.5080e-07 -1.9656  0.049341 *
PAY_AMT5  -4.0341e-07  2.5160e-07 -1.6034  0.108855
PAY_AMT6  -2.4504e-07  1.8319e-07 -1.3376  0.181016
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

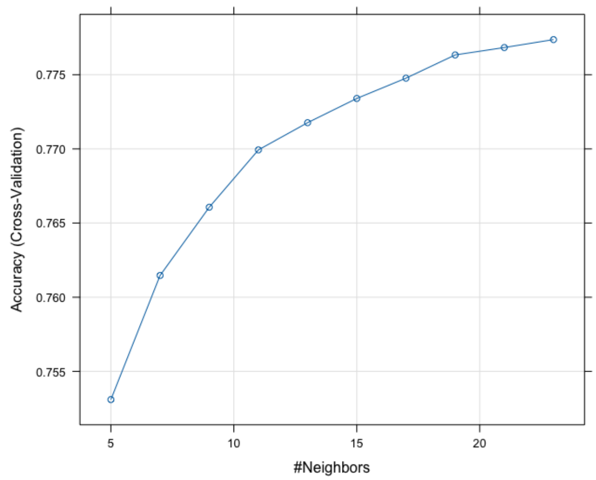
The marginal effects for probit is very similar to logit. Positive value represents an increase in the probability of default (for example PAY_0 to PAY_6).

Prediction Model:
Dataset size = 30000
Predictors = 21
Cross validation number of folds = 5
tune length = 10

1. **KNN Model:** It gives an accuracy of 77.74% with the final value of k = 23.

```
k-Nearest Neighbors

30000 samples
   21 predictor
    2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 24000, 24000, 24000, 24000, 24000
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   5  0.7531000  0.11605658
   7  0.7614667  0.11173715
   9  0.7660667  0.10847952
  11  0.7699333  0.10798225
  13  0.7717667  0.10205788
  15  0.7734000  0.10084118
  17  0.7747667  0.09925668
  19  0.7763333  0.09844912
  21  0.7768333  0.09391628
  23  0.7773667  0.09114417

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 23.
```

Plotting accuracy vs number of neighbors



2. **Naïve Bayes Model:** It gives an accuracy of 79.56% and the optimal model is chosen with laplace = 0, usekernel = TRUE and adjust = 1.

```
Naive Bayes

30000 samples
   21 predictor
    2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 23999, 24001, 24000, 24000, 24000
Resampling results across tuning parameters:

  usekernel  Accuracy   Kappa
  FALSE      0.7098011  0.3106027
   TRUE      0.7956000  0.1835265

Tuning parameter 'laplace' was held constant at a value of 0
Tuning parameter 'adjust' was
 held constant at a value of 1
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were laplace = 0, usekernel = TRUE and adjust = 1.
```

Thus, based on the accuracy, Naïve Bayes will be the optimal model for predicting default for credit card clients.