MET CS 699 Data Mining
Final Summarized Report 2

Classification for Credit Card Fraud Detection

By:
RAJSHRI TOTLA

Table of Contents

Summary

1. Originally, we used weka's **Percentage Split** in which the split was 70%-30% balanced credit card dataset.
2. Secondly, used weka's **Supplied Test Set** Feature in which for training I used 100% balanced dataset and 30% original dataset.

Combinations Tried For Classifier Models

Sr. No.	TRAIN	TEST	WEKA FEATURE	ATTRIBUTE SELECTION
1	Balanced 100% Dataset	Balanced 100% Dataset	**Percentage Split**	Balanced 100% Dataset
2	Balanced 100% Dataset	Unbalanced 30% Dataset	**Supplied Test Set**	Balanced 100% Dataset
3	Balanced Train Dataset (Made from 70% Original)	Unbalanced 30% Dataset (exclusive of train dataset)	**Supplied Test Set**	Unbalanced Train Dataset (Made from 70% Original)
4	Balanced Train Dataset (Made from 70% Original)	Unbalanced 30% Dataset (exclusive of train dataset)	**Supplied Test Set**	Balanced Train Dataset (Made from 70% Original)
1: Results Originally Submitted 2: Results don't changed so decided to redo whole project 3 & 4: Results are in Excel File				

3. Finally, the result mentioned in excel and screenshots in this report are after I did the whole project from beginning:
    a. Performing pre-processing using R again as mentioned in previous report. (including Split and Sampling)
    b. Performed Attribute Selection from
        i. Unbalanced Train Dataset
        ii. Balanced Train Dataset
    c. Ran 5 classifier models for each
    d. Total: 4*5*2 = 40 Models
4. Attribute Selection Results
   Interestingly attributes are repeating for the highlighted ones:

Information Gain Ranking Filter

Balanced Dataset : 16,27,23,14,26,25
Unbalanced Dataset :**15,18,13,11,12,17**

OneR feature evaluator

Balanced Dataset: **18,13,15,12,11,17**

Unbalanced Dataset: **13,18,17,15,12,11**

Gain Ratio feature evaluator

Balanced Dataset: **18,15,13,11,12,17**

Unbalanced Dataset: 11,18,23,8,15,19

Correlation Ranking Filter

Balanced Dataset: 15,5,12,13,11,17

Unbalanced Dataset: 18,15,13,11,17,4

5. Now to generate classifiers:
   Training Data: Balanced Train Dataset (Made from 70% of original dataset)
   Test Data: Unbalanced original dataset other 30% (70% in train)

RCODE

```
library('ROSE') #Sampling Package
setwd("") # Set working directory
data<- read.csv("creditcard.csv",header = TRUE) #read original file

# to work in weka change class variable to categorical
data$Class[which(data$Class==1)] <- "yes"
data$Class[which(data$Class==0)] <- "no"
table(data$Class) # check the class distibution

# perfor test-train split
smp_siz = floor(0.70*nrow(data)) # sample size creation
set.seed(3)   # set seed to ensure you always have same random numbers generated
# train data index
train_ind = sample(seq_len(nrow(data)),size = smp_siz, replace = FALSE)  # Randomly identifies the rows equal to sample size
train = data[train_ind,] #creates the training dataset with row numbers stored in train_ind
table(train$Class) # check the class distibution
write.csv(train, "traincreditcard.csv") # save data

test=data[-train_ind,]  # creates test data
table(test$Class) # check the class distibution
write.csv(test, "testcreditcard.csv") # save data

# sampling performed on train data
data <- as.matrix(data)
train1 <- as.data.frame(train)
balanced_data <- ovun.sample(Class ~ ., data = train1, method = "both",N = nrow(train1)) # performed hybrid sampling
table(balanced_data$data$Class)
write.csv(balanced_data$data, "balancedtraincreditcard.csv") # save data
```

Attribute Selection

1. Information Gain Ranking Filter - Balanced Dataset

```
=== Attribute Selection on all input data ===

Search Method:
	Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 31 Class):
	Information Gain Ranking Filter

Ranked attributes:
 0.999   16 V15
 0.999   27 V26
 0.999   23 V22
 0.999   14 V13
 0.999   26 V25
 0.998   25 V24
 0.998   20 V19
 0.998   19 V18
 0.998   21 V20
 0.998   24 V23
 0.998   10 V9
 0.997   18 V17
 0.996   17 V16
 0.996    2 V1
 0.996   13 V12
 0.996    9 V8
 0.995   29 V28
 0.995    6 V5
 0.995   12 V11
 0.994   28 V27
 0.993   15 V14
 0.993   11 V10
 0.993    3 V2
 0.993    5 V4
 0.992   22 V21
 0.99     1 Time
 0.988    8 V7
 0.984    7 V6
 0.982    4 V3
 0.763   30 Amount

Selected attributes: 16,27,23,14,26,25,20,19,21,24,10,18,17,2,13,9,29,6,12,28,15,11,3,5,22,1,8,7,4,30 : 30
```

2. Information Gain Ranking Filter - Unbalanced Dataset

```
=== Attribute Selection on all input data ===

Search Method:
	Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 31 Class):
	Information Gain Ranking Filter

Ranked attributes:
 0.0118406   15 V14
 0.0118262   18 V17
 0.0108356   13 V12
 0.0106566   11 V10
 0.00978     12 V11
 0.0088374   17 V16
 0.0070626    5 V4
 0.0069322    4 V3
 0.0060476   10 V9
 0.0059665   19 V18
 0.0058219    8 V7
 0.0045742    3 V2
 0.0034498   22 V21
 0.0032144    7 V6
 0.0032013   28 V27
 0.0031911    6 V5
 0.0026272    2 V1
 0.0026012    9 V8
 0.0022539   29 V28
 0.0015825   20 V19
 0.0015727   30 Amount
 0.0012676   21 V20
 0.001068     1 Time
 0.0008594   24 V23
 0.000381    23 V22
 0.000354    26 V25
 0.0001126   25 V24
 0.0000772   14 V13
 0           16 V15
 0           27 V26

Selected attributes: 15,18,13,11,12,17,5,4,10,19,8,3,22,7,28,6,2,9,29,20,30,21,1,24,23,26,25,14,16,27 : 30
```

3. OneR feature evaluator - Balanced Dataset

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 31 Class):
        OneR feature evaluator.

        Using 10 fold cross validation for evaluating attributes.
        Minimum bucket size for OneR: 6

Ranked attributes:
99.927   18 V17
99.904   13 V12
99.899   15 V14
99.897   12 V11
99.884   11 V10
99.878   17 V16
99.871   19 V18
99.865   10 V9
99.843   23 V22
99.842   25 V24
99.841   14 V13
99.838   26 V25
99.836    5 V4
99.833   16 V15
99.833   27 V26
99.832    9 V8
99.82    24 V23
99.819   20 V19
99.816   21 V20
99.813   29 V28
99.807    6 V5
99.797    4 V3
99.796    7 V6
99.791    3 V2
99.789    8 V7
99.788    2 V1
99.785   22 V21
99.773   28 V27
99.768    1 Time
93.464   30 Amount

Selected attributes: 18,13,15,12,11,17,19,10,23,25,14,26,5,16,27,9,24,20,21,29,6,4,7,3,8,2,22,28,1,30 : 30
```

4. OneR feature evaluator - Unbalanced Dataset

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 31 Class):
        OneR feature evaluator.

        Using 10 fold cross validation for evaluating attributes.
        Minimum bucket size for OneR: 6

Ranked attributes:
99.903   13 V12
99.902   18 V17
99.881   17 V16
99.874   15 V14
99.873   12 V11
99.873   11 V10
99.864   19 V18
99.857   10 V9
99.838    4 V3
99.837    3 V2
99.836    8 V7
99.831    9 V8
99.829   22 V21
99.829    1 Time
99.829   28 V27
99.829    6 V5
99.827   25 V24
99.827   27 V26
99.827   16 V15
99.827   26 V25
99.827   30 Amount
99.827   20 V19
99.827   21 V20
99.827   23 V22
99.827   14 V13
99.827   24 V23
99.827   29 V28
99.826    7 V6
99.826    2 V1
99.819    5 V4

Selected attributes: 13,18,17,15,12,11,19,10,4,3,8,9,22,1,28,6,25,27,16,26,30,20,21,23,14,24,29,7,2,5 : 30
```

5. Gain Ratio feature evaluator - Balanced Dataset

```
=== Attribute Selection on all input data ===

Search Method:
	Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 31 Class):
	Gain Ratio feature evaluator

Ranked attributes:
 0.199    18 V17
 0.198    15 V14
 0.184    13 V12
 0.175    11 V10
 0.167    12 V11
 0.156    17 V16
 0.153     5 V4
 0.149     4 V3
 0.135     8 V7
 0.132    10 V9
 0.131    19 V18
 0.126     3 V2
 0.123    22 V21
 0.122     7 V6
 0.121     6 V5
 0.121     2 V1
 0.12     28 V27
 0.119    29 V28
 0.118     9 V8
 0.115     1 Time
 0.115    20 V19
 0.114    21 V20
 0.113    24 V23
 0.112    27 V26
 0.112    25 V24
 0.112    26 V25
 0.111    14 V13
 0.111    16 V15
 0.111    23 V22
 0.105    30 Amount

Selected attributes: 18,15,13,11,12,17,5,4,8,10,19,3,22,7,6,2,28,29,9,1,20,21,24,27,25,26,14,16,23,30 : 30
```

6. Gain Ratio feature evaluator - Unbalanced Dataset

```
=== Attribute Selection on all input data ===

Search Method:
	Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 31 Class):
	Gain Ratio feature evaluator

Ranked attributes:
 0.078839   11 V10
 0.033006   18 V17
 0.031096   23 V22
 0.017316    8 V7
 0.016823   15 V14
 0.010765   19 V18
 0.009834   13 V12
 0.00953    12 V11
 0.009175   17 V16
 0.006261    5 V4
 0.006071    6 V5
 0.005725    3 V2
 0.005684   10 V9
 0.005278    7 V6
 0.005208    4 V3
 0.00455    28 V27
 0.003266   22 V21
 0.002701    9 V8
 0.002454   26 V25
 0.001648    2 V1
 0.001463   29 V28
 0.001245   21 V20
 0.001079   20 V19
 0.001066   30 Amount
 0.00086    24 V23
 0.000686    1 Time
 0.000159   25 V24
 0.000118   14 V13
 0          16 V15
 0          27 V26

Selected attributes: 11,18,23,8,15,19,13,12,17,5,6,3,10,7,4,28,22,9,26,2,29,21,20,30,24,1,25,14,16,27 : 30
```

7. Correlation Ranking Filter - Balanced Dataset

```
=== Attribute Selection on all input data ===

Search Method:
	Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 31 Class):
	Correlation Ranking Filter
Ranked attributes:
 0.7558   15 V14
 0.7068    5 V4
 0.6856   12 V11
 0.6852   13 V12
 0.624    11 V10
 0.5887   17 V16
 0.5628    4 V3
 0.556    18 V17
 0.5409   10 V9
 0.4832    3 V2
 0.4712    8 V7
 0.4582   19 V18
 0.4293    2 V1
 0.3848    7 V6
 0.3782    6 V5
 0.2504   20 V19
 0.1796    1 Time
 0.1453   21 V20
 0.1369   22 V21
 0.1009   28 V27
 0.0997   25 V24
 0.0899   29 V28
 0.0623   26 V25
 0.0578   30 Amount
 0.0561   14 V13
 0.0501   27 V26
 0.0467   24 V23
 0.0463    9 V8
 0.046    16 V15
 0.0353   23 V22

Selected attributes: 15,5,12,13,11,17,4,18,10,3,8,19,2,7,6,20,1,21,22,28,25,29,26,30,14,27,24,9,16,23 : 30
```

8. Correlation Ranking Filter - Unbalanced Dataset

```
=== Attribute Selection on all input data ===

Search Method:
	Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 31 Class):
	Correlation Ranking Filter
Ranked attributes:
 0.326481   18 V17
 0.302544   15 V14
 0.260593   13 V12
 0.216883   11 V10
 0.196539   17 V16
 0.192961    4 V3
 0.187257    8 V7
 0.154876   12 V11
 0.133447    5 V4
 0.111485   19 V18
 0.101347    2 V1
 0.097733   10 V9
 0.094974    6 V5
 0.091289    3 V2
 0.043643    7 V6
 0.040413   22 V21
 0.034783   20 V19
 0.02009    21 V20
 0.019875    9 V8
 0.01758    28 V27
 0.012323    1 Time
 0.009536   29 V28
 0.007221   25 V24
 0.005632   30 Amount
 0.00457    14 V13
 0.004455   27 V26
 0.004223   16 V15
 0.003308   26 V25
 0.002685   24 V23
 0.000805   23 V22

Selected attributes: 18,15,13,11,17,4,8,12,5,19,2,10,6,3,7,22,20,21,9,28,1,29,25,30,14,27,16,26,24,23 : 30
```

Classifier Models

1. Random Forest

1. Information Gain Ranking Filter - Balanced Dataset

```
=== Summary ===

Correctly Classified Instances       85384               99.9309 %
Incorrectly Classified Instances        59                0.0691 %
Kappa statistic                          0.801
Mean absolute error                      0.0018
Root mean squared error                  0.0251
Relative absolute error                  0.3648 %
Root relative squared error              5.0152 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.217    1.000      1.000    1.000      0.801    0.941     1.000     no
                 0.783    0.000    0.821      0.783    0.801      0.801    0.941     0.811     yes
Weighted Avg.    0.999    0.217    0.999      0.999    0.999      0.801    0.941     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 85265    26 |     a = no
    33   119 |     b = yes
```

2. Information Gain Ranking Filter - Unbalanced Dataset

```
=== Summary ===

Correctly Classified Instances       85390               99.938  %
Incorrectly Classified Instances        53                0.062  %
Kappa statistic                          0.8224
Mean absolute error                      0.0017
Root mean squared error                  0.0248
Relative absolute error                  0.3472 %
Root relative squared error              4.9597 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.191    1.000      1.000    1.000      0.823    0.938     1.000     no
                 0.809    0.000    0.837      0.809    0.823      0.823    0.938     0.804     yes
Weighted Avg.    0.999    0.190    0.999      0.999    0.999      0.823    0.938     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 85267    24 |     a = no
    29   123 |     b = yes
```

3. OneR feature evaluator - Balanced Dataset

```
=== Summary ===

Correctly Classified Instances       85390               99.938  %
Incorrectly Classified Instances        53                0.062  %
Kappa statistic                          0.8224
Mean absolute error                      0.0017
Root mean squared error                  0.0248
Relative absolute error                  0.3472 %
Root relative squared error              4.9597 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.191    1.000      1.000    1.000      0.823    0.938     1.000     no
                 0.809    0.000    0.837      0.809    0.823      0.823    0.938     0.804     yes
Weighted Avg.    0.999    0.190    0.999      0.999    0.999      0.823    0.938     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 85267    24 |     a = no
    29   123 |     b = yes
```

4. OneR feature evaluator - Unbalanced Dataset

```
=== Summary ===

Correctly Classified Instances       85390               99.938  %
Incorrectly Classified Instances        53                0.062  %
Kappa statistic                          0.8224
Mean absolute error                      0.0017
Root mean squared error                  0.0248
Relative absolute error                  0.3472 %
Root relative squared error              4.9597 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.191    1.000      1.000    1.000      0.823    0.938     1.000     no
                 0.809    0.000    0.837      0.809    0.823      0.823    0.938     0.804     yes
Weighted Avg.    0.999    0.190    0.999      0.999    0.999      0.823    0.938     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 85267    24 |     a = no
    29   123 |     b = yes
```

5. Gain Ratio feature evaluator - Balanced Dataset

```
=== Summary ===

Correctly Classified Instances       85390               99.938  %
Incorrectly Classified Instances        53                0.062  %
Kappa statistic                          0.8224
Mean absolute error                      0.0017
Root mean squared error                  0.0248
Relative absolute error                  0.3472 %
Root relative squared error              4.9597 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.191    1.000      1.000    1.000      0.823    0.938     1.000     no
                 0.809    0.000    0.837      0.809    0.823      0.823    0.938     0.804     yes
Weighted Avg.    0.999    0.190    0.999      0.999    0.999      0.823    0.938     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 85267    24 |     a = no
    29   123 |     b = yes
```

6. Gain Ratio feature evaluator - Unbalanced Dataset

```
=== Summary ===

Correctly Classified Instances       85368               99.9122 %
Incorrectly Classified Instances        75                0.0878 %
Kappa statistic                          0.7187
Mean absolute error                      0.0028
Root mean squared error                  0.0299
Relative absolute error                  0.5506 %
Root relative squared error              5.9872 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.368    0.999      1.000    1.000      0.726    0.914     1.000     no
                 0.632    0.000    0.835      0.632    0.719      0.726    0.914     0.672     yes
Weighted Avg.    0.999    0.368    0.999      0.999    0.999      0.726    0.914     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 85272    19 |     a = no
    56    96 |     b = yes
```

7. Correlation Ranking Filter - Balanced Dataset

```
=== Summary ===

Correctly Classified Instances       85392               99.9403 %
Incorrectly Classified Instances        51                0.0597 %
Kappa statistic                          0.8256
Mean absolute error                      0.0016
Root mean squared error                  0.0242
Relative absolute error                  0.3291 %
Root relative squared error              4.8487 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.204    1.000      1.000    1.000      0.826    0.941     1.000     no
                 0.796    0.000    0.858      0.796    0.826      0.826    0.941     0.804     yes
Weighted Avg.    0.999    0.204    0.999      0.999    0.999      0.826    0.941     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 85271    20 |     a = no
    31   121 |     b = yes
```

8. Correlation Ranking Filter - Unbalanced Dataset

```
=== Summary ===

Correctly Classified Instances       85384               99.9309 %
Incorrectly Classified Instances        59                0.0691 %
Kappa statistic                          0.801
Mean absolute error                      0.0018
Root mean squared error                  0.025
Relative absolute error                  0.3563 %
Root relative squared error              4.9946 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.217    1.000      1.000    1.000      0.801    0.945     1.000     no
                 0.783    0.000    0.821      0.783    0.801      0.801    0.945     0.791     yes
Weighted Avg.    0.999    0.217    0.999      0.999    0.999      0.801    0.945     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 85265    26 |     a = no
    33   119 |     b = yes
```

2. J48

1. Information Gain Ranking Filter - Balanced Dataset

```
=== Summary ===

Correctly Classified Instances       85166               99.6758 %
Incorrectly Classified Instances       277                0.3242 %
Kappa statistic                          0.4649
Mean absolute error                      0.0032
Root mean squared error                  0.0564
Relative absolute error                  0.6431 %
Root relative squared error             11.2921 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.997    0.204    1.000      0.997    0.998      0.511    0.897     1.000     no
                 0.796    0.003    0.330      0.796    0.466      0.511    0.897     0.470     yes
Weighted Avg.    0.997    0.204    0.998      0.997    0.997      0.511    0.897     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 85045   246 |     a = no
    31   121 |     b = yes
```

2. Information Gain Ranking Filter - Unbalanced Dataset

```
=== Summary ===

Correctly Classified Instances       85253               99.7776 %
Incorrectly Classified Instances       190                0.2224 %
Kappa statistic                          0.5691
Mean absolute error                      0.0022
Root mean squared error                  0.0469
Relative absolute error                  0.4427 %
Root relative squared error              9.3846 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.998    0.171    1.000      0.998    0.999      0.599    0.914     1.000     no
                 0.829    0.002    0.434      0.829    0.570      0.599    0.914     0.532     yes
Weighted Avg.    0.998    0.171    0.999      0.998    0.998      0.599    0.914     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 85127   164 |     a = no
    26   126 |     b = yes
```

3. OneR feature evaluator - Balanced Dataset

```
=== Summary ===

Correctly Classified Instances       85253               99.7776 %
Incorrectly Classified Instances       190                0.2224 %
Kappa statistic                          0.5691
Mean absolute error                      0.0022
Root mean squared error                  0.0469
Relative absolute error                  0.4427 %
Root relative squared error              9.3846 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.998    0.171    1.000      0.998   0.999      0.599    0.914     1.000     no
                 0.829    0.002    0.434      0.829   0.570      0.599    0.914     0.532     yes
Weighted Avg.    0.998    0.171    0.999      0.998   0.998      0.599    0.914     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 85127   164 |     a = no
    26   126 |     b = yes
```

4. OneR feature evaluator - Unbalanced Dataset

```
=== Summary ===

Correctly Classified Instances       85253               99.7776 %
Incorrectly Classified Instances       190                0.2224 %
Kappa statistic                          0.5691
Mean absolute error                      0.0022
Root mean squared error                  0.0469
Relative absolute error                  0.4427 %
Root relative squared error              9.3846 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.998    0.171    1.000      0.998   0.999      0.599    0.914     1.000     no
                 0.829    0.002    0.434      0.829   0.570      0.599    0.914     0.532     yes
Weighted Avg.    0.998    0.171    0.999      0.998   0.998      0.599    0.914     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 85127   164 |     a = no
    26   126 |     b = yes
```

5. Gain Ratio feature evaluator - Balanced Dataset

```
=== Summary ===

Correctly Classified Instances       85253               99.7776 %
Incorrectly Classified Instances       190                0.2224 %
Kappa statistic                          0.5691
Mean absolute error                      0.0022
Root mean squared error                  0.0469
Relative absolute error                  0.4427 %
Root relative squared error              9.3846 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.998    0.171    1.000      0.998    0.999      0.599    0.914     1.000     no
                 0.829    0.002    0.434      0.829    0.570      0.599    0.914     0.532     yes
Weighted Avg.    0.998    0.171    0.999      0.998    0.998      0.599    0.914     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 85127   164 |     a = no
    26   126 |     b = yes
```

6. Gain Ratio feature evaluator - Unbalanced Dataset

```
=== Summary ===

Correctly Classified Instances       85026               99.512  %
Incorrectly Classified Instances       417                0.488  %
Kappa statistic                          0.3267
Mean absolute error                      0.0048
Root mean squared error                  0.0693
Relative absolute error                  0.9694 %
Root relative squared error             13.8722 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.996    0.329    0.999      0.996    0.998      0.380    0.834     0.999     no
                 0.671    0.004    0.217      0.671    0.329      0.380    0.834     0.292     yes
Weighted Avg.    0.995    0.328    0.998      0.995    0.996      0.380    0.834     0.998

=== Confusion Matrix ===

     a     b   <-- classified as
 84924   367 |     a = no
    50   102 |     b = yes
```

7. Correlation Ranking Filter - Balanced Dataset

```
=== Summary ===

Correctly Classified Instances       85262               99.7882 %
Incorrectly Classified Instances       181                0.2118 %
Kappa statistic                          0.581
Mean absolute error                      0.0021
Root mean squared error                  0.0458
Relative absolute error                  0.4217 %
Root relative squared error              9.1591 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.998    0.171    1.000      0.998    0.999      0.609    0.914     1.000     no
                 0.829    0.002    0.448      0.829    0.582      0.609    0.914     0.532     yes
Weighted Avg.    0.998    0.171    0.999      0.998    0.998      0.609    0.914     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 85136   155 |     a = no
    26   126 |     b = yes
```

8. Correlation Ranking Filter - Unbalanced Dataset

```
=== Summary ===

Correctly Classified Instances       85195               99.7097 %
Incorrectly Classified Instances       248                0.2903 %
Kappa statistic                          0.5028
Mean absolute error                      0.0029
Root mean squared error                  0.0535
Relative absolute error                  0.5773 %
Root relative squared error             10.7095 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.997    0.171    1.000      0.997    0.999      0.547    0.914     1.000     no
                 0.829    0.003    0.362      0.829    0.504      0.547    0.914     0.567     yes
Weighted Avg.    0.997    0.171    0.999      0.997    0.998      0.547    0.914     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 85069   222 |     a = no
    26   126 |     b = yes
```

3. Naïve Bayes

1. Information Gain Ranking Filter - Balanced Dataset

```
=== Summary ===

Correctly Classified Instances       82498               96.5533 %
Incorrectly Classified Instances      2945                3.4467 %
Kappa statistic                          0.0803
Mean absolute error                      0.0426
Root mean squared error                  0.1745
Relative absolute error                  8.5306 %
Root relative squared error             34.9146 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.966    0.118    1.000      0.966    0.982      0.192    0.951     1.000     no
                 0.882    0.034    0.044      0.882    0.083      0.192    0.952     0.168     yes
Weighted Avg.    0.966    0.118    0.998      0.966    0.981      0.192    0.951     0.998

=== Confusion Matrix ===

     a     b   <-- classified as
 82364  2927 |     a = no
    18   134 |     b = yes
```

2. Information Gain Ranking Filter - Unbalanced Dataset

```
=== Summary ===

Correctly Classified Instances       83967               98.2725 %
Incorrectly Classified Instances      1476                1.7275 %
Kappa statistic                          0.147
Mean absolute error                      0.023
Root mean squared error                  0.1209
Relative absolute error                  4.6027 %
Root relative squared error             24.1806 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.983    0.145    1.000      0.983    0.991      0.262    0.965     1.000     no
                 0.855    0.017    0.082      0.855    0.150      0.262    0.965     0.575     yes
Weighted Avg.    0.983    0.145    0.998      0.983    0.990      0.262    0.965     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 83837  1454 |     a = no
    22   130 |     b = yes
```

3. OneR feature evaluator - Balanced Dataset

```
=== Summary ===

Correctly Classified Instances       83967               98.2725 %
Incorrectly Classified Instances      1476                1.7275 %
Kappa statistic                          0.147
Mean absolute error                      0.023
Root mean squared error                  0.1209
Relative absolute error                  4.6027 %
Root relative squared error             24.1806 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.983    0.145    1.000      0.983    0.991      0.262    0.965     1.000     no
                 0.855    0.017    0.082      0.855    0.150      0.262    0.965     0.575     yes
Weighted Avg.    0.983    0.145    0.998      0.983    0.990      0.262    0.965     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 83837  1454 |     a = no
    22   130 |     b = yes
```

4. OneR feature evaluator - Unbalanced Dataset

```
=== Summary ===

Correctly Classified Instances       83967               98.2725 %
Incorrectly Classified Instances      1476                1.7275 %
Kappa statistic                          0.147
Mean absolute error                      0.023
Root mean squared error                  0.1209
Relative absolute error                  4.6027 %
Root relative squared error             24.1806 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.983    0.145    1.000      0.983    0.991      0.262    0.965     1.000     no
                 0.855    0.017    0.082      0.855    0.150      0.262    0.965     0.575     yes
Weighted Avg.    0.983    0.145    0.998      0.983    0.990      0.262    0.965     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 83837  1454 |     a = no
    22   130 |     b = yes
```

5. Gain Ratio feature evaluator - Balanced Dataset

```
=== Summary ===

Correctly Classified Instances       83967               98.2725 %
Incorrectly Classified Instances      1476                1.7275 %
Kappa statistic                          0.147
Mean absolute error                      0.023
Root mean squared error                  0.1209
Relative absolute error                  4.6027 %
Root relative squared error             24.1806 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.983    0.145    1.000      0.983    0.991      0.262    0.965     1.000     no
                 0.855    0.017    0.082      0.855    0.150      0.262    0.965     0.575     yes
Weighted Avg.    0.983    0.145    0.998      0.983    0.990      0.262    0.965     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 83837  1454 |     a = no
    22   130 |     b = yes
```

6. Gain Ratio feature evaluator - Unbalanced Dataset

```
=== Summary ===

Correctly Classified Instances       82695               96.7838 %
Incorrectly Classified Instances      2748                3.2162 %
Kappa statistic                          0.076
Mean absolute error                      0.0443
Root mean squared error                  0.1672
Relative absolute error                  8.8684 %
Root relative squared error             33.4456 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.968    0.224    1.000      0.968    0.984      0.175    0.943     1.000     no
                 0.776    0.032    0.042      0.776    0.079      0.175    0.943     0.086     yes
Weighted Avg.    0.968    0.223    0.998      0.968    0.982      0.175    0.943     0.998

=== Confusion Matrix ===

     a     b   <-- classified as
 82577  2714 |     a = no
    34   118 |     b = yes
```

7. Correlation Ranking Filter - Balanced Dataset

```
=== Summary ===

Correctly Classified Instances       83756               98.0256 %
Incorrectly Classified Instances      1687                1.9744 %
Kappa statistic                          0.1316
Mean absolute error                      0.0248
Root mean squared error                  0.1301
Relative absolute error                  4.9538 %
Root relative squared error             26.0168 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.980    0.138    1.000      0.980    0.990      0.247    0.969     1.000     no
                 0.862    0.020    0.073      0.862    0.134      0.247    0.969     0.341     yes
Weighted Avg.    0.980    0.138    0.998      0.980    0.988      0.247    0.969     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 83625  1666 |     a = no
    21   131 |     b = yes
```

8. Correlation Ranking Filter - Unbalanced Dataset

```
=== Summary ===

Correctly Classified Instances       84043               98.3615 %
Incorrectly Classified Instances      1400                1.6385 %
Kappa statistic                          0.1549
Mean absolute error                      0.024
Root mean squared error                  0.1195
Relative absolute error                  4.793  %
Root relative squared error             23.9104 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.984    0.138    1.000      0.984    0.992      0.270    0.977     1.000     no
                 0.862    0.016    0.087      0.862    0.158      0.270    0.977     0.515     yes
Weighted Avg.    0.984    0.138    0.998      0.984    0.990      0.270    0.977     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 83912  1379 |     a = no
    21   131 |     b = yes
```

4. Multilayer Perceptron

1. Information Gain Ranking Filter - Balanced Dataset

```
=== Summary ===

Correctly Classified Instances       82302               96.3239 %
Incorrectly Classified Instances      3141                3.6761 %
Kappa statistic                          0.0755
Mean absolute error                      0.0714
Root mean squared error                  0.1473
Relative absolute error                 14.2744 %
Root relative squared error             29.4605 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.963    0.118    1.000      0.963    0.981      0.186    0.949     1.000     no
                 0.882    0.037    0.041      0.882    0.079      0.186    0.949     0.687     yes
Weighted Avg.    0.963    0.118    0.998      0.963    0.980      0.186    0.949     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 82168  3123 |     a = no
    18   134 |     b = yes
```

2. Information Gain Ranking Filter - Unbalanced Dataset

```
=== Summary ===

Correctly Classified Instances       83087               97.2426 %
Incorrectly Classified Instances      2356                2.7574 %
Kappa statistic                          0.1005
Mean absolute error                      0.0613
Root mean squared error                  0.1563
Relative absolute error                 12.2554 %
Root relative squared error             31.2695 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.973    0.105    1.000      0.973    0.986      0.218    0.943     1.000     no
                 0.895    0.027    0.055      0.895    0.104      0.218    0.943     0.348     yes
Weighted Avg.    0.972    0.105    0.998      0.972    0.984      0.218    0.943     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 82951  2340 |     a = no
    16   136 |     b = yes
```

3. OneR feature evaluator - Balanced Dataset

```
=== Summary ===

Correctly Classified Instances       83087               97.2426 %
Incorrectly Classified Instances      2356                2.7574 %
Kappa statistic                          0.1005
Mean absolute error                      0.0613
Root mean squared error                  0.1563
Relative absolute error                 12.2554 %
Root relative squared error             31.2695 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.973    0.105    1.000      0.973    0.986      0.218    0.943     1.000     no
                 0.895    0.027    0.055      0.895    0.104      0.218    0.943     0.348     yes
Weighted Avg.    0.972    0.105    0.998      0.972    0.984      0.218    0.943     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 82951  2340 |     a = no
    16   136 |     b = yes
```

4. OneR feature evaluator - Unbalanced Dataset

```
=== Summary ===

Correctly Classified Instances       83087               97.2426 %
Incorrectly Classified Instances      2356                2.7574 %
Kappa statistic                          0.1005
Mean absolute error                      0.0613
Root mean squared error                  0.1563
Relative absolute error                 12.2554 %
Root relative squared error             31.2695 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.973    0.105    1.000      0.973    0.986      0.218    0.943     1.000     no
                 0.895    0.027    0.055      0.895    0.104      0.218    0.943     0.348     yes
Weighted Avg.    0.972    0.105    0.998      0.972    0.984      0.218    0.943     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 82951  2340 |     a = no
    16   136 |     b = yes
```

5. Gain Ratio feature evaluator - Balanced Dataset

```
=== Summary ===

Correctly Classified Instances       83087               97.2426 %
Incorrectly Classified Instances      2356                2.7574 %
Kappa statistic                          0.1005
Mean absolute error                      0.0613
Root mean squared error                  0.1563
Relative absolute error                 12.2554 %
Root relative squared error             31.2695 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.973    0.105    1.000      0.973    0.986      0.218    0.943     1.000     no
                 0.895    0.027    0.055      0.895    0.104      0.218    0.943     0.348     yes
Weighted Avg.    0.972    0.105    0.998      0.972    0.984      0.218    0.943     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 82951  2340 |     a = no
    16   136 |     b = yes
```

6. Gain Ratio feature evaluator - Unbalanced Dataset

```
=== Summary ===

Correctly Classified Instances       81877               95.8265 %
Incorrectly Classified Instances      3566                4.1735 %
Kappa statistic                          0.0677
Mean absolute error                      0.1544
Root mean squared error                  0.2181
Relative absolute error                 30.883  %
Root relative squared error             43.6374 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.958    0.105    1.000      0.958    0.979      0.177    0.969     1.000     no
                 0.895    0.042    0.037      0.895    0.071      0.177    0.969     0.583     yes
Weighted Avg.    0.958    0.105    0.998      0.958    0.977      0.177    0.969     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 81741  3550 |     a = no
    16   136 |     b = yes
```

7. Correlation Ranking Filter - Balanced Dataset

```
=== Summary ===

Correctly Classified Instances       81948               95.9096 %
Incorrectly Classified Instances      3495                4.0904 %
Kappa statistic                          0.0636
Mean absolute error                      0.0878
Root mean squared error                  0.188
Relative absolute error                 17.5696 %
Root relative squared error             37.6194 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.959    0.178    1.000      0.959    0.979      0.164    0.920     1.000     no
                 0.822    0.041    0.035      0.822    0.067      0.164    0.921     0.099     yes
Weighted Avg.    0.959    0.177    0.998      0.959    0.977      0.164    0.920     0.998

=== Confusion Matrix ===

     a     b   <-- classified as
 81823  3468 |     a = no
    27   125 |     b = yes
```

8. Correlation Ranking Filter - Unbalanced Dataset

```
=== Summary ===

Correctly Classified Instances       82324               96.3496 %
Incorrectly Classified Instances      3119                3.6504 %
Kappa statistic                          0.0749
Mean absolute error                      0.0655
Root mean squared error                  0.1476
Relative absolute error                 13.0985 %
Root relative squared error             29.5242 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.964    0.132    1.000      0.964    0.981      0.184    0.971     1.000     no
                 0.868    0.036    0.041      0.868    0.078      0.184    0.971     0.687     yes
Weighted Avg.    0.963    0.131    0.998      0.963    0.980      0.184    0.971     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 82192  3099 |     a = no
    20   132 |     b = yes
```

5. Simple Logistic

1. Information Gain Ranking Filter - Balanced Dataset

```
=== Summary ===

Correctly Classified Instances       83210               97.3866 %
Incorrectly Classified Instances      2233                2.6134 %
Kappa statistic                          0.1042
Mean absolute error                      0.1276
Root mean squared error                  0.1812
Relative absolute error                 25.5279 %
Root relative squared error             36.2419 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.974    0.118    1.000      0.974    0.987      0.221    0.954     1.000     no
                 0.882    0.026    0.057      0.882    0.107      0.221    0.954     0.679     yes
Weighted Avg.    0.974    0.118    0.998      0.974    0.985      0.221    0.954     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 83076  2215 |     a = no
    18   134 |     b = yes
```

2. Information Gain Ranking Filter - Unbalanced Dataset

```
=== Summary ===

Correctly Classified Instances       80949               94.7404 %
Incorrectly Classified Instances      4494                5.2596 %
Kappa statistic                          0.0522
Mean absolute error                      0.1581
Root mean squared error                  0.2234
Relative absolute error                 31.6296 %
Root relative squared error             44.69   %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.948    0.132    1.000      0.948    0.973      0.152    0.958     1.000     no
                 0.868    0.052    0.029      0.868    0.055      0.152    0.958     0.617     yes
Weighted Avg.    0.947    0.131    0.998      0.947    0.971      0.152    0.958     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 80817  4474 |     a = no
    20   132 |     b = yes
```

3. OneR feature evaluator - Balanced Dataset

```
=== Summary ===

Correctly Classified Instances       80949               94.7404 %
Incorrectly Classified Instances      4494                5.2596 %
Kappa statistic                          0.0522
Mean absolute error                      0.1581
Root mean squared error                  0.2234
Relative absolute error                 31.6296 %
Root relative squared error             44.69   %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.948    0.132    1.000      0.948    0.973      0.152    0.958     1.000     no
                 0.868    0.052    0.029      0.868    0.055      0.152    0.958     0.617     yes
Weighted Avg.    0.947    0.131    0.998      0.947    0.971      0.152    0.958     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 80817  4474 |     a = no
    20   132 |     b = yes
```

4. OneR feature evaluator - Unbalanced Dataset

```
=== Summary ===

Correctly Classified Instances       80949               94.7404 %
Incorrectly Classified Instances      4494                5.2596 %
Kappa statistic                          0.0522
Mean absolute error                      0.1581
Root mean squared error                  0.2234
Relative absolute error                 31.6296 %
Root relative squared error             44.69   %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.948    0.132    1.000      0.948    0.973      0.152    0.958     1.000     no
                 0.868    0.052    0.029      0.868    0.055      0.152    0.958     0.617     yes
Weighted Avg.    0.947    0.131    0.998      0.947    0.971      0.152    0.958     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 80817  4474 |     a = no
    20   132 |     b = yes
```

5. Gain Ratio feature evaluator - Balanced Dataset

```
=== Summary ===

Correctly Classified Instances       80949               94.7404 %
Incorrectly Classified Instances      4494                5.2596 %
Kappa statistic                          0.0522
Mean absolute error                      0.1581
Root mean squared error                  0.2234
Relative absolute error                 31.6296 %
Root relative squared error             44.69   %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.948    0.132    1.000      0.948    0.973      0.152    0.958     1.000     no
                 0.868    0.052    0.029      0.868    0.055      0.152    0.958     0.617     yes
Weighted Avg.    0.947    0.131    0.998      0.947    0.971      0.152    0.958     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 80817  4474 |     a = no
    20   132 |     b = yes
```

6. Gain Ratio feature evaluator - Unbalanced Dataset

```
=== Summary ===

Correctly Classified Instances       83415               97.6265 %
Incorrectly Classified Instances      2028                2.3735 %
Kappa statistic                          0.0997
Mean absolute error                      0.2915
Root mean squared error                  0.3083
Relative absolute error                 58.309  %
Root relative squared error             61.6684 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.977    0.237    1.000      0.977    0.988      0.201    0.929     1.000     no
                 0.763    0.023    0.055      0.763    0.103      0.201    0.929     0.497     yes
Weighted Avg.    0.976    0.236    0.998      0.976    0.986      0.201    0.929     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 83299  1992 |     a = no
    36   116 |     b = yes
```

7. Correlation Ranking Filter - Balanced Dataset

```
=== Summary ===

Correctly Classified Instances       81948               95.9096 %
Incorrectly Classified Instances      3495                4.0904 %
Kappa statistic                          0.0636
Mean absolute error                      0.0878
Root mean squared error                  0.188
Relative absolute error                 17.5696 %
Root relative squared error             37.6194 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.959    0.178    1.000      0.959    0.979      0.164    0.920     1.000     no
                 0.822    0.041    0.035      0.822    0.067      0.164    0.921     0.099     yes
Weighted Avg.    0.959    0.177    0.998      0.959    0.977      0.164    0.920     0.998

=== Confusion Matrix ===

     a     b   <-- classified as
 81823  3468 |     a = no
    27   125 |     b = yes
```

8. Correlation Ranking Filter - Unbalanced Dataset

```
=== Summary ===

Correctly Classified Instances       81868               95.8159 %
Incorrectly Classified Instances      3575                4.1841 %
Kappa statistic                          0.068
Mean absolute error                      0.1242
Root mean squared error                  0.1967
Relative absolute error                 24.8437 %
Root relative squared error             39.3482 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.958    0.099    1.000      0.958    0.979      0.178    0.978     1.000     no
                 0.901    0.042    0.037      0.901    0.071      0.178    0.978     0.665     yes
Weighted Avg.    0.958    0.099    0.998      0.958    0.977      0.178    0.978     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 81731  3560 |     a = no
    15   137 |     b = yes
```

2nd Approach example

Tried a few models but no better results so restarted again.

One of the example:

TRAIN : Approach 1 balanced Dataset (100%) + TEST: Unbalanced Dataset (30%)

J48 + Info Gain Attribute Selection

```
=== Summary ===

Correctly Classified Instances       84860               99.3177 %
Incorrectly Classified Instances       583                0.6823 %
Kappa statistic                          0.3566
Mean absolute error                      0.0068
Root mean squared error                  0.0818
Relative absolute error                  1.3603 %
Root relative squared error             16.4479 %
Total Number of Instances            85443

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.993    0.000    1.000      0.993    0.997      0.466    0.998     1.000     no
                 1.000    0.007    0.218      1.000    0.359      0.466    0.998     0.385     yes
Weighted Avg.    0.993    0.000    0.999      0.993    0.995      0.466    0.998     0.999

=== Confusion Matrix ===

     a     b   <-- classified as
 84697   583 |     a = no
     0   163 |     b = yes
```

Conclusion

After trying multiple approaches and starting the project from scratch, classifier accuracy still gives approx 99% for most, error need to be rectify. Possible reason of error coud be:

- **Sampling error:** In R on trying sampling only hybrid sampling it able to balnace data, no other sampling or strata sampling in JMP also not working. Possible some different language needs to be used for sampling .
- **Wrong Methodology:** While building the classifier some wrong procedure is being followed.
- One more **point to be considered** is that the attribute values are PCA components, numeric values which are leading to a perfect model. If we had known which attribute signifies what we could have made better choices, e.g. If we know its Address1/ Mobile number's PCA attribute then we could have skip that as that may not help in building good model.
- Option could be to try all possible sets to get the best model which will require a lot time frame as more than 100 model could be possible considering set of 6 attributes and 4 classifier algorithms.