Network Intrusion Detection

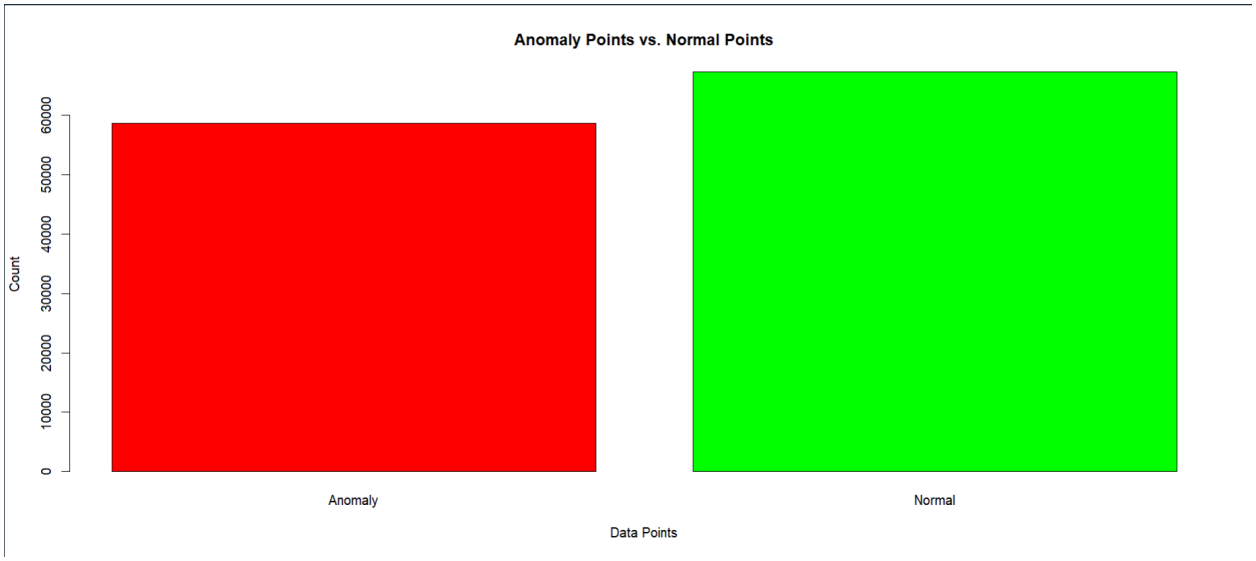
Dataset : NSL- KDD dataset.
Columns : 42 (41 features and 1 target variable [class])
Rows : 1,25,873 datapoints.

The dataset is divided into **train** and **test** as **75%** and **25%** respectively.

The Class column has **58,630 points of Anomaly** and **67,343 points of Normal.**

.For all the 3 algorithms, **K-Fold Cross Validation** with number = **10** is applied.

Seed = 123



	Accuracy	Precision	Recall	F1 score	Specificity
RandomForest	0.9988569	0.9994534	0.9980896	0.9987711	0.9995248
SVM	0.9837737	0.9822719	0.9828751	0.9825734	0.984556
XGBoost	0.9988251	0.9992487	0.9982261	0.9987372	0.9993466

RandomForest

Confusion Matrix:

Confusion Matrix and Statistics

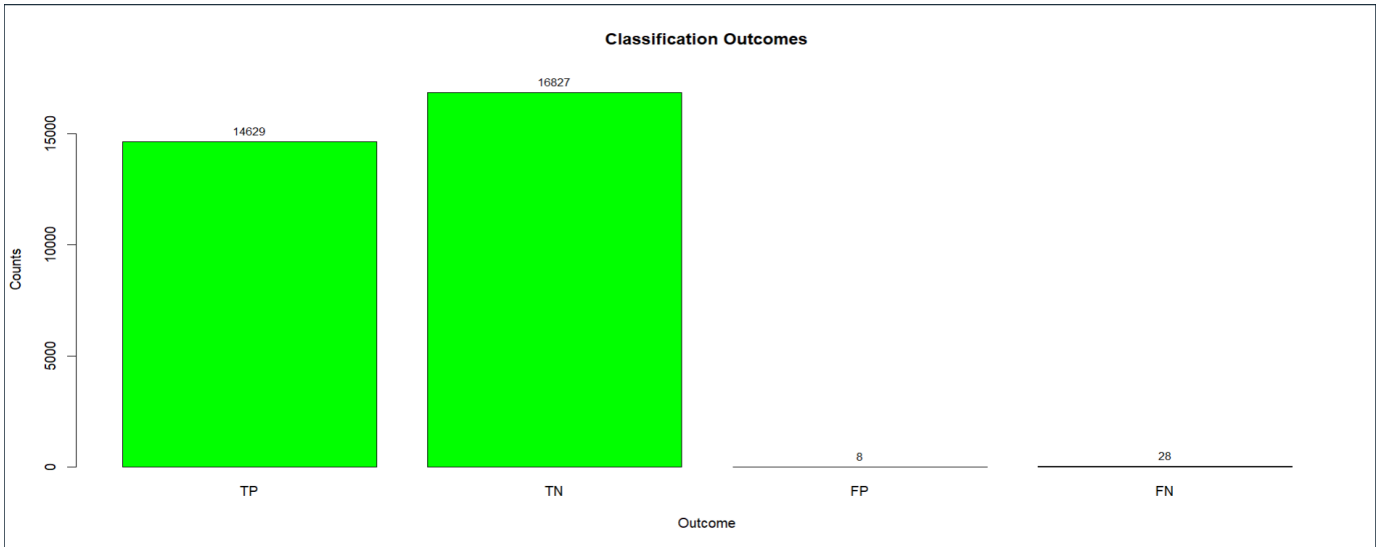
	Reference	
Prediction	anomaly	normal
anomaly	14629	8
normal	28	16827

Training Accuracy:

Resampling results across tuning parameters:

mtry	Accuracy	Kappa
25	0.9990157	0.9980218
30	0.9991109	0.9982132
35	0.9990792	0.9981494

Visualization of above Confusion Matrix(Class Anomaly is treated as positive class**):**



SVM

Confusion Matrix:

Confusion Matrix and Statistics

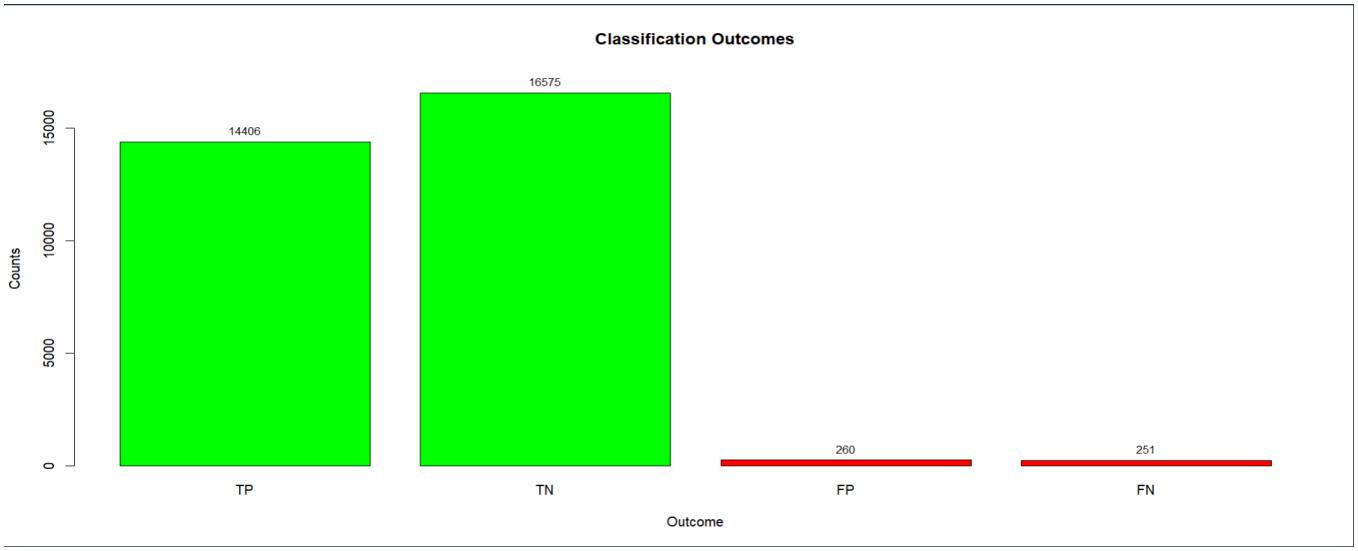
	Reference	
Prediction	anomaly	normal
anomaly	14406	260
normal	251	16575

Training Accuracy:

Resampling results across tuning parameters:

C	Accuracy	Kappa
0.25	0.9789058	0.9576195
0.50	0.9816259	0.9630862
1.00	0.9837322	0.9673157

Visualization of above Confusion Matrix(Class Anomaly is treated as positive class**):**



XGBoost

Confusion Matrix:

Confusion Matrix and Statistics

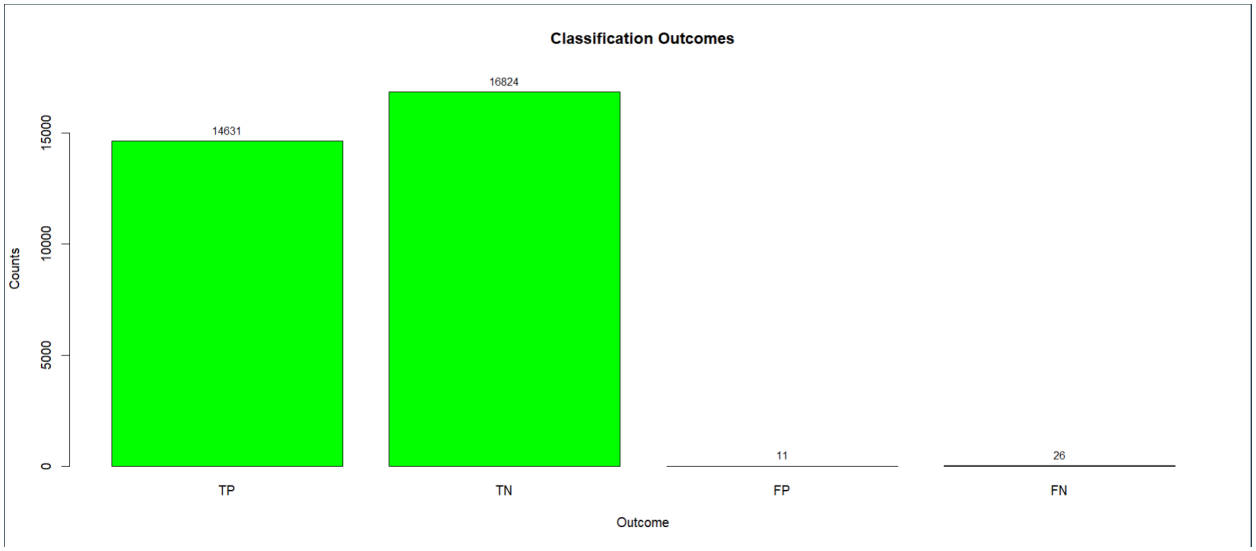
	Reference	
Prediction	anomaly	normal
anomaly	144631	11
normal	26	16824

Training Accuracy:

Resampling results:

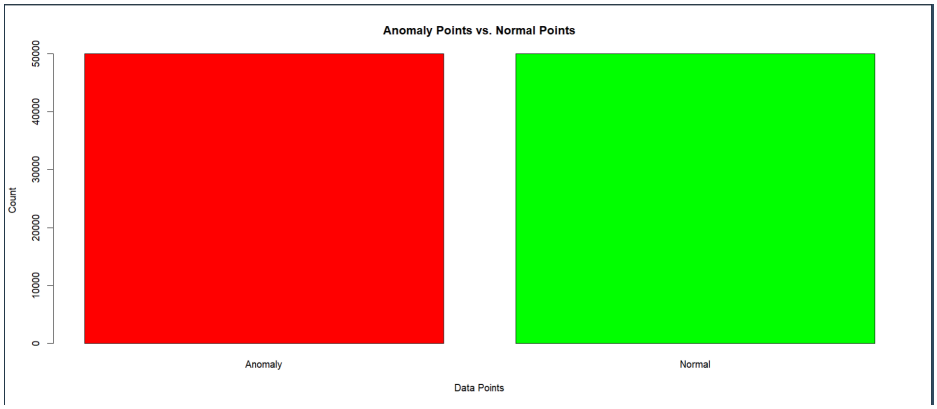
Accuracy	Kappa
0.9988146	0.9976176

Visualization of above Confusion Matrix(Class Anomaly is treated as positive class**):**



To solve the problem of **class Imbalance,** we are **under sampling 50k data points** of **each class**. We have **two** classes and hence **total 100k data points**. The dataset is divided into **train and test as 75% and 25% respectively.**

Columns : 42 (41 features + 1 target variable)
Rows : 100k
Seed = 123
The Class column has 50k points of Anomaly and 50k points of Normal.
For all the 3 algorithms, K-Fold Cross Validation with number = 10 is applied.



	Accuracy	Precision	Recall	F1 score	Specificity
RandomForest	0.9994	0.9998399	0.99896	0.9993997	0.99984
SVM	0.98344	0.9814372	0.98552	0.9834744	0.98136
XGBoost	0.9992	0.9996797	0.99872	0.9991996	0.99968

RandomForest

Confusion Matrix:

Confusion Matrix and Statistics

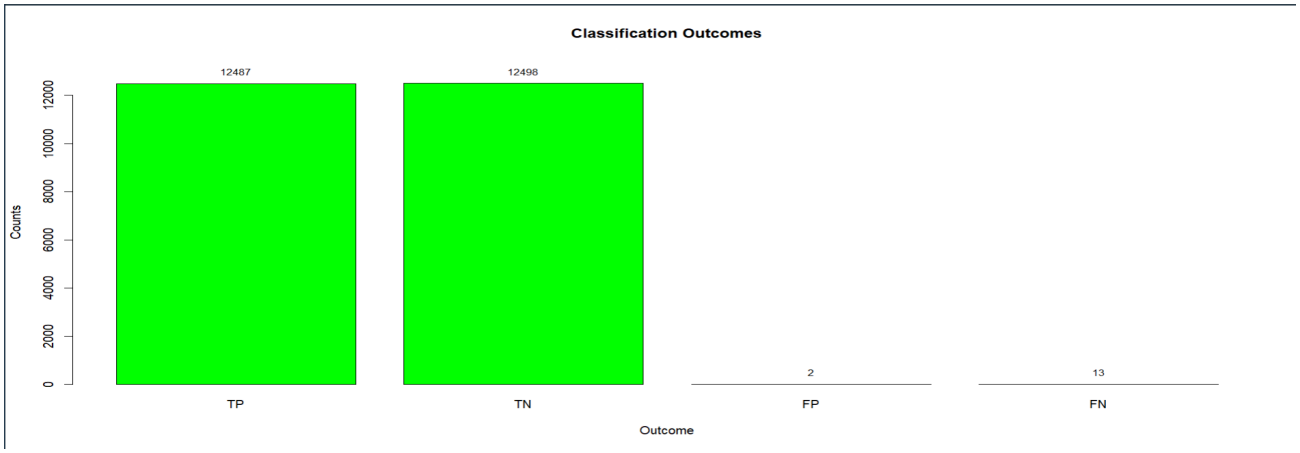
	Reference	
Prediction	anomaly	normal
anomaly	12487	2
normal	13	12498

Training Accuracy :

Resampling results across tuning parameters:

mtry	Accuracy	Kappa
25	0.9581867	0.9163733
30	0.9987067	0.9974133
35	0.9984267	0.9968533

Visualization of above Confusion Matrix(Class Anomaly is treated as positive class**):**



SVM

Confusion Matrix:

Confusion Matrix and Statistics

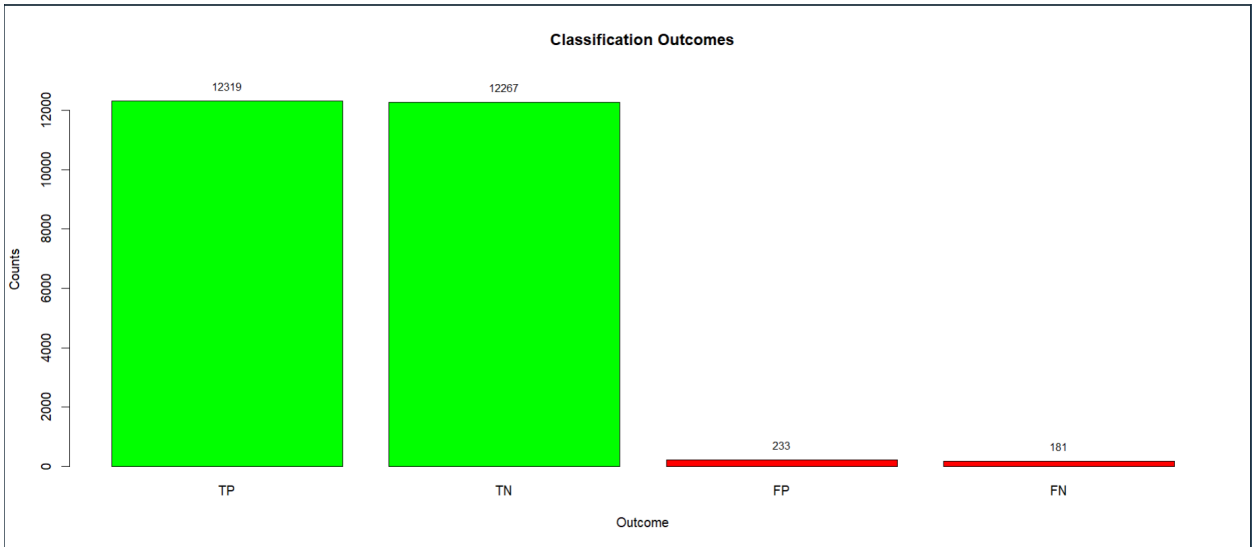
	Reference	
Prediction	anomaly	normal
anomaly	12319	233
normal	181	12267

Training Accuracy:

Resampling results across tuning parameters:

C	Accuracy	Kappa
0.25	0.9788267	0.9576533
0.50	0.9817733	0.9635467
1.00	0.9838800	0.9677600

Visualization of above Confusion Matrix(Class Anomaly is treated as positive class**):**



XGBoost

Confusion Matrix:

Confusion Matrix and Statistics

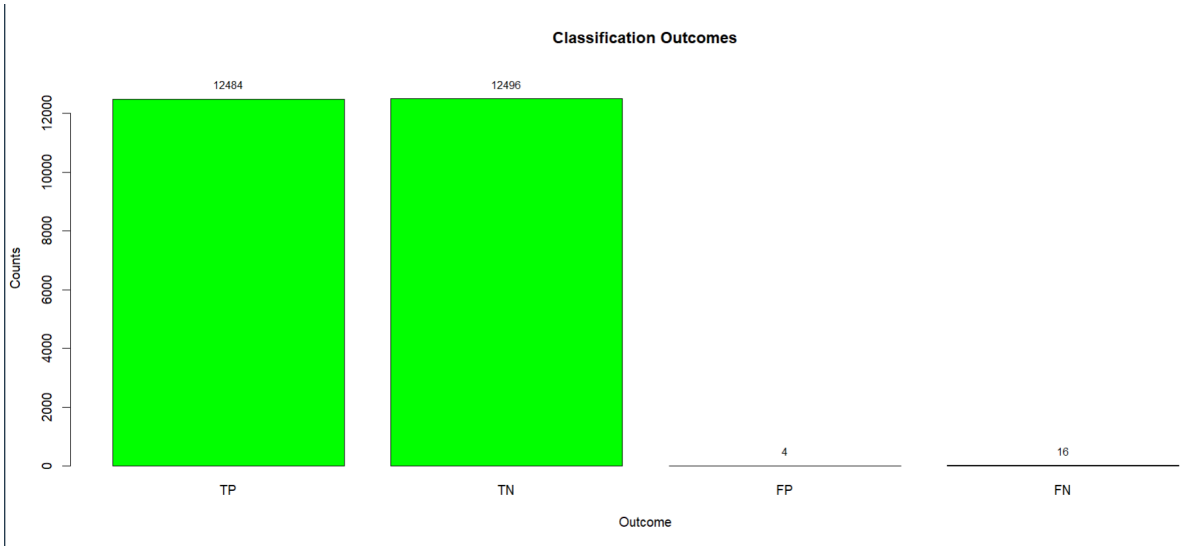
	Reference	
Prediction	anomaly	normal
anomaly	12484	4
normal	16	12496

Training Accuracy:

Resampling results:

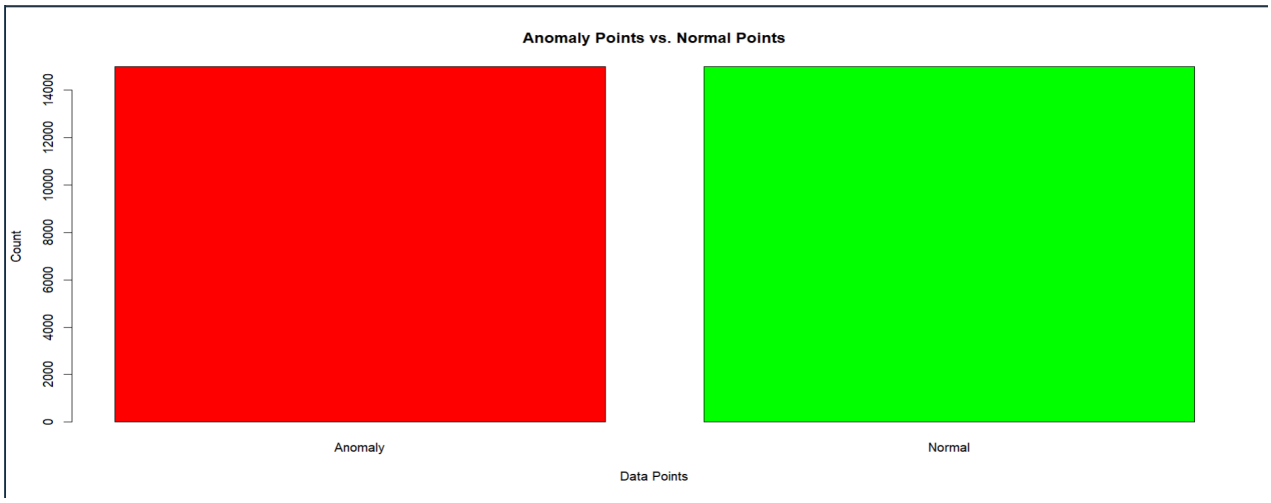
Accuracy	Kappa
0.9986	0.9972

Visualization of above Confusion Matrix(Class Anomaly is treated as positive class**):**



As the **time required for computation was significantly high** i.e. 3 hours and more, we decided to choose **15k points of Anomaly and 15k points of Normal** and divided the dataset into 75% and 25% as train and test dataset.While training the model, **K-fold cross validation** with **number = 10** is applied.

Columns : 42 (41 features + 1 target variable)
Rows : 30k
Seed = 123
The Class column has 15k points of Anomaly and 15k points of Normal.
For all the 3 algorithms, K-Fold Cross Validation with number = 10 is applied.



Seed = 123

	Accuracy	Precision	Recall	F1 Score	Specificity
RandomForest	0.9986667	0.9989328	0.9984	0.9986663	0.9989333
SVM	0.9785333	0.9736078	0.9837333	0.9786444	0.9733333
XGBoost	0.9988	0.9991994	0.9984	0.9987995	0.9992

RandomForest

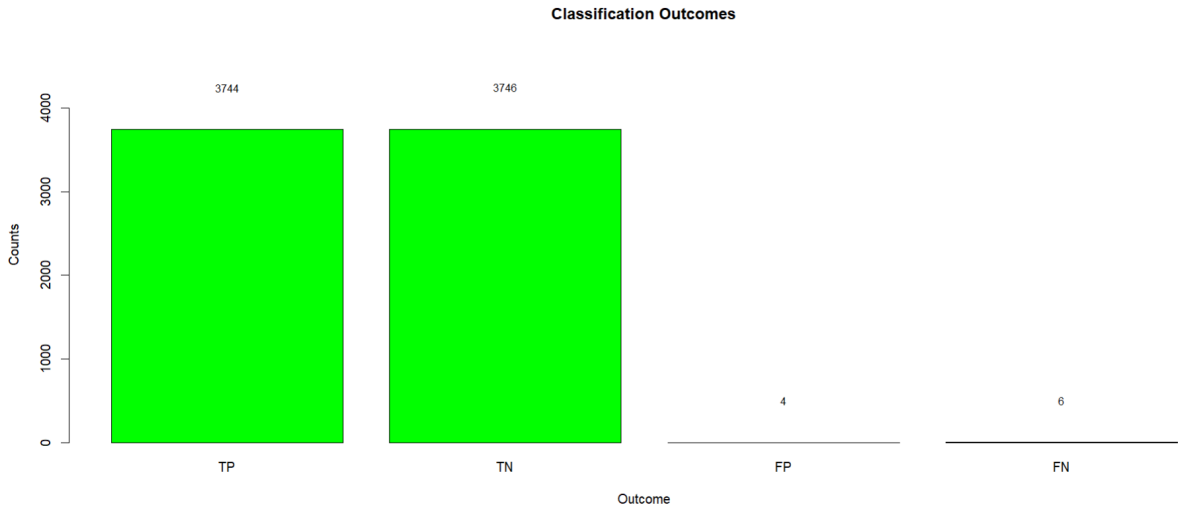
Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction anomaly normal
   anomaly    3744      4
   normal        6   3746
```

Training Accuracy:

```
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
  25    0.9974667  0.9949333
  30    0.9973778  0.9947556
  35    0.9974667  0.9949333

Accuracy was used to select the optimal model using the largest value.
```

Visualization of above Confusion Matrix(Class Anomaly is treated as positive class**):**



SVM

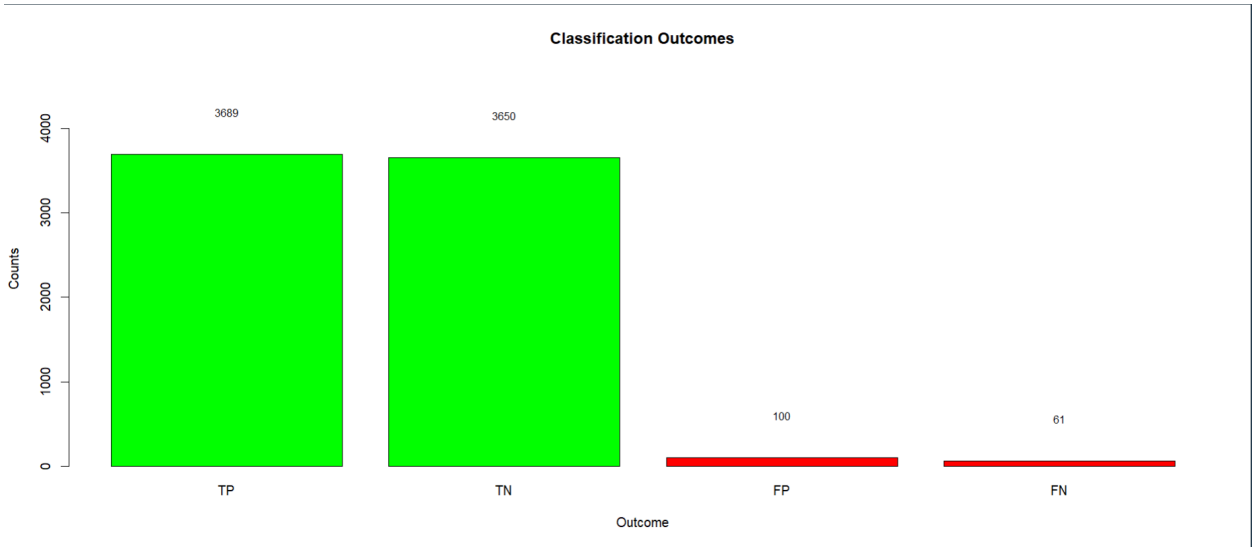
Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction anomaly normal
   anomaly    3689    100
   normal       61   3650
```

Training Accuracy:

```
Resampling results across tuning parameters:

  C     Accuracy   Kappa
  0.25  0.9660000  0.9320000
  0.50  0.9745778  0.9491556
  1.00  0.9782667  0.9565333
```

Visualization of above Confusion Matrix(Class Anomaly is treated as positive class**):**



XGboost

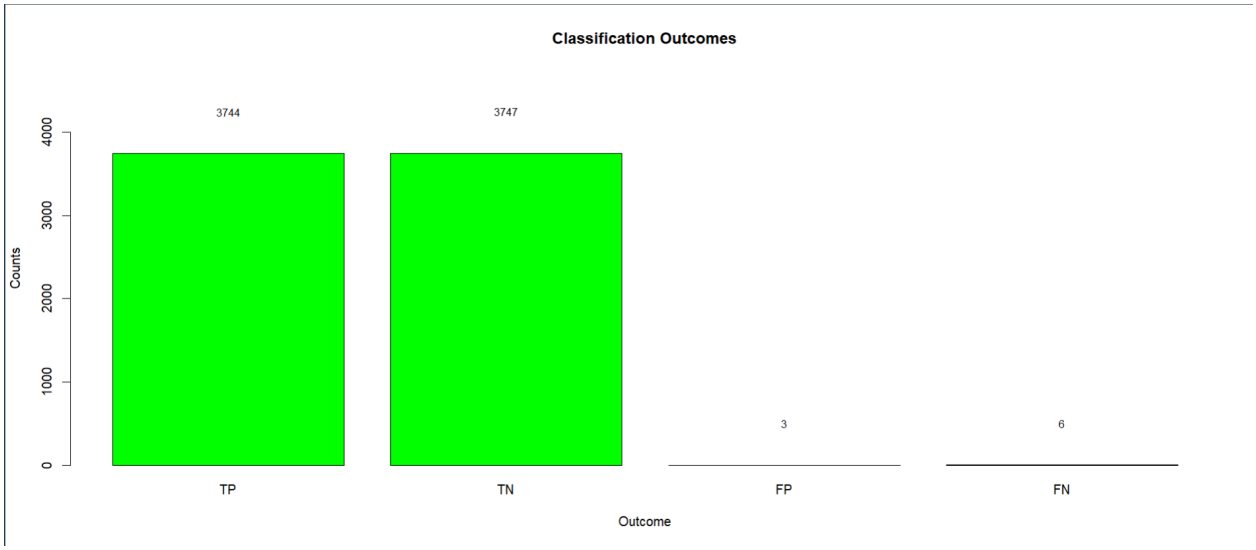
Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction anomaly normal
   anomaly    3744      3
   normal        6   3747
```

Training Accuracy:

```
Resampling results:

  Accuracy   Kappa
  0.9979556  0.9959111
```

Visualization of above Confusion Matrix(Class Anomaly is treated as positive class**):**



Random Forest with ntree = 100 and mtry = 6.

No change in SVM, same radial kernel.

For XGBoost, no hyperparameter tunning was done during this experiment.

	Accuracy	Precision	Recall	F1 Score	Specificity
RandomForest	0.9932	0.9962436	0.9901333	0.9931791	0.9962667
SVM	0.9784	0.9736008	0.9834667	0.9785089	0.9733333
XGBoost	0.9982667	0.9983996	0.9981333	0.9982664	0.9984

RandomForest

Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction anomaly normal
   anomaly    3713     14
   normal       37   3736
```

Training Accuracy:

```
Resampling results:

  Accuracy   Kappa
  0.9937778  0.9875556

Tuning parameter 'mtry' was held constant at a value of 6
```

SVM

Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction anomaly normal
   anomaly    3688    100
   normal       62   3650
```

Training Accuracy:

```
Resampling results across tuning parameters:

  C     Accuracy   Kappa
  0.25  0.9655111  0.9310222
  0.50  0.9744000  0.9488000
  1.00  0.9777778  0.9555556
```

XGBoost

Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction anomaly normal
   anomaly    3743      6
   normal        7   3744
```

Training Accuracy:

eta	max_depth	colsample_bytree	subsample	nrounds	Accuracy	Kappa
0.4	3	0.8	1.00	150	0.9978222	0.9956444

Mean Decrease Gini Impurity Feature Selection

Total 30k points(15k of each class) were chosen and training and testing was divided into 75% and 25%.Value of seed was kept constant as 123.The model is trained on 75% of datapoints and then MeanDecrease Gini is found:

	MeanDecreaseGini
service	1952.9359126
src_bytes	1778.2118461
dst_bytes	1326.6837022
flag	826.0949744
dst_host_srv_count	592.4396401
same_srv_rate	547.4305553
logged_in	437.6990512
diff_srv_rate	433.0022501
dst_host_same_srv_rate	370.0280634
dst_host_diff_srv_rate	310.2916077
count	295.2385992
protocol_type	277.5183756
dst_host_same_src_port_rate	254.7980372
srv_serror_rate	202.5105939
dst_host_srv_serror_rate	182.0805532
dst_host_srv_diff_host_rate	176.1586802
dst_host_serror_rate	167.0581727
serror_rate	145.3938250
srv_count	143.3283980
dst_host_count	130.5141005
hot	101.0416961
dst_host_srv_rerror_rate	88.9539206
dst_host_rerror_rate	73.6724363
num_compromised	66.9153748
srv_rerror_rate	59.0308193
rerror_rate	49.7847376
duration	45.4308300
srv_diff_host_rate	40.4312622
wrong_fragment	18.2599225
is_guest_login	10.8252574
num_root	2.3072706
num_file_creations	1.5195035
root_shell	1.2269975
num_access_files	0.5674964
num_failed_logins	0.5433294
num_shells	0.4134241
su_attempted	0.2202016
land	0.1949419
urgent	0.0000000
num_outbound_cmds	0.0000000
is_host_login	0.0000000

Result after choosing 1st **10 features** from above table:-

	Accuracy	Precision	Recall	F1 score	Specificity
RandomForest	0.9970667	0.9960085	0.9981333	0.9970698	0.996
SVM	0.9846667	0.9766588	0.9930667	0.9847944	0.9762667
XGBoost	0.9973333	0.9965389	0.9981333	0.9973355	0.9965333

RandomForest

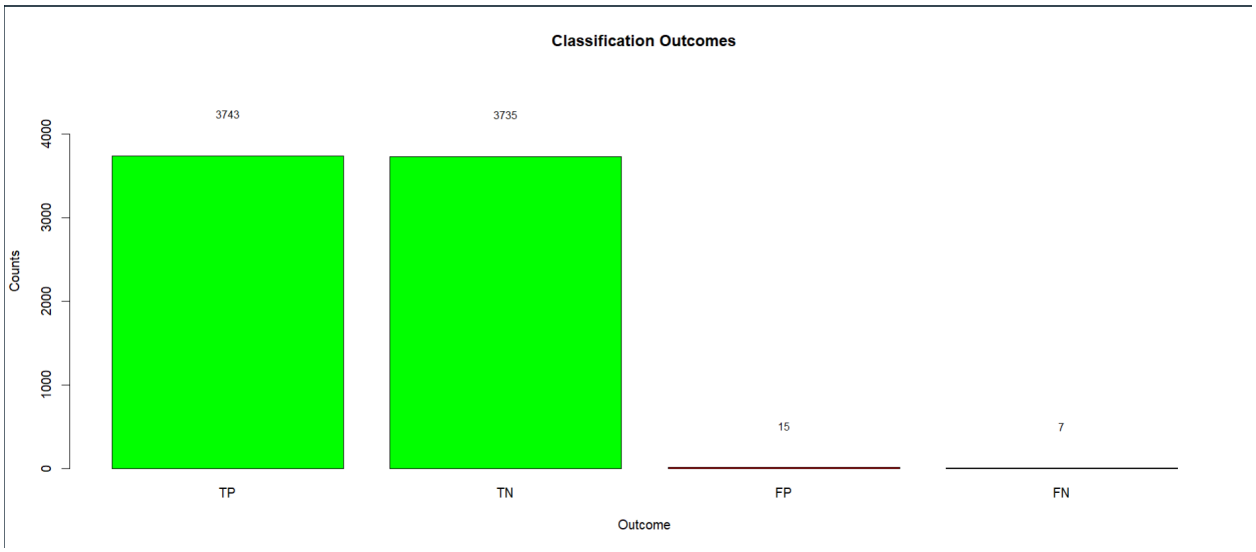
Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction anomaly normal
   anomaly    3743     15
   normal        7   3735
```

Training Accuracy:

```
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
  25    0.9967556  0.9935111
  30    0.9969333  0.9938667
  35    0.9969333  0.9938667

Accuracy was used to select the optimal model using the largest value.
```

Visualization of above Confusion Matrix(Class Anomaly is treated as positive class**):**



SVM

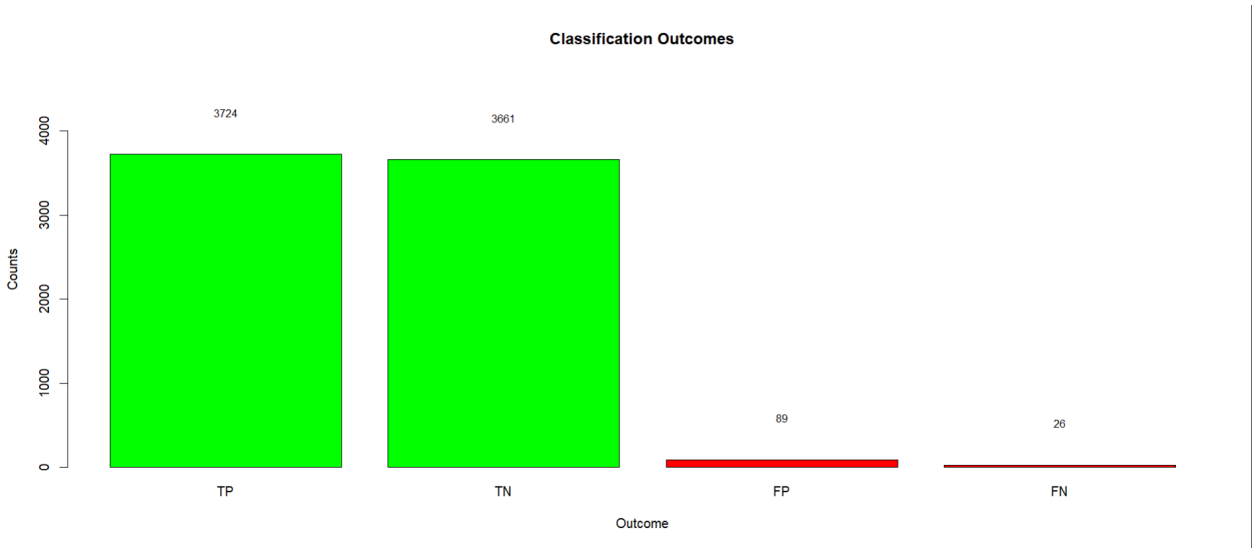
Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction anomaly normal
   anomaly    3724     89
   normal       26   3661
```

Training Accuracy:

```
Resampling results across tuning parameters

  C     Accuracy   Kappa
  0.25  0.9753333  0.9506667
  0.50  0.9799111  0.9598222
  1.00  0.9825778  0.9651556
```

Visualization of above Confusion Matrix(Class Anomaly is treated as positive class**):**



XGBoost

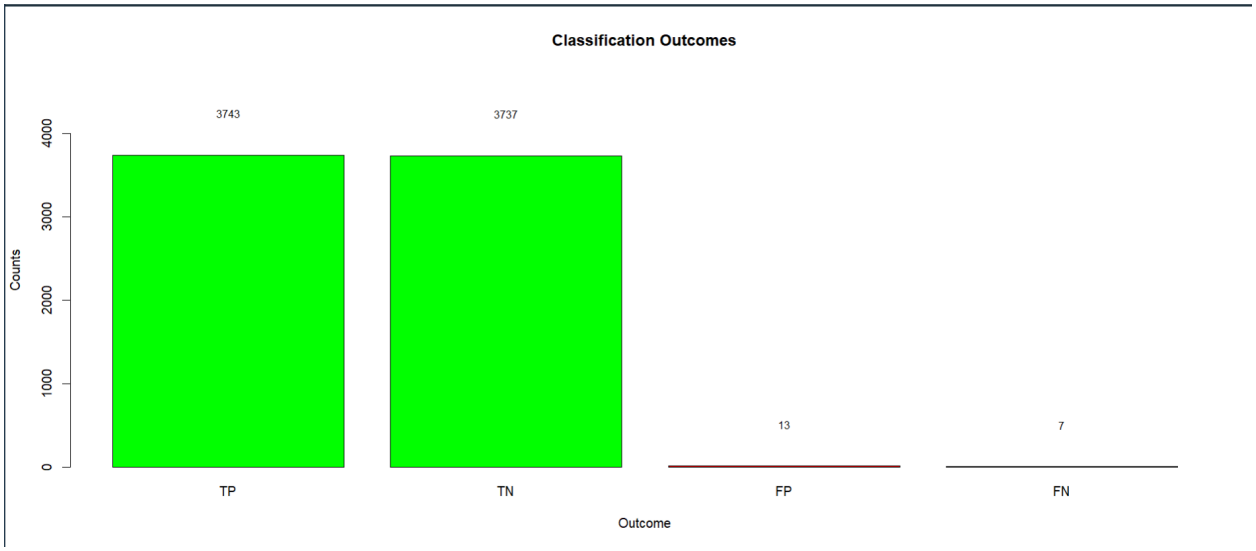
Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction anomaly normal
   anomaly    3743     13
   normal        7   3737
```

Training Accuracy:

```
Resampling results:

  Accuracy   Kappa
  0.9971556  0.9943111
```

Visualization of above Confusion Matrix(Class Anomaly is treated as positive class**):**



Result after choosing 1st **15 features** from above table:-

	Accuracy	Precision	Recall	F1score	Specificity
RandomForest	0.9984	0.9984	0.9984	0.9984	0.9984
SVM	0.97	0.9595828	0.9813333	0.9703362	0.9586667
XGBoost	0.9984	0.9989322	0.9978667	0.9983991	0.9989333

RandomForest

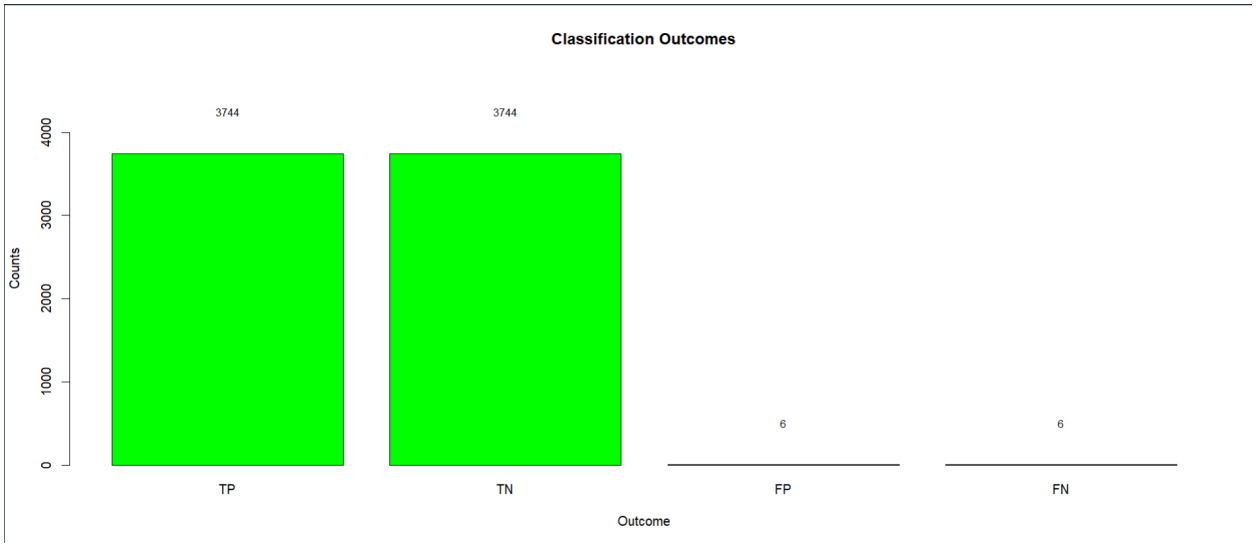
Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction anomaly normal
   anomaly    3744      6
   normal        6   3744
```

Training Accuracy:

```
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
  25    0.9971556  0.9943111
  30    0.9971556  0.9943111
  35    0.9972889  0.9945778

Accuracy was used to select the optimal model using the largest value.
```

Visualization of above Confusion Matrix(Class Anomaly is treated as positive class**):**



SVM

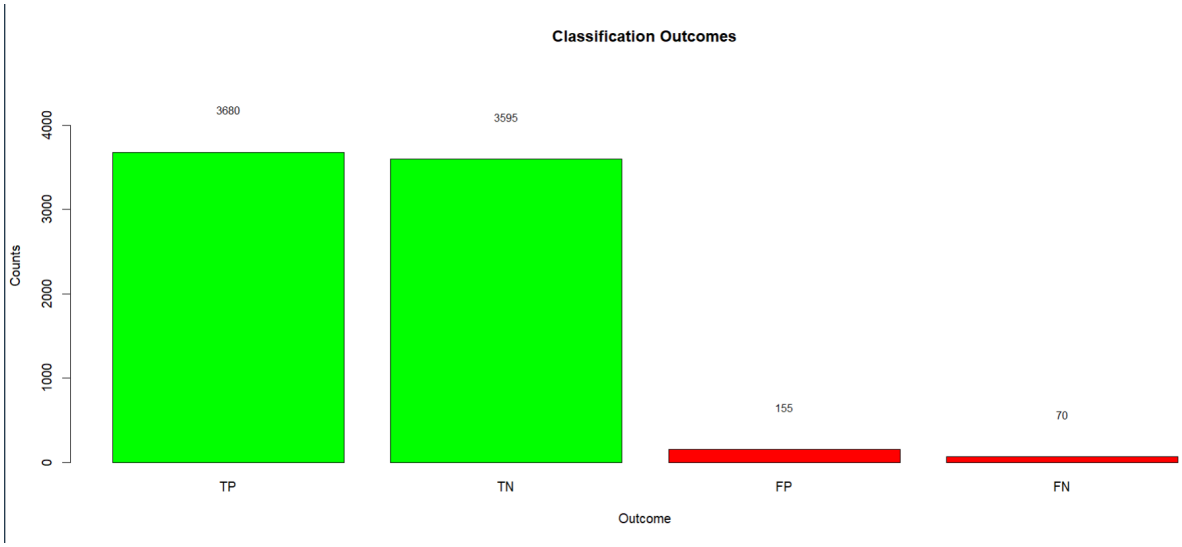
Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction anomaly normal
   anomaly    3680    155
   normal       70   3595
```

Training Accuracy:

```
Resampling results across tuning parameters:

  C     Accuracy   Kappa
  0.25  0.9644000  0.9288000
  0.50  0.9680444  0.9360889
  1.00  0.9713778  0.9427556
```

Visualization of above Confusion Matrix(Class Anomaly is treated as positive class**):**



XGBoost

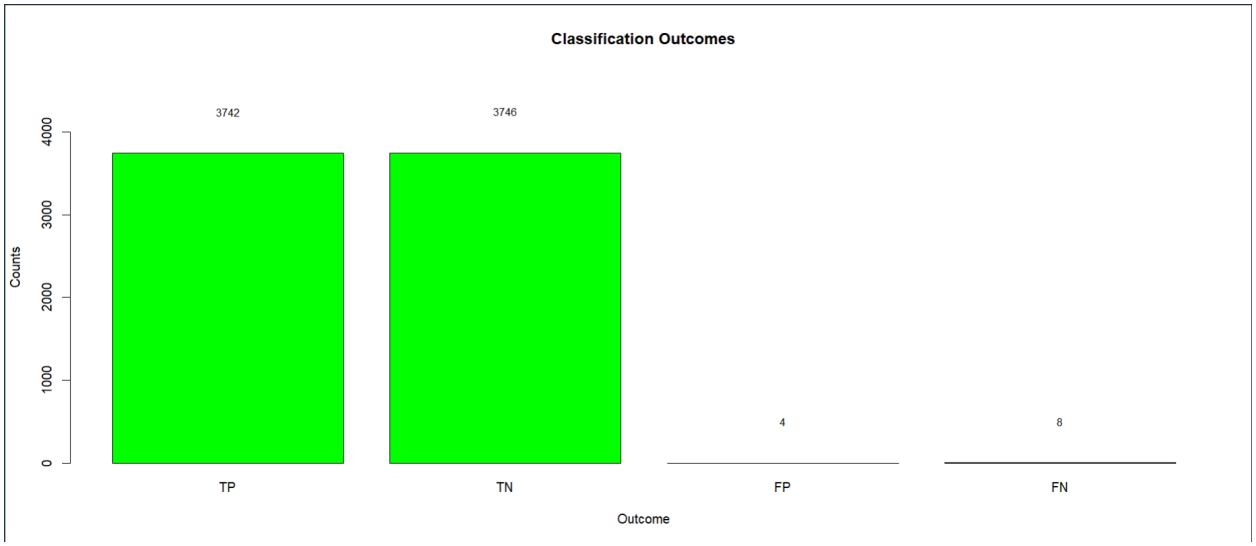
Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction anomaly normal
   anomaly    3742      4
   normal        8   3746
```

Training Accuracy:

```
Resampling results:

  Accuracy  Kappa
  0.9976    0.9952
```

Visualization of above Confusion Matrix(Class Anomaly is treated as positive class**):**



Result after choosing 1st **20 features** from above table:-

	Accuracy	Precision	Recall	F1score	Specificity
RandomForest	0.9986667	0.9989328	0.9984	0.9986663	0.9989333
SVM	0.9793333	0.9736495	0.9853333	0.9794566	0.9733333
XGBoost	0.9984	0.998666	0.9981333	0.9983996	0.9986667

RandomForest

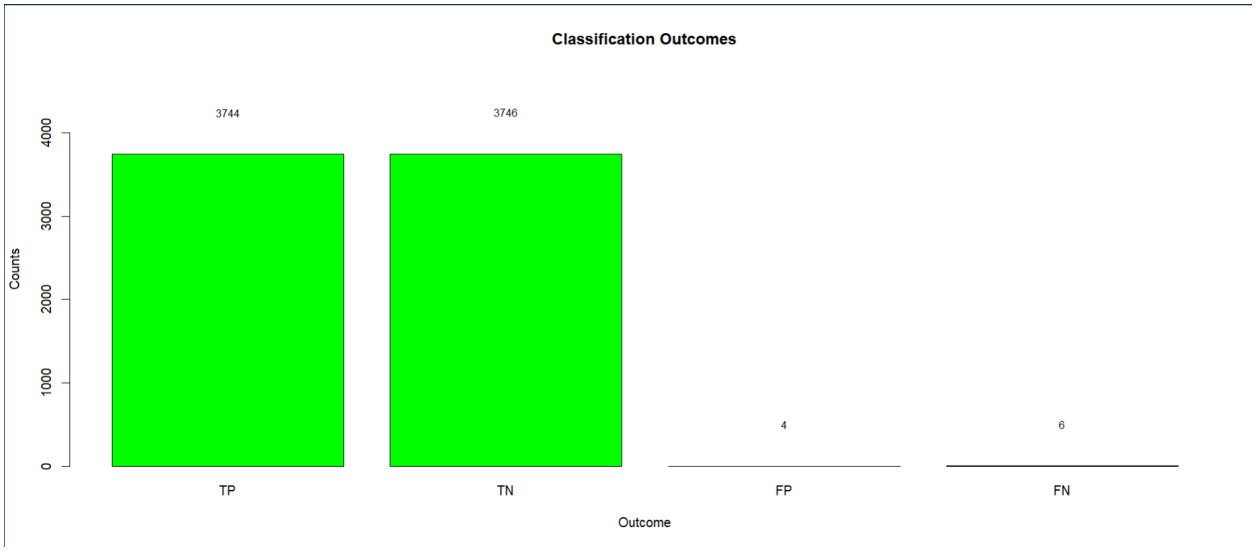
Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction anomaly normal
   anomaly    3744      4
   normal        6   3746
```

Training Accuracy:

```
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
  25    0.9972444  0.9944889
  30    0.9972889  0.9945778
  35    0.9973333  0.9946667

Accuracy was used to select the optimal model using the largest value.
```

Visualization of above Confusion Matrix(Class Anomaly is treated as positive class**):**



SVM

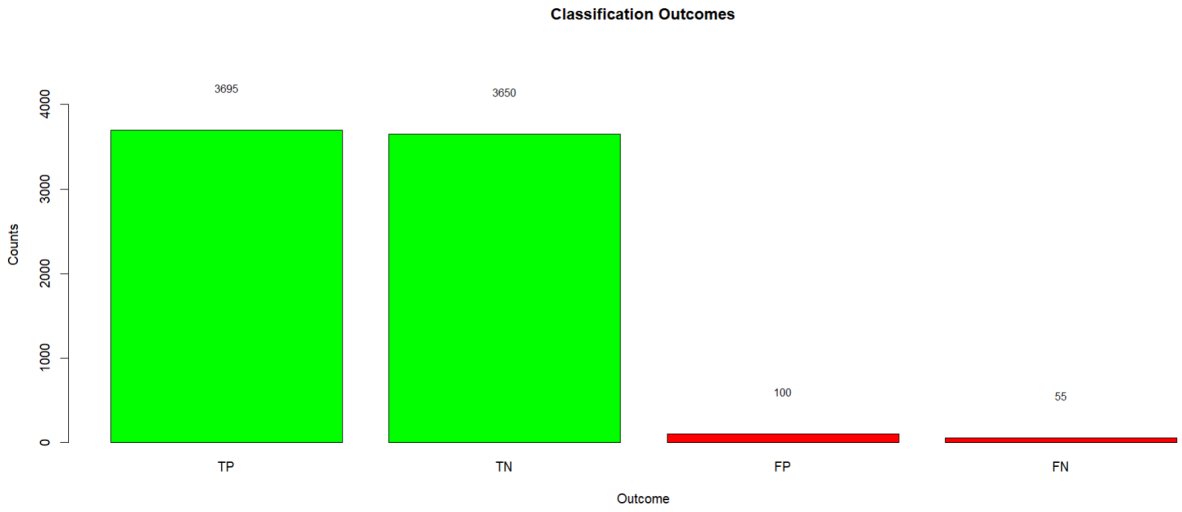
Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction anomaly normal
   anomaly    3695    100
   normal       55   3650
```

Training Accuracy:

```
Resampling results across tuning parameters:

  C     Accuracy   Kappa
  0.25  0.9744000  0.9488000
  0.50  0.9779556  0.9559111
  1.00  0.9802667  0.9605333
```

Visualization of above Confusion Matrix(Class Anomaly is treated as positive class**):**



XGBoost

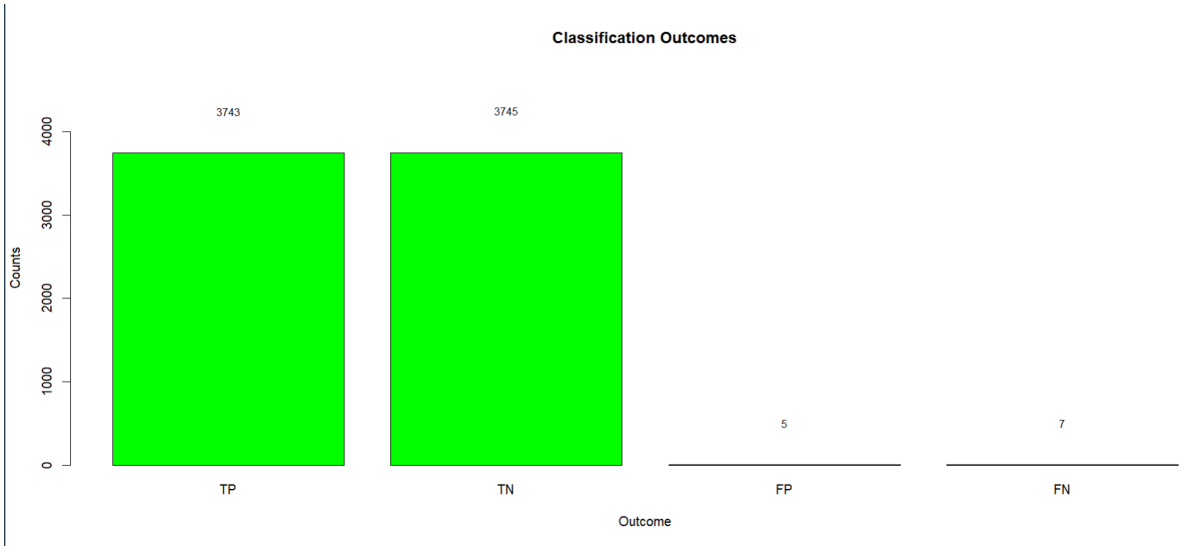
Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction anomaly normal
   anomaly    3743      5
   normal        7   3745
```

Training Accuracy:

```
Resampling results:

  Accuracy   Kappa
  0.9975556  0.9951111
```

Visualization of above Confusion Matrix(Class Anomaly is treated as positive class**):**



Result after choosing 1st **25 features** from above table:-

	Accuracy	Precision	Recall	F1score	Specificity
RandomForest	0.9984	0.998666	0.9981333	0.9983996	0.9986667
SVM	0.9796	0.9736634	0.9858667	0.979727	0.9733333
XGBoost	0.9982667	0.9986656	0.9978667	0.998266	0.9986667

RandomForest

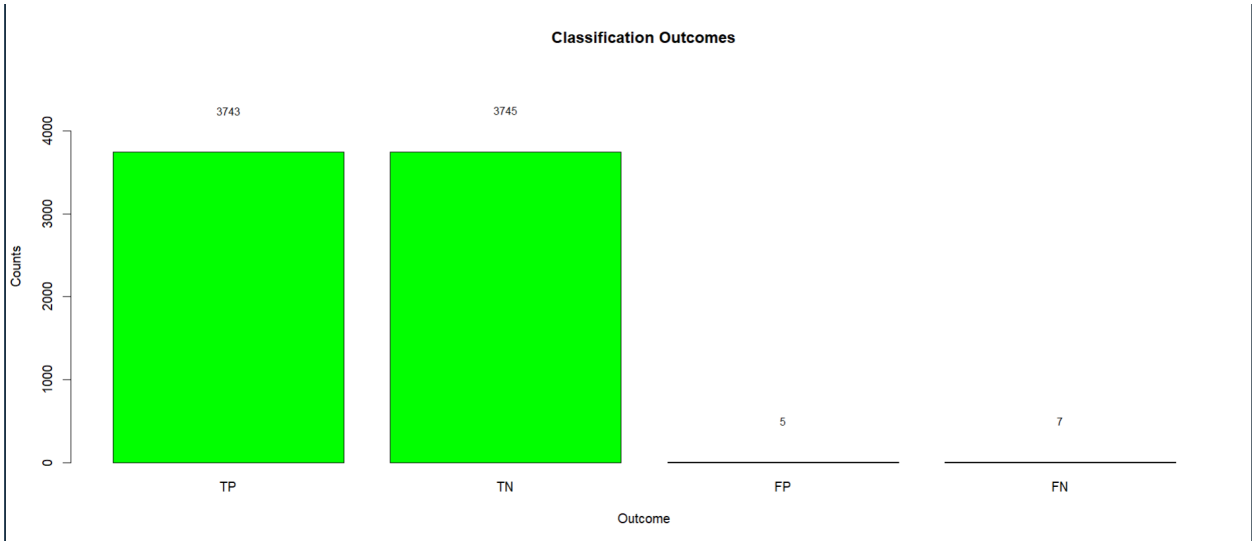
Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction anomaly normal
   anomaly    3743      5
   normal        7   3745
```

Training Accuracy:

```
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
  25    0.9973778  0.9947556
  30    0.9975556  0.9951111
  35    0.9975111  0.9950222

Accuracy was used to select the optimal model using the largest value.
```

Visualization of above Confusion Matrix(Class Anomaly is treated as positive class):



SVM

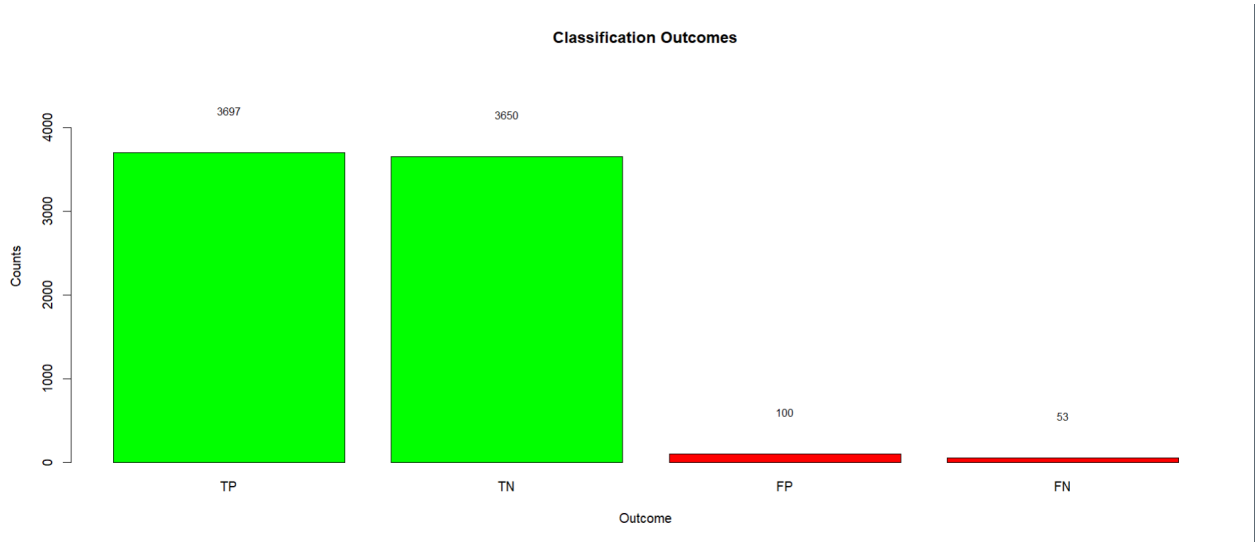
Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction anomaly normal
   anomaly    3697    100
   normal       53   3650
```

Training Accuracy:

```
Resampling results across tuning parameters:

  C     Accuracy   Kappa
  0.25  0.9745333  0.9490667
  0.50  0.9777778  0.9555556
  1.00  0.9799111  0.9598222
```

Visualization of above Confusion Matrix(Class Anomaly is treated as positive class**):**



XGBoost

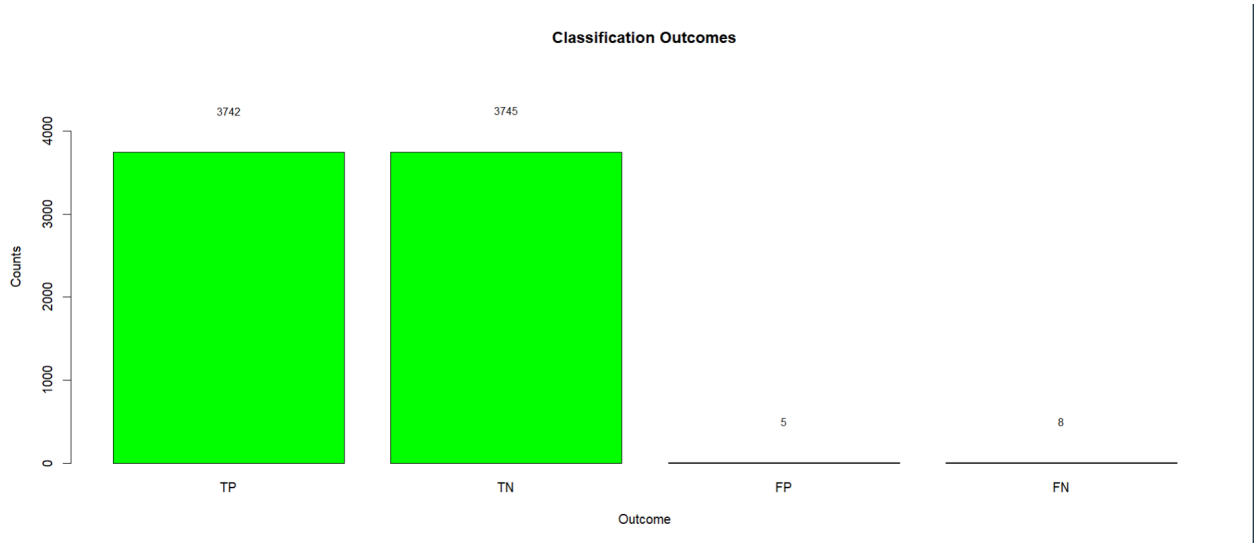
Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction anomaly normal
   anomaly    3742      5
   normal        8   3745
```

Training Accuracy:

```
Resampling results:

  Accuracy   Kappa
  0.9974667  0.9949333
```

Visualization of above Confusion Matrix(Class Anomaly is treated as positive class**):**



Result after choosing 1st **30 features** from above table:-

	Accuracy	Precision	Recall	F1score	Specificity
RandomForest	0.9986667	0.9989328	0.9984	0.9986663	0.9989333
SVM	0.9784	0.9736008	0.9834667	0.9785089	0.9733333
XGBoost	0.9985333	0.9991989	0.9978667	0.9985324	0.9992

RandomForest

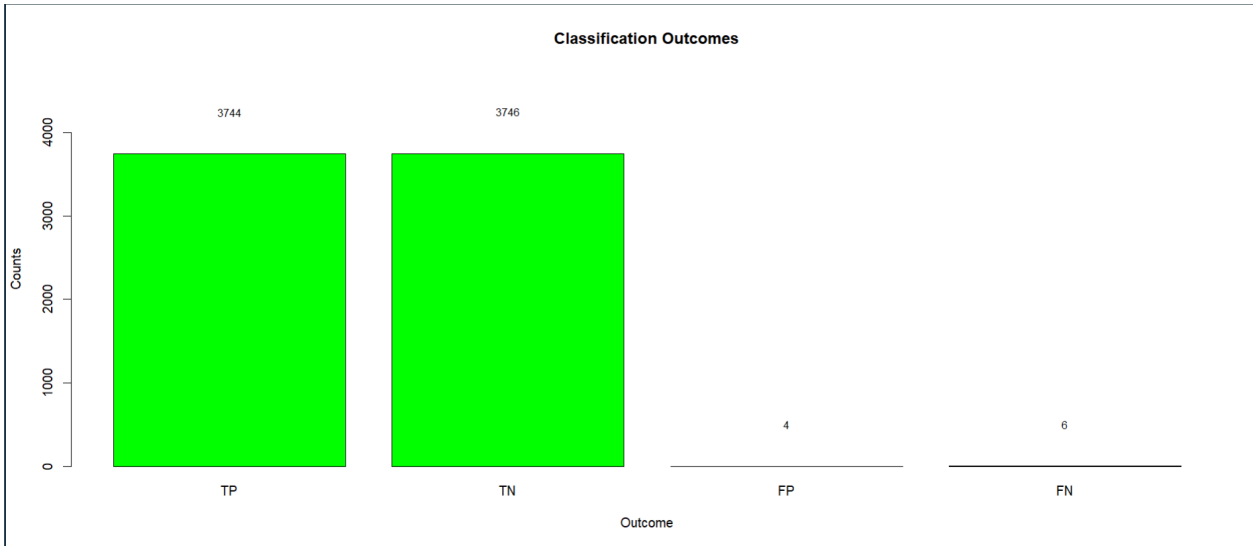
Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction anomaly normal
   anomaly    3744      4
   normal        6   3746
```

Training Accuracy:

```
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
  25    0.9972000  0.9944000
  30    0.9972000  0.9944000
  35    0.9972889  0.9945778

Accuracy was used to select the optimal model using the largest value.
```

Visualization of above Confusion Matrix(Class Anomaly is treated as positive class**):**



SVM

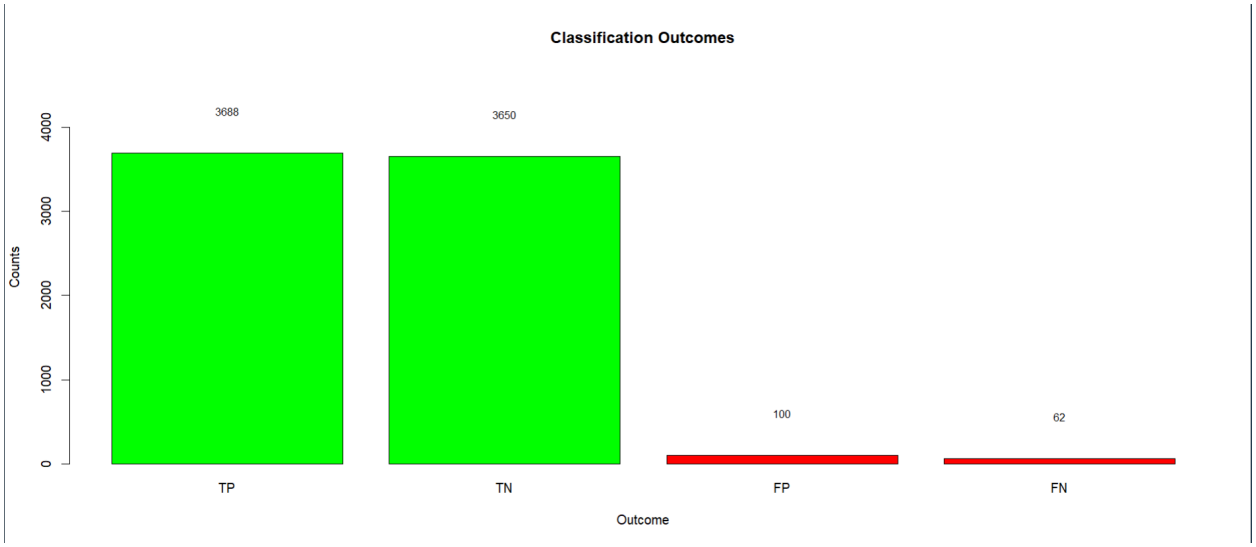
Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction anomaly normal
   anomaly    3688    100
   normal       62   3650
```

Training Accuracy:

```
Resampling results across tuning parameters:

  C     Accuracy   Kappa
  0.25  0.9661333  0.9322667
  0.50  0.9736000  0.9472000
  1.00  0.9782222  0.9564444
```

Visualization of above Confusion Matrix(Class Anomaly is treated as positive class**):**



XGBoost

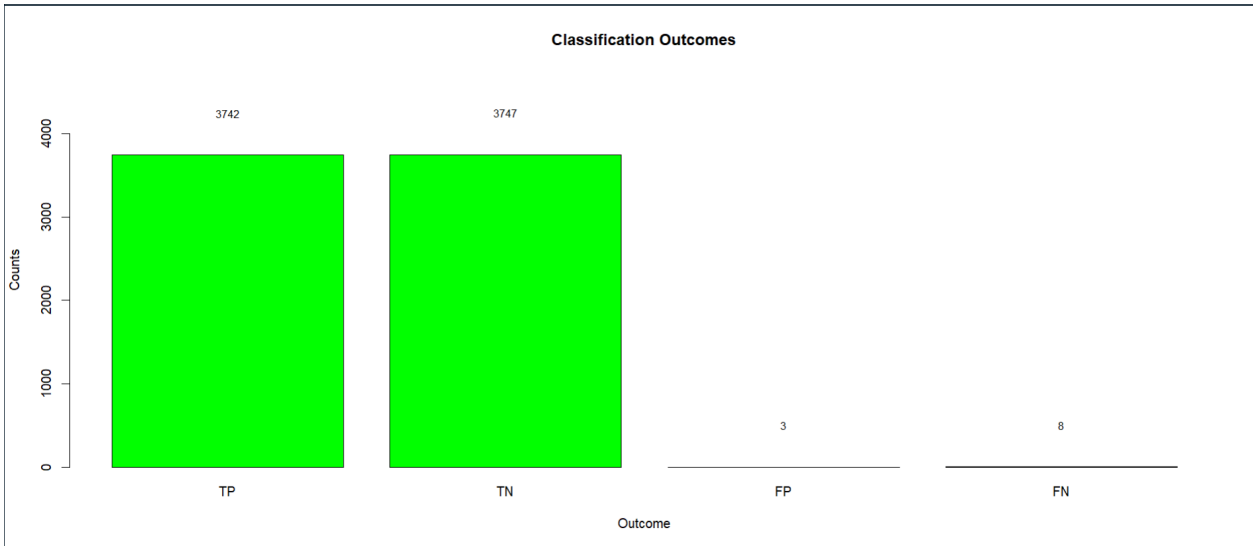
Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction anomaly normal
   anomaly    3742      3
   normal        8   3747
```

Training Accuracy:

```
Resampling results:

  Accuracy   Kappa
  0.9978222  0.9956444
```

Visualization of above Confusion Matrix(Class Anomaly is treated as positive class**):**



Result after choosing 1st **35 features** from above table:-

	Accuracy	Precision	Recall	F1score	Specificity
RandomForest	0.9988	0.9991994	0.9984	0.9987995	0.9992
SVM	0.9785333	0.9738579	0.9834667	0.9786387	0.9736
XGBoost	0.9984	0.998666	0.9981333	0.9983996	0.9986667

RandomForest

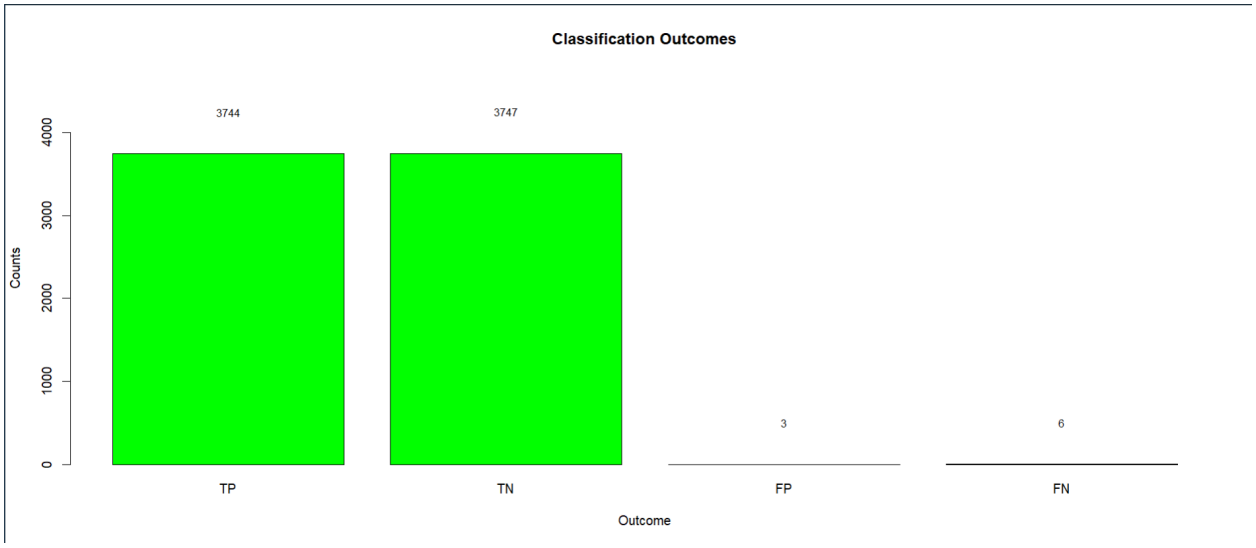
Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction anomaly normal
   anomaly    3744      3
   normal        6   3747
```

Training Accuracy:

```
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
  25    0.9976444  0.9952889
  30    0.9975556  0.9951111
  35    0.9976444  0.9952889

Accuracy was used to select the optimal model using the largest value.
```

Visualization of above Confusion Matrix(Class Anomaly is treated as positive class**):**



SVM

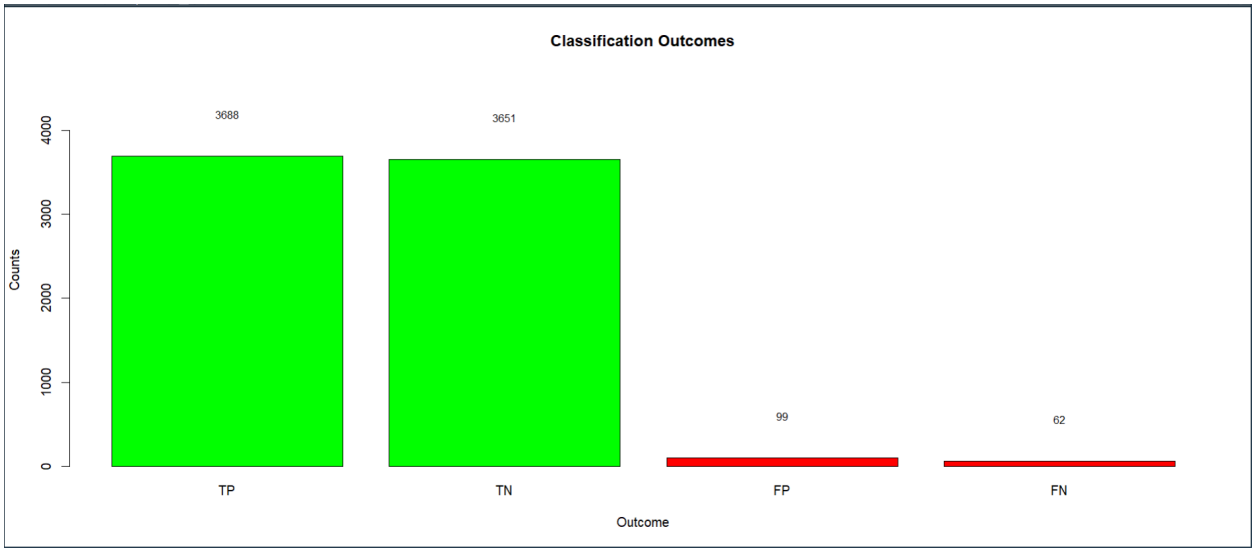
Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction anomaly normal
   anomaly    3688     99
   normal       62   3651
```

Training Accuracy:

```
Resampling results across tuning parameters

  C     Accuracy   Kappa
  0.25  0.9661333  0.9322667
  0.50  0.9747111  0.9494222
  1.00  0.9786222  0.9572444
```

Visualization of above Confusion Matrix(Class Anomaly is treated as positive class**):**



XGBoost

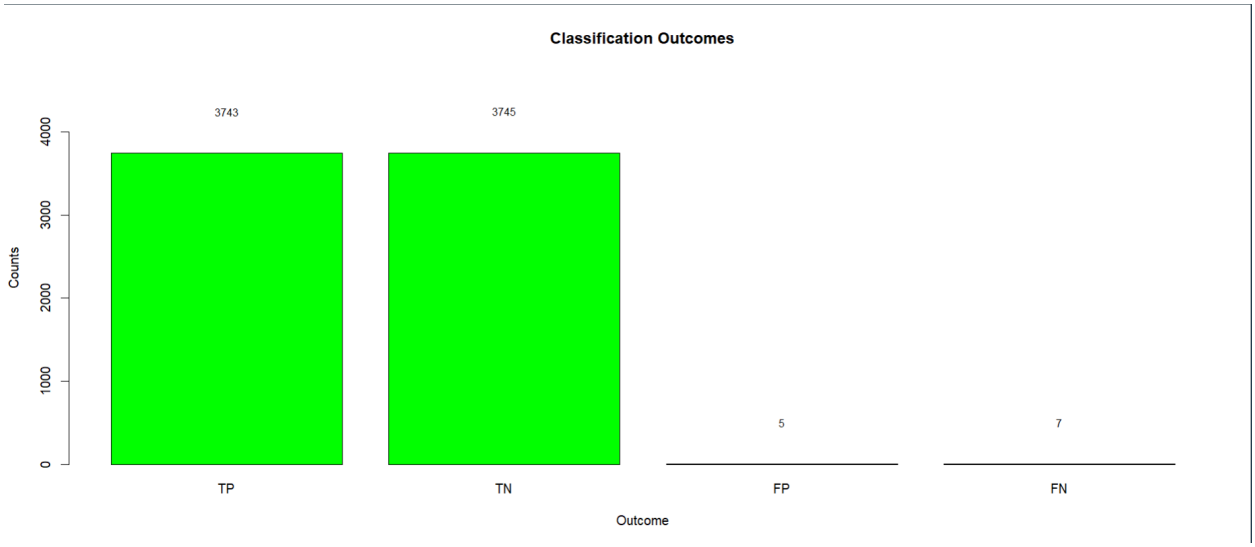
Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction anomaly normal
   anomaly    3743      5
   normal        7   3745
```

Training Accuracy:

```
Resampling results:

  Accuracy   Kappa
  0.9974667  0.9949333
```

Visualization of above Confusion Matrix(Class Anomaly is treated as positive class**):**



Results after choosing top 10 features with ntree = 100 and mtry = 6.No change in Hyperparameter in XGBoost(Same like before):-

	Accuracy	Precision	Recall	F1score	Specificity
RandomForest	0.9773333	0.9877384	0.9666667	0.9770889	0.988
SVM	0.9846667	0.9766588	0.9930667	0.9847944	0.9762667
XGBoost	0.9974667	0.9968043	0.9981333	0.9974684	0.9968

RandomForest

Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction anomaly normal
   anomaly    3625     45
   normal      125   3705
```

Training Accuracy:

```
Resampling results:

  Accuracy   Kappa
  0.9780444  0.9560889

Tuning parameter 'mtry' was held constant at a value of 6
```

SVM

Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction anomaly normal
   anomaly    3724     89
   normal       26   3661
```

Training Accuracy:

```
Resampling results across tuning parameters:

  C     Accuracy   Kappa
  0.25  0.9753333  0.9506667
  0.50  0.9799111  0.9598222
  1.00  0.9827556  0.9655111
```

XGBoost

Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction anomaly normal
   anomaly    3743     12
   normal        7   3738
```

Training Accuracy:

```
Resampling results:

  Accuracy   Kappa
  0.9974222  0.9948444
```