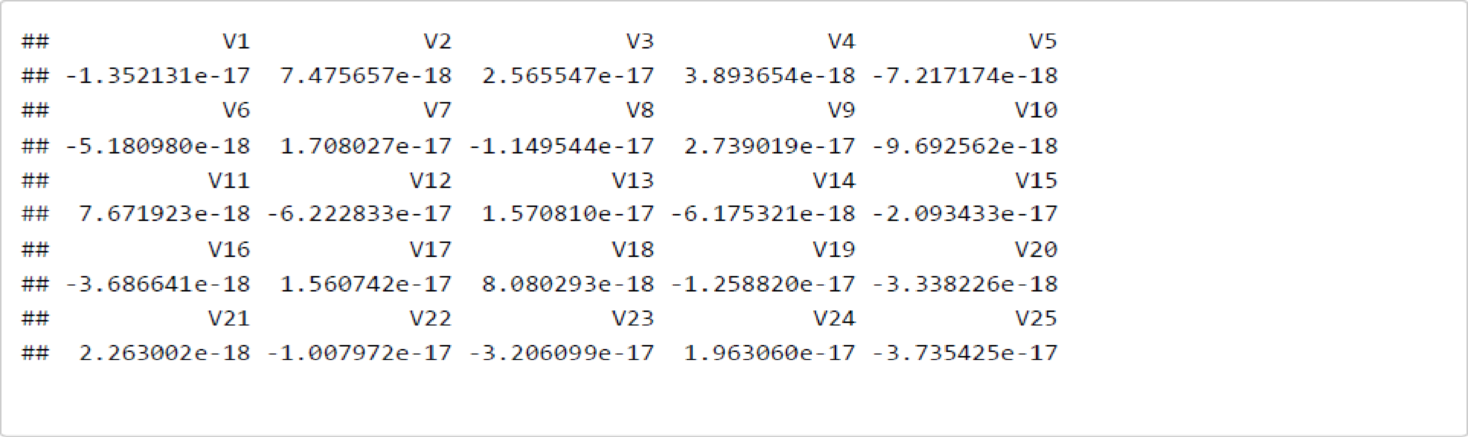
Spam Classification

Abirami Sabbani
3/9/2022

Designing a classifier with smallest error rate using different methods with various preprocessed data.

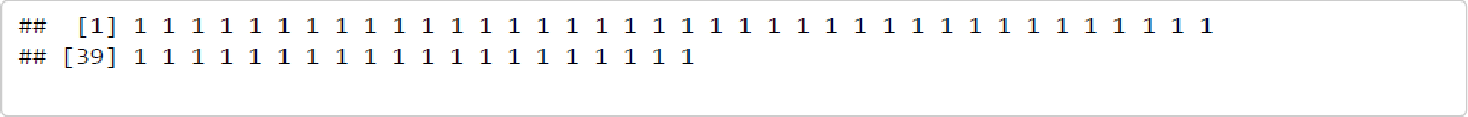
1. Standardized the columns so that they all have zero mean and unit variance.

Sample columns of standardized train csv.



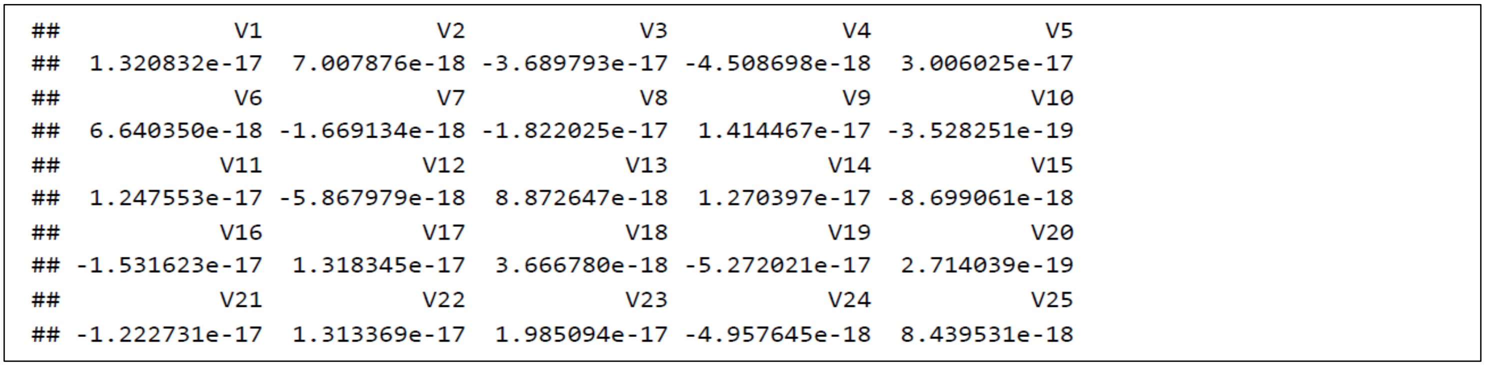
```
##            V1            V2            V3            V4            V5
## -1.352131e-17  7.475657e-18  2.565547e-17  3.893654e-18 -7.217174e-18
##            V6            V7            V8            V9           V10
## -5.180980e-18  1.708027e-17 -1.149544e-17  2.739019e-17 -9.692562e-18
##           V11           V12           V13           V14           V15
##  7.671923e-18 -6.222833e-17  1.570810e-17 -6.175321e-18 -2.093433e-17
##           V16           V17           V18           V19           V20
## -3.686641e-18  1.560742e-17  8.080293e-18 -1.258820e-17 -3.338226e-18
##           V21           V22           V23           V24           V25
##  2.263002e-18 -1.007972e-17 -3.206099e-17  1.963060e-17 -3.735425e-17
```

Unit variance for standardized train.



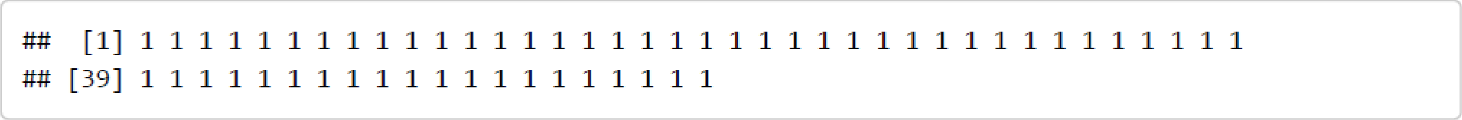
```
##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [39] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

Sample columns of standardized test csv.



```
##            V1            V2            V3            V4            V5
##  1.320832e-17  7.007876e-18 -3.689793e-17 -4.508698e-18  3.006025e-17
##            V6            V7            V8            V9           V10
##  6.640350e-18 -1.669134e-18 -1.822025e-17  1.414467e-17 -3.528251e-19
##           V11           V12           V13           V14           V15
##  1.247553e-17 -5.867979e-18  8.872647e-18  1.270397e-17 -8.699061e-18
##           V16           V17           V18           V19           V20
## -1.531623e-17  1.318345e-17  3.666780e-18 -5.272021e-17  2.714039e-19
##           V21           V22           V23           V24           V25
## -1.222731e-17  1.313369e-17  1.985094e-17 -4.957645e-18  8.439531e-18
```

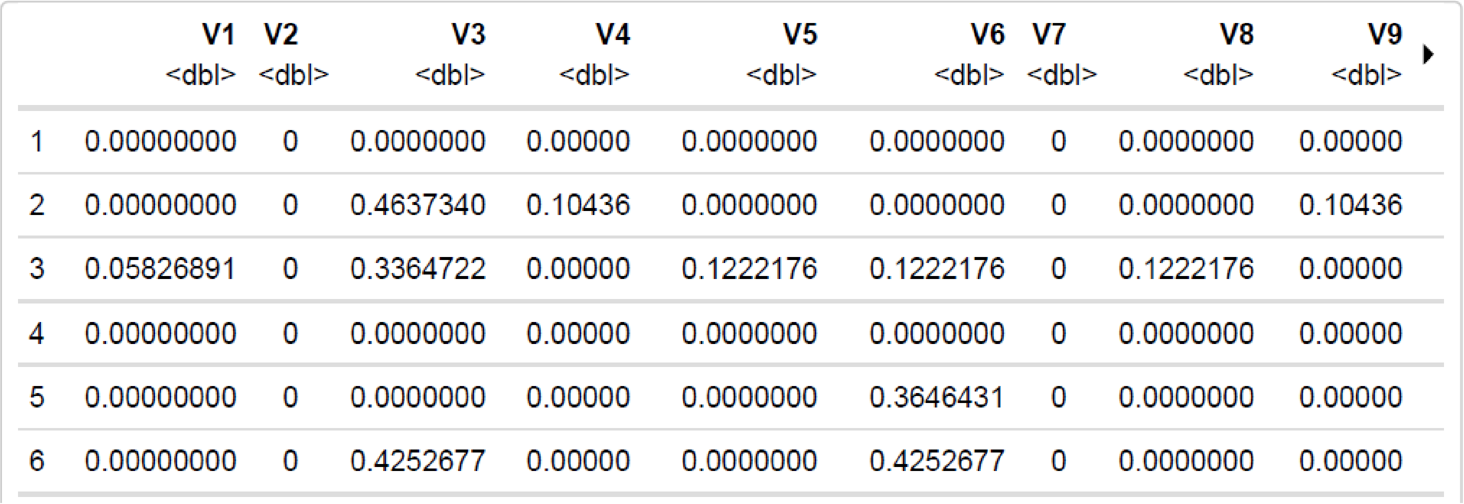
Unit variance for standardized test.



```
##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [39] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

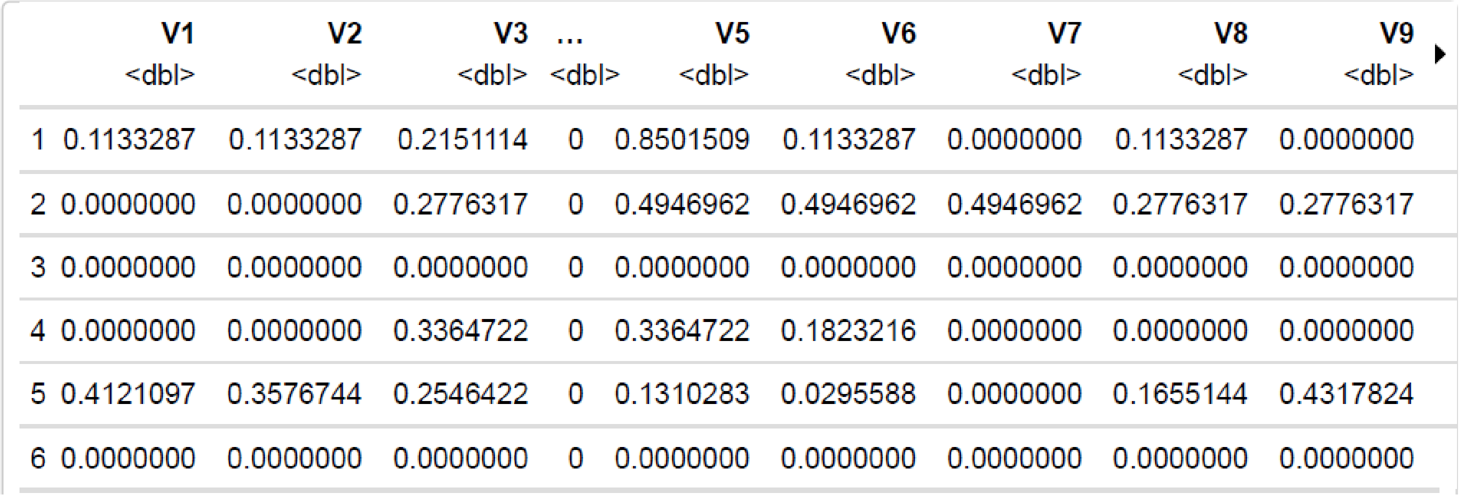
2. Transforming the features using log.

Sample columns of log train csv.



	V1	V2	V3	V4	V5	V6	V7	V8	V9
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	0.00000000	0	0.0000000	0.00000	0.0000000	0.0000000	0	0.0000000	0.00000
2	0.00000000	0	0.4637340	0.10436	0.0000000	0.0000000	0	0.0000000	0.10436
3	0.05826891	0	0.3364722	0.00000	0.1222176	0.1222176	0	0.1222176	0.00000
4	0.00000000	0	0.0000000	0.00000	0.0000000	0.0000000	0	0.0000000	0.00000
5	0.00000000	0	0.0000000	0.00000	0.0000000	0.3646431	0	0.0000000	0.00000
6	0.00000000	0	0.4252677	0.00000	0.0000000	0.4252677	0	0.0000000	0.00000

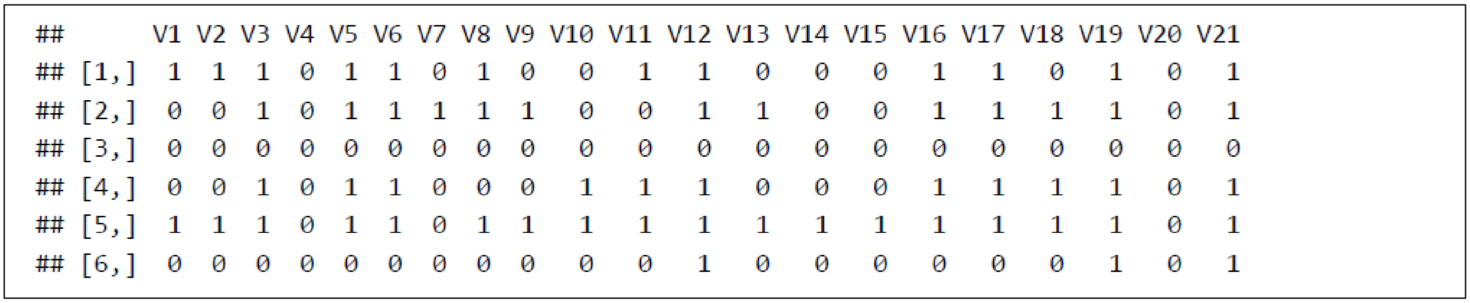
Sample columns of log test csv.



	V1	V2	V3	…	V5	V6	V7	V8	V9
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	0.1133287	0.1133287	0.2151114	0	0.8501509	0.1133287	0.0000000	0.1133287	0.0000000
2	0.0000000	0.0000000	0.2776317	0	0.4946962	0.4946962	0.4946962	0.2776317	0.2776317
3	0.0000000	0.0000000	0.0000000	0	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
4	0.0000000	0.0000000	0.3364722	0	0.3364722	0.1823216	0.0000000	0.0000000	0.0000000
5	0.4121097	0.3576744	0.2546422	0	0.1310283	0.0295588	0.0000000	0.1655144	0.4317824
6	0.0000000	0.0000000	0.0000000	0	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000

3. Discretizing each feature.

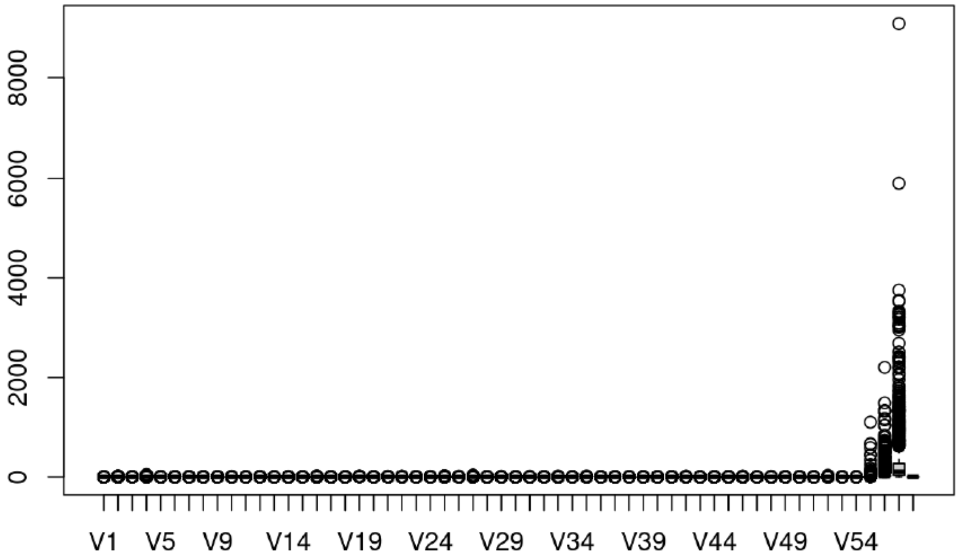
Sample of discretized test csv columns.



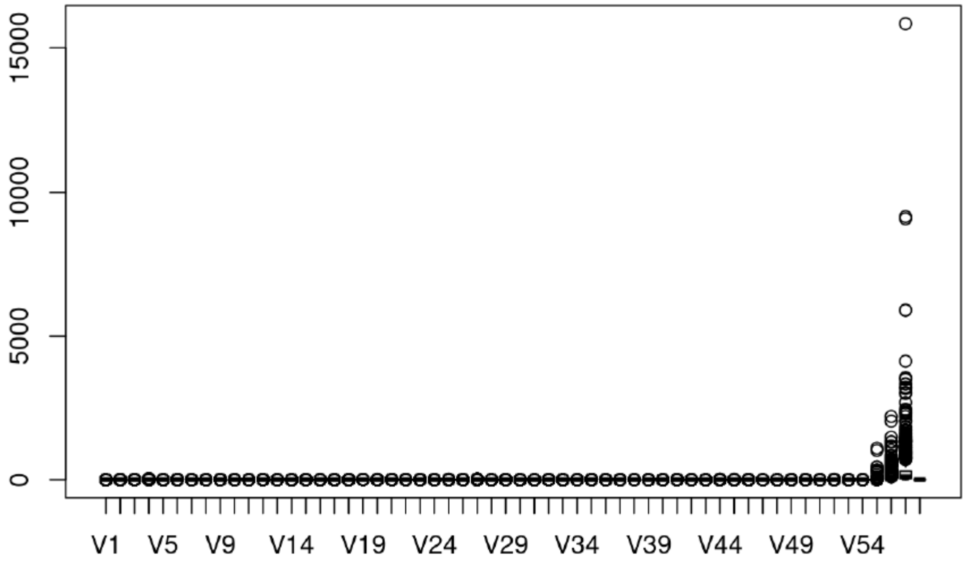
```
##      V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20 V21
## [1,]  1  1  1  0  1  1  0  1  0   0   1   1   0   0   0   1   1   0   1   0   1
## [2,]  0  0  1  0  1  1  1  1  1   0   0   1   1   0   0   1   1   1   1   0   1
## [3,]  0  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0   0   0   0
## [4,]  0  0  1  0  1  1  0  0  0   1   1   1   0   0   0   1   1   1   1   0   1
## [5,]  1  1  1  0  1  1  0  1  1   1   1   1   1   1   1   1   1   1   1   0   1
## [6,]  0  0  0  0  0  0  0  0  0   0   0   1   0   0   0   0   0   0   1   0   1
```

Visualization for original train and test data

```
boxplot(train)
```

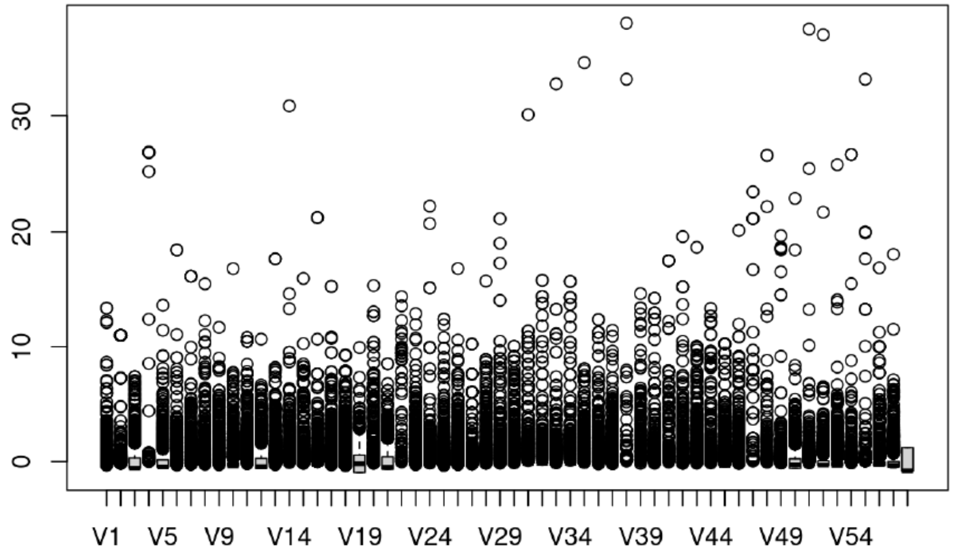


```
boxplot(test)
```

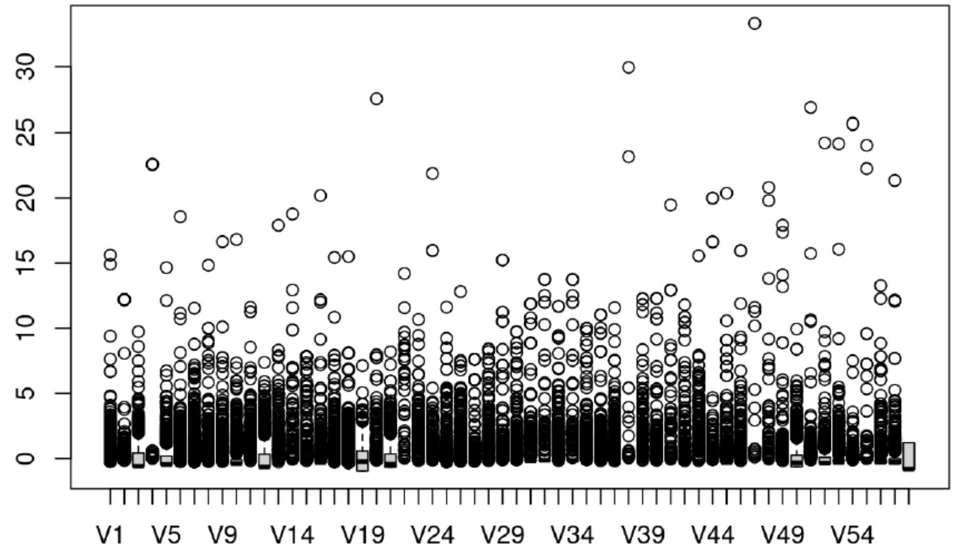


Visualization for standardized train and test data

```
boxplot(stan_train)
```

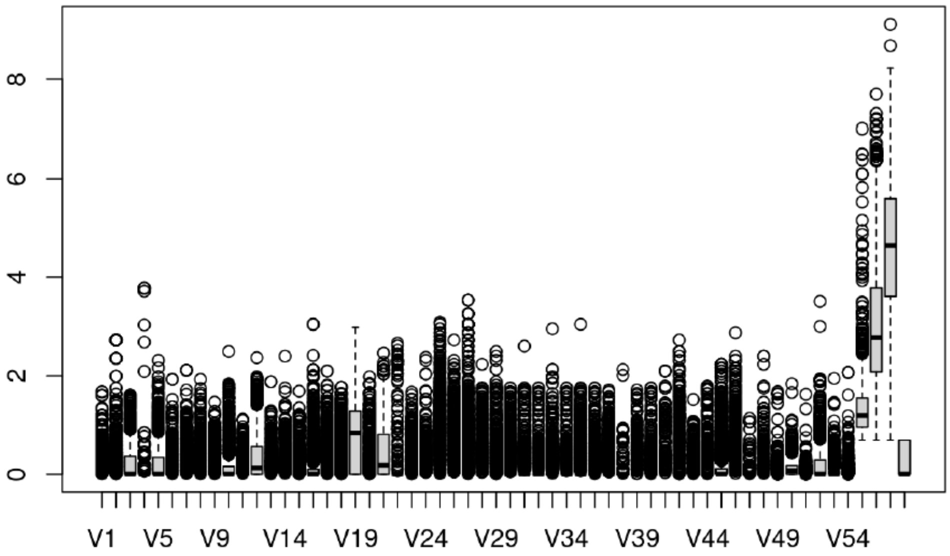


```
boxplot(stan_test)
```

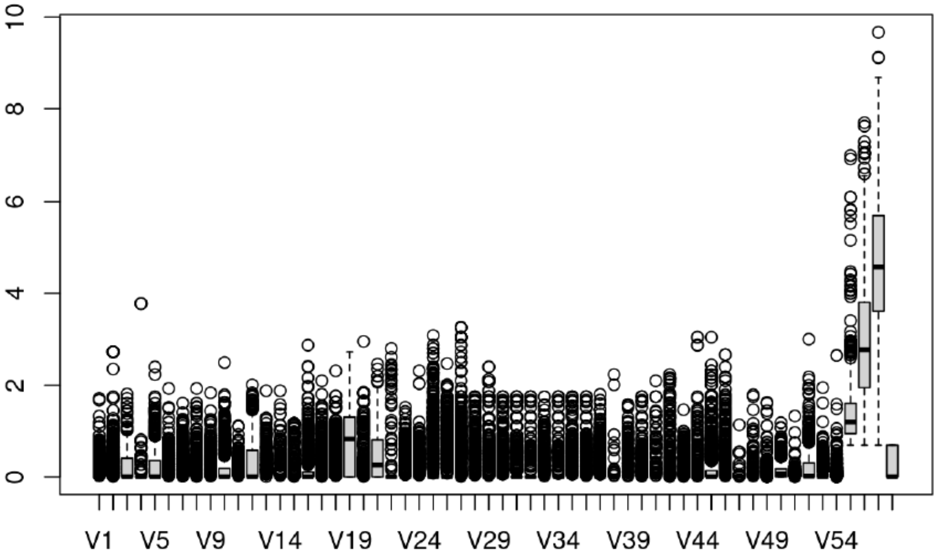


Visualization for log transformed train and test data

```
boxplot(log_train)
```

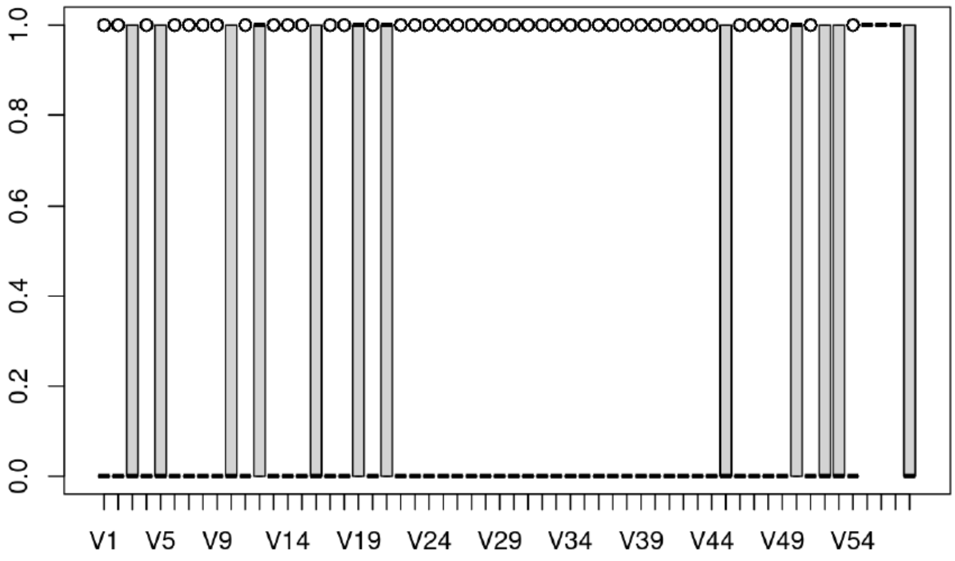


```
boxplot(log_test)
```

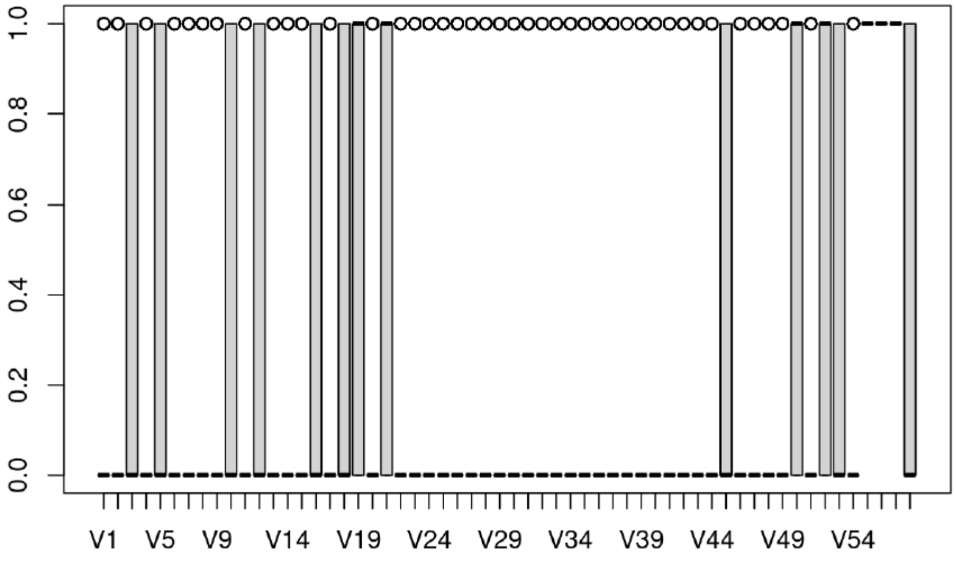


Visualization for discretized train and test data

```
boxplot(I_train)
```

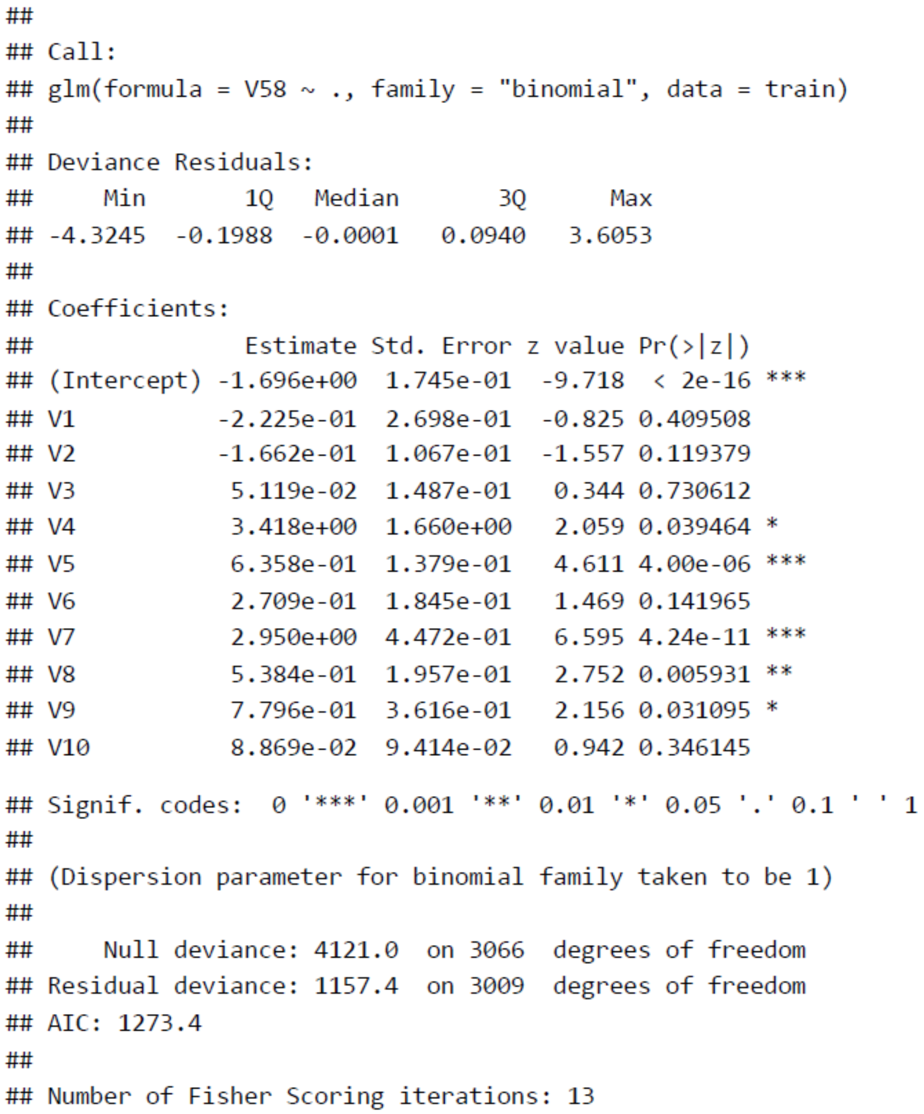


```
boxplot(I_test)
```



Since the train and test datasets have a different amount of data, the scale is different but the ratios are about the same. Also the log transformation feature shows a high variance for features 56 and 57, but it is not as noticeable when the feature is standardized.

4. Linear Regression on Original train and test data.

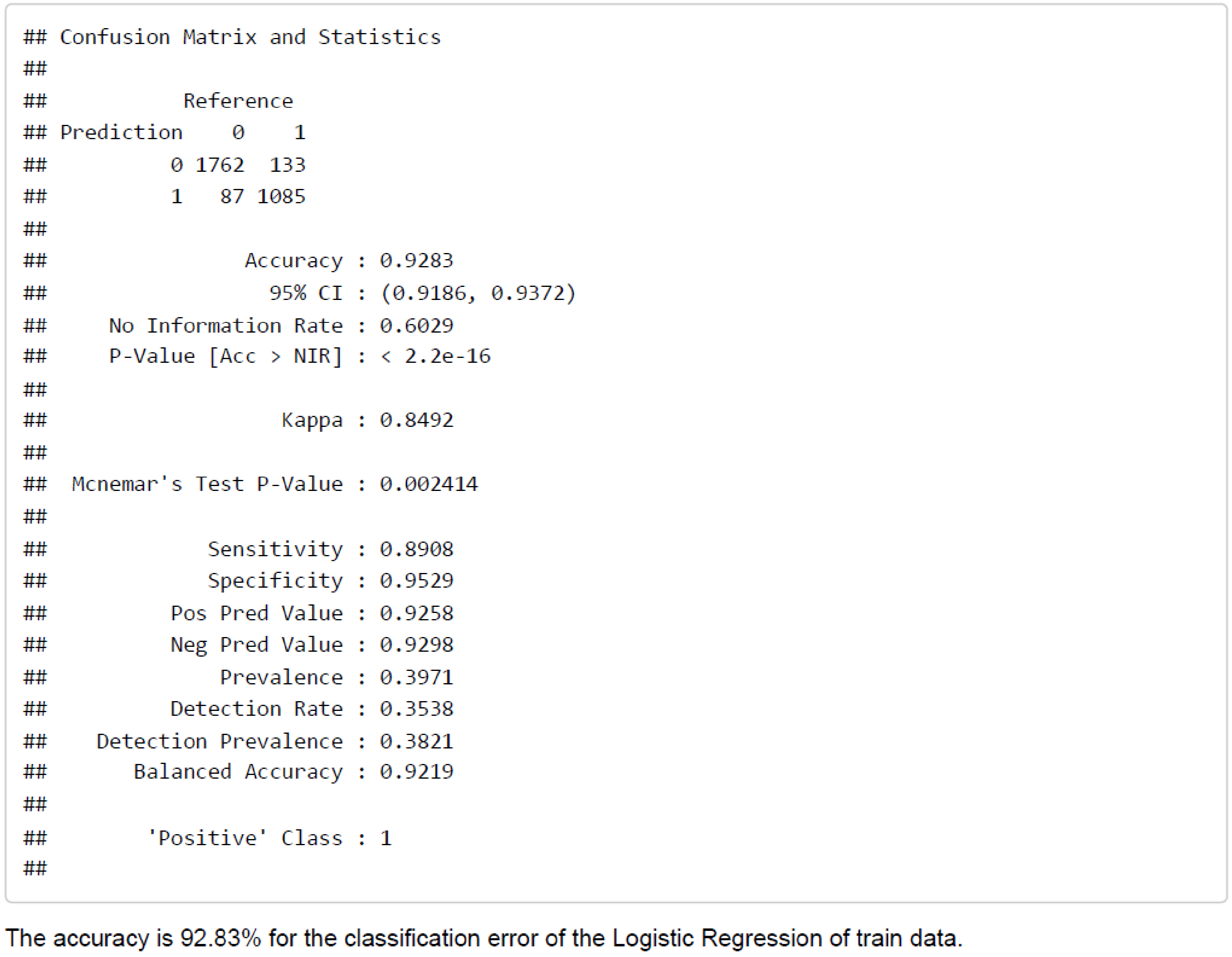


```
##
## Call:
## glm(formula = V58 ~ ., family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.3245  -0.1988  -0.0001   0.0940   3.6053
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.696e+00  1.745e-01  -9.718  < 2e-16 ***
## V1          -2.225e-01  2.698e-01  -0.825 0.409508
## V2          -1.662e-01  1.067e-01  -1.557 0.119379
## V3           5.119e-02  1.487e-01   0.344 0.730612
## V4           3.418e+00  1.660e+00   2.059 0.039464 *
## V5           6.358e-01  1.379e-01   4.611 4.00e-06 ***
## V6           2.709e-01  1.845e-01   1.469 0.141965
## V7           2.950e+00  4.472e-01   6.595 4.24e-11 ***
## V8           5.384e-01  1.957e-01   2.752 0.005931 **
## V9           7.796e-01  3.616e-01   2.156 0.031095 *
## V10          8.869e-02  9.414e-02   0.942 0.346145

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4121.0  on 3066  degrees of freedom
## Residual deviance: 1157.4  on 3009  degrees of freedom
## AIC: 1273.4
##
## Number of Fisher Scoring iterations: 13
```

From the summary, the result indicates that features: 4,5, 7, 8, 9, 11, 16, 17, 19, 20, 21, 23, 25, 27, 42, 44, 45, 46, 47, 49, 52, 52, 55, 56, and 57 are statistically significant because their p-values are less then 0.05

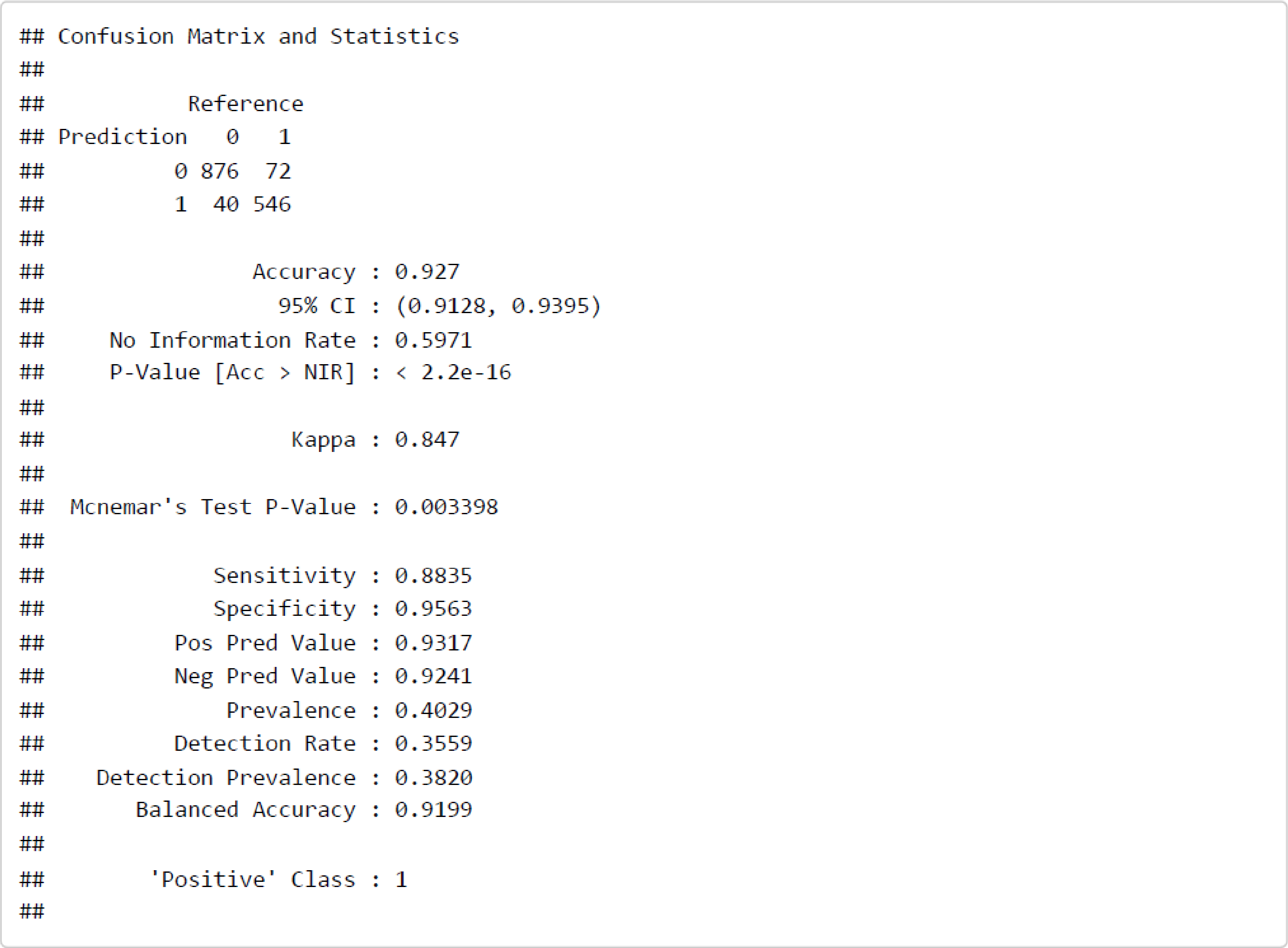
Confusion Matrix for Logistic Regression Original train data



```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1762  133
##          1   87 1085
##
##                Accuracy : 0.9283
##                  95% CI : (0.9186, 0.9372)
##     No Information Rate : 0.6029
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8492
##
##  Mcnemar's Test P-Value : 0.002414
##
##             Sensitivity : 0.8908
##             Specificity : 0.9529
##          Pos Pred Value : 0.9258
##          Neg Pred Value : 0.9298
##              Prevalence : 0.3971
##          Detection Rate : 0.3538
##    Detection Prevalence : 0.3821
##       Balanced Accuracy : 0.9219
##
##        'Positive' Class : 1
##
```

The accuracy is 92.83% for the classification error of the Logistic Regression of train data.

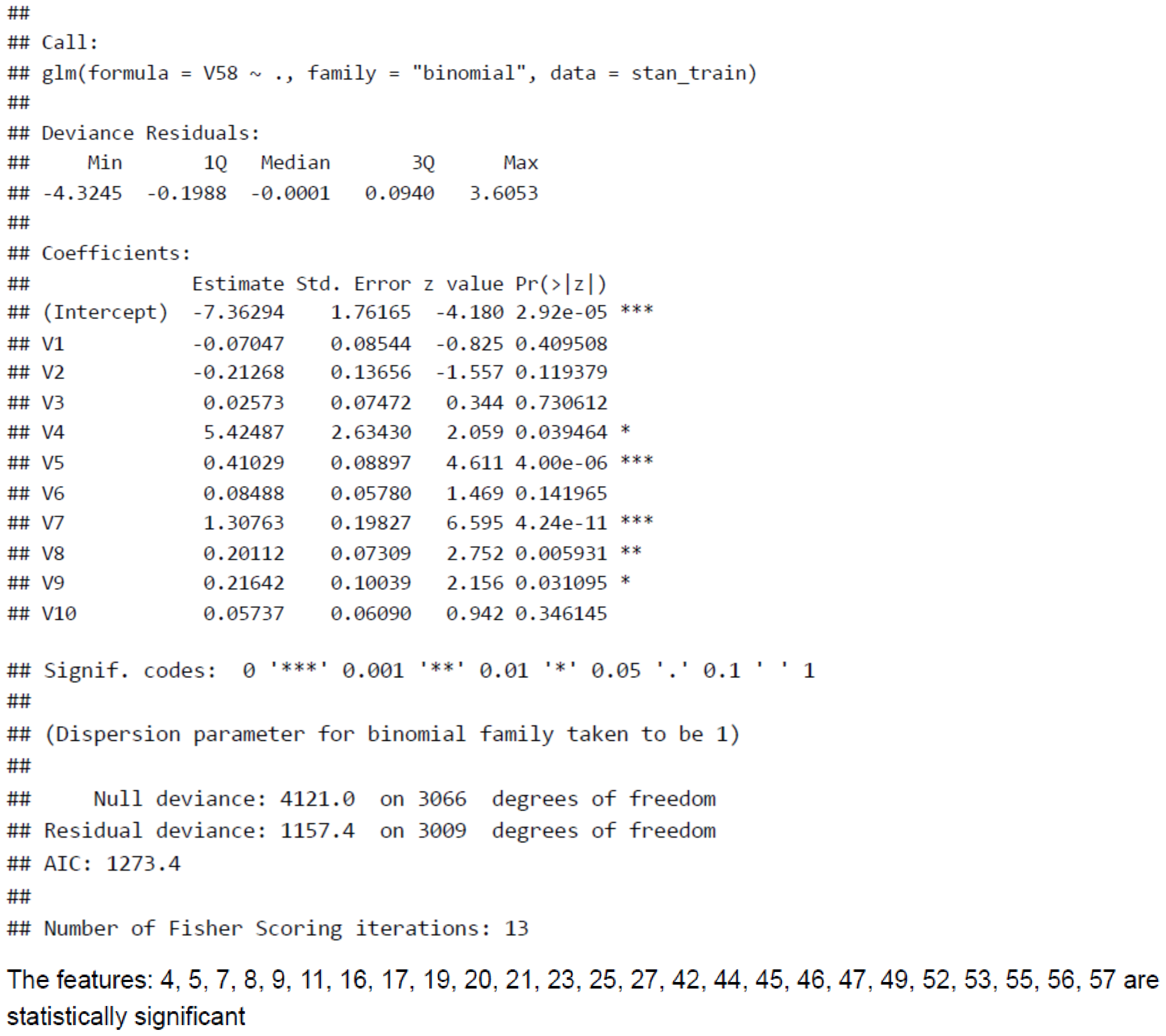
Confusion Matrix for Logistic Regression Original test data



```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 876  72
##          1  40 546
##
##                Accuracy : 0.927
##                  95% CI : (0.9128, 0.9395)
##     No Information Rate : 0.5971
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.847
##
##  Mcnemar's Test P-Value : 0.003398
##
##             Sensitivity : 0.8835
##             Specificity : 0.9563
##          Pos Pred Value : 0.9317
##          Neg Pred Value : 0.9241
##              Prevalence : 0.4029
##          Detection Rate : 0.3559
##    Detection Prevalence : 0.3820
##       Balanced Accuracy : 0.9199
##
##        'Positive' Class : 1
##
```

The accuracy is 92.7% for the classification error of the Logistic Regression of test data.

Linear Regression on Standardized Train and Test Data

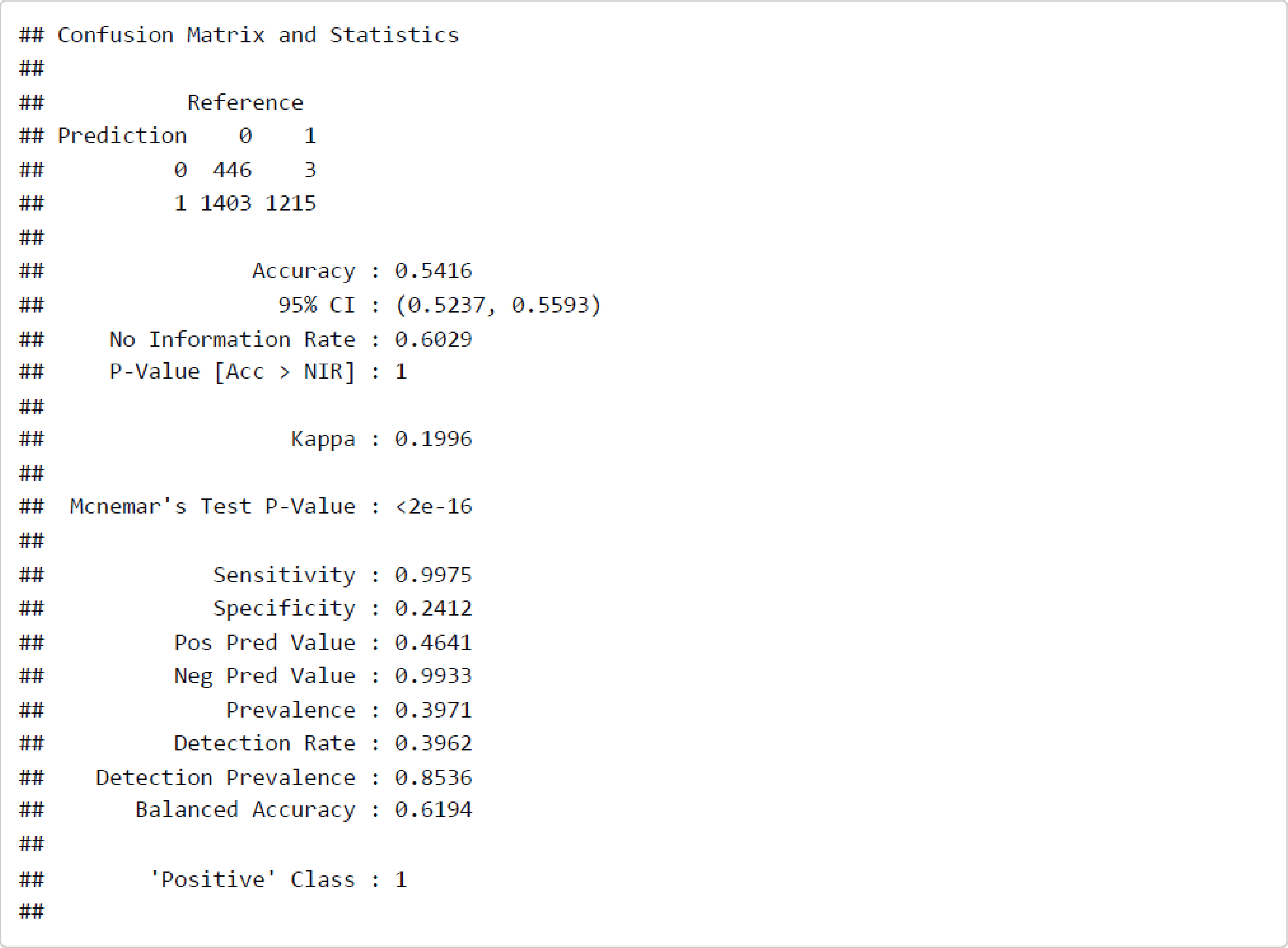


```
##
## Call:
## glm(formula = V58 ~ ., family = "binomial", data = stan_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.3245  -0.1988  -0.0001   0.0940   3.6053
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -7.36294    1.76165  -4.180 2.92e-05 ***
## V1           -0.07047    0.08544  -0.825 0.409508
## V2           -0.21268    0.13656  -1.557 0.119379
## V3            0.02573    0.07472   0.344 0.730612
## V4            5.42487    2.63430   2.059 0.039464 *
## V5            0.41029    0.08897   4.611 4.00e-06 ***
## V6            0.08488    0.05780   1.469 0.141965
## V7            1.30763    0.19827   6.595 4.24e-11 ***
## V8            0.20112    0.07309   2.752 0.005931 **
## V9            0.21642    0.10039   2.156 0.031095 *
## V10           0.05737    0.06090   0.942 0.346145

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4121.0  on 3066  degrees of freedom
## Residual deviance: 1157.4  on 3009  degrees of freedom
## AIC: 1273.4
##
## Number of Fisher Scoring iterations: 13
```

The features: 4, 5, 7, 8, 9, 11, 16, 17, 19, 20, 21, 23, 25, 27, 42, 44, 45, 46, 47, 49, 52, 53, 55, 56, 57 are statistically significant

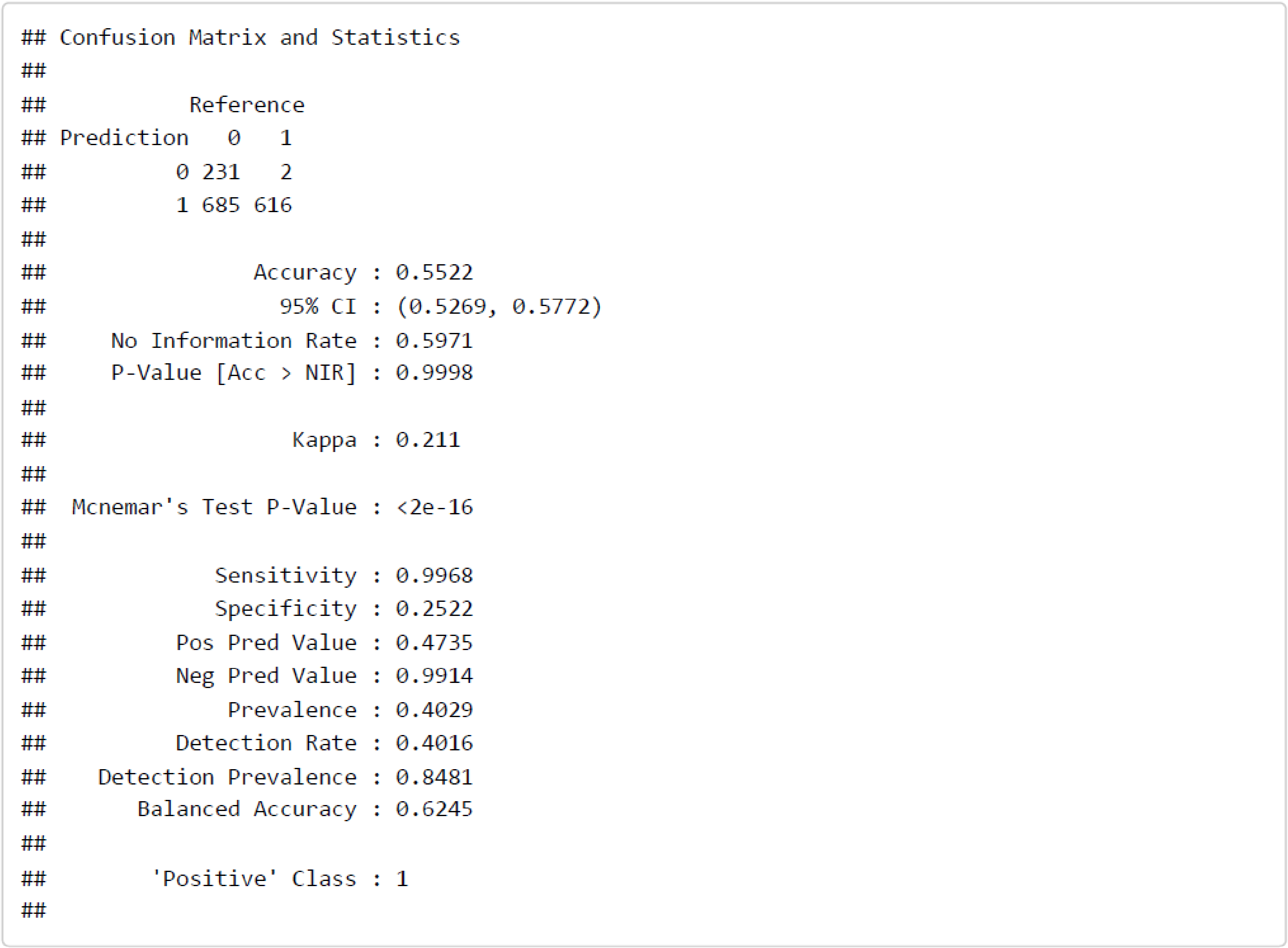
Confusion Matrix for Logistic Regression Standardized train data



```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction    0    1
##          0  446    3
##          1 1403 1215
## 
##                Accuracy : 0.5416
##                  95% CI : (0.5237, 0.5593)
##     No Information Rate : 0.6029
##     P-Value [Acc > NIR] : 1
## 
##                   Kappa : 0.1996
## 
##  Mcnemar's Test P-Value : <2e-16
## 
##             Sensitivity : 0.9975
##             Specificity : 0.2412
##          Pos Pred Value : 0.4641
##          Neg Pred Value : 0.9933
##              Prevalence : 0.3971
##          Detection Rate : 0.3962
##    Detection Prevalence : 0.8536
##       Balanced Accuracy : 0.6194
## 
##        'Positive' Class : 1
## 
```

The accuracy is 54.16% for the classification error of the Logistic Regression of standardized train data.

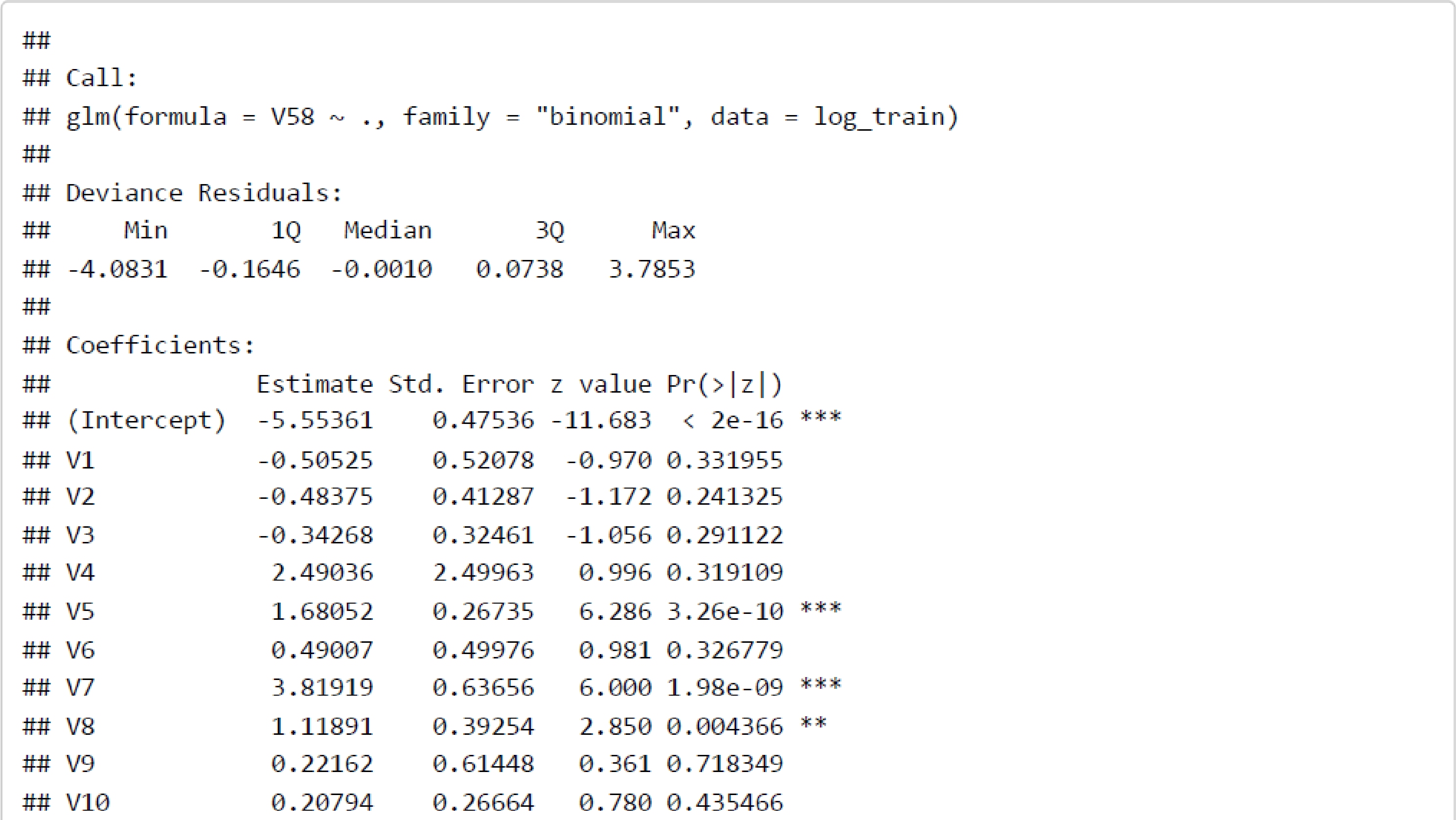
Confusion Matrix for Logistic Regression Standardized test data



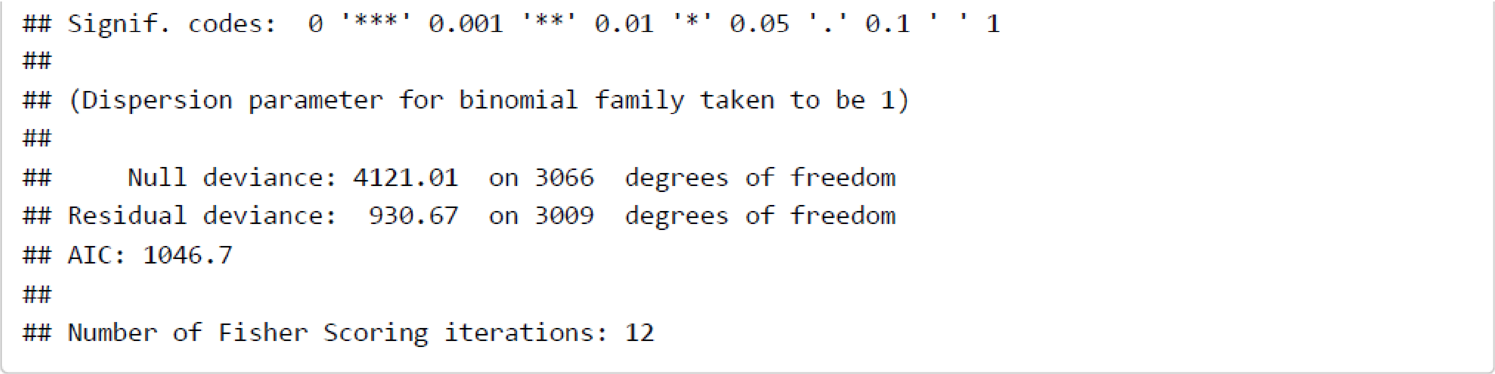
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 231   2
##          1 685 616
##
##                Accuracy : 0.5522
##                  95% CI : (0.5269, 0.5772)
##     No Information Rate : 0.5971
##     P-Value [Acc > NIR] : 0.9998
##
##                   Kappa : 0.211
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.9968
##             Specificity : 0.2522
##          Pos Pred Value : 0.4735
##          Neg Pred Value : 0.9914
##              Prevalence : 0.4029
##          Detection Rate : 0.4016
##    Detection Prevalence : 0.8481
##       Balanced Accuracy : 0.6245
##
##        'Positive' Class : 1
##
```

The accuracy is 55.22% for the classification error of the Logistic Regression of standardized test data.

Linear Regression on Log Transformation train and test data.



```
##
## Call:
## glm(formula = V58 ~ ., family = "binomial", data = log_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.0831  -0.1646  -0.0010   0.0738   3.7853
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -5.55361    0.47536 -11.683  < 2e-16 ***
## V1           -0.50525    0.52078  -0.970 0.331955
## V2           -0.48375    0.41287  -1.172 0.241325
## V3           -0.34268    0.32461  -1.056 0.291122
## V4            2.49036    2.49963   0.996 0.319109
## V5            1.68052    0.26735   6.286 3.26e-10 ***
## V6            0.49007    0.49976   0.981 0.326779
## V7            3.81919    0.63656   6.000 1.98e-09 ***
## V8            1.11891    0.39254   2.850 0.004366 **
## V9            0.22162    0.61448   0.361 0.718349
## V10           0.20794    0.26664   0.780 0.435466
```



```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4121.01  on 3066  degrees of freedom
## Residual deviance:  930.67  on 3009  degrees of freedom
## AIC: 1046.7
##
## Number of Fisher Scoring iterations: 12
```

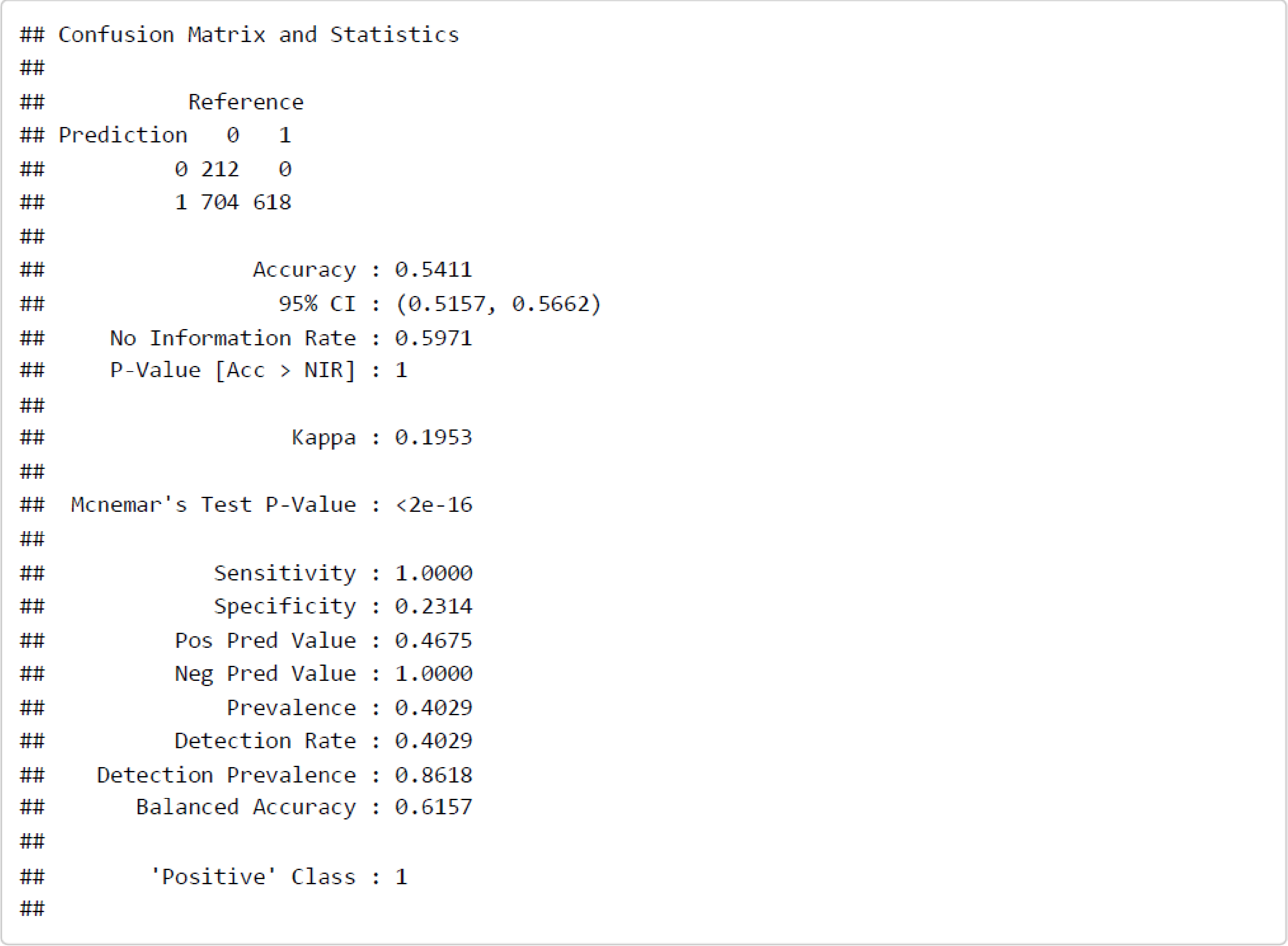
The features: 5, 7, 8, 11, 13, 16, 17, 20, 21, 23, 24, 25, 27, 28, 33, 35, 37, 42, 43, 45, 46, 49, 52, 53, 57 are statistically significant

Confusion Matrix for Logistic Regression Log transformation train data

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0  424    2
##          1 1425 1216
##
##                Accuracy : 0.5347
##                  95% CI : (0.5169, 0.5525)
##     No Information Rate : 0.6029
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1898
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.9984
##             Specificity : 0.2293
##          Pos Pred Value : 0.4604
##          Neg Pred Value : 0.9953
##              Prevalence : 0.3971
##          Detection Rate : 0.3965
##    Detection Prevalence : 0.8611
##       Balanced Accuracy : 0.6138
##
##        'Positive' Class : 1
##
```

The accuracy is 53.47% for the classification error of the Logistic Regression of log transformation of train data.

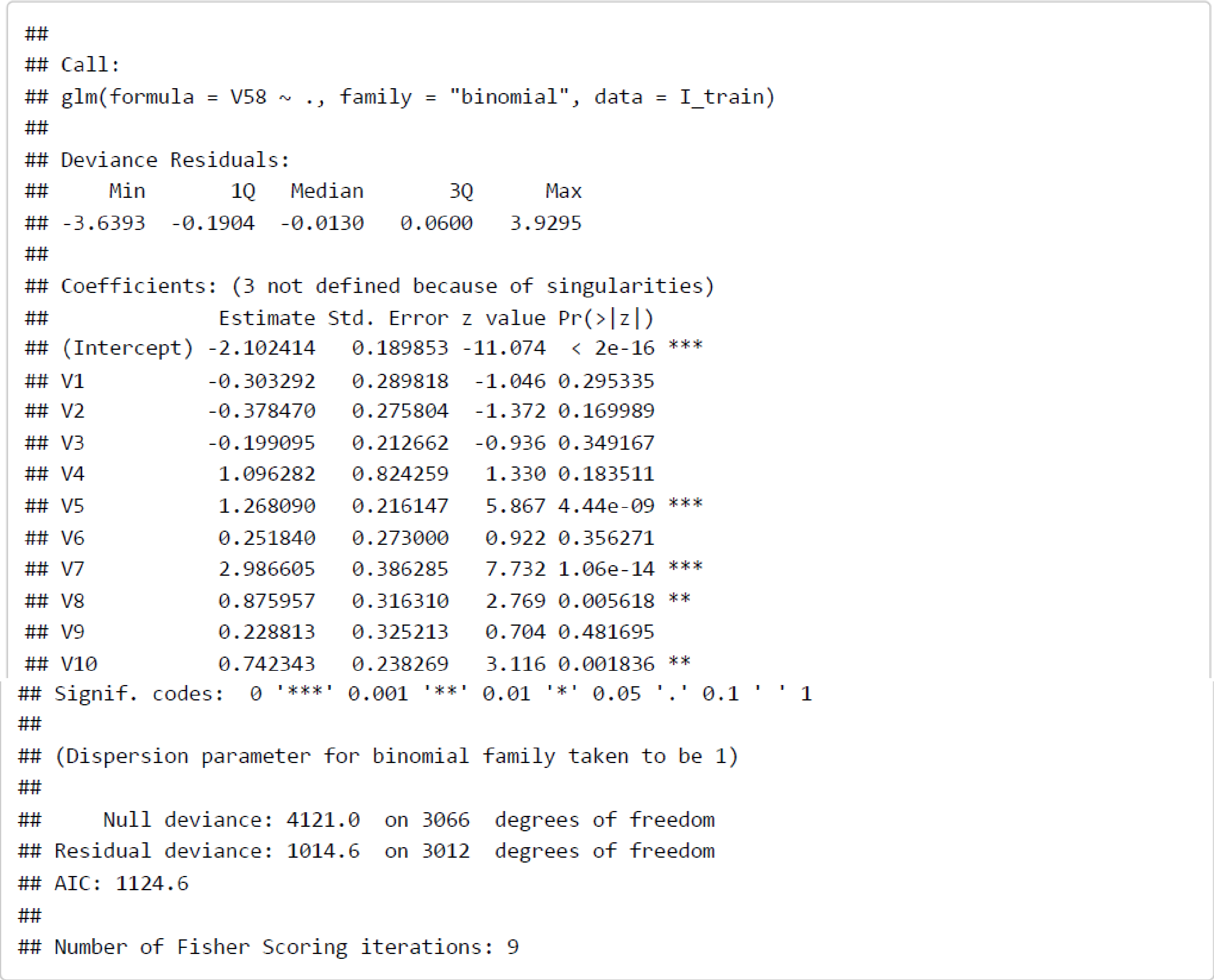
Confusion Matrix for Logistic Regression Log transformation test data.



```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 212   0
##          1 704 618
##
##                Accuracy : 0.5411
##                  95% CI : (0.5157, 0.5662)
##     No Information Rate : 0.5971
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1953
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 1.0000
##             Specificity : 0.2314
##          Pos Pred Value : 0.4675
##          Neg Pred Value : 1.0000
##              Prevalence : 0.4029
##          Detection Rate : 0.4029
##    Detection Prevalence : 0.8618
##       Balanced Accuracy : 0.6157
##
##        'Positive' Class : 1
##
```

The accuracy is 54.11% for the classification error of the Logistic Regression of log transformation of test data.

Logistic Regression on Discretized train and test data



```
##
## Call:
## glm(formula = V58 ~ ., family = "binomial", data = I_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6393  -0.1904  -0.0130   0.0600   3.9295
##
## Coefficients: (3 not defined because of singularities)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.102414   0.189853 -11.074  < 2e-16 ***
## V1          -0.303292   0.289818  -1.046 0.295335
## V2          -0.378470   0.275804  -1.372 0.169989
## V3          -0.199095   0.212662  -0.936 0.349167
## V4           1.096282   0.824259   1.330 0.183511
## V5           1.268090   0.216147   5.867 4.44e-09 ***
## V6           0.251840   0.273000   0.922 0.356271
## V7           2.986605   0.386285   7.732 1.06e-14 ***
## V8           0.875957   0.316310   2.769 0.005618 **
## V9           0.228813   0.325213   0.704 0.481695
## V10          0.742343   0.238269   3.116 0.001836 **
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4121.0  on 3066  degrees of freedom
## Residual deviance: 1014.6  on 3012  degrees of freedom
## AIC: 1124.6
##
## Number of Fisher Scoring iterations: 9
```

The features: 5, 7, 8, 10, 11, 13, 14, 15, 16, 17, 18, 20, 21, 23, 24, 25, 27, 28, 37, 42, 43, 44, 45, 46, 48, 52, 53, 54 are statistically significant. Also have features 55, 56, 57 as NA in the summary function because these features are singularities, meaning that their respective columns are either all 0s or all 1s so cannot get a p-value from it.

Confusion Matrix for Logistic Regression Discretized train data

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction    0    1
##          0 1741  193
##          1  108 1025
##                                           
##                Accuracy : 0.9019          
##                  95% CI : (0.8908, 0.9122)
##     No Information Rate : 0.6029          
##     P-Value [Acc > NIR] : < 2.2e-16       
##                                           
##                   Kappa : 0.7926          
##                                           
##  Mcnemar's Test P-Value : 1.287e-06       
##                                           
##             Sensitivity : 0.8415          
##             Specificity : 0.9416          
##          Pos Pred Value : 0.9047          
##          Neg Pred Value : 0.9002          
##              Prevalence : 0.3971          
##          Detection Rate : 0.3342          
##    Detection Prevalence : 0.3694          
##       Balanced Accuracy : 0.8916          
##                                           
##        'Positive' Class : 1               
## 
```

The accuracy is 90.19% for the classification error of the Logistic Regression of discretize transformation of train data.

Confusion Matrix for Logistic Regression Discretized test data

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 861 103
##          1  55 515
##
##                Accuracy : 0.897
##                  95% CI : (0.8807, 0.9118)
##     No Information Rate : 0.5971
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.7832
##
##  Mcnemar's Test P-Value : 0.0001847
##
##             Sensitivity : 0.8333
##             Specificity : 0.9400
##          Pos Pred Value : 0.9035
##          Neg Pred Value : 0.8932
##              Prevalence : 0.4029
##          Detection Rate : 0.3357
##    Detection Prevalence : 0.3716
##       Balanced Accuracy : 0.8866
##
##        'Positive' Class : 1
##
```

The accuracy is 89.7% for the classification error of the Logistic Regression of discretize transformation of test data.

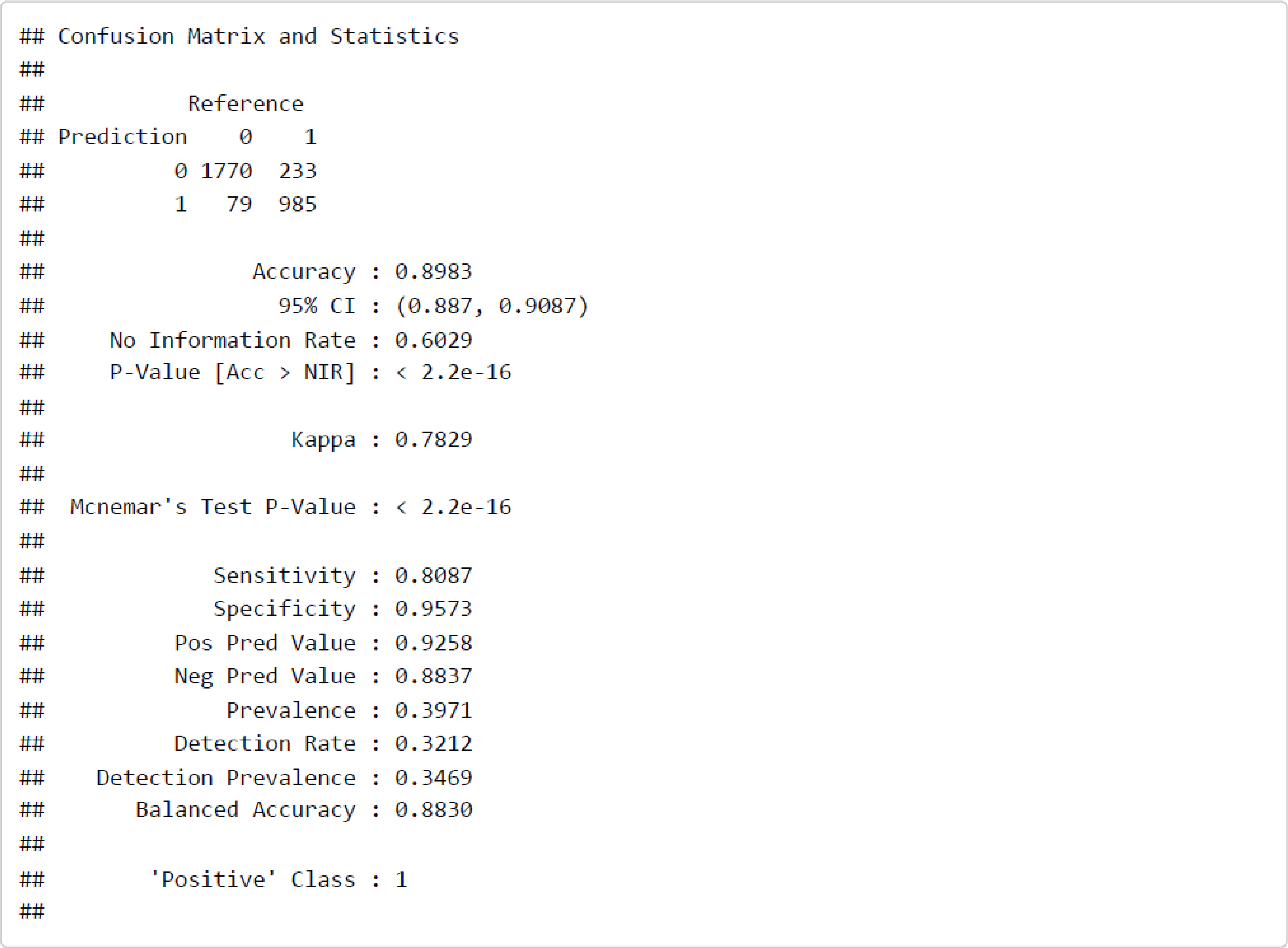
Classification Accuracies for training and testing datasets.

```
##       lr original lr standardized    lr log      lr I
## train   0.9282687       0.5415716 0.5347245 0.9018585
## test    0.9269883       0.5521512 0.5410691 0.8970013
```

5. Applying both linear and quadratic discriminant analysis methods to the standardized data, and the log transformed data.

LDA for standardized train and test Data

Confusion Matrix for LDA Standardized train data



```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction    0    1
##          0 1770  233
##          1   79  985
## 
##                Accuracy : 0.8983
##                  95% CI : (0.887, 0.9087)
##     No Information Rate : 0.6029
##     P-Value [Acc > NIR] : < 2.2e-16
## 
##                   Kappa : 0.7829
## 
##  Mcnemar's Test P-Value : < 2.2e-16
## 
##             Sensitivity : 0.8087
##             Specificity : 0.9573
##          Pos Pred Value : 0.9258
##          Neg Pred Value : 0.8837
##              Prevalence : 0.3971
##          Detection Rate : 0.3212
##    Detection Prevalence : 0.3469
##       Balanced Accuracy : 0.8830
## 
##        'Positive' Class : 1
## 
```

The accuracy is 89.83% for the classification error of the lda standardized train data.

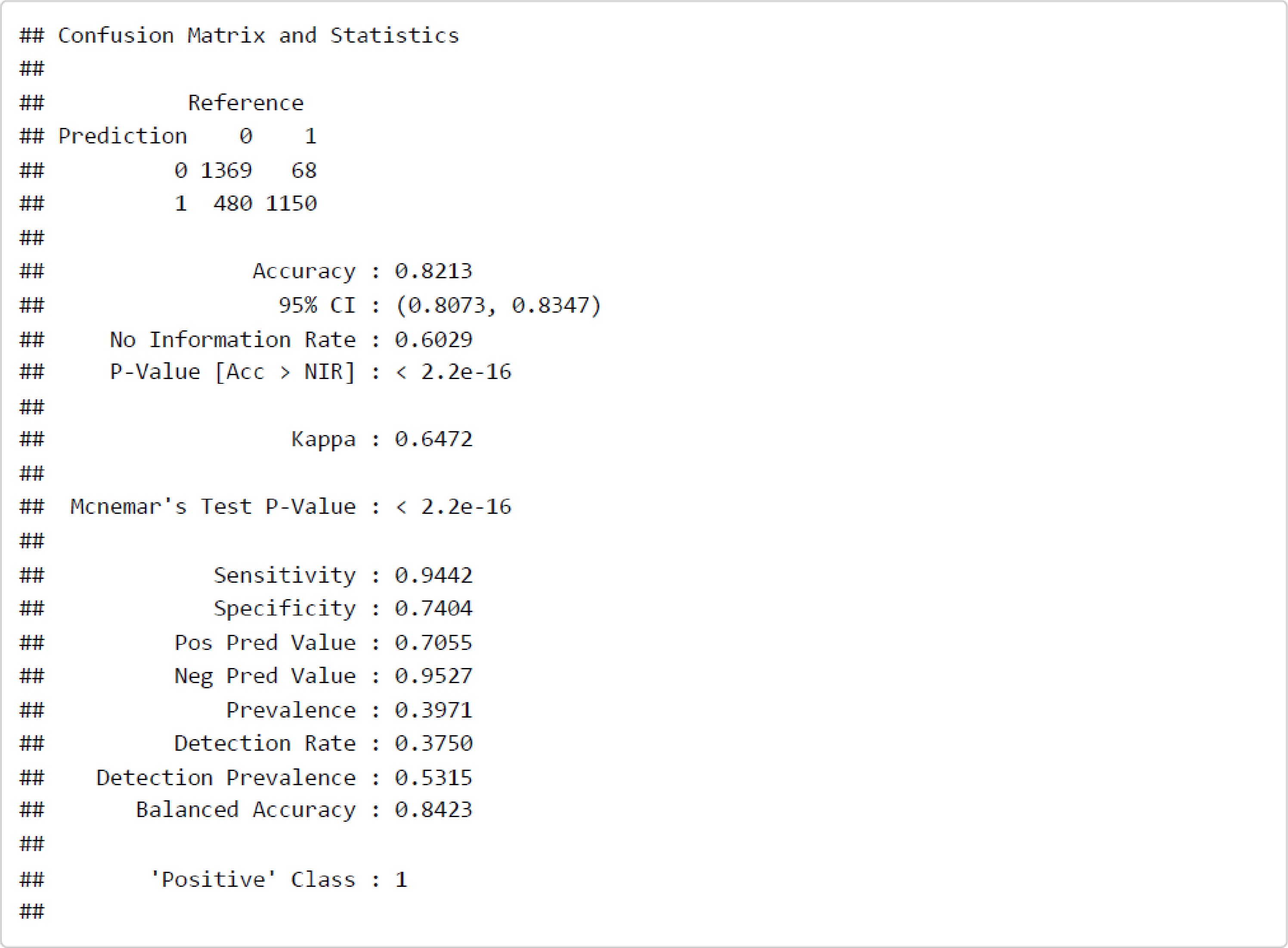
Confusion Matrix for LDA Standardized test data

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction   0   1
##          0 873 115
##          1  43 503
## 
##                Accuracy : 0.897           
##                  95% CI : (0.8807, 0.9118)
##     No Information Rate : 0.5971          
##     P-Value [Acc > NIR] : < 2.2e-16       
## 
##                   Kappa : 0.7818          
## 
##  Mcnemar's Test P-Value : 1.619e-08       
## 
##             Sensitivity : 0.8139          
##             Specificity : 0.9531          
##          Pos Pred Value : 0.9212          
##          Neg Pred Value : 0.8836          
##              Prevalence : 0.4029          
##          Detection Rate : 0.3279          
##    Detection Prevalence : 0.3559          
##       Balanced Accuracy : 0.8835          
## 
##        'Positive' Class : 1               
## 
```

The accuracy is 89.7% for the classification error of the lda standardized test data.

QDA for standardized train and test data

Confusion Matrix for QDA Standardized train data



```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction    0    1
##          0 1369   68
##          1  480 1150
##                                           
##                Accuracy : 0.8213          
##                  95% CI : (0.8073, 0.8347)
##     No Information Rate : 0.6029          
##     P-Value [Acc > NIR] : < 2.2e-16       
##                                           
##                   Kappa : 0.6472          
##                                           
##  Mcnemar's Test P-Value : < 2.2e-16       
##                                           
##             Sensitivity : 0.9442          
##             Specificity : 0.7404          
##          Pos Pred Value : 0.7055          
##          Neg Pred Value : 0.9527          
##              Prevalence : 0.3971          
##          Detection Rate : 0.3750          
##    Detection Prevalence : 0.5315          
##       Balanced Accuracy : 0.8423          
##                                           
##        'Positive' Class : 1               
## 
```

The accuracy is 82.13% for the classification error of the qda standardized train data.

Confusion Matrix for QDA Standardized test data

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction   0   1
##          0 673  25
##          1 243 593
##                                           
##                Accuracy : 0.8253          
##                  95% CI : (0.8053, 0.844) 
##     No Information Rate : 0.5971          
##     P-Value [Acc > NIR] : < 2.2e-16       
##                                           
##                   Kappa : 0.6566          
##                                           
##  Mcnemar's Test P-Value : < 2.2e-16       
##                                           
##             Sensitivity : 0.9595          
##             Specificity : 0.7347          
##          Pos Pred Value : 0.7093          
##          Neg Pred Value : 0.9642          
##              Prevalence : 0.4029          
##          Detection Rate : 0.3866          
##    Detection Prevalence : 0.5450          
##       Balanced Accuracy : 0.8471          
##                                           
##        'Positive' Class : 1               
## 
```

The accuracy is 82.53% for the classification error of the qda standardized test data.

LDA for log transformation train and test data

Confusion Matrix for LDA Log transformed train data

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1795  131
##          1   54 1087
##
##                Accuracy : 0.9397
##                  95% CI : (0.9307, 0.9478)
##     No Information Rate : 0.6029
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8727
##
##  Mcnemar's Test P-Value : 2.302e-08
##
##             Sensitivity : 0.8924
##             Specificity : 0.9708
##          Pos Pred Value : 0.9527
##          Neg Pred Value : 0.9320
##              Prevalence : 0.3971
##          Detection Rate : 0.3544
##    Detection Prevalence : 0.3720
##       Balanced Accuracy : 0.9316
##
##        'Positive' Class : 1
##
```

The accuracy is 93.48% for the classification error of the lda log train data.

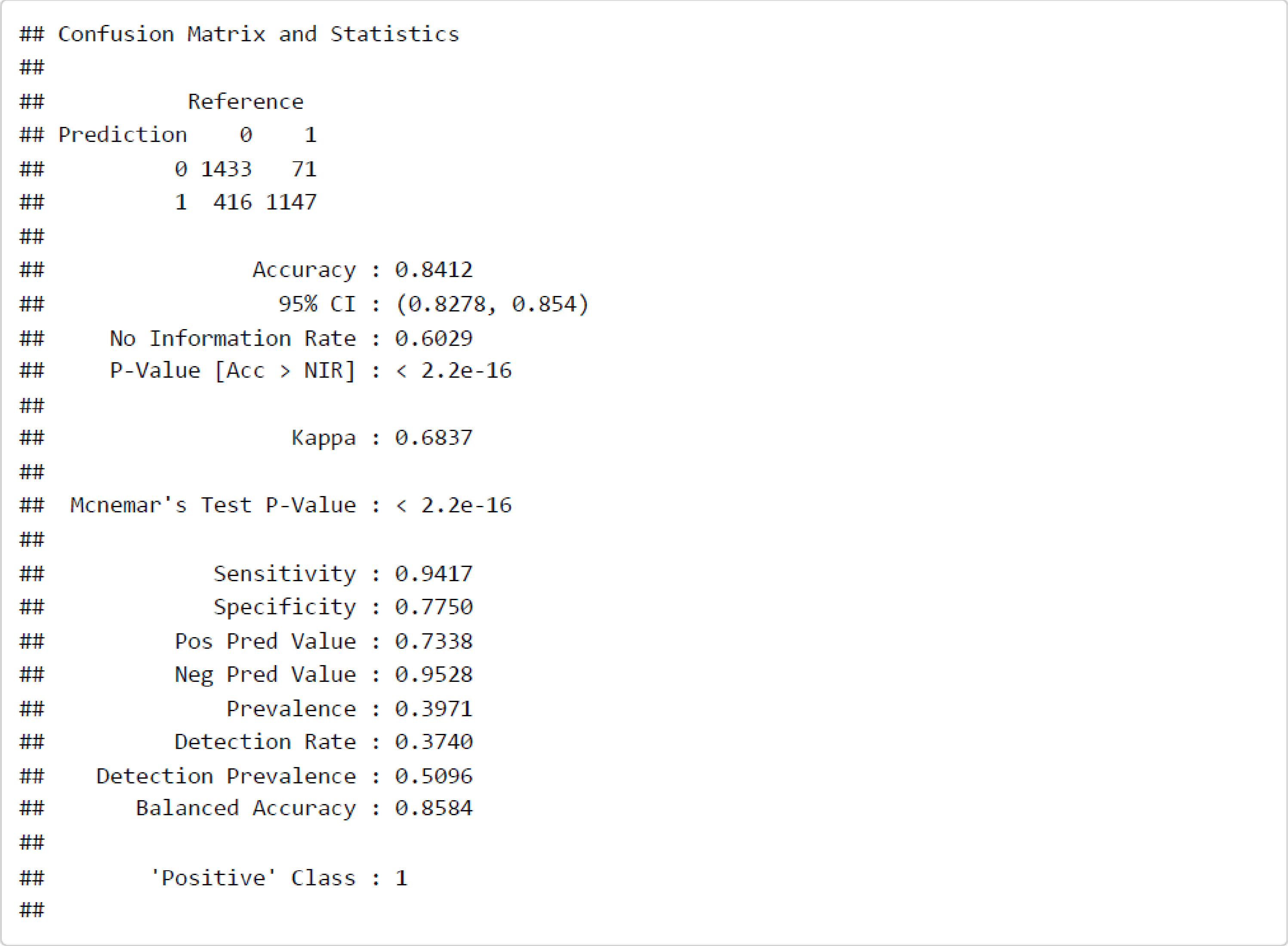
Confusion Matrix for LDA Log transformed test data

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 885  69
##          1  31 549
##
##                Accuracy : 0.9348
##                  95% CI : (0.9213, 0.9466)
##     No Information Rate : 0.5971
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8631
##
##  Mcnemar's Test P-Value : 0.0002156
##
##             Sensitivity : 0.8883
##             Specificity : 0.9662
##          Pos Pred Value : 0.9466
##          Neg Pred Value : 0.9277
##              Prevalence : 0.4029
##          Detection Rate : 0.3579
##    Detection Prevalence : 0.3781
##       Balanced Accuracy : 0.9273
##
##        'Positive' Class : 1
##
```

The accuracy is 93.48% for the classification error of the lda log test data.

QDA for log transformation train and test data

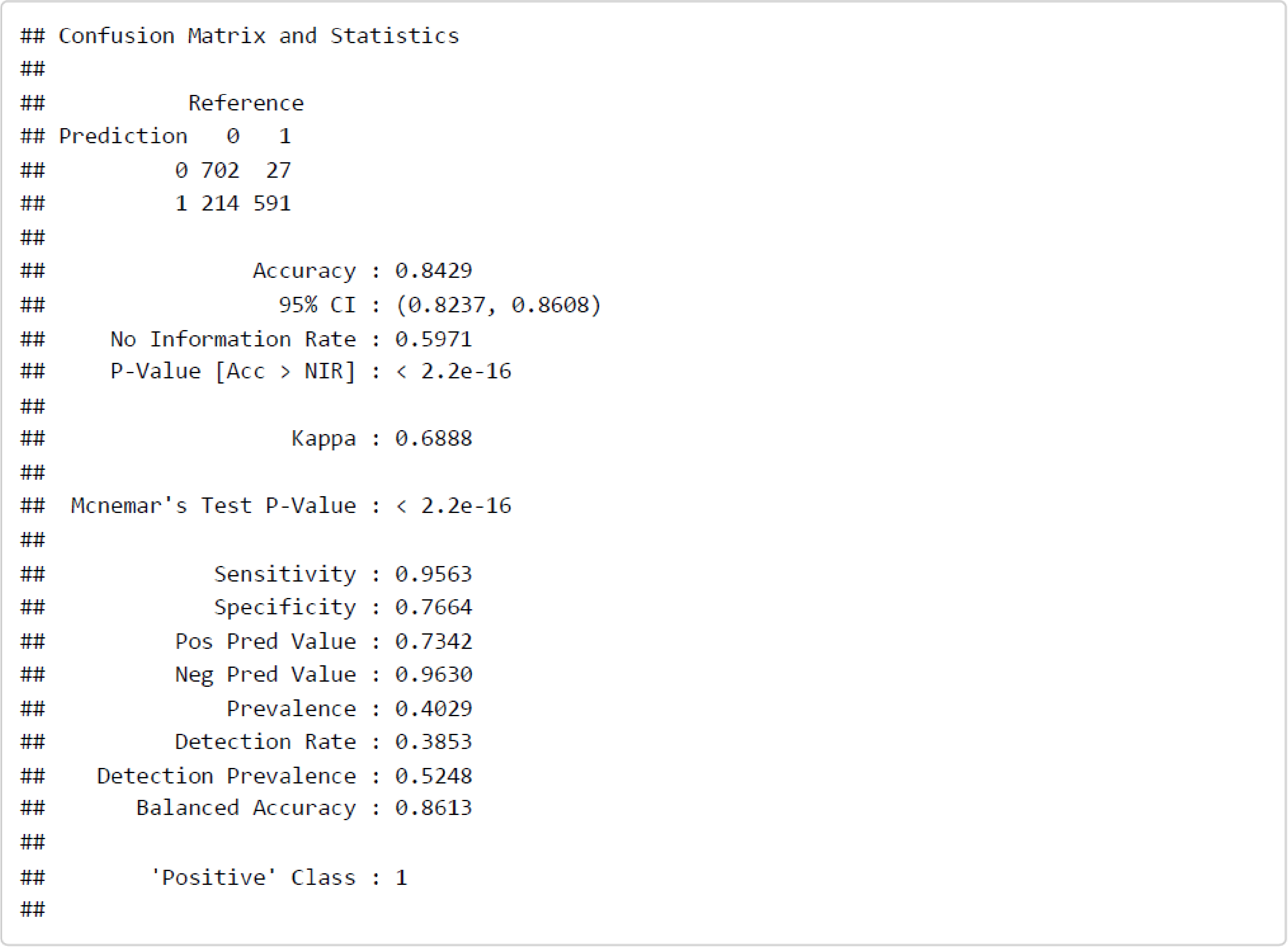
Confusion Matrix for QDA Log transformed train data



```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1433   71
##          1  416 1147
##
##                Accuracy : 0.8412
##                  95% CI : (0.8278, 0.854)
##     No Information Rate : 0.6029
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6837
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9417
##             Specificity : 0.7750
##          Pos Pred Value : 0.7338
##          Neg Pred Value : 0.9528
##              Prevalence : 0.3971
##          Detection Rate : 0.3740
##    Detection Prevalence : 0.5096
##       Balanced Accuracy : 0.8584
##
##        'Positive' Class : 1
##
```

The accuracy is 84.12% for the classification error of the qda log train data.

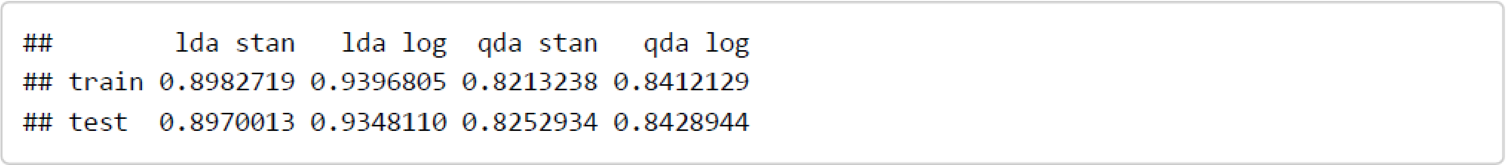
Confusion Matrix for QDA Log transformed test data



```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 702  27
##          1 214 591
##
##                Accuracy : 0.8429
##                  95% CI : (0.8237, 0.8608)
##     No Information Rate : 0.5971
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6888
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9563
##             Specificity : 0.7664
##          Pos Pred Value : 0.7342
##          Neg Pred Value : 0.9630
##              Prevalence : 0.4029
##          Detection Rate : 0.3853
##    Detection Prevalence : 0.5248
##       Balanced Accuracy : 0.8613
##
##        'Positive' Class : 1
##
```

The accuracy is 84.29% for the classification error of the qda log test data.

Accuracies for LDA and QDA of train and test



```
##        lda stan   lda log  qda stan   qda log
## train 0.8982719 0.9396805 0.8213238 0.8412129
## test  0.8970013 0.9348110 0.8252934 0.8428944
```

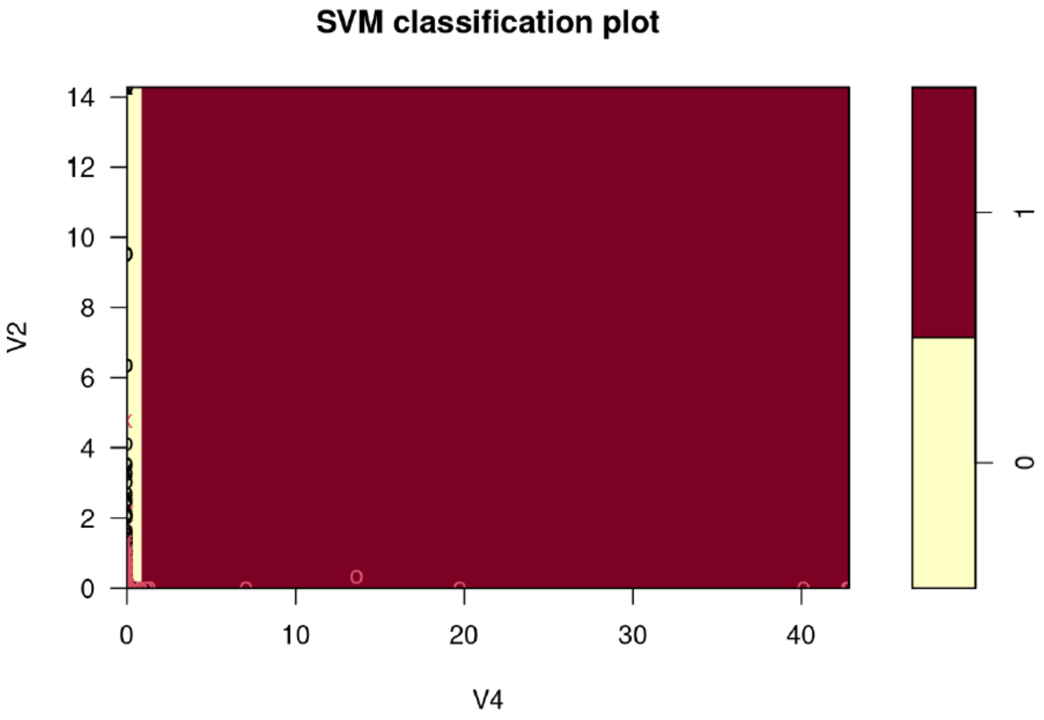
For all the above, LDA and QDA for standardized and log transformed data on both test and train data sets, the LDA performed better than the QDA.

6. Applying linear and nonlinear support vector machine classifiers to each version of the data.

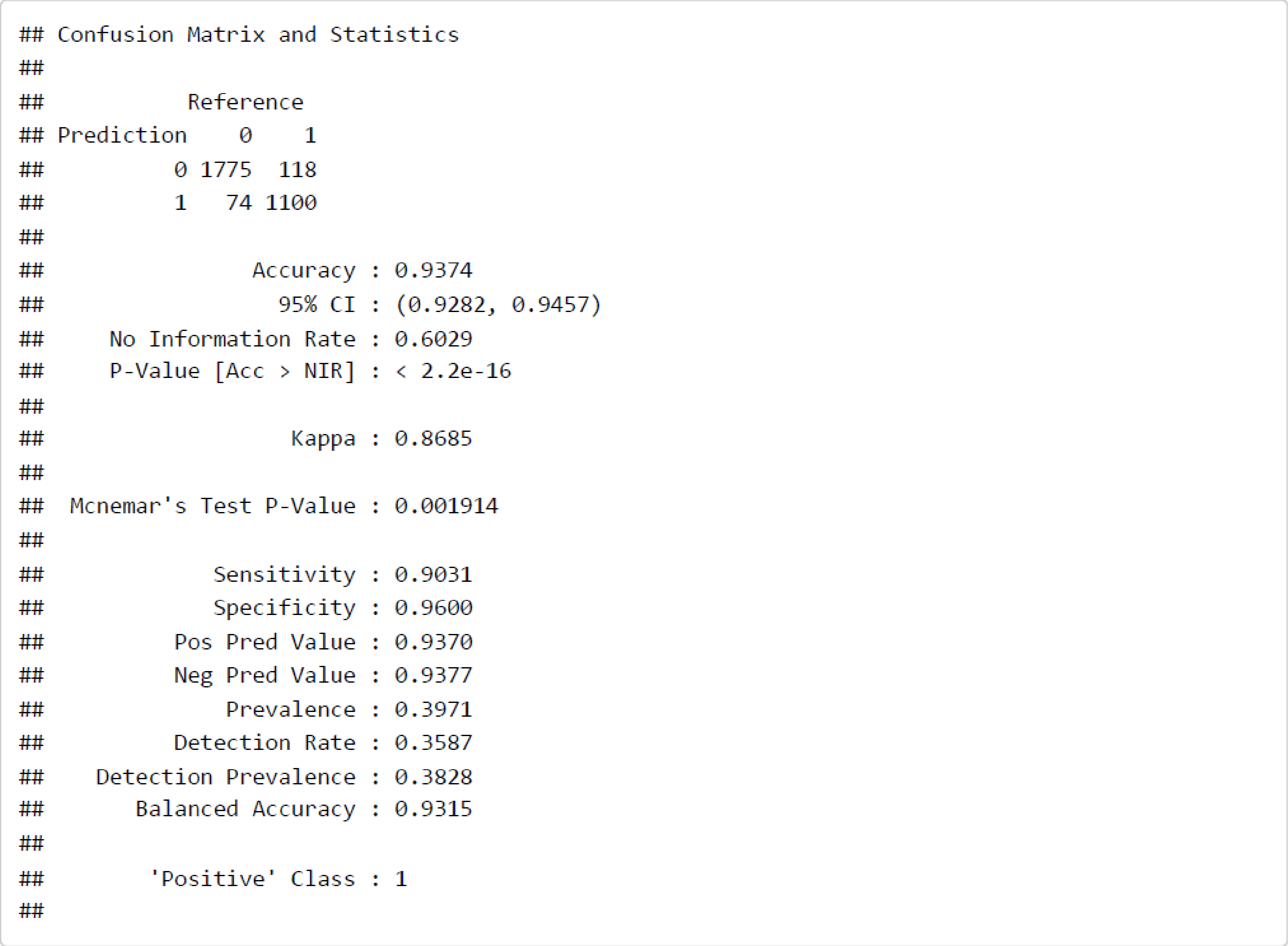
Linear SVM for original data

SVM parameter training for Original data

```
##
## Parameter tuning of 'svm':
##
## - sampling method: 10-fold cross validation
##
## - best parameters:
##  cost
##    10
##
## - best performance: 0.07238402
##
## - Detailed performance results:
##    cost      error dispersion
## 1 1e-03 0.10954844 0.02793485
## 2 1e-02 0.08151200 0.01646617
## 3 1e-01 0.07760001 0.01898033
## 4 1e+00 0.07466841 0.02357943
## 5 5e+00 0.07271189 0.02376119
## 6 1e+01 0.07238402 0.02318753
```



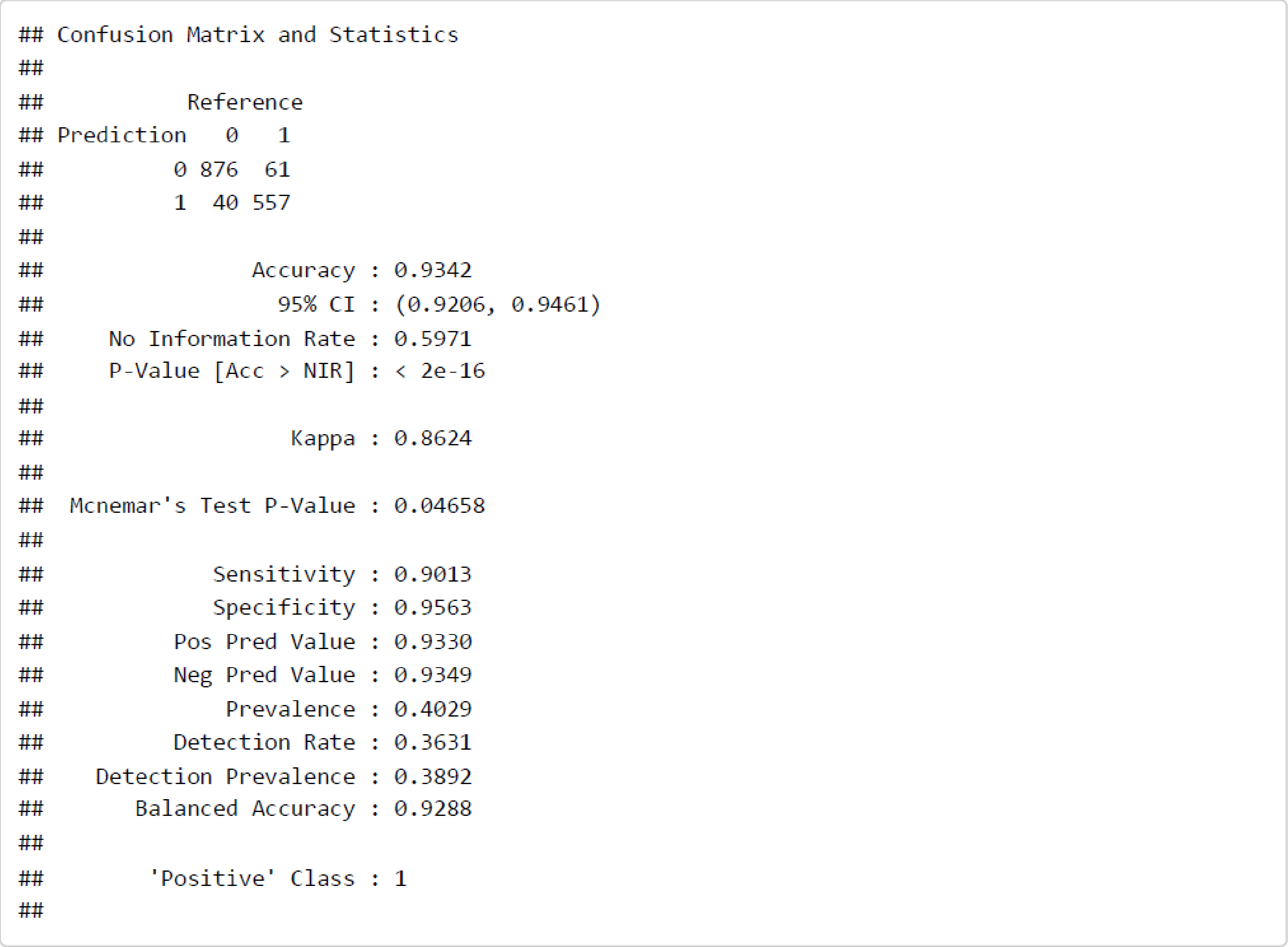
Confusion Matrix for Linear SVM Classifier for Original train data



```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction    0    1
##          0 1775  118
##          1   74 1100
##                                           
##                Accuracy : 0.9374          
##                  95% CI : (0.9282, 0.9457)
##     No Information Rate : 0.6029          
##     P-Value [Acc > NIR] : < 2.2e-16       
##                                           
##                   Kappa : 0.8685          
##                                           
##  Mcnemar's Test P-Value : 0.001914        
##                                           
##             Sensitivity : 0.9031          
##             Specificity : 0.9600          
##          Pos Pred Value : 0.9370          
##          Neg Pred Value : 0.9377          
##              Prevalence : 0.3971          
##          Detection Rate : 0.3587          
##    Detection Prevalence : 0.3828          
##       Balanced Accuracy : 0.9315          
##                                           
##        'Positive' Class : 1               
## 
```

The accuracy is 93.74% for the linear svm classifier of the train data.

Confusion Matrix for Linear SVM Classifier for Original test data



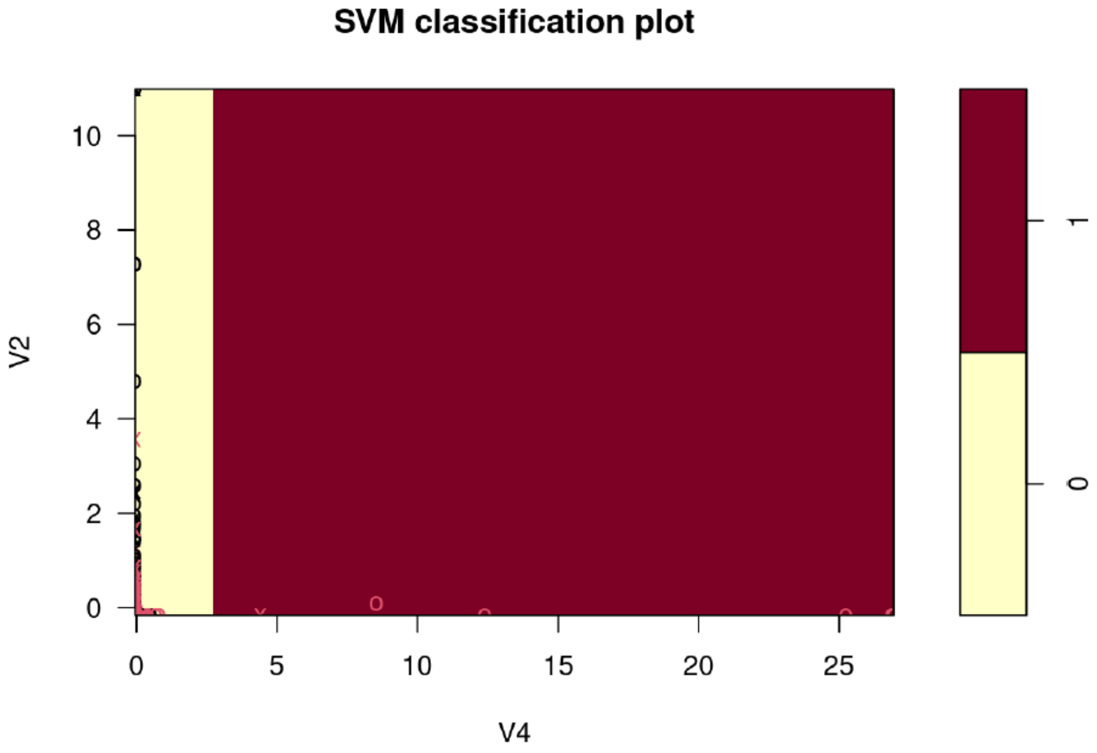
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 876  61
##          1  40 557
##
##                Accuracy : 0.9342
##                  95% CI : (0.9206, 0.9461)
##     No Information Rate : 0.5971
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.8624
##
##  Mcnemar's Test P-Value : 0.04658
##
##             Sensitivity : 0.9013
##             Specificity : 0.9563
##          Pos Pred Value : 0.9330
##          Neg Pred Value : 0.9349
##              Prevalence : 0.4029
##          Detection Rate : 0.3631
##    Detection Prevalence : 0.3892
##       Balanced Accuracy : 0.9288
##
##        'Positive' Class : 1
##
```

The accuracy is 93.42% for the linear svm classifier of the test data.

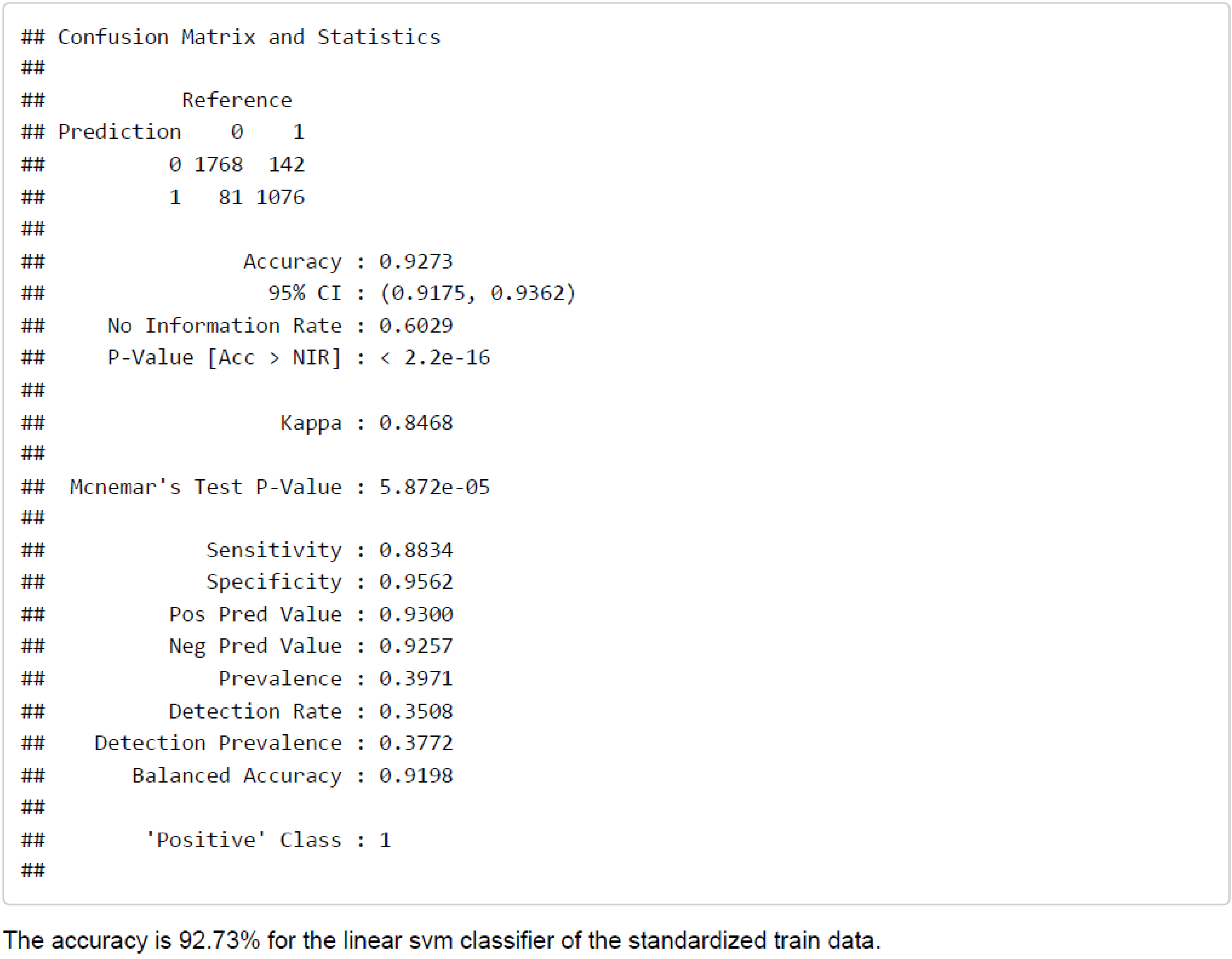
Linear SVM for standardized data

Linear SVM parameter training for Standardized data

```
##
## Parameter tuning of 'svm':
##
## - sampling method: 10-fold cross validation
##
## - best parameters:
##  cost
##     1
##
## - best performance: 0.07368802
##
## - Detailed performance results:
##    cost      error dispersion
## 1 1e-03 0.11020204 0.01214033
## 2 1e-02 0.08444892 0.01600749
## 3 1e-01 0.07825360 0.01521989
## 4 1e+00 0.07368802 0.01279367
## 5 5e+00 0.07401482 0.01435890
## 6 1e+01 0.07368802 0.01163016
```



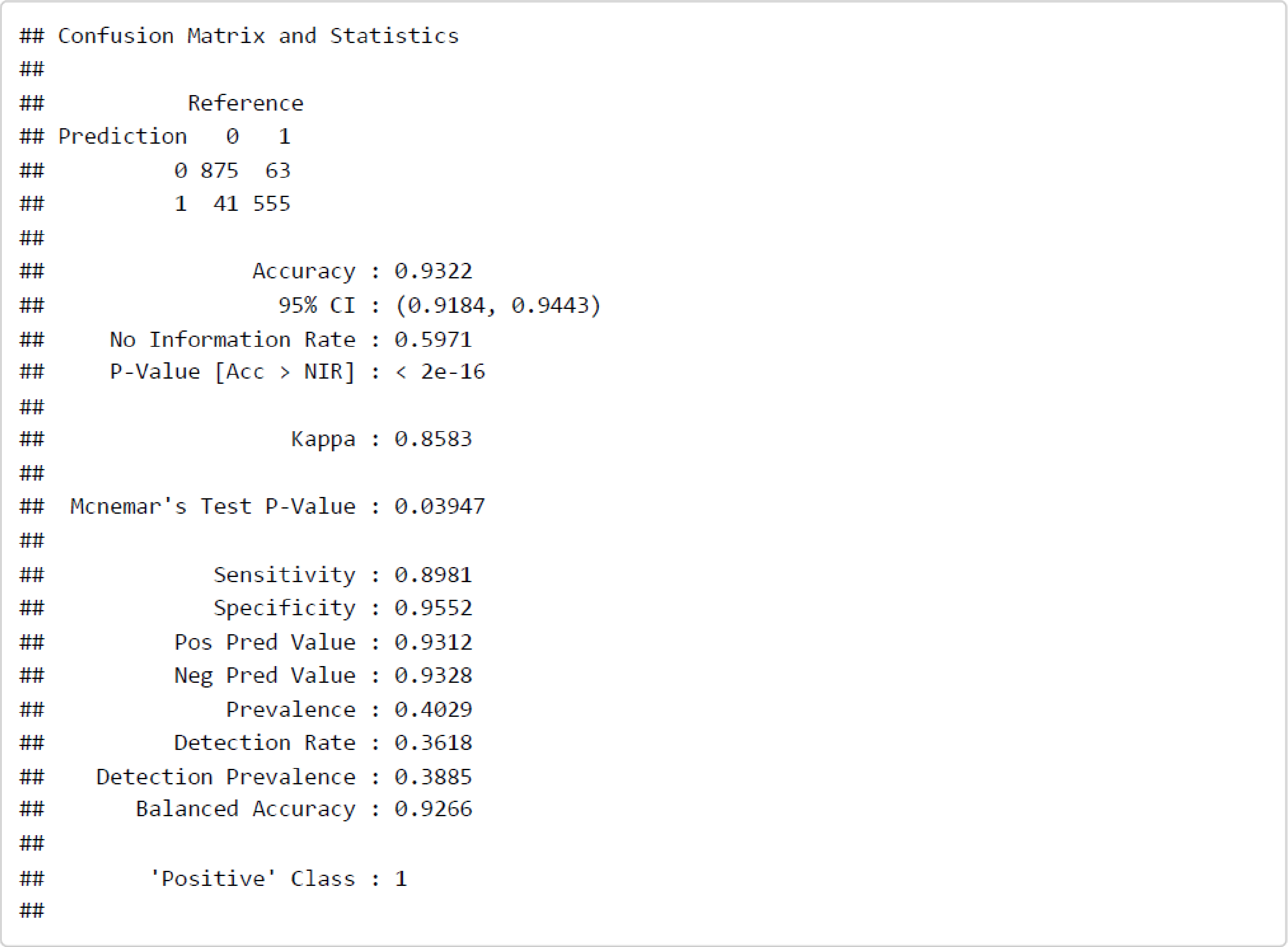
Confusion Matrix for Linear SVM Classifier for Standardized train data



```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1768  142
##          1   81 1076
##
##                Accuracy : 0.9273
##                  95% CI : (0.9175, 0.9362)
##     No Information Rate : 0.6029
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8468
##
##  Mcnemar's Test P-Value : 5.872e-05
##
##             Sensitivity : 0.8834
##             Specificity : 0.9562
##          Pos Pred Value : 0.9300
##          Neg Pred Value : 0.9257
##              Prevalence : 0.3971
##          Detection Rate : 0.3508
##    Detection Prevalence : 0.3772
##       Balanced Accuracy : 0.9198
##
##        'Positive' Class : 1
##
```

The accuracy is 92.73% for the linear svm classifier of the standardized train data.

Confusion Matrix for Linear SVM Classifier for Standardized test data



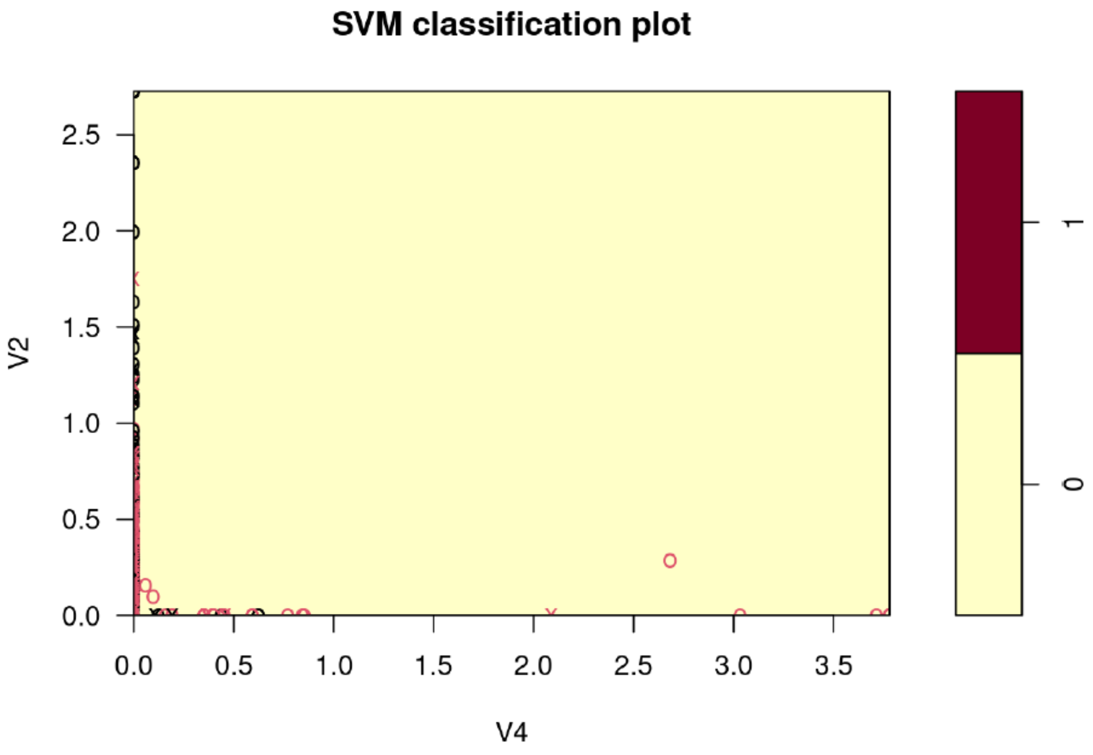
```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction   0   1
##          0 875  63
##          1  41 555
## 
##                Accuracy : 0.9322
##                  95% CI : (0.9184, 0.9443)
##     No Information Rate : 0.5971
##     P-Value [Acc > NIR] : < 2e-16
## 
##                   Kappa : 0.8583
## 
##  Mcnemar's Test P-Value : 0.03947
## 
##             Sensitivity : 0.8981
##             Specificity : 0.9552
##          Pos Pred Value : 0.9312
##          Neg Pred Value : 0.9328
##              Prevalence : 0.4029
##          Detection Rate : 0.3618
##    Detection Prevalence : 0.3885
##       Balanced Accuracy : 0.9266
## 
##        'Positive' Class : 1
## 
```

The accuracy is 93.22% for the linear svm classifier of the standardized test data.

Linear SVM for log transformed data

Linear SVM Parameter tuning for Log transformed data

```
##
## Parameter tuning of 'svm':
##
## - sampling method: 10-fold cross validation
##
## - best parameters:
##  cost
##  0.01
##
## - best performance: 0.05771327
##
## - Detailed performance results:
##    cost      error  dispersion
## 1 1e-03 0.06912563 0.011083682
## 2 1e-02 0.05771327 0.008587184
## 3 1e-01 0.06325499 0.011116892
## 4 1e+00 0.06455898 0.010968706
## 5 5e+00 0.06325605 0.010471460
## 6 1e+01 0.06325605 0.010471460
```



Confusion Matrix for Linear SVM Log transformed train data

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction    0    1
##          0 1777  109
##          1   72 1109
##                                           
##                Accuracy : 0.941           
##                  95% CI : (0.9321, 0.9491)
##     No Information Rate : 0.6029          
##     P-Value [Acc > NIR] : < 2.2e-16       
##                                           
##                   Kappa : 0.8761          
##                                           
##  Mcnemar's Test P-Value : 0.007454        
##                                           
##             Sensitivity : 0.9105          
##             Specificity : 0.9611          
##          Pos Pred Value : 0.9390          
##          Neg Pred Value : 0.9422          
##              Prevalence : 0.3971          
##          Detection Rate : 0.3616          
##    Detection Prevalence : 0.3851          
##       Balanced Accuracy : 0.9358          
##                                           
##        'Positive' Class : 1               
## 
```

The linear svm accuracy for the log transformed train data is 94.10%.

Confusion Matrix for Linear SVM Log transformed test data

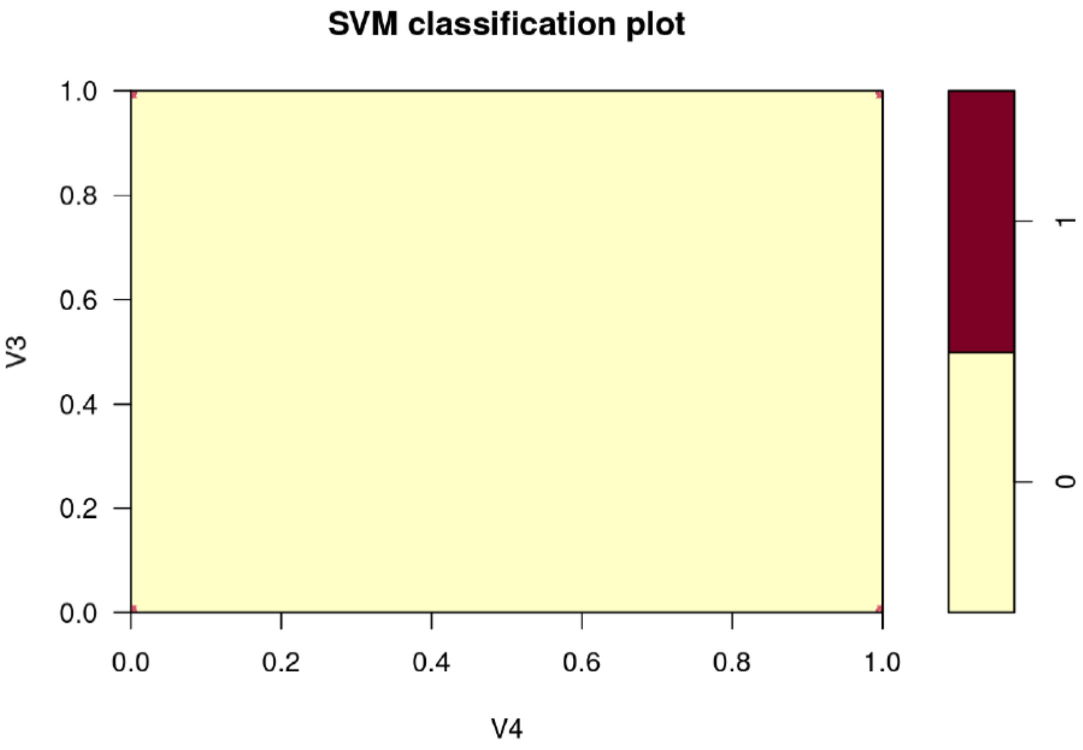
```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction   0   1
##          0 880  53
##          1  36 565
## 
##                Accuracy : 0.942
##                  95% CI : (0.9291, 0.9532)
##     No Information Rate : 0.5971
##     P-Value [Acc > NIR] : < 2e-16
## 
##                   Kappa : 0.8789
## 
##  Mcnemar's Test P-Value : 0.08989
## 
##             Sensitivity : 0.9142
##             Specificity : 0.9607
##          Pos Pred Value : 0.9401
##          Neg Pred Value : 0.9432
##              Prevalence : 0.4029
##          Detection Rate : 0.3683
##    Detection Prevalence : 0.3918
##       Balanced Accuracy : 0.9375
## 
##        'Positive' Class : 1
## 
```

The linear svm accuracy for the log transformed test data is 94.20%.

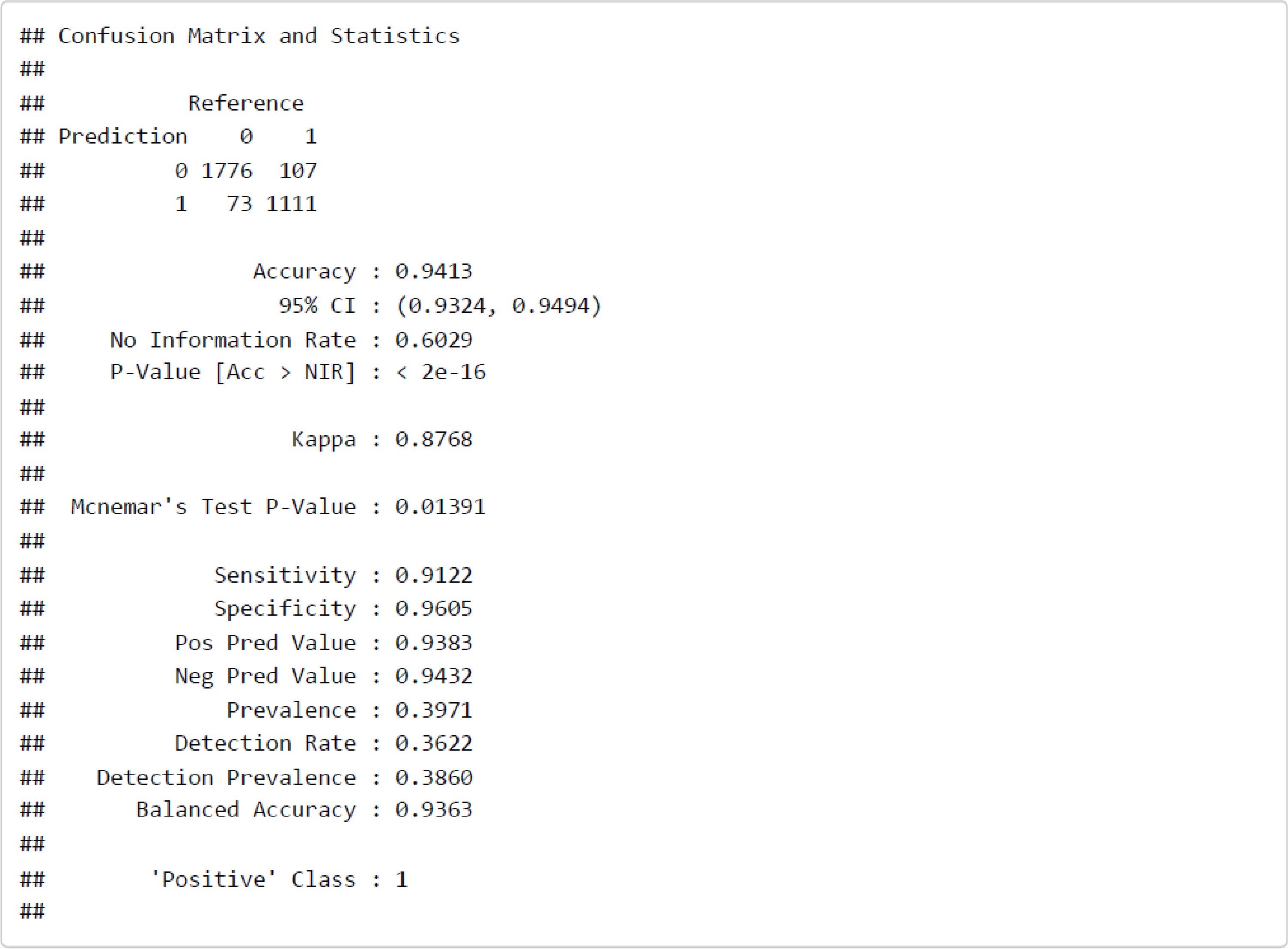
Linear SVM for Discretized I data

Linear SVM Parameter tuning for Discretized data

```
## 
## Parameter tuning of 'svm':
## 
## - sampling method: 10-fold cross validation 
## 
## - best parameters:
##  cost
##     1
## 
## - best performance: 0.06813779 
## 
## - Detailed performance results:
##    cost      error dispersion
## 1 1e-03 0.13075089 0.01864518
## 2 1e-02 0.07629708 0.01152786
## 3 1e-01 0.07042005 0.01380397
## 4 1e+00 0.06813779 0.01330741
## 5 5e+00 0.07075004 0.01535053
## 6 1e+01 0.07140257 0.01426862
```



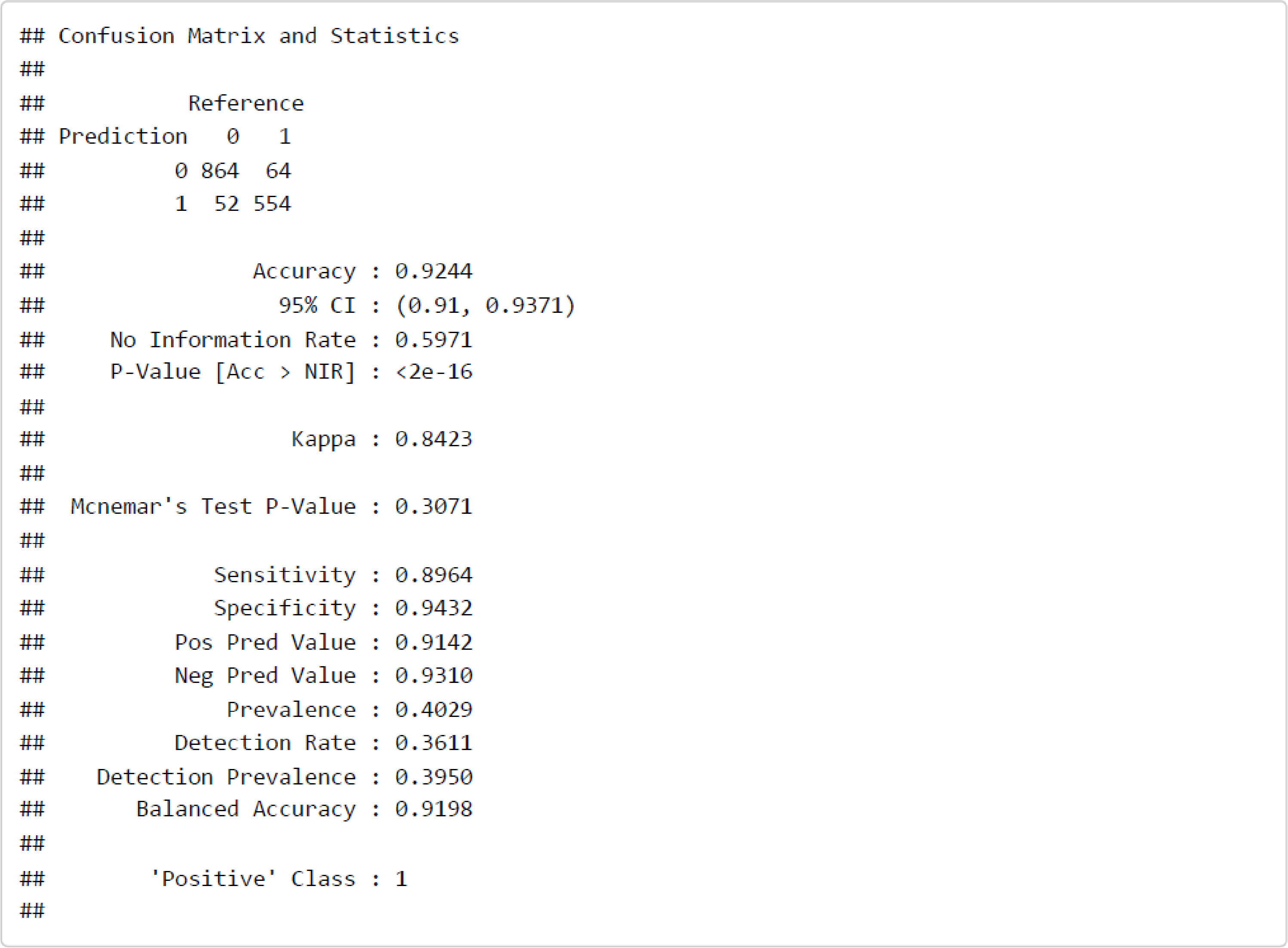
Confusion Matrix for Linear SVM Discretized train data



```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction    0    1
##          0 1776  107
##          1   73 1111
##                                           
##                Accuracy : 0.9413          
##                  95% CI : (0.9324, 0.9494)
##     No Information Rate : 0.6029          
##     P-Value [Acc > NIR] : < 2e-16         
##                                           
##                   Kappa : 0.8768          
##                                           
##  Mcnemar's Test P-Value : 0.01391         
##                                           
##             Sensitivity : 0.9122          
##             Specificity : 0.9605          
##          Pos Pred Value : 0.9383          
##          Neg Pred Value : 0.9432          
##              Prevalence : 0.3971          
##          Detection Rate : 0.3622          
##    Detection Prevalence : 0.3860          
##       Balanced Accuracy : 0.9363          
##                                           
##        'Positive' Class : 1               
## 
```

The linear SVM accuracy for the discretized train data is 94.13%.

Confusion Matrix for Linear SVM Discretized test data



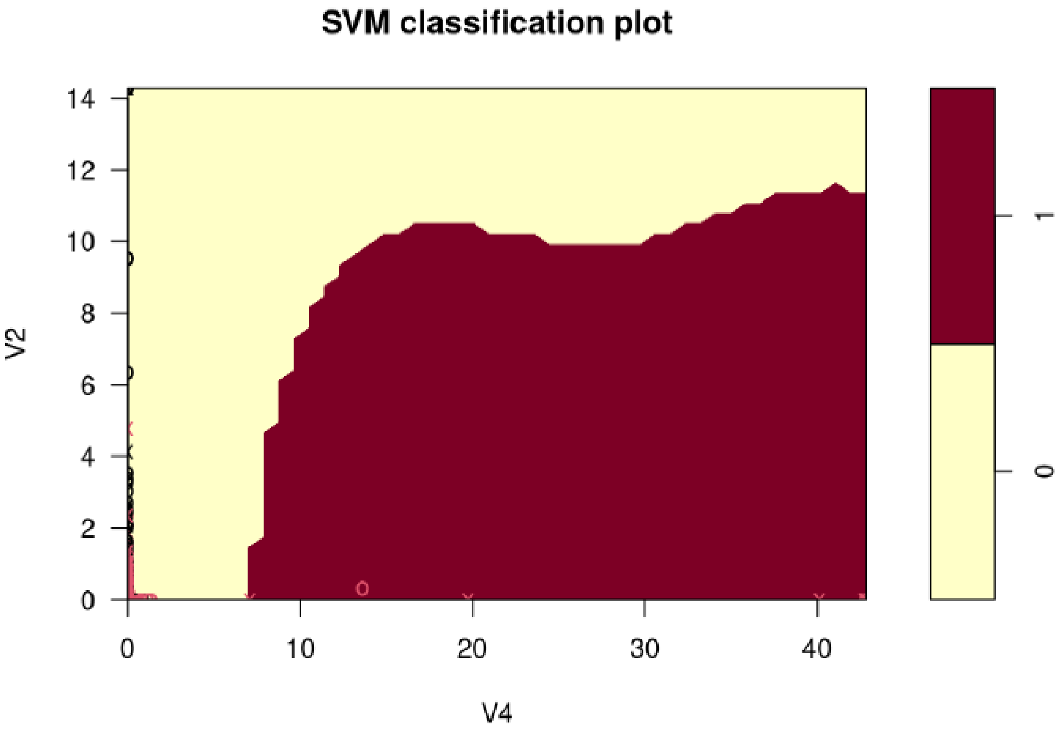
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 864  64
##          1  52 554
##
##                Accuracy : 0.9244
##                  95% CI : (0.91, 0.9371)
##     No Information Rate : 0.5971
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.8423
##
##  Mcnemar's Test P-Value : 0.3071
##
##             Sensitivity : 0.8964
##             Specificity : 0.9432
##          Pos Pred Value : 0.9142
##          Neg Pred Value : 0.9310
##              Prevalence : 0.4029
##          Detection Rate : 0.3611
##    Detection Prevalence : 0.3950
##       Balanced Accuracy : 0.9198
##
##        'Positive' Class : 1
##
```

The linear SVM accuracy for the discretized test data is 92.44%.

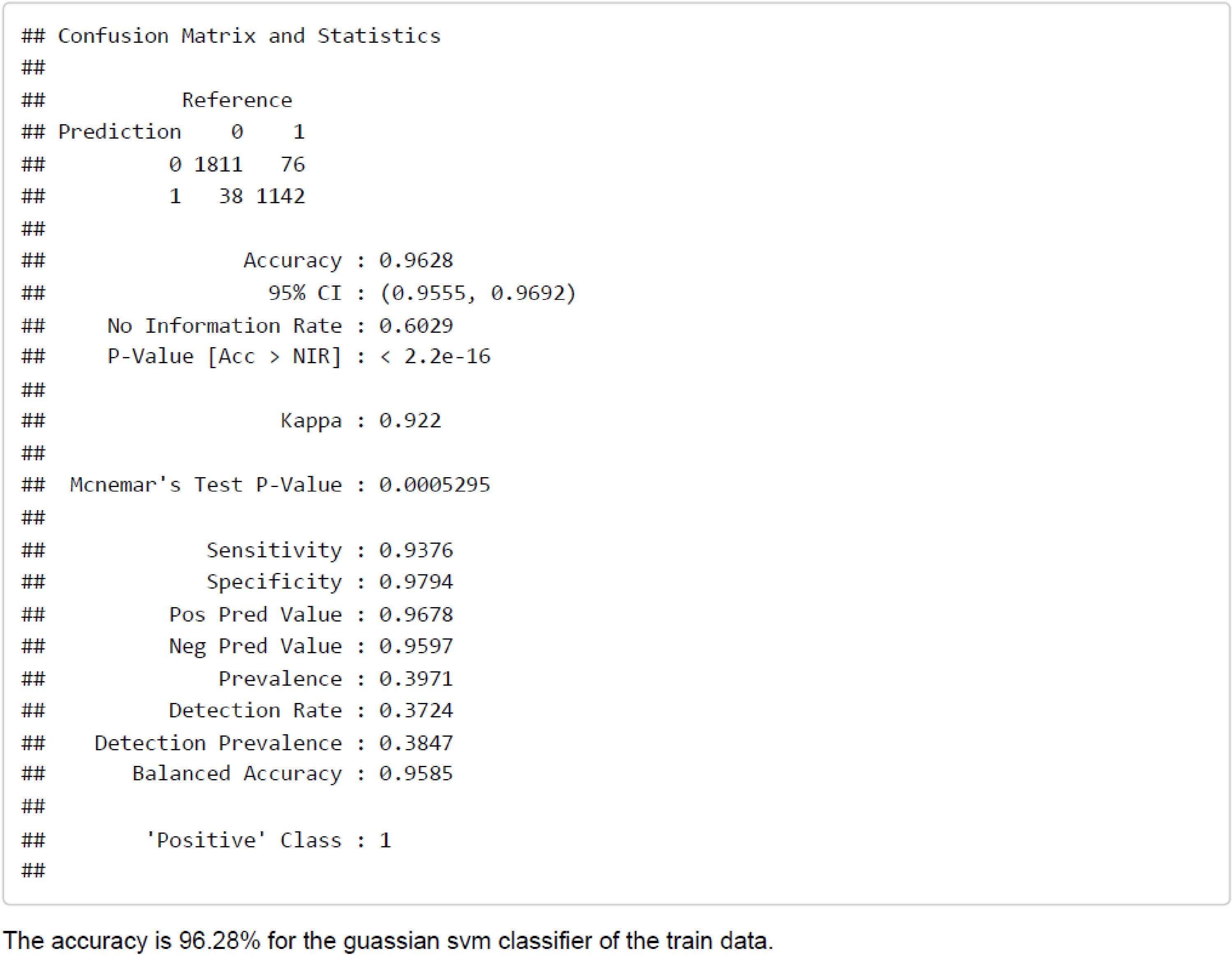
Gaussian SVM for original data

Gaussian SVM Parameter tuning for Original data

```
##
## Parameter tuning of 'svm':
##
## - sampling method: 10-fold cross validation
##
## - best parameters:
##  cost gamma
##    10  0.01
##
## - best performance: 0.05803687
##
## - Detailed performance results:
##      cost gamma      error  dispersion
## 1  1e-03 0.001 0.39712056 0.028732231
## 2  1e-02 0.001 0.39712056 0.028732231
## 3  1e-01 0.001 0.20738541 0.027120756
## 4  1e+00 0.001 0.08771476 0.018238630
## 5  5e+00 0.001 0.07532946 0.016637956
## 6  1e+01 0.001 0.07108535 0.015684356
## 7  1e-03 0.010 0.39712056 0.028732231
## 8  1e-02 0.010 0.37332077 0.036749828
## 9  1e-01 0.010 0.08934449 0.020821958
## 10 1e+00 0.010 0.06651445 0.013668092
```



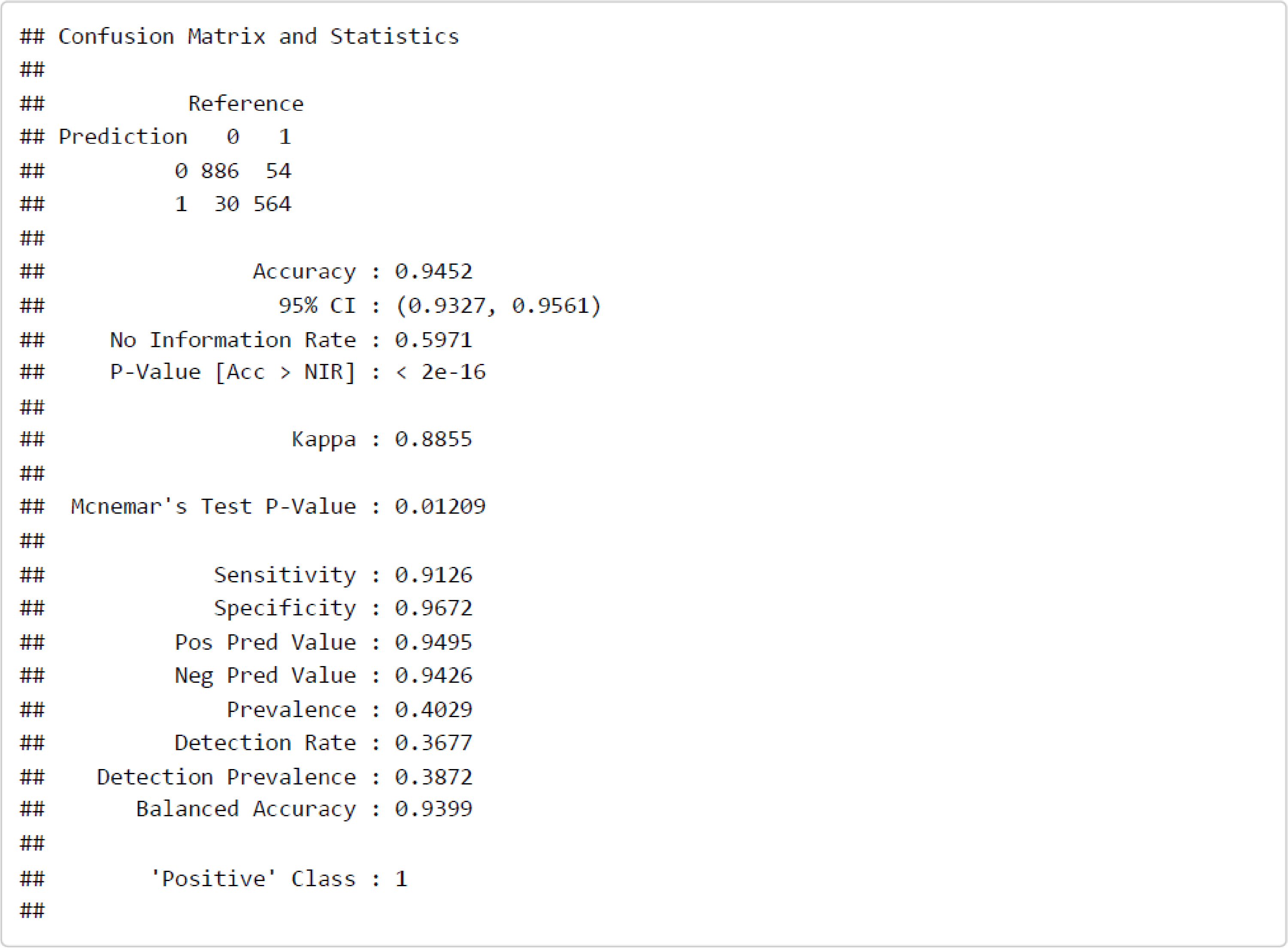
Confusion Matrix for Gaussian SVM Original train data



```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction    0    1
##          0 1811   76
##          1   38 1142
##
##                Accuracy : 0.9628
##                  95% CI : (0.9555, 0.9692)
##     No Information Rate : 0.6029
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.922
##
##  Mcnemar's Test P-Value : 0.0005295
##
##             Sensitivity : 0.9376
##             Specificity : 0.9794
##          Pos Pred Value : 0.9678
##          Neg Pred Value : 0.9597
##              Prevalence : 0.3971
##          Detection Rate : 0.3724
##    Detection Prevalence : 0.3847
##       Balanced Accuracy : 0.9585
##
##        'Positive' Class : 1
##
```

The accuracy is 96.28% for the guassian svm classifier of the train data.

Confusion Matrix for Gaussian SVM Original test data

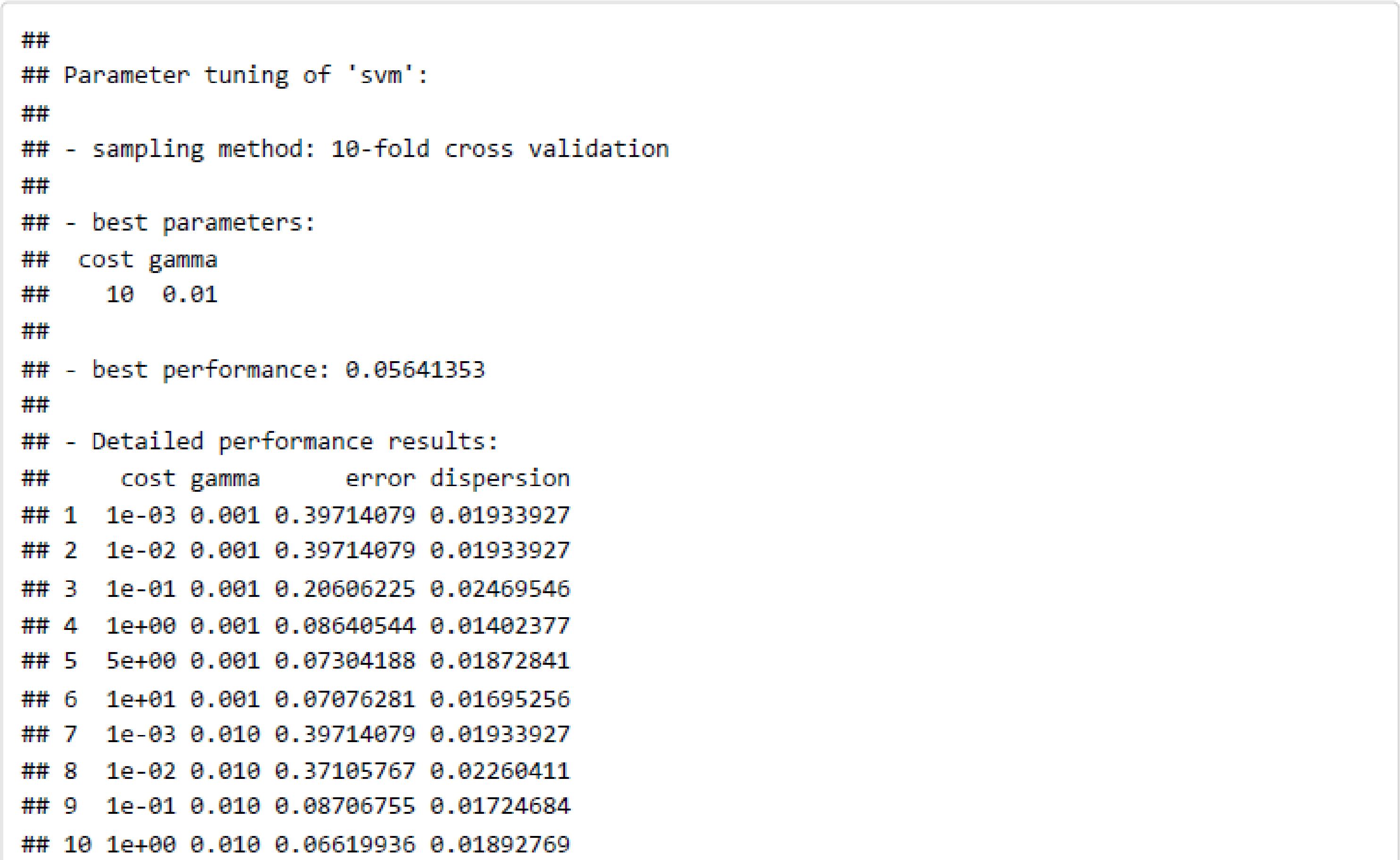


```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 886  54
##          1  30 564
##
##                Accuracy : 0.9452
##                  95% CI : (0.9327, 0.9561)
##     No Information Rate : 0.5971
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.8855
##
##  Mcnemar's Test P-Value : 0.01209
##
##             Sensitivity : 0.9126
##             Specificity : 0.9672
##          Pos Pred Value : 0.9495
##          Neg Pred Value : 0.9426
##              Prevalence : 0.4029
##          Detection Rate : 0.3677
##    Detection Prevalence : 0.3872
##       Balanced Accuracy : 0.9399
##
##        'Positive' Class : 1
##
```

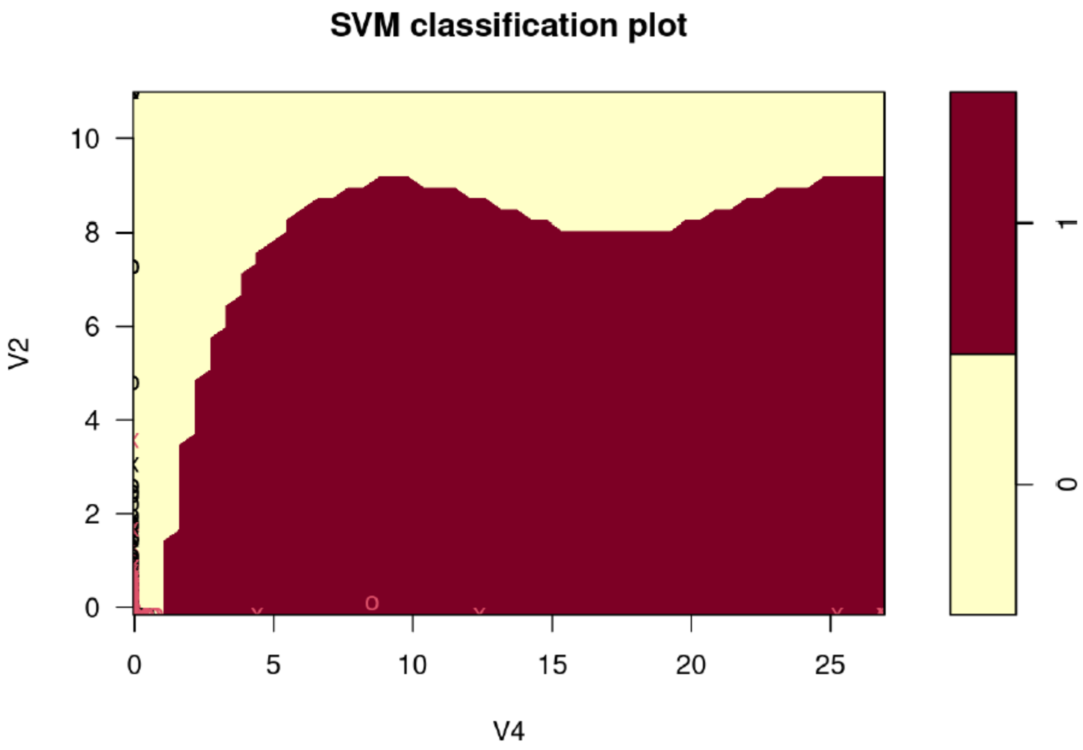
The accuracy is 94.52% for the Guassian svm classifier of the test data.

Gaussian SVM for standardized data

Gaussian SVM Parameter tuning for Standardized data



```
##
## Parameter tuning of 'svm':
##
## - sampling method: 10-fold cross validation
##
## - best parameters:
##  cost gamma
##    10  0.01
##
## - best performance: 0.05641353
##
## - Detailed performance results:
##      cost gamma      error dispersion
## 1  1e-03 0.001 0.39714079 0.01933927
## 2  1e-02 0.001 0.39714079 0.01933927
## 3  1e-01 0.001 0.20606225 0.02469546
## 4  1e+00 0.001 0.08640544 0.01402377
## 5  5e+00 0.001 0.07304188 0.01872841
## 6  1e+01 0.001 0.07076281 0.01695256
## 7  1e-03 0.010 0.39714079 0.01933927
## 8  1e-02 0.010 0.37105767 0.02260411
## 9  1e-01 0.010 0.08706755 0.01724684
## 10 1e+00 0.010 0.06619936 0.01892769
```



Confusion Matrix for Gaussian SVM Standardized train data

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1811   76
##          1   38 1142
##
##                Accuracy : 0.9628
##                  95% CI : (0.9555, 0.9692)
##     No Information Rate : 0.6029
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.922
##
##  Mcnemar's Test P-Value : 0.0005295
##
##             Sensitivity : 0.9376
##             Specificity : 0.9794
##          Pos Pred Value : 0.9678
##          Neg Pred Value : 0.9597
##              Prevalence : 0.3971
##          Detection Rate : 0.3724
##    Detection Prevalence : 0.3847
##       Balanced Accuracy : 0.9585
##
##        'Positive' Class : 1
##
```

The accuracy is 96.28% for the guassian svm classifier of the standardized train data.

Confusion Matrix for Gaussian SVM Standardized test data

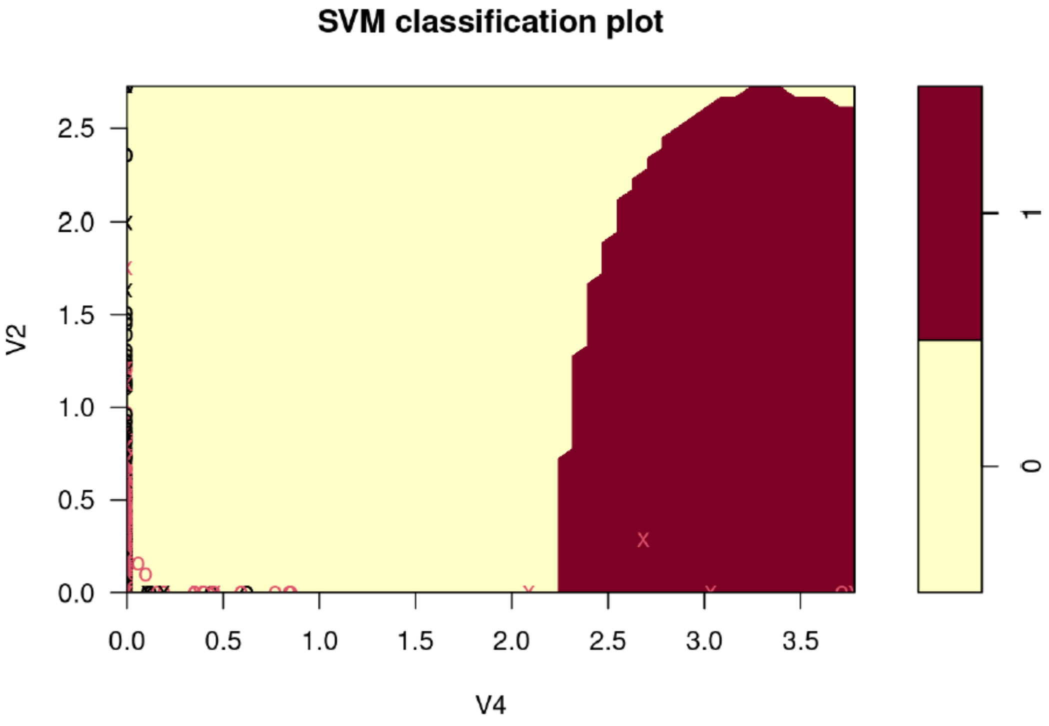
```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction   0   1
##          0 884  58
##          1  32 560
## 
##                Accuracy : 0.9413
##                  95% CI : (0.9284, 0.9526)
##     No Information Rate : 0.5971
##     P-Value [Acc > NIR] : < 2.2e-16
## 
##                   Kappa : 0.8772
## 
##  Mcnemar's Test P-Value : 0.008408
## 
##             Sensitivity : 0.9061
##             Specificity : 0.9651
##          Pos Pred Value : 0.9459
##          Neg Pred Value : 0.9384
##              Prevalence : 0.4029
##          Detection Rate : 0.3651
##    Detection Prevalence : 0.3859
##       Balanced Accuracy : 0.9356
## 
##        'Positive' Class : 1
## 
```

The accuracy is 94.13% for the gaussian svm classifier of the standardized test data.

Gaussian SVM for log transformed data

Gaussian SCM Parameter tuning for Log transformed data

```
##
## Parameter tuning of 'svm':
##
## - sampling method: 10-fold cross validation
##
## - best parameters:
##  cost gamma
##    10  0.01
##
## - best performance: 0.04631262
##
## - Detailed performance results:
##     cost gamma      error dispersion
## 1  1e-03 0.001 0.39716208 0.03006158
## 2  1e-02 0.001 0.39716208 0.03006158
## 3  1e-01 0.001 0.10599412 0.02086794
## 4  1e+00 0.001 0.06163910 0.01477330
## 5  5e+00 0.001 0.06130804 0.01484794
## 6  1e+01 0.001 0.05902791 0.01530110
## 7  1e-03 0.010 0.39716208 0.03006158
## 8  1e-02 0.010 0.17871027 0.02581543
## 9  1e-01 0.010 0.06359882 0.01707959
## 10 1e+00 0.010 0.05772391 0.01431630
```

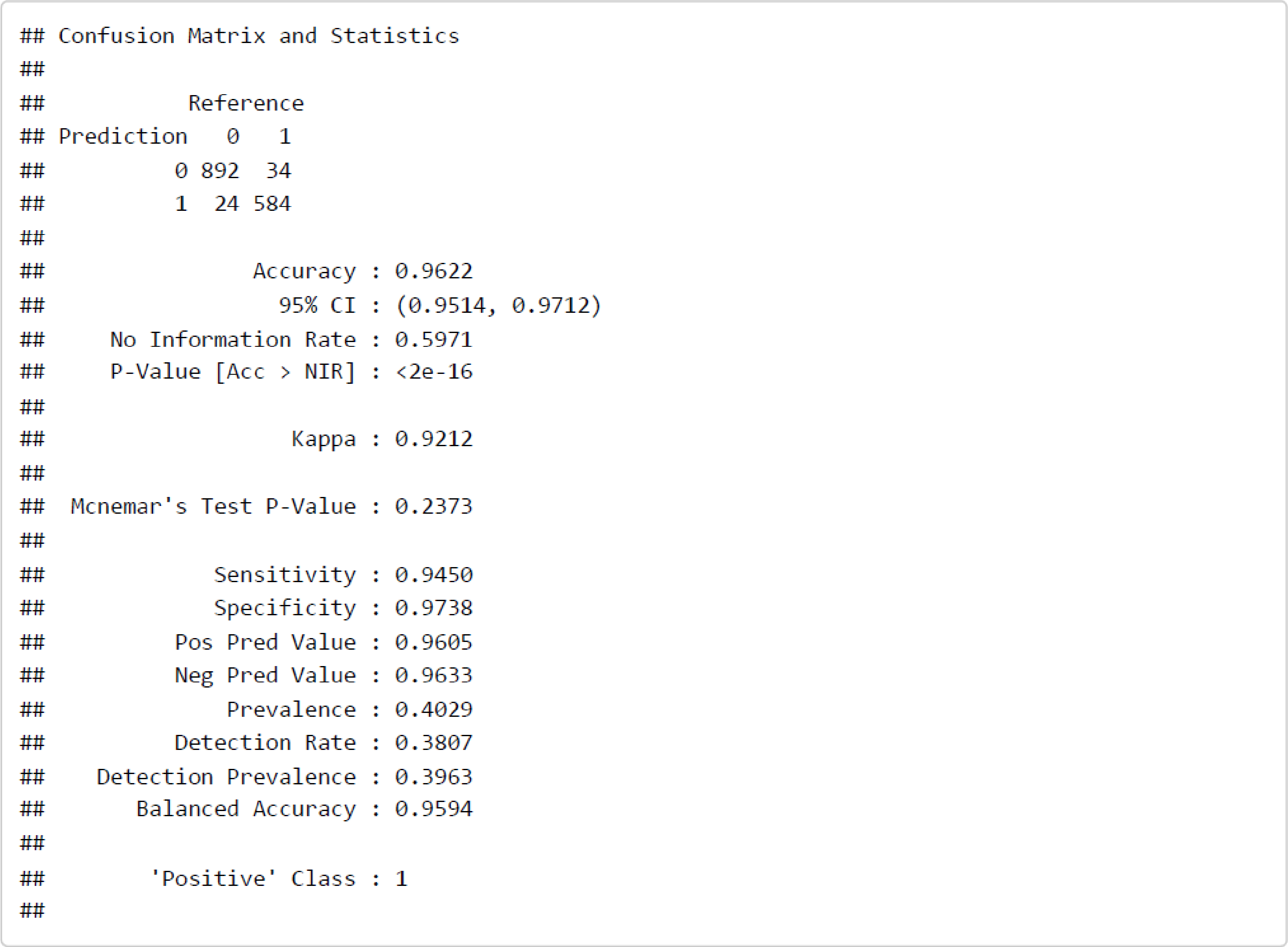


Confusion Matrix for Gaussian SVM Log transformed train data

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction    0    1
##          0 1826   45
##          1   23 1173
##                                          
##                Accuracy : 0.9778         
##                  95% CI : (0.972, 0.9827)
##     No Information Rate : 0.6029         
##     P-Value [Acc > NIR] : < 2e-16        
##                                          
##                   Kappa : 0.9536         
##                                          
##  Mcnemar's Test P-Value : 0.01088        
##                                          
##             Sensitivity : 0.9631         
##             Specificity : 0.9876         
##          Pos Pred Value : 0.9808         
##          Neg Pred Value : 0.9759         
##              Prevalence : 0.3971         
##          Detection Rate : 0.3825         
##    Detection Prevalence : 0.3900         
##       Balanced Accuracy : 0.9753         
##                                          
##        'Positive' Class : 1              
## 
```

The accuracy is 97.78% for the gaussian svm classifier of the log transformed train data.

Confusion Matrix for Gaussian SVM Log transformed test data

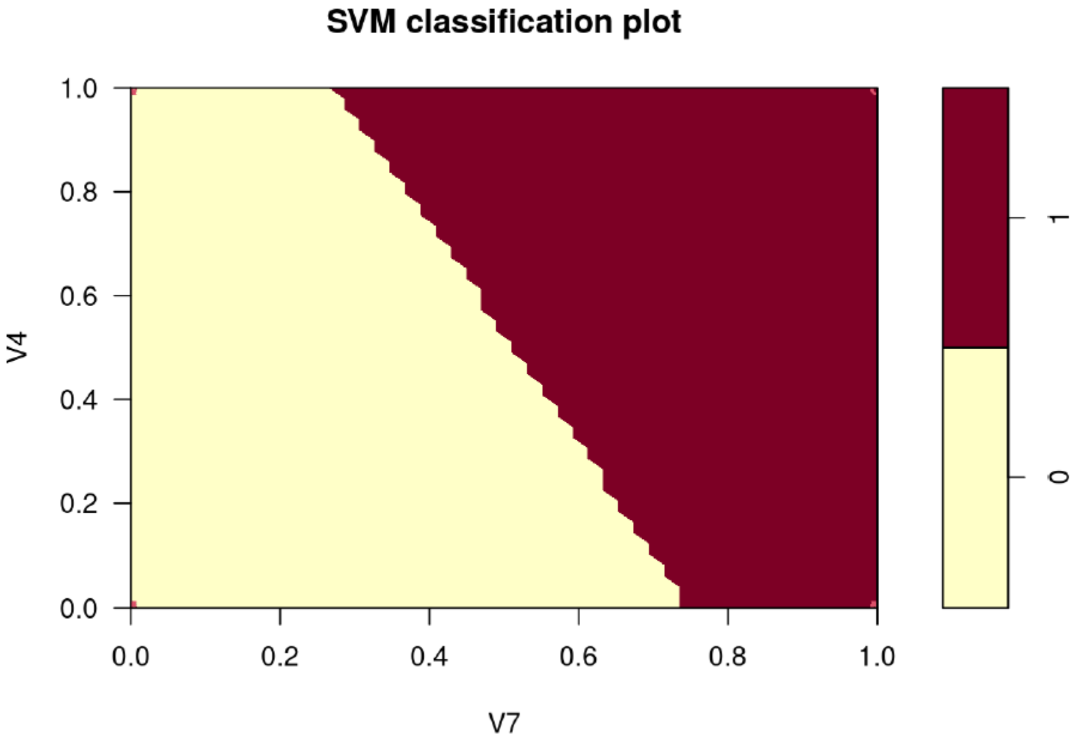


```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 892  34
##          1  24 584
##
##                Accuracy : 0.9622
##                  95% CI : (0.9514, 0.9712)
##     No Information Rate : 0.5971
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9212
##
##  Mcnemar's Test P-Value : 0.2373
##
##             Sensitivity : 0.9450
##             Specificity : 0.9738
##          Pos Pred Value : 0.9605
##          Neg Pred Value : 0.9633
##              Prevalence : 0.4029
##          Detection Rate : 0.3807
##    Detection Prevalence : 0.3963
##       Balanced Accuracy : 0.9594
##
##        'Positive' Class : 1
##
```

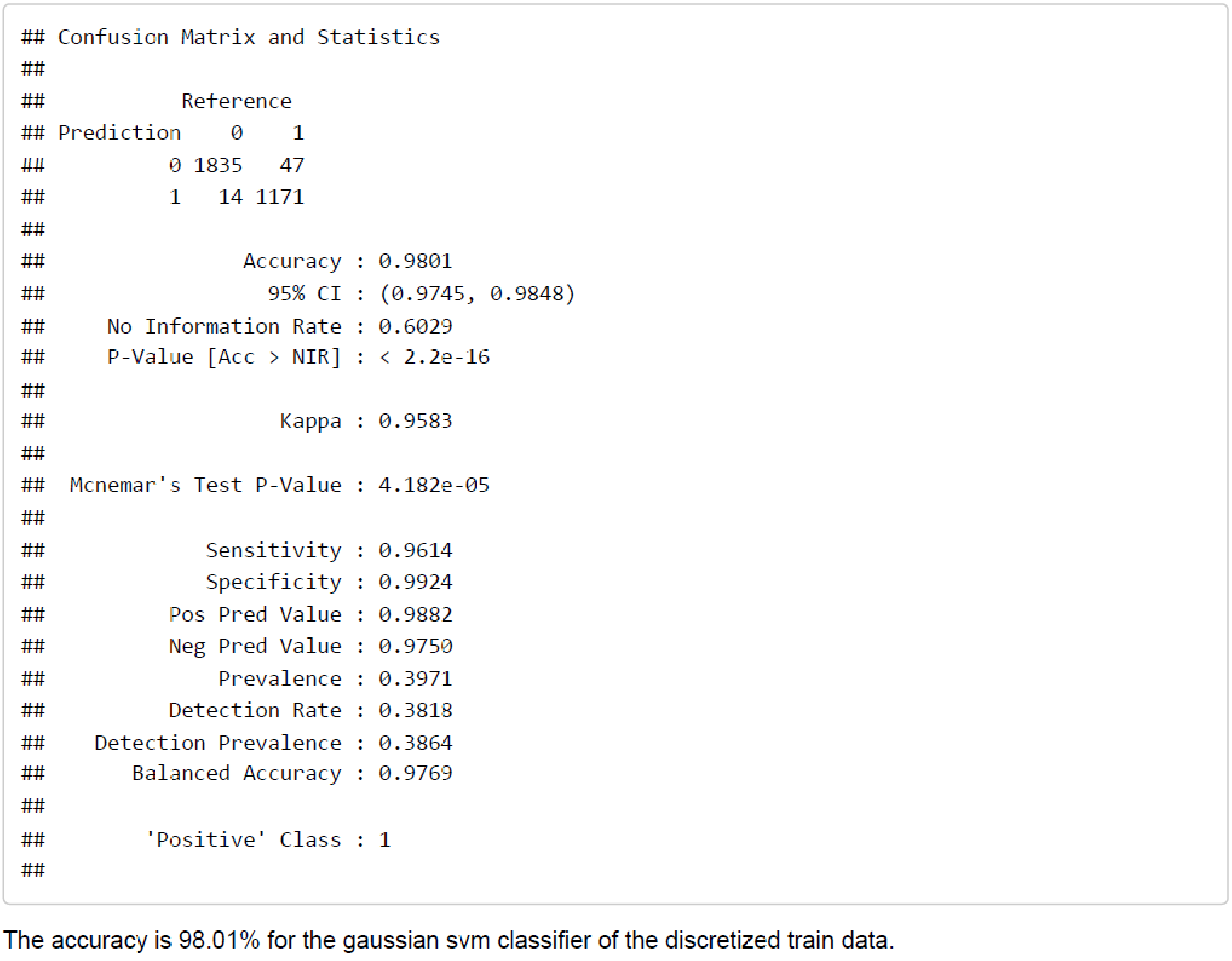
The accuracy is 96.22% for the gaussian svm classifier of the log transform test data.

Gaussian kernel for discretized train and test data

```
##
## Parameter tuning of 'svm':
##
## - sampling method: 10-fold cross validation
##
## - best parameters:
##  cost gamma
##    10   0.1
##
## - best performance: 0.04760597
##
## - Detailed performance results:
##      cost gamma      error  dispersion
## 1  1e-03 0.001 0.39715463 0.032287980
## 2  1e-02 0.001 0.39715463 0.032287980
## 3  1e-01 0.001 0.34173426 0.037491000
## 4  1e+00 0.001 0.10824977 0.021114051
## 5  5e+00 0.001 0.07532094 0.014461172
## 6  1e+01 0.001 0.07075749 0.016062712
## 7  1e-03 0.010 0.39715463 0.032287980
## 8  1e-02 0.010 0.37955015 0.036116047
## 9  1e-01 0.010 0.10759831 0.021679645
## 10 1e+00 0.010 0.07010496 0.016364826
```



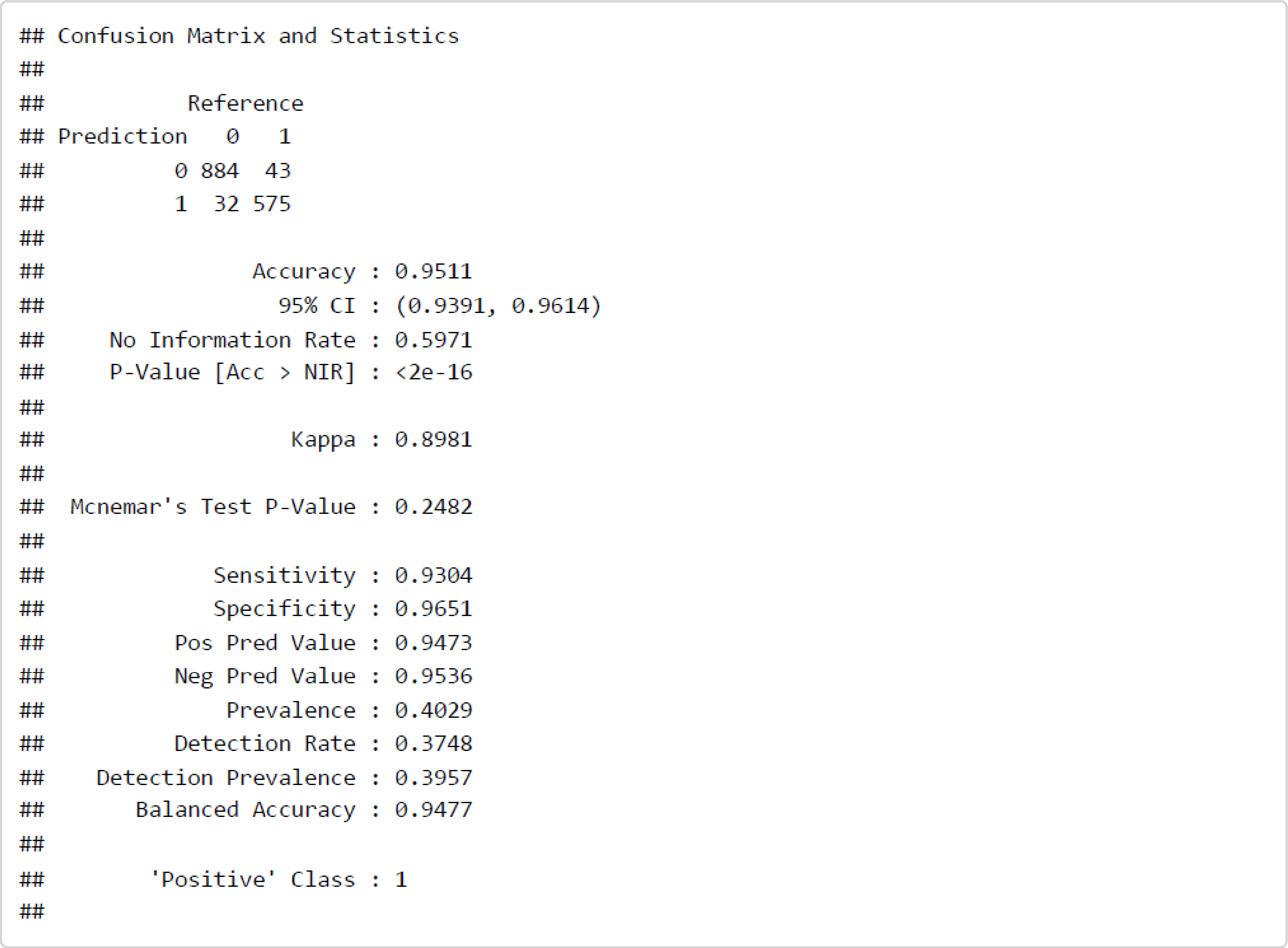
Confusion Matrix for Gaussian SVM Discretized train data



```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1835   47
##          1   14 1171
##
##                Accuracy : 0.9801
##                  95% CI : (0.9745, 0.9848)
##     No Information Rate : 0.6029
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9583
##
##  Mcnemar's Test P-Value : 4.182e-05
##
##             Sensitivity : 0.9614
##             Specificity : 0.9924
##          Pos Pred Value : 0.9882
##          Neg Pred Value : 0.9750
##              Prevalence : 0.3971
##          Detection Rate : 0.3818
##    Detection Prevalence : 0.3864
##       Balanced Accuracy : 0.9769
##
##        'Positive' Class : 1
##
```

The accuracy is 98.01% for the gaussian svm classifier of the discretized train data.

Confusion Matrix for Gaussian SVM Discretized test data



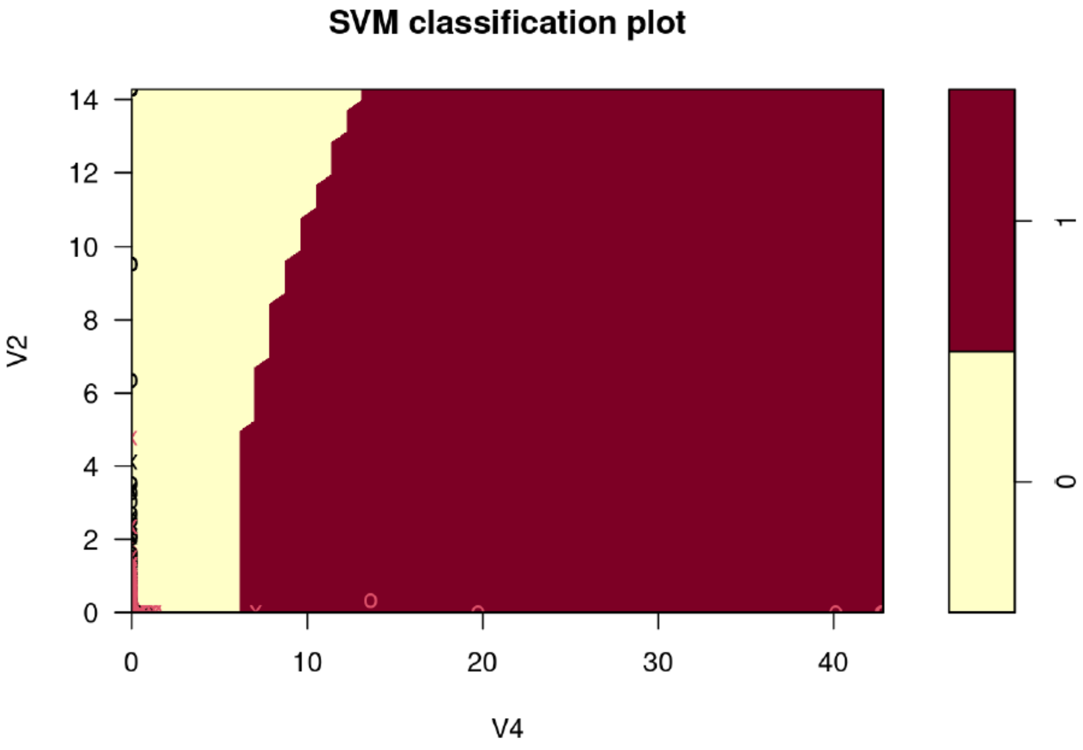
```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction   0   1
##          0 884  43
##          1  32 575
##                                           
##                Accuracy : 0.9511          
##                  95% CI : (0.9391, 0.9614)
##     No Information Rate : 0.5971          
##     P-Value [Acc > NIR] : <2e-16          
##                                           
##                   Kappa : 0.8981          
##                                           
##  Mcnemar's Test P-Value : 0.2482          
##                                           
##             Sensitivity : 0.9304          
##             Specificity : 0.9651          
##          Pos Pred Value : 0.9473          
##          Neg Pred Value : 0.9536          
##              Prevalence : 0.4029          
##          Detection Rate : 0.3748          
##    Detection Prevalence : 0.3957          
##       Balanced Accuracy : 0.9477          
##                                           
##        'Positive' Class : 1               
## 
```

The accuracy is 95.11% for the gaussian svm classifier of the discretized test data.

Polynomial SVM for original data

Polynomial SVM Parameter tuning for Original data

```
##
## Parameter tuning of 'svm':
##
## - sampling method: 10-fold cross validation
##
## - best parameters:
##  cost degree
##    10      2
##
## - best performance: 0.07727747
##
## - Detailed performance results:
##      cost degree      error dispersion
## 1  1e-03      2 0.39648187 0.03019684
## 2  1e-02      2 0.37268527 0.02878043
## 3  1e-01      2 0.28268293 0.01654857
## 4  1e+00      2 0.15944200 0.02207160
## 5  5e+00      2 0.08477997 0.02268680
## 6  1e+01      2 0.07727747 0.02105691
## 7  1e-03      3 0.39224628 0.03142381
## 8  1e-02      3 0.36714462 0.02979141
## 9  1e-01      3 0.30289966 0.02241404
## 10 1e+00      3 0.23280322 0.01616452
```

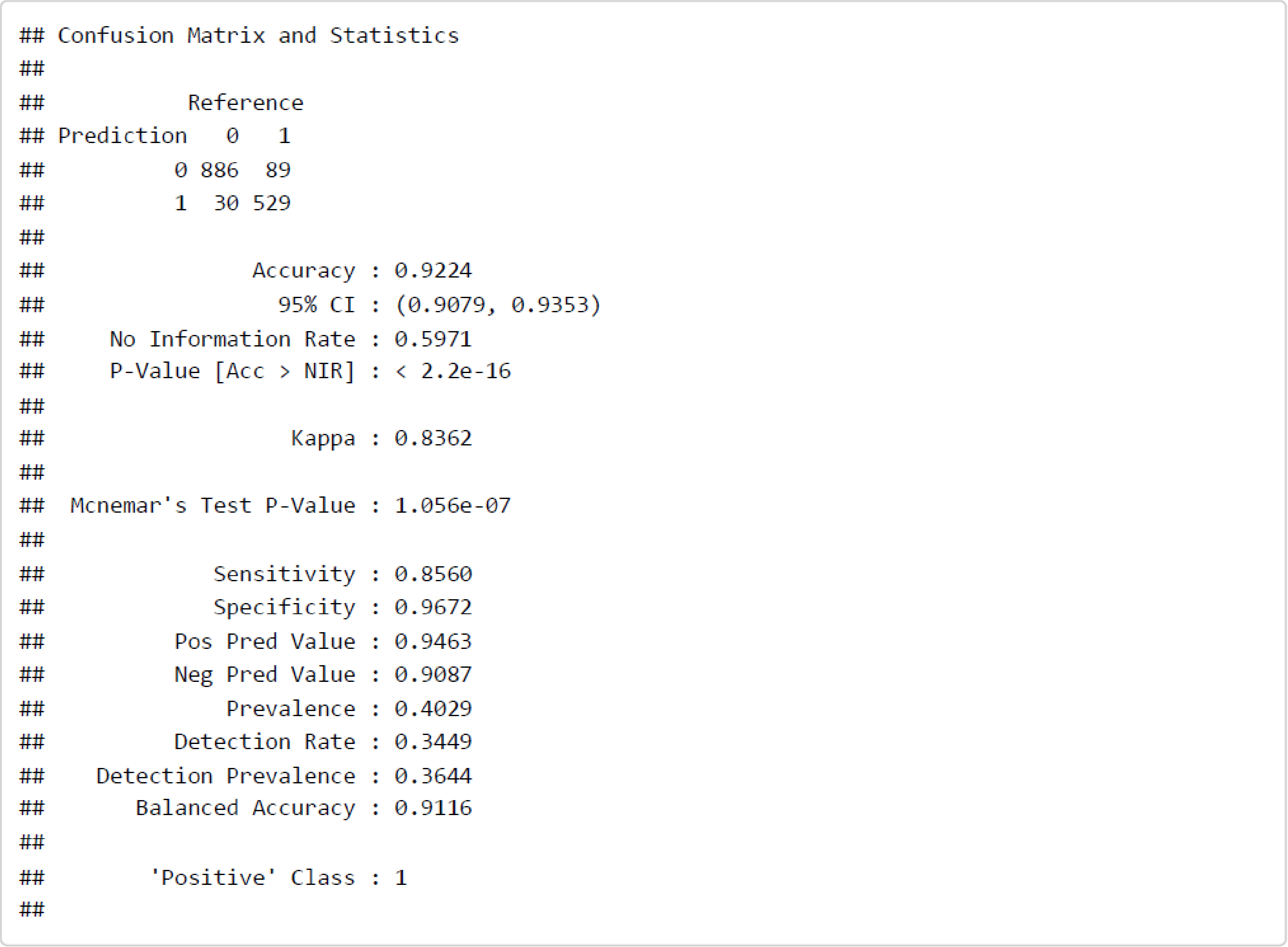


Confusion Matrix for Polynomial SVM Original train data

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction    0    1
##          0 1815  105
##          1   34 1113
##                                           
##                Accuracy : 0.9547          
##                  95% CI : (0.9467, 0.9618)
##     No Information Rate : 0.6029          
##     P-Value [Acc > NIR] : < 2.2e-16       
##                                           
##                   Kappa : 0.9044          
##                                           
##  Mcnemar's Test P-Value : 2.897e-09       
##                                           
##             Sensitivity : 0.9138          
##             Specificity : 0.9816          
##          Pos Pred Value : 0.9704          
##          Neg Pred Value : 0.9453          
##              Prevalence : 0.3971          
##          Detection Rate : 0.3629          
##    Detection Prevalence : 0.3740          
##       Balanced Accuracy : 0.9477          
##                                           
##        'Positive' Class : 1               
## 
```

The accuracy is 95.47% for the polynomial svm classifier of the train data.

Confusion Matrix for Polynomial SVM Original test data



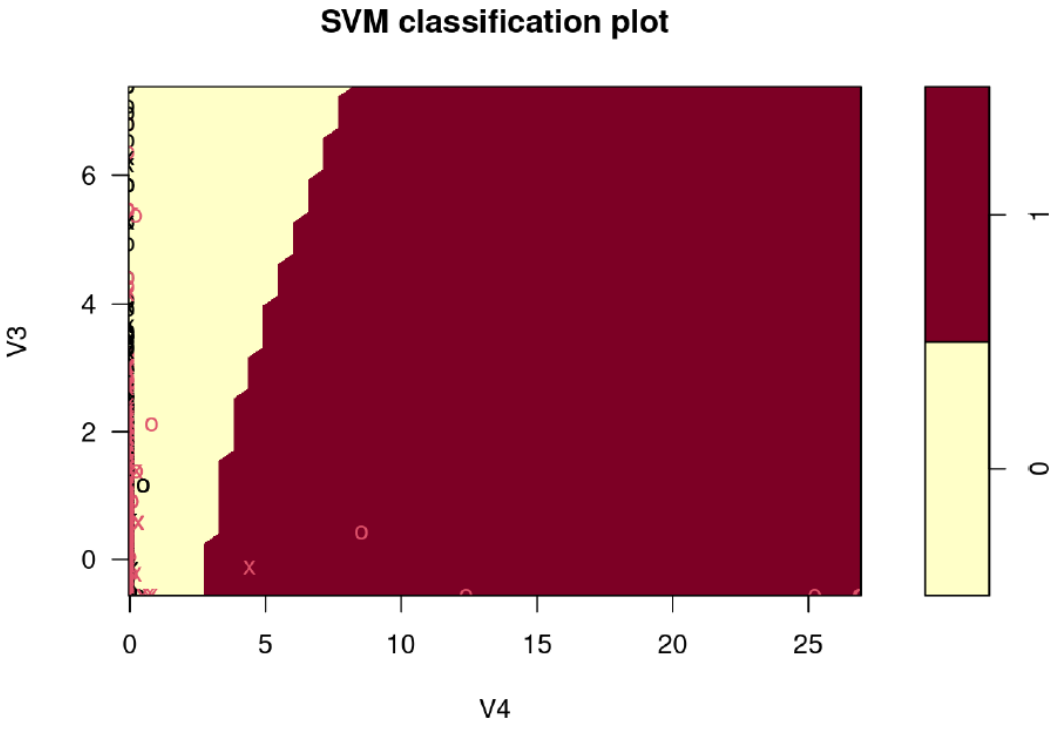
```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction   0   1
##          0 886  89
##          1  30 529
##                                           
##                Accuracy : 0.9224          
##                  95% CI : (0.9079, 0.9353)
##     No Information Rate : 0.5971          
##     P-Value [Acc > NIR] : < 2.2e-16       
##                                           
##                   Kappa : 0.8362          
##                                           
##  Mcnemar's Test P-Value : 1.056e-07       
##                                           
##             Sensitivity : 0.8560          
##             Specificity : 0.9672          
##          Pos Pred Value : 0.9463          
##          Neg Pred Value : 0.9087          
##              Prevalence : 0.4029          
##          Detection Rate : 0.3449          
##    Detection Prevalence : 0.3644          
##       Balanced Accuracy : 0.9116          
##                                           
##        'Positive' Class : 1               
## 
```

The accuracy is 92.24% for the polynomial svm classifier of the test data.

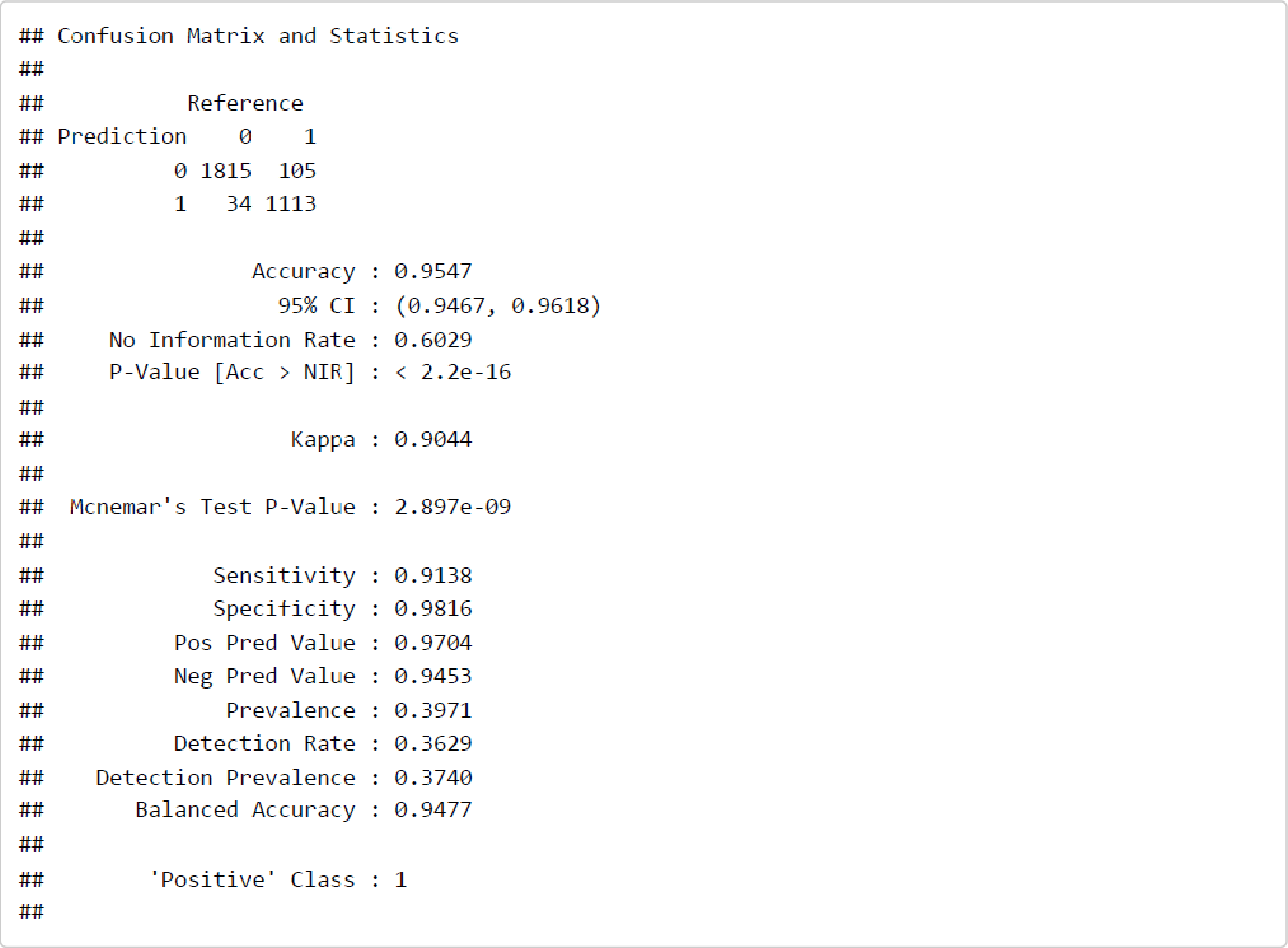
Polynomial SVM for standardized data

Polynomial SVM Parameter tuning for standardized data

```
##
## Parameter tuning of 'svm':
##
## - sampling method: 10-fold cross validation
##
## - best parameters:
##  cost degree
##    10      2
##
## - best performance: 0.07597347
##
## - Detailed performance results:
##      cost degree      error  dispersion
## 1   1e-03      2 0.39680654 0.019944036
## 2   1e-02      2 0.37430862 0.022744849
## 3   1e-01      2 0.28432011 0.017143964
## 4   1e+00      2 0.15715548 0.028329825
## 5   5e+00      2 0.08379958 0.010700926
## 6   1e+01      2 0.07597347 0.009386065
## 7   1e-03      3 0.39321922 0.021701704
## 8   1e-02      3 0.36745864 0.022015277
## 9   1e-01      3 0.30192140 0.018837322
## 10 1e+00      3 0.23215069 0.024551946
```



Confusion Matrix for Polynomial SVM standardized train data



```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction    0    1
##          0 1815  105
##          1   34 1113
## 
##                Accuracy : 0.9547
##                  95% CI : (0.9467, 0.9618)
##     No Information Rate : 0.6029
##     P-Value [Acc > NIR] : < 2.2e-16
## 
##                   Kappa : 0.9044
## 
##  Mcnemar's Test P-Value : 2.897e-09
## 
##             Sensitivity : 0.9138
##             Specificity : 0.9816
##          Pos Pred Value : 0.9704
##          Neg Pred Value : 0.9453
##              Prevalence : 0.3971
##          Detection Rate : 0.3629
##    Detection Prevalence : 0.3740
##       Balanced Accuracy : 0.9477
## 
##        'Positive' Class : 1
## 
```

The accuracy is 95.47% for the polynomial svm classifier of the standardized train data.

Confusion Matrix for Polynomial SVM standardized test data

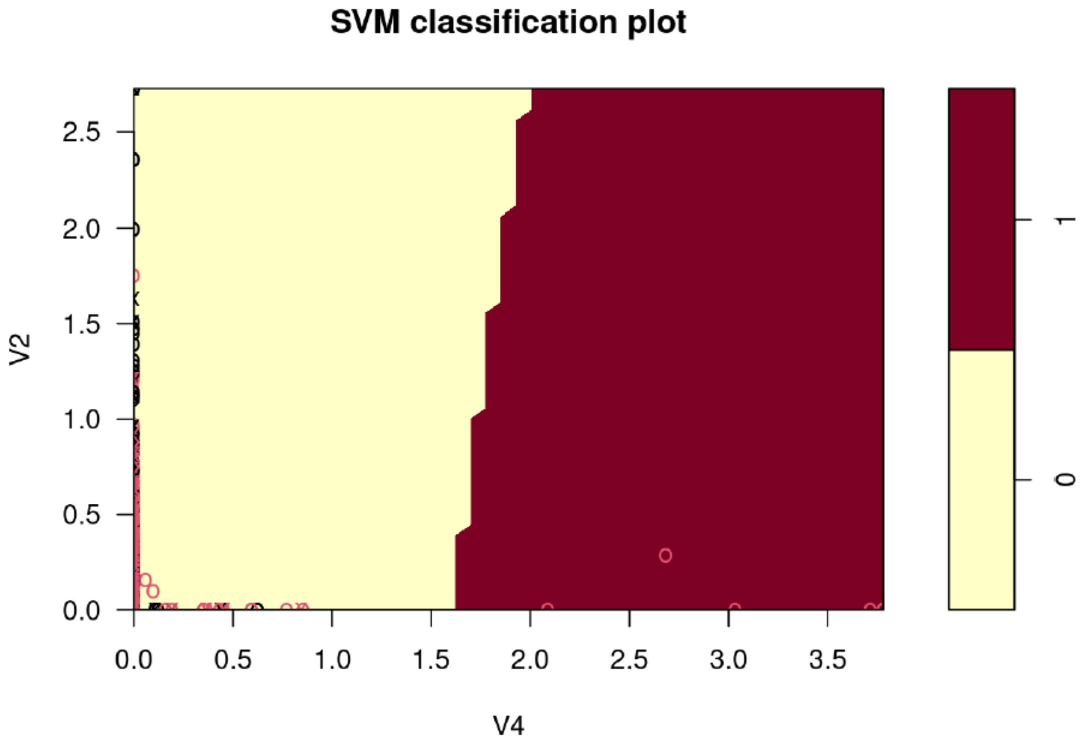
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 884  90
##          1  32 528
##
##                Accuracy : 0.9205
##                  95% CI : (0.9058, 0.9335)
##     No Information Rate : 0.5971
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8321
##
##  Mcnemar's Test P-Value : 2.462e-07
##
##             Sensitivity : 0.8544
##             Specificity : 0.9651
##          Pos Pred Value : 0.9429
##          Neg Pred Value : 0.9076
##              Prevalence : 0.4029
##          Detection Rate : 0.3442
##    Detection Prevalence : 0.3651
##       Balanced Accuracy : 0.9097
##
##        'Positive' Class : 1
##
```

The accuracy is 92.05% for the polynomial svm classifier of the standardized test data.

Polynomial SVM for the log transformed data

Polynomial SVM Parameter turning for Log transformed data

```
##
## Parameter tuning of 'svm':
##
## - sampling method: 10-fold cross validation
##
## - best parameters:
##  cost degree
##     5      2
##
## - best performance: 0.05673501
##
## - Detailed performance results:
##     cost degree      error dispersion
## 1  1e-03      2 0.39714185 0.03020173
## 2  1e-02      2 0.36127717 0.03201246
## 3  1e-01      2 0.20282515 0.03251248
## 4  1e+00      2 0.07435120 0.01789309
## 5  5e+00      2 0.05673501 0.01203088
## 6  1e+01      2 0.05868728 0.01159128
## 7  1e-03      3 0.39551212 0.03025857
## 8  1e-02      3 0.35214707 0.03316304
## 9  1e-01      3 0.24521726 0.03498523
## 10 1e+00      3 0.10336271 0.02053258
```

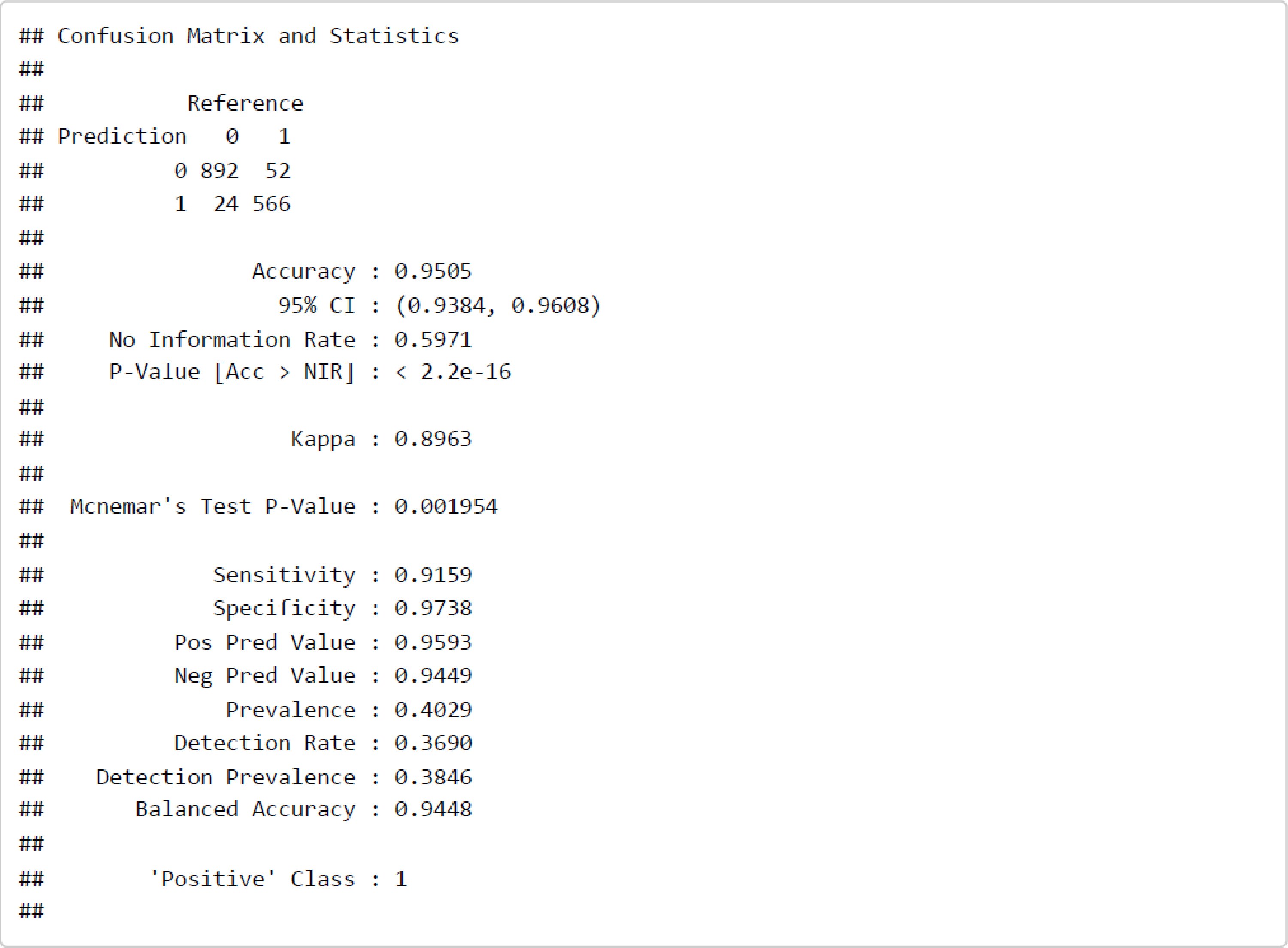


Confusion Matrix for Polynomial SVM Log transformed train data

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction    0    1
##          0 1827   48
##          1   22 1170
## 
##                Accuracy : 0.9772
##                  95% CI : (0.9713, 0.9822)
##     No Information Rate : 0.6029
##     P-Value [Acc > NIR] : < 2.2e-16
## 
##                   Kappa : 0.9522
## 
##  Mcnemar's Test P-Value : 0.002807
## 
##             Sensitivity : 0.9606
##             Specificity : 0.9881
##          Pos Pred Value : 0.9815
##          Neg Pred Value : 0.9744
##              Prevalence : 0.3971
##          Detection Rate : 0.3815
##    Detection Prevalence : 0.3887
##       Balanced Accuracy : 0.9743
## 
##        'Positive' Class : 1
## 
```

The accuracy is 97.72% for the polynomial svm classifier of the log transformed train data.

Confusion Matrix for Polynomial SVM Log transformed test data



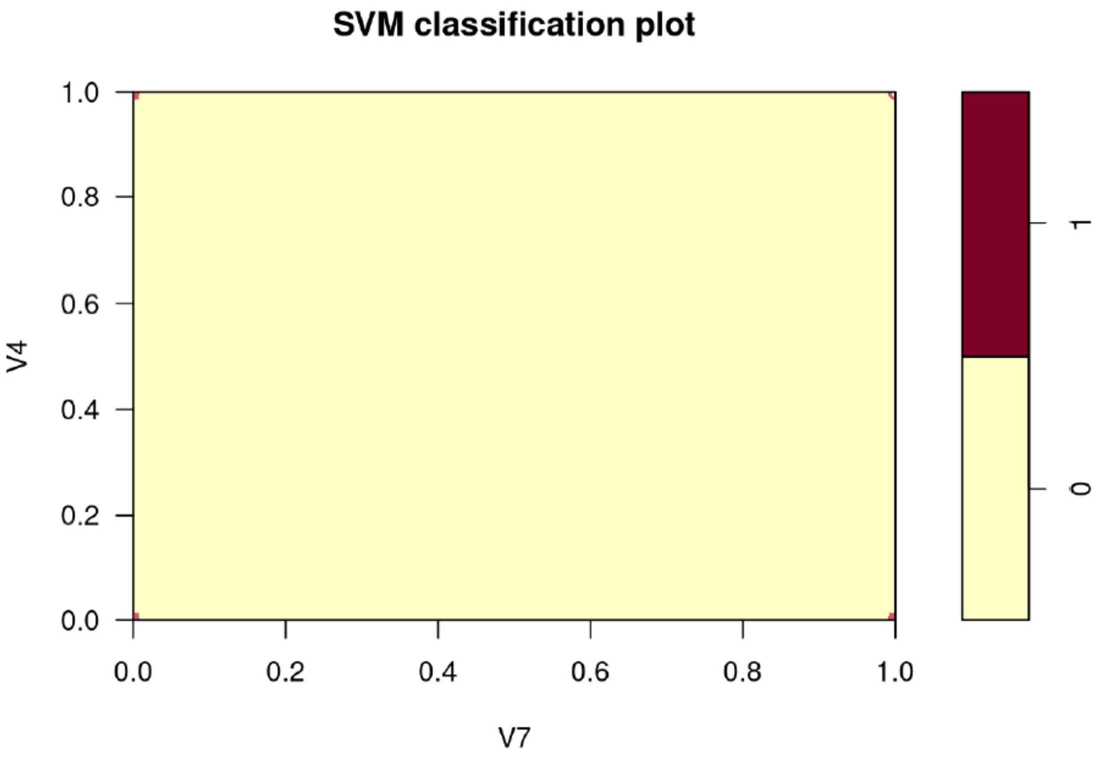
```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction   0   1
##          0 892  52
##          1  24 566
## 
##                Accuracy : 0.9505
##                  95% CI : (0.9384, 0.9608)
##     No Information Rate : 0.5971
##     P-Value [Acc > NIR] : < 2.2e-16
## 
##                   Kappa : 0.8963
## 
##  Mcnemar's Test P-Value : 0.001954
## 
##             Sensitivity : 0.9159
##             Specificity : 0.9738
##          Pos Pred Value : 0.9593
##          Neg Pred Value : 0.9449
##              Prevalence : 0.4029
##          Detection Rate : 0.3690
##    Detection Prevalence : 0.3846
##       Balanced Accuracy : 0.9448
## 
##        'Positive' Class : 1
## 
```

The accuracy is 95.05% for the polynomial svm classifier of the log transformed test data.

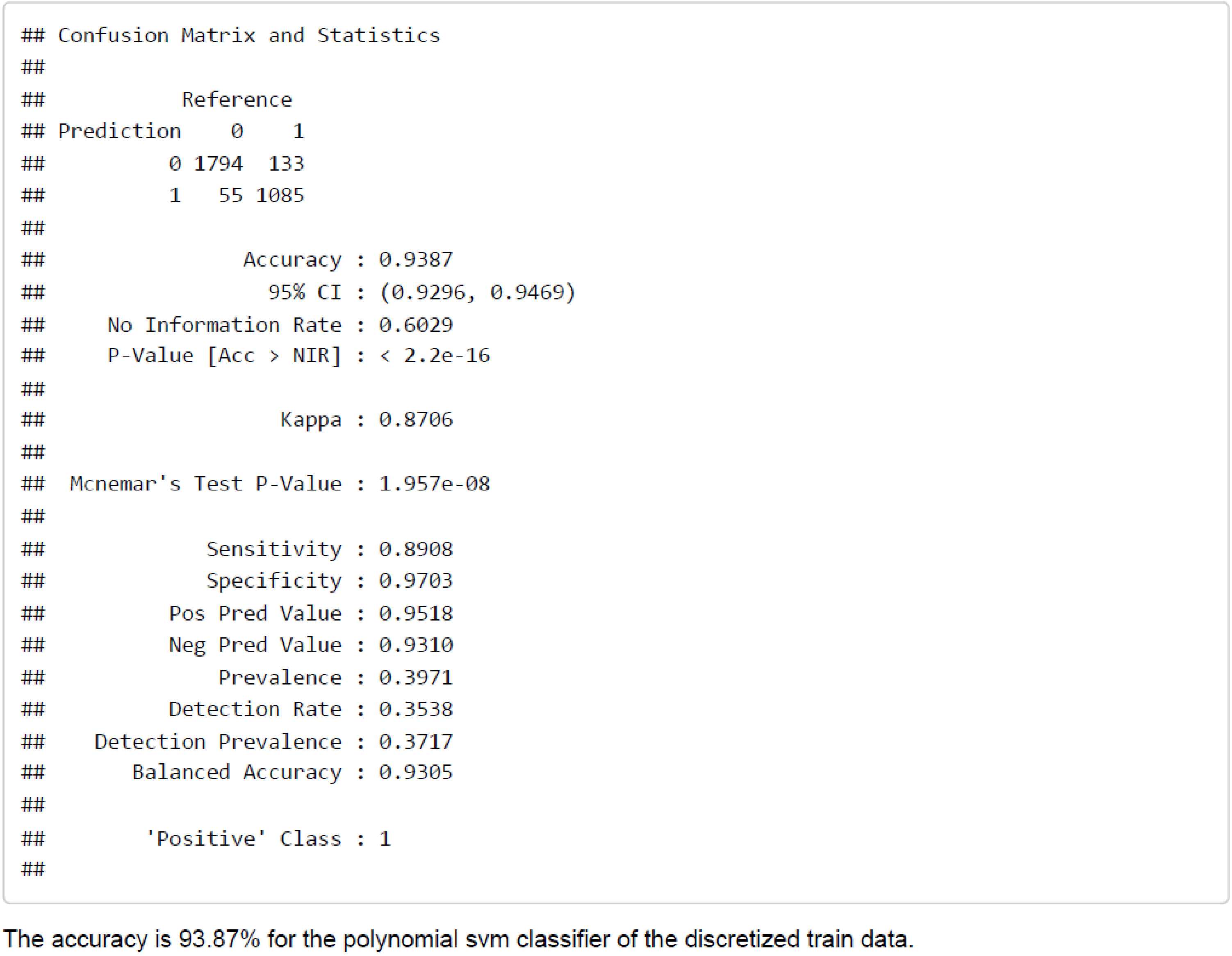
Polynomial SVM of the discretized data

Polynomial SVM Parameter tuning for Discretized data

```
##
## Parameter tuning of 'svm':
##
## - sampling method: 10-fold cross validation
##
## - best parameters:
##  cost degree
##    10      2
##
## - best performance: 0.06618871
##
## - Detailed performance results:
##       cost degree      error  dispersion
## 1  1e-03      2 0.39712908 0.022693219
## 2  1e-02      2 0.39712908 0.022693219
## 3  1e-01      2 0.19824360 0.024454789
## 4  1e+00      2 0.10988269 0.016296517
## 5  5e+00      2 0.07825147 0.011902827
## 6  1e+01      2 0.06618871 0.009971212
## 7  1e-03      3 0.39712908 0.022693219
## 8  1e-02      3 0.39712908 0.022693219
## 9  1e-01      3 0.32735943 0.016134043
## 10 1e+00      3 0.17313768 0.021098926
```



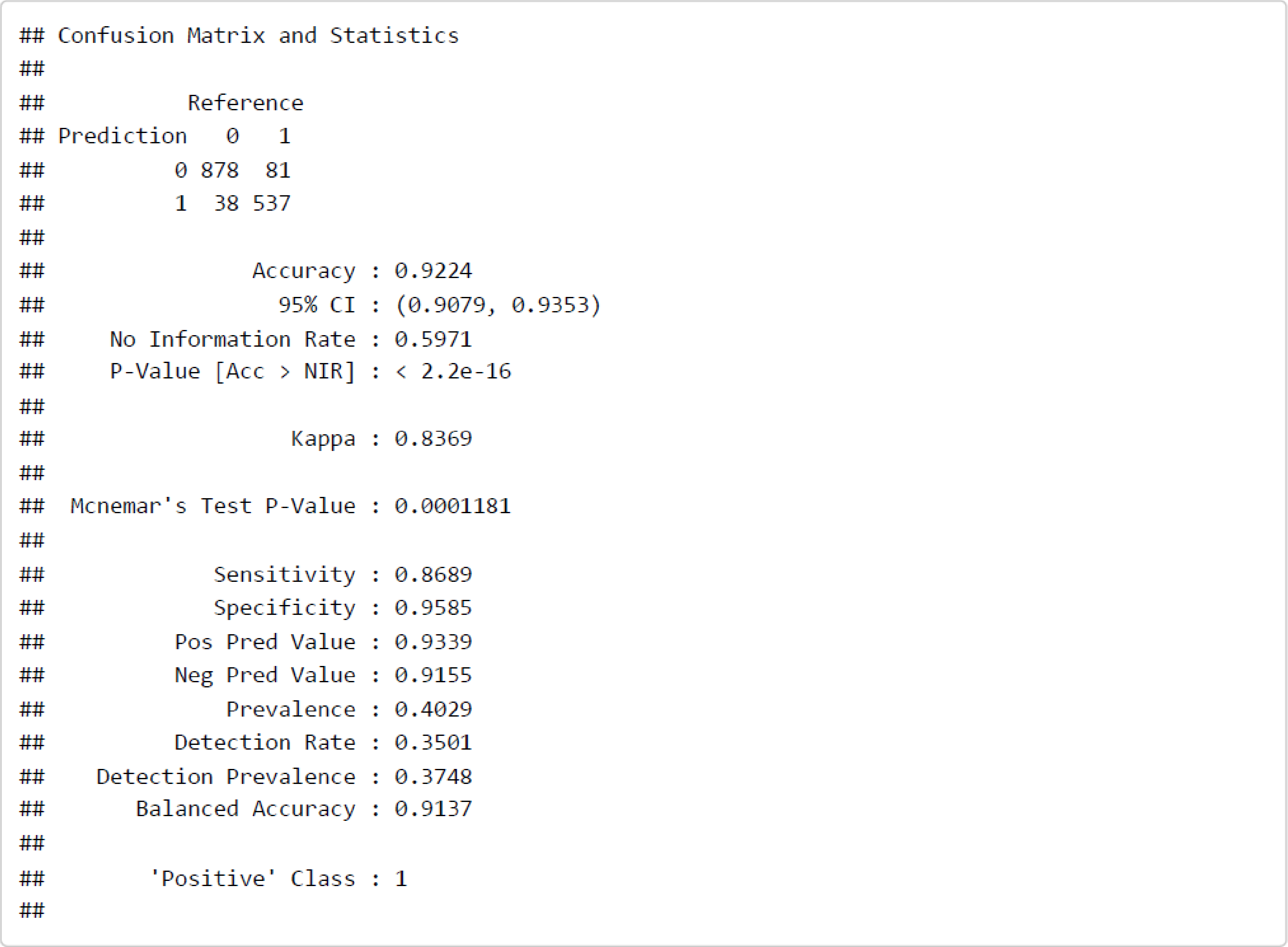
Confusion Matrix for Polynomial SVM Discretized train data



```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction    0    1
##          0 1794  133
##          1   55 1085
##                                           
##                Accuracy : 0.9387          
##                  95% CI : (0.9296, 0.9469)
##     No Information Rate : 0.6029          
##     P-Value [Acc > NIR] : < 2.2e-16       
##                                           
##                   Kappa : 0.8706          
##                                           
##  Mcnemar's Test P-Value : 1.957e-08       
##                                           
##             Sensitivity : 0.8908          
##             Specificity : 0.9703          
##          Pos Pred Value : 0.9518          
##          Neg Pred Value : 0.9310          
##              Prevalence : 0.3971          
##          Detection Rate : 0.3538          
##    Detection Prevalence : 0.3717          
##       Balanced Accuracy : 0.9305          
##                                           
##        'Positive' Class : 1               
## 
```

The accuracy is 93.87% for the polynomial svm classifier of the discretized train data.

Confusion Matrix for Polynomial SVM Discretized test data



```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 878  81
##          1  38 537
##
##                Accuracy : 0.9224
##                  95% CI : (0.9079, 0.9353)
##     No Information Rate : 0.5971
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8369
##
##  Mcnemar's Test P-Value : 0.0001181
##
##             Sensitivity : 0.8689
##             Specificity : 0.9585
##          Pos Pred Value : 0.9339
##          Neg Pred Value : 0.9155
##              Prevalence : 0.4029
##          Detection Rate : 0.3501
##    Detection Prevalence : 0.3748
##       Balanced Accuracy : 0.9137
##
##        'Positive' Class : 1
##
```

The accuracy is 92.24% for the polynomial svm classifier of the discretized test data.

7. A report of classification errors using different methods and different preprocessed data.

For Logistic Regression table,

```
##       lr original lr standardized    lr log      lr I
## train   0.9282687       0.5415716 0.5347245 0.9018585
## test    0.9269883       0.5521512 0.5410691 0.8970013
```

For LDA and QDA table,

```
##        lda stan   lda log  qda stan   qda log
## train 0.8982719 0.9396805 0.8213238 0.8412129
## test  0.8970013 0.9348110 0.8252934 0.8428944
```

For the SVM table,

```
##                       linear  gaussian polynomial
## train              0.9373981 0.9628301  0.9546788
## standardized train 0.9272905 0.9628301  0.9546788
## log train          0.9409847 0.9778285  0.9771764
## I train            0.9413107 0.9801109  0.9387023
## test               0.9341591 0.9452412  0.9224250
## standardized test  0.9322034 0.9413299  0.9204694
## log test           0.9419817 0.9621904  0.9504563
## I test             0.9243807 0.9511082  0.9224250
```

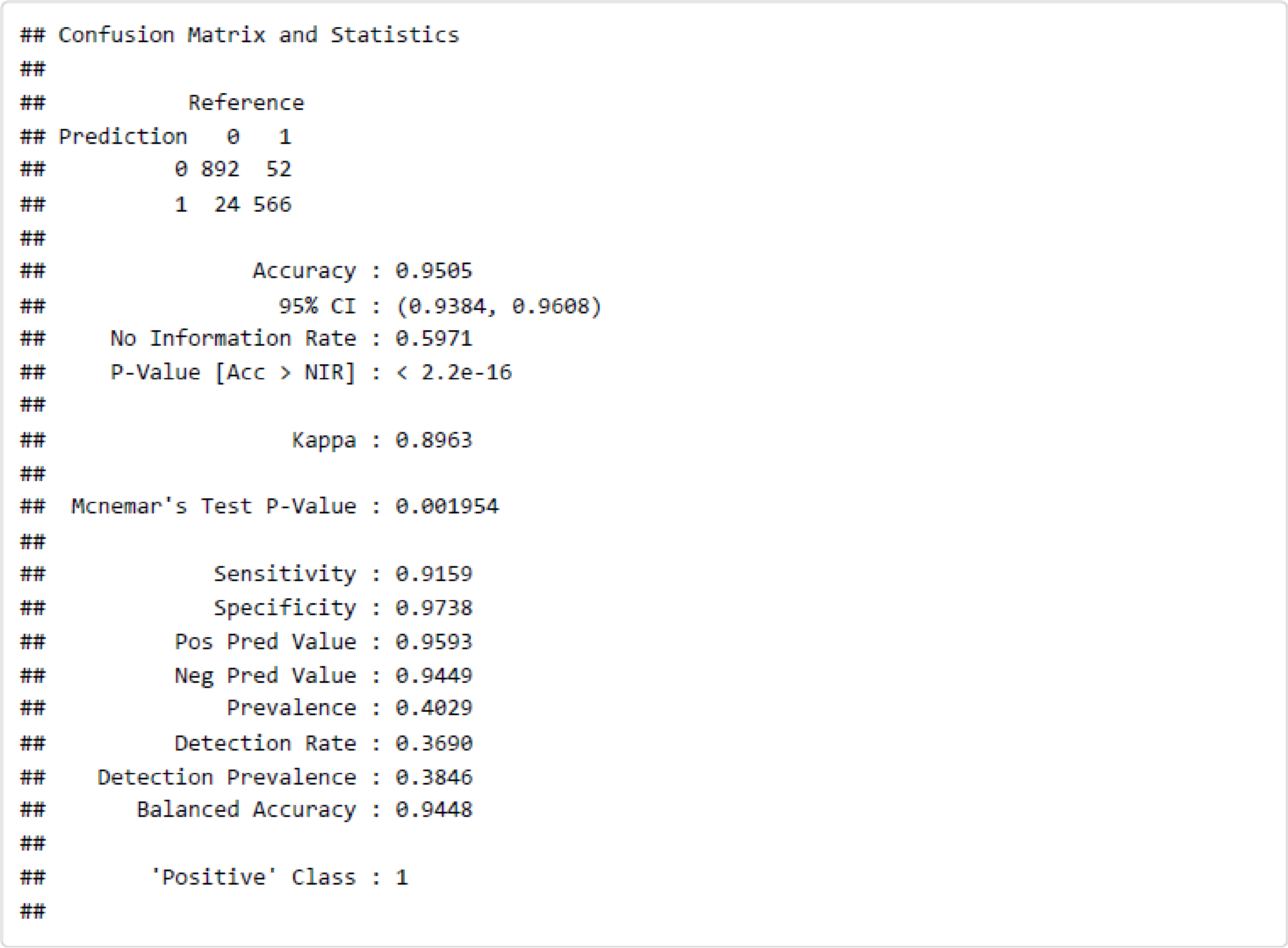
8. Designed a classifier with test error rate as small as possible using a single method with properly chosen tuning parameter and a combination of several methods.

The log transformation Gaussian SVM classifier has the best test accuracy rate of 96.22% based on the table.
Combine PCA with the Gaussian SVM classifier for the log transformed to get the smallest test error rate.
Continue single method tuning with more precise parameters.
Fine tune it even more to achieve a smaller test error.

SVM Parameter tuning for Gaussian Log transformation data

```
## 
## Parameter tuning of 'svm':
## 
## - sampling method: 10-fold cross validation 
## 
## - best parameters:
##  cost  gamma
##  9.34 0.0162
## 
## - best performance: 0.04205787 
## 
## - Detailed performance results:
##    cost  gamma      error dispersion
## 1  9.30 0.0162 0.04238466 0.01320277
## 2  9.31 0.0162 0.04238466 0.01320277
## 3  9.32 0.0162 0.04238466 0.01320277
## 4  9.33 0.0162 0.04238466 0.01320277
## 5  9.34 0.0162 0.04205787 0.01324044
## 6  9.30 0.0163 0.04205787 0.01324044
## 7  9.31 0.0163 0.04205787 0.01324044
## 8  9.32 0.0163 0.04205787 0.01324044
## 9  9.33 0.0163 0.04205787 0.01324044
## 10 9.34 0.0163 0.04205787 0.01324044
## 11 9.30 0.0164 0.04205787 0.01324044
## 12 9.31 0.0164 0.04205787 0.01324044
## 13 9.32 0.0164 0.04205787 0.01324044
## 14 9.33 0.0164 0.04205787 0.01324044
## 15 9.34 0.0164 0.04205787 0.01324044
## 16 9.30 0.0165 0.04238360 0.01372470
## 17 9.31 0.0165 0.04271040 0.01367972
## 18 9.32 0.0165 0.04271040 0.01367972
## 19 9.33 0.0165 0.04271040 0.01367972
## 20 9.34 0.0165 0.04271040 0.01367972
## 21 9.30 0.0166 0.04271040 0.01367972
## 22 9.31 0.0166 0.04271040 0.01367972
## 23 9.32 0.0166 0.04271040 0.01367972
## 24 9.33 0.0166 0.04271040 0.01367972
## 25 9.34 0.0166 0.04271040 0.01367972
```

Confusion Matrix for Gaussian Log transformation test data



```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 892  52
##          1  24 566
##
##                Accuracy : 0.9505
##                  95% CI : (0.9384, 0.9608)
##     No Information Rate : 0.5971
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8963
##
##  Mcnemar's Test P-Value : 0.001954
##
##             Sensitivity : 0.9159
##             Specificity : 0.9738
##          Pos Pred Value : 0.9593
##          Neg Pred Value : 0.9449
##              Prevalence : 0.4029
##          Detection Rate : 0.3690
##    Detection Prevalence : 0.3846
##       Balanced Accuracy : 0.9448
##
##        'Positive' Class : 1
##
```

Tuned 20 times to make the test error rate smaller. In conclusion, the optimal parameters for the gaussian classifier on the log transformation data is roughly: cost = 9.565, or cost = 9.575, or cost =9.56, gamma = 0.017, or gamma = 0.0175 for an accuracy of 96.61%, an improvement of 0.4% compared to the original tuning parameter.