Jialin Shi – Prediction about APS Failure

Weka Report

For Train:

Train dataset: 60000, class= ‘neg’ | ‘pos’
Dataset is imbalanced.
Use Weka to train a Logical Model Tree, with cv=5 folds.

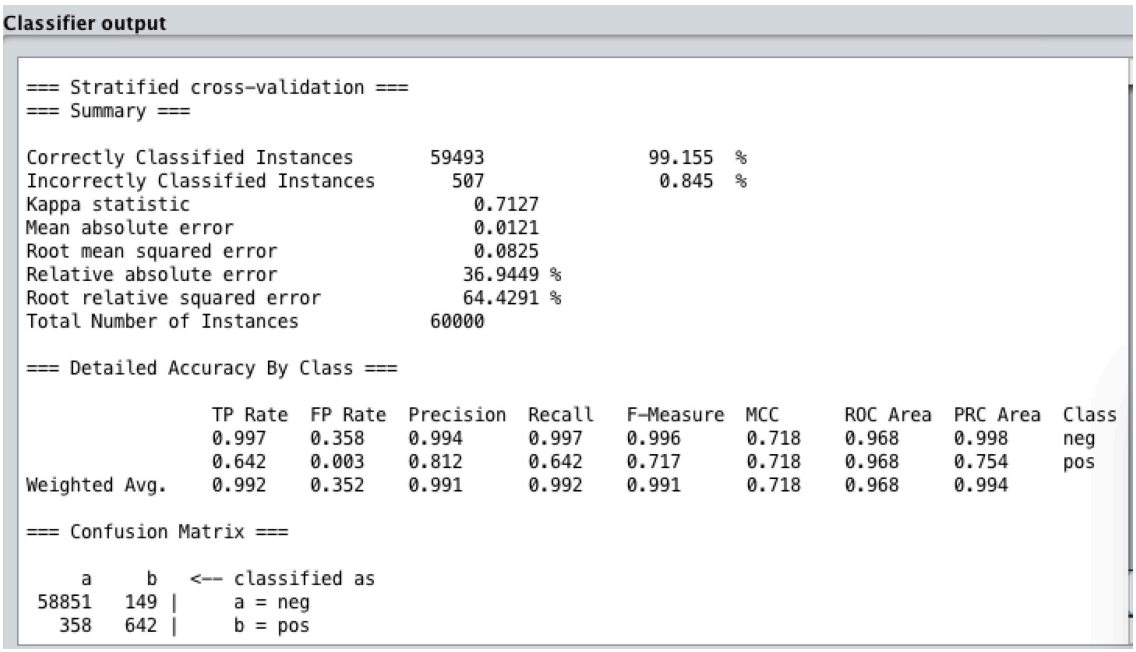
Train Report:

Model score

Misclassification rate = 0.00845
LMT Accuracy = 0.99155

Confusion Matrix

FP rate for ‘neg’ = 0.358 | (358 out of 1000 ‘pos’ class are misclassified as ‘neg’)
Recall for ‘pos’ = 0.642 | only 64.2% ‘pos’ are classified as ‘pos’



Classifier output

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       59493               99.155  %
Incorrectly Classified Instances       507                0.845  %
Kappa statistic                          0.7127
Mean absolute error                      0.0121
Root mean squared error                  0.0825
Relative absolute error                 36.9449 %
Root relative squared error             64.4291 %
Total Number of Instances            60000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.997    0.358    0.994      0.997    0.996      0.718    0.968     0.998     neg
                 0.642    0.003    0.812      0.642    0.717      0.718    0.968     0.754     pos
Weighted Avg.    0.992    0.352    0.991      0.992    0.991      0.718    0.968     0.994

=== Confusion Matrix ===

     a     b   <-- classified as
 58851   149 |     a = neg
   358   642 |     b = pos
```

For Test:

Test dataset: 16000, class= ‘neg’ | ‘pos’
Re-evaluate the model with test dataset.

Test Report:

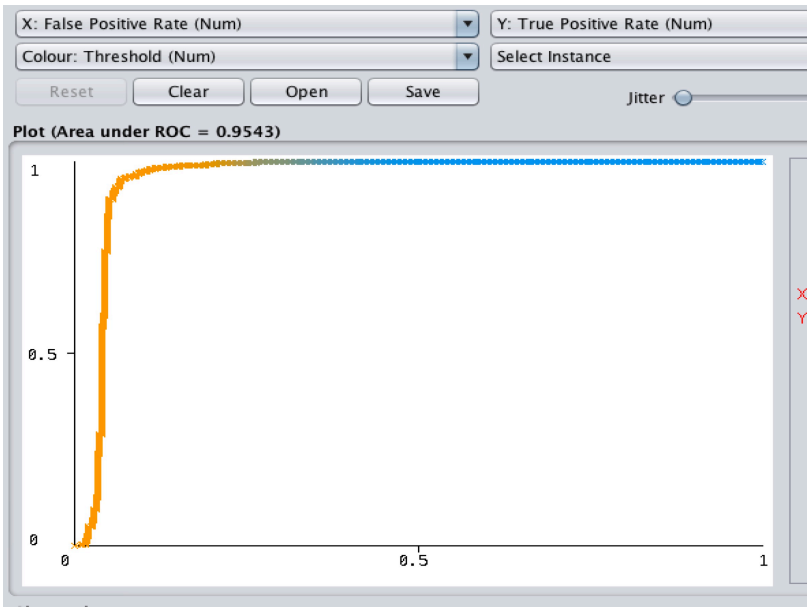
Model score

Misclassification rate = 0.0095
LMT Accuracy = 0.9905

Confusion Matrix

FP rate for ‘neg’ = 0.291 | (109 out of 375 ‘pos’ class are misclassified as ‘neg’)
Recall for ‘pos’ = 0.709 | only 70.9% ‘pos’ are classified as ‘pos’
Auc for test = 0.9543

```
=== Summary ===

Correctly Classified Instances       15849               99.0563 %
Incorrectly Classified Instances       151                0.9437 %
Kappa statistic                          0.7741
Mean absolute error                      0.0121
Root mean squared error                  0.0896
Total Number of Instances            16000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.997    0.291    0.993      0.997    0.995      0.778    0.954     0.995     neg
                 0.709    0.003    0.864      0.709    0.779      0.778    0.954     0.809     pos
Weighted Avg.    0.991    0.284    0.990      0.991    0.990      0.778    0.954     0.991

=== Confusion Matrix ===

     a     b   <-- classified as
 15583    42 |     a = neg
   109   266 |     b = pos
```



Conclusion for 2-(e):

Train has lower misclassification rate (I think it is because train has much bigger dataset and the data in imbalanced.)

However, train has higher FP rate for ‘neg’ class, and lower recall for ‘pos’ class.

Test Auc is lower than train Auc.

SMOTE

For Train:

Use Smote to up sample train to 120000.

Train dataset: 120000, class= ‘neg’ | ‘pos’

Dataset is now balanced.

Use Weka to train a Logical Model Tree, with cv=5 folds.

Train Report:

Model score

Misclassification rate = 0.0199

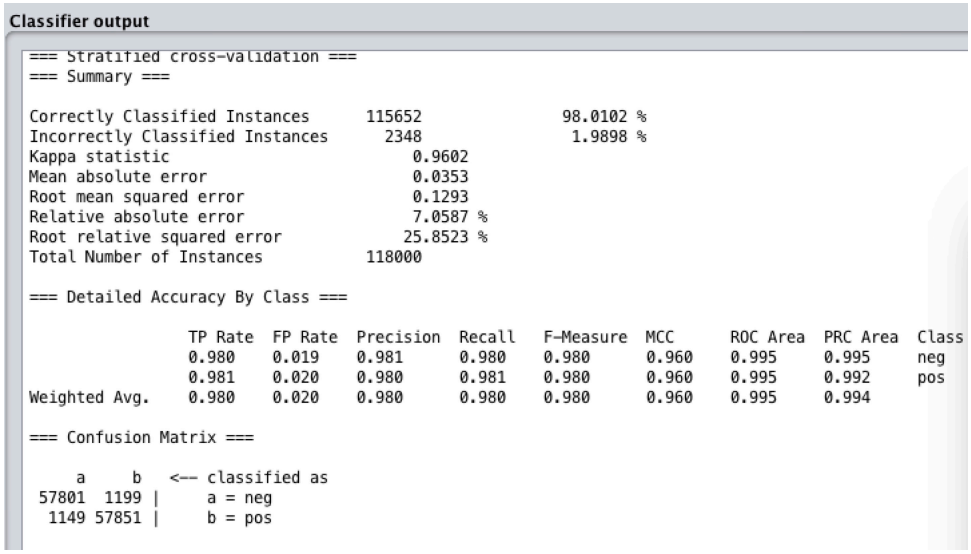
LMT Accuracy = 0.98

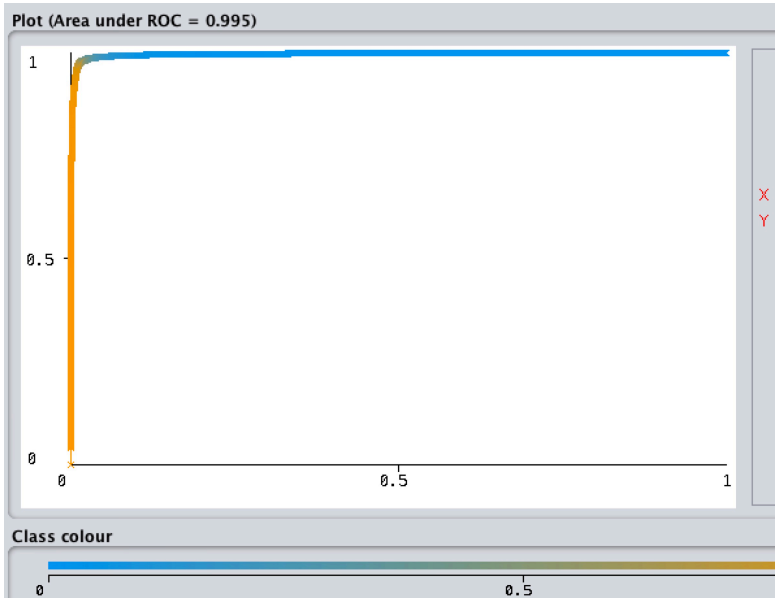
Confusion Matrix

FP rate for ‘neg’ = 0.019 | (1149 out of 59000 ‘pos’ class are misclassified as ‘neg’)

Recall for ‘pos’ = 0.981 | 98.1% ‘pos’ are classified as ‘pos’

Auc=0.995





For Test:

Test dataset: 16000, class= ‘neg’ | ‘pos’

Reevaluate the model with test dataset.

Test Report:

Model score

Misclassification rate = 0.01356

LMT Accuracy = 0.9864

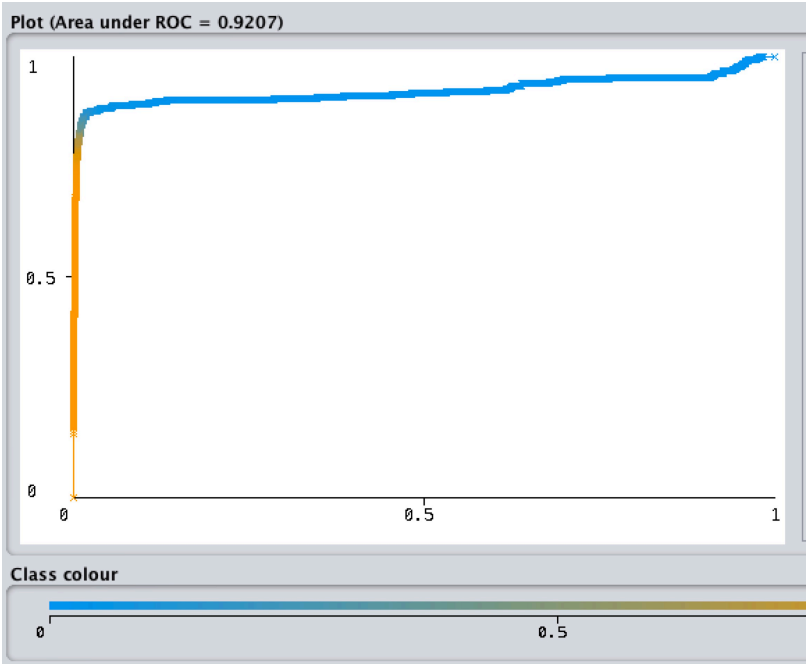
Confusion Matrix

FP rate for ‘neg’ = 0.176 | (66 out of 375 ‘pos’ class are misclassified as ‘neg’)

Recall for ‘pos’ = 0.824 | 82.4% ‘pos’ are classified as ‘pos’

Auc for test = 0.9207

```
=== Summary ===

Correctly Classified Instances       15783               98.6437 %
Incorrectly Classified Instances       217                1.3562 %
Kappa statistic                          0.7332
Mean absolute error                      0.0214
Root mean squared error                  0.1095
Total Number of Instances            16000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.990    0.176    0.996      0.990    0.993      0.737    0.920     0.995     neg
                 0.824    0.010    0.672      0.824    0.740      0.737    0.921     0.705     pos
Weighted Avg.    0.986    0.172    0.988      0.986    0.987      0.737    0.920     0.988

=== Confusion Matrix ===

     a     b   <-- classified as
 15474   151 |     a = neg
    66   309 |     b = pos
```



Conclusion for 2-(f):

Uncompensated data has lower misclassification rate and higher AUC.
However, imbalanced dataset has higher FP rate for 'neg' class, and lower recall for 'pos' class. This is because the model sees too few minor class data.

After using SMOTE to balance the data, FP rate for 'neg' class becomes lower and recall for 'pos' becomes higher, which means that the minor class 'pos' has been classified more accurately than imbalanced data.