Assignment #3

Sima Shafaei

April 9

Imbalance Data:

We have an imbalance dataset which contains 573518 samples in class 0 and 21694 samples in class 1. Figure 1 shows the distribution of target value in this dataset.

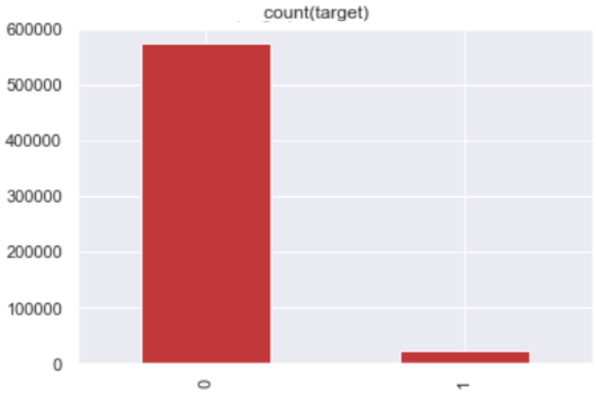


Figure 1: In the initial dataset there are 21694 cases with the value of target =0 and 573518 cases with the value of target =1

We want to use a logistic classification on this dataset. Following sampling methods applied on this dataset to improve the classification results.

1. Random Under-sampling

Random under-sampling involves randomly selecting examples from the majority class and deleting them from the training dataset. In the random under-sampling, the majority class instances are discarded at random until a more balanced distribution is reached. Figure 2 shows the distribution of target column after random under sampling. As it is shown, after under sampling we have 21694 cases in class 0 and 21694 cases in class 1.

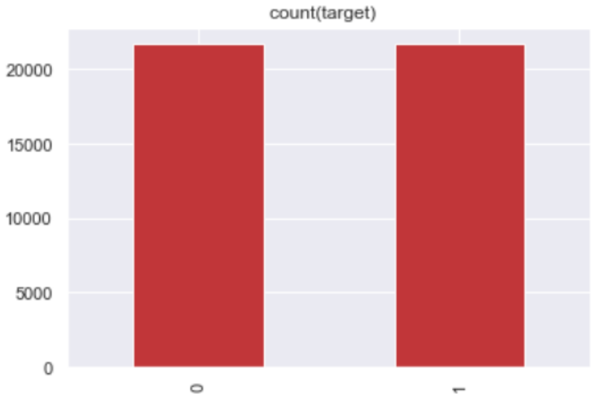


Figure 2: after under sampling we have 21694 cases with the value of target =0 and 21694 cases with the value of target =1

The result of logistic classification on this dataset is as follows:

- Accuracy: 0.50
- Precision: 0.50
- Recall: 0.56

Figure 3 shows the confusion matrix of logistic classification over dataset after using random under sampling method. As it is obvious, it cannot act better than assigning samples to classes randomly. In fact, under sampling can destroy initial distribution of data so it cannot perform very well on this dataset.

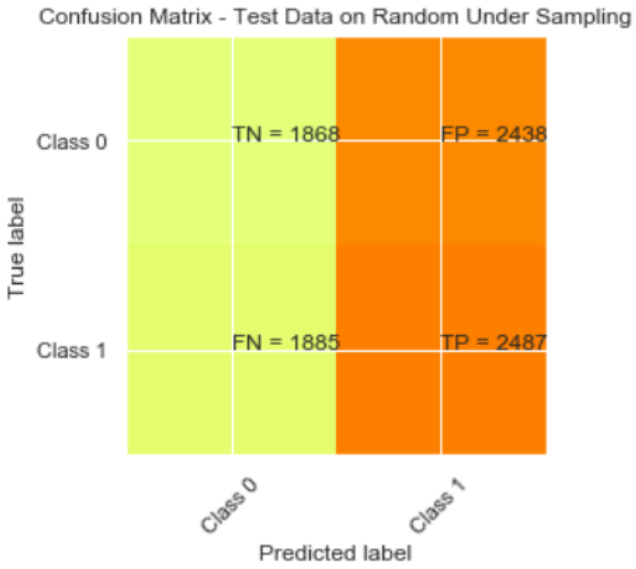


Figure 3: confusion matrix of logistic classification after random under sampling on initial dataset

2. Random Over Sampling

Random oversampling involves randomly duplicating examples from the minority class and adding them to the training dataset. Examples from the training dataset are selected randomly with replacement. This means that examples from the minority class can be chosen and added to the new "more balanced" training dataset multiple times; they are selected from the original training dataset, added to the new training dataset, and then returned or "replaced" in the original dataset, allowing them to be selected again. Figure 4 shows the distribution of target value after random over sampling. As it is shown, after over sampling, we have 537518 cases in class 0 and 537518 cases in class 1.

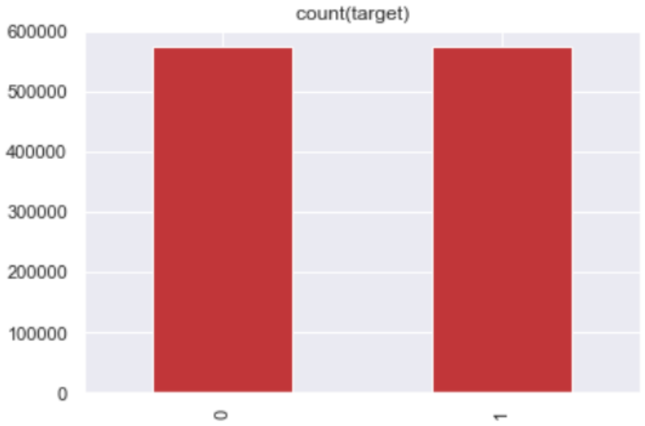


Figure 4: after random over sampling we have 573518 cases with the value of target =0 and 573518 cases with the value of target =1

The result of logistic classification on this dataset is as follows:

- Accuracy: 0.50
- Precision: 0.50
- Recall: 0.66

Figure 5 shows the confusion matrix of logistic classification over dataset after using random over sampling method. This method again acts like a random classification. The reason is that over sampling can cause overfitting.

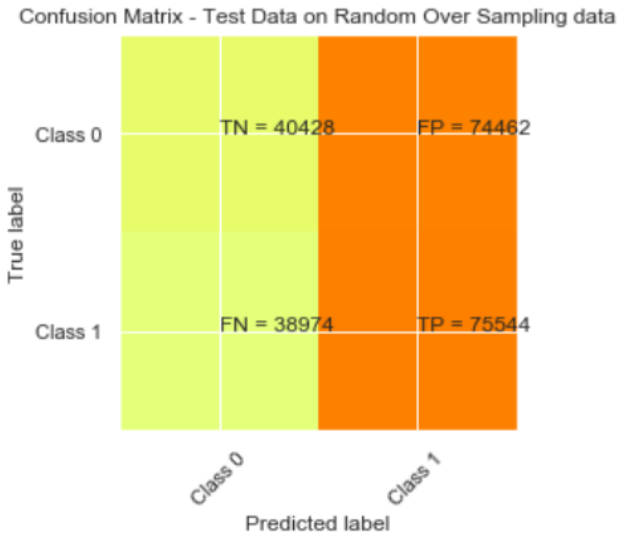


Figure 5: confusion matrix of logistic classification after random over sampling on initial dataset

3. Under-sampling: Tomek links

Tomek proposed an effective method that considers samples near the borderline. Given two instances a and b belonging to different classes and are separated by a distance d(a,b), the pair (a, b) is called a Tomek link if there is no instance c such that d(a,c) < d(a,b) or d(b,c) < d(a,b). Instances participating in Tomek links are either borderline or noise so both are removed. Figure 6 shows the distribution of target column after under sampling using Tomek link method. As it is shown, after under sampling we have 561433 cases with the value of target =0 and 21694 cases with the value of target =1.

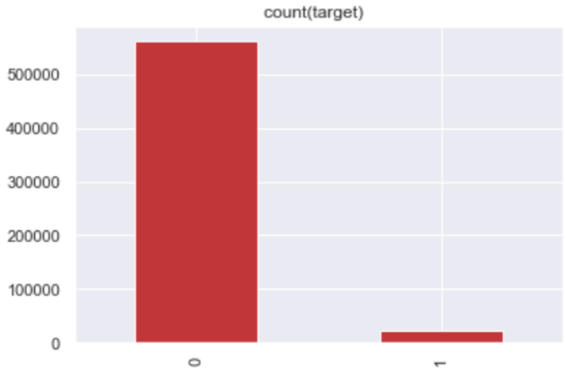


Figure 6: after under sampling using Tomek Link method we have 561433 cases with the value of target =0 and 21694 cases with the value of target =1

The result of logistic classification on this dataset is as follows:

- Accuracy: 0.96
- Precision: 0.0
- Recall: 0.0

Figure 7 shows the confusion matrix of logistic classification over dataset after using under sampling using Tomek link method. As we can see this method cannot solve the problem of imbalance data. We have still an imbalance data after sampling. Therefore, the accuracy is high but the reason is that all data are classified as class 0.

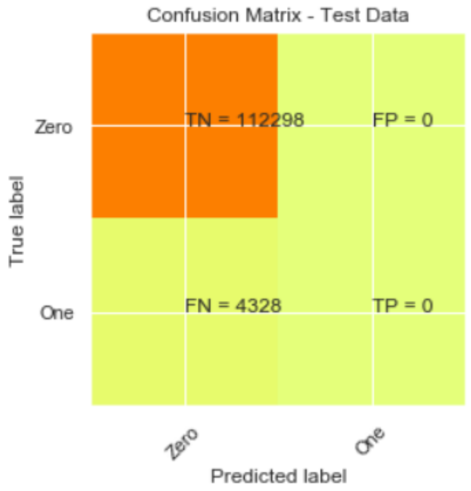


Figure 7: confusion matrix of logistic classification after under sampling using Tomek-link method on initial dataset

4. Over-sampling: SMOTE

SMOTE (synthetic minority oversampling technique) is one of the most commonly used oversampling methods to solve the imbalance problem. It aims to balance class distribution by randomly increasing minority class examples by replicating them. SMOTE synthesizes new minority instances between existing minority instances. Figure 8 shows the distribution of target column after over sampling using SMOTE method. As it is shown, after over sampling, we have 573518 cases with the value of target =0 and 573518 cases with the value of target =1

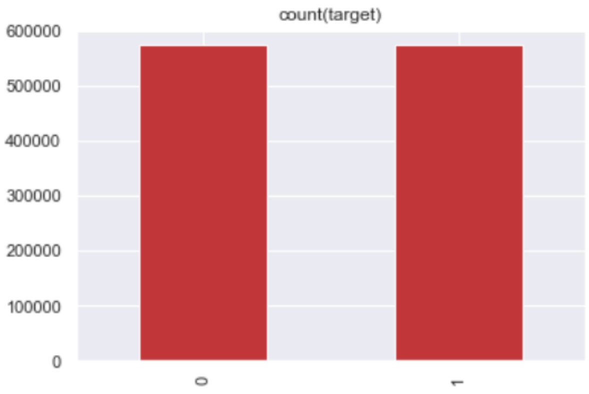


Figure 8: after over sampling using SMOTE method we have 573518 cases with the value of target =0 and 573518 cases with the value of target =1

The result of logistic classification on this dataset is as follows:

- Accuracy: 0.51
- Precision: 0.51
- Recall: 0.66

Figure 9 shows the confusion matrix of logistic classification over dataset after using SMOTE over sampling. This method performs a little better than random over sampling but the result is not much different and needs more improvement.

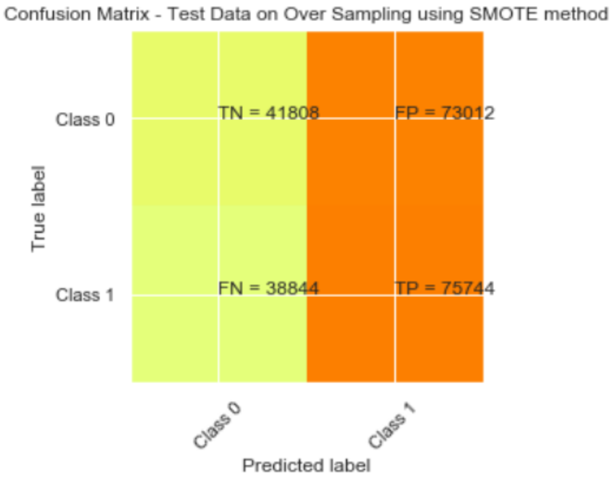


Figure 9: confusion matrix of logistic classification after over sampling using SMOTE method on initial dataset

5. Under-sampling: Cluster Centroids

This method clusters of majority class and replace that cluster with centroid of that cluster. So we under-sample majority class by forming clusters and replacing it with cluster centroids. Because my computer was not able to run this method on initial dataset, I selected 20% of this dataset but kept the ratio of class 0 and class 1. In this dataset we have 114712 samples in class 0 and 4331 samples in class 1. Figure 10 shows the distribution of target value in this dataset.

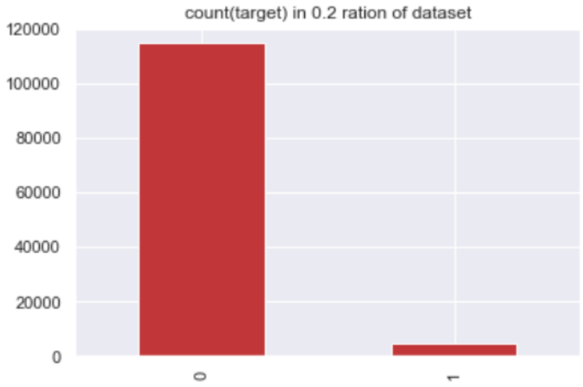


Figure 10: We select 0.1 dataset and keep the ratio of Class 0 and Class 1

Figure 11 shows the distribution of target value after under sampling using Cluster Centroid method. As it is shown, after under sampling we have 4331 cases in class 0 and 4331 cases in class 1.

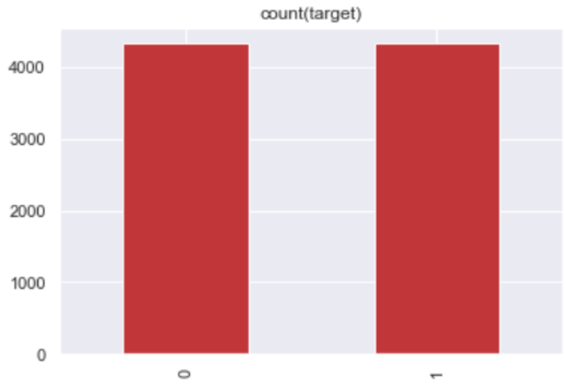


Figure 11: after under sampling using Cluster Centroid method we have 4331 cases with the value of target =0 and 4331 cases with the value of target =1

The result of logistic classification on this dataset is as follows:

- Accuracy: 0.50
- Precision: 0.50
- Recall: 1

Figure 12 shows the confusion matrix of logistic classification over dataset after using cluster centroid under sampling. To be able to compare this method with other ones we used the same dataset for smote over sampling which was one of our best methods. Figure 13 shows the result of classification using smote over sampling method. As we can see SMOTE can perform better but the results are not much different. The result of classification on SMOTE sampling was as follows:

- Accuracy: 54%
- Precision: 54%
- recall:48%

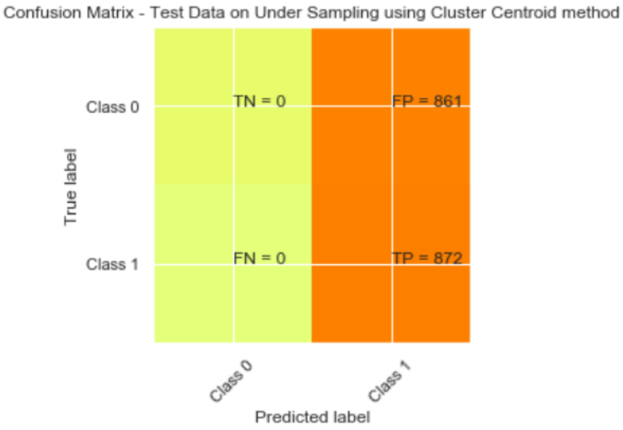


Figure 12: confusion matrix of logistic classification after under sampling using Cluster Centroid method on 20% of initial dataset

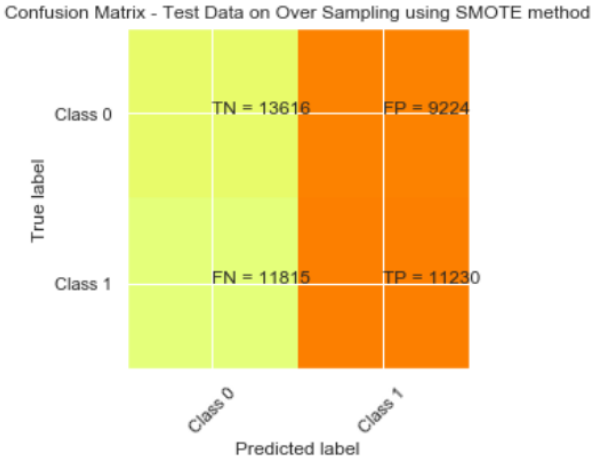


Figure 13: confusion matrix of logistic classification after SMOTE over sampling on 20% of initial dataset

6. Over-sampling->Under-sampling: SMOTE-Tomek

Here we first use SMOTE method for oversampling data and then apply Tomek link method to remove boundary samples and make the classes more separable. Figure 14Figure 8 shows the distribution of target column after over sampling using these methods. As it is shown, after sampling, we have 555420 cases in class 0 and 555420 cases in class 1

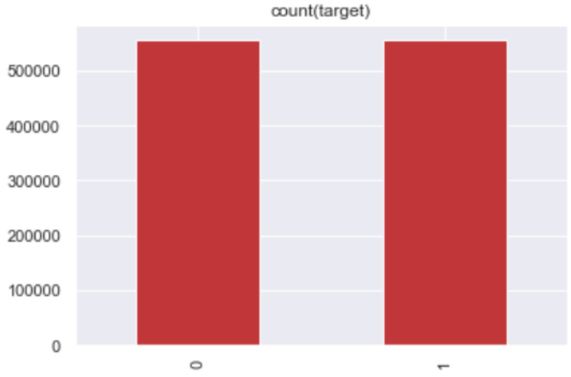


Figure 14: after sampling using SMOTE and Tokem-Link methods we have 555420 cases with the value of target =0 and 555420 cases with the value of target =1

The result of logistic classification on this dataset is as follows:

- Accuracy: 0.51
- Precision: 0.51
- Recall: 0.67

Figure 15 shows the confusion matrix of logistic classification over dataset after using SMOTE over sampling. This method performs a little better than random over sampling but again the result is not much different.

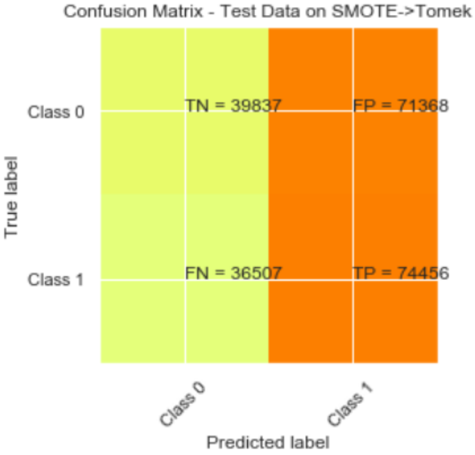


Figure 15: confusion matrix of logistic classification after SMOTE and Tokem sampling on initial dataset

XGBooster Classifier

Because the results obtained by logistic classification were very similar, we used XGBooster as a classifier to be more confident about the results and to be able to compare sampling methods. Figure 16 shows confusion matrix of XGBooster classifier after "random under sampling","random over sampling", "TomkenLink under sampling", "SMOTE over sampling" and "SMOTE->Tomek-Link sampling" on our initial dataset

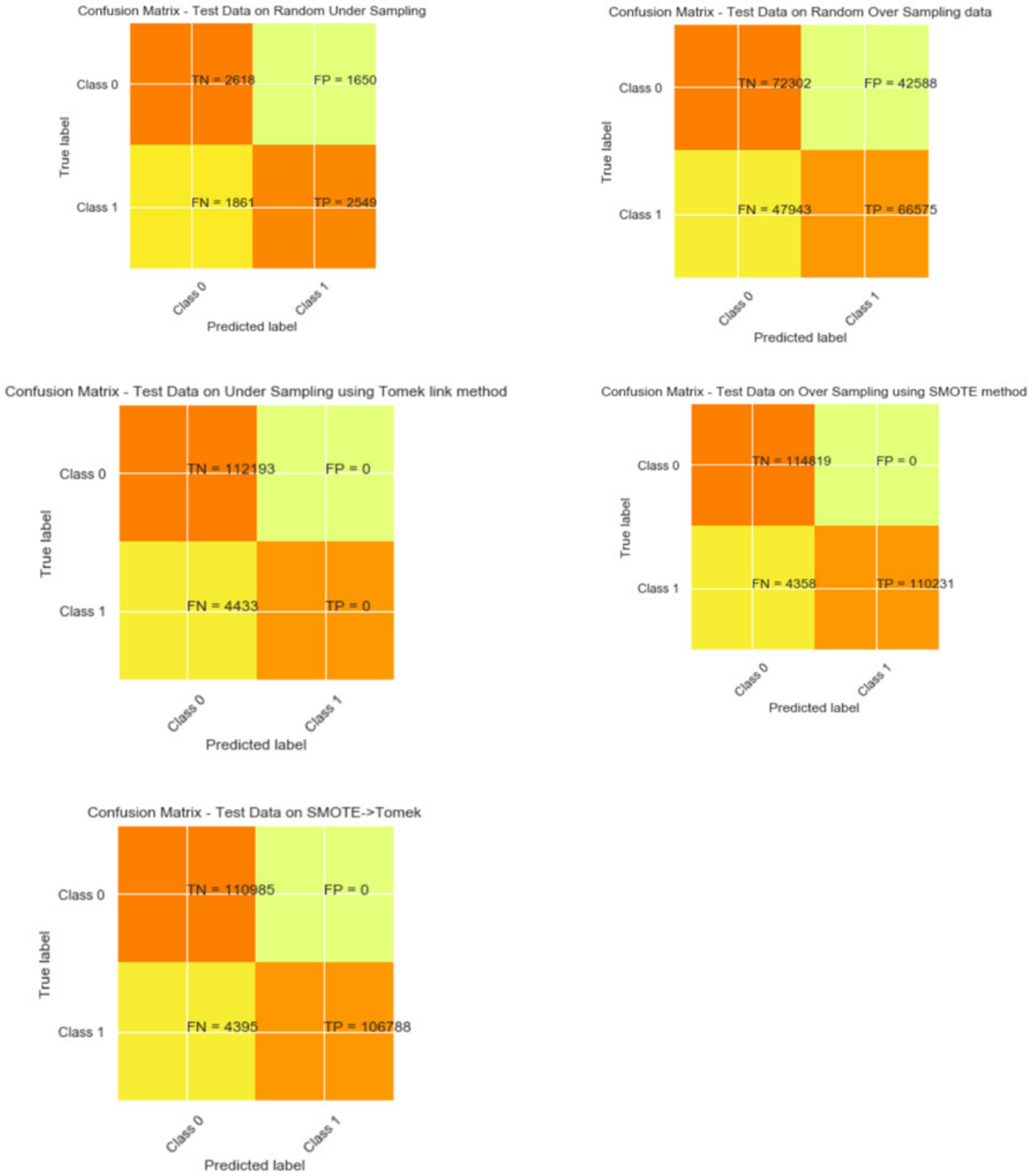
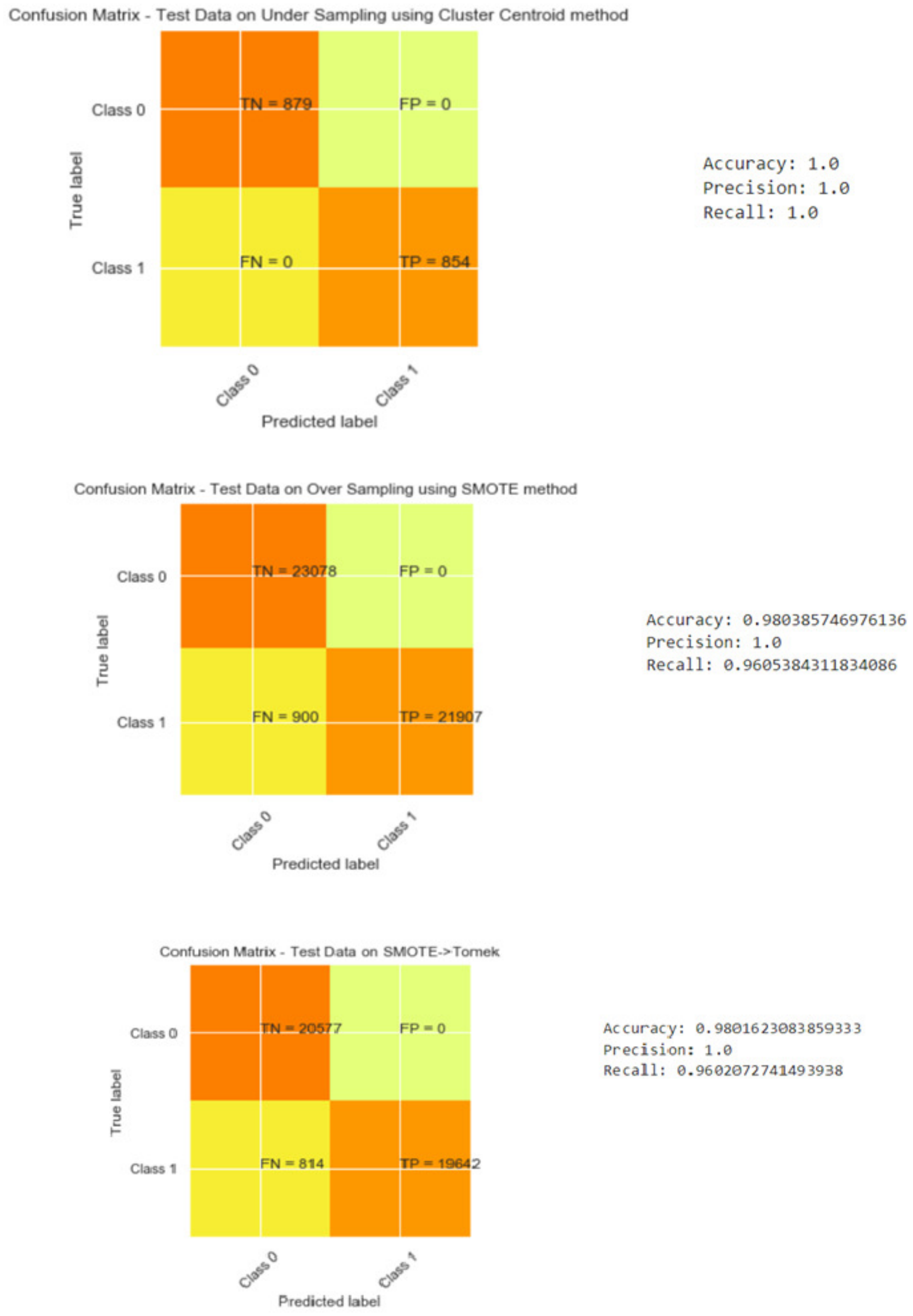


Figure 16: confusion matrix of XGBoost classification for random under sampling, random over sampling, TomkenLink under sampling, SMOTE over sampling and SMOTE->Tomek-Link sampling on initial dataset

Figure 17 represents confusion matrix of XGBooster classifier after "cluster centroid under sampling","SMOTE over sampling", " and "SMOTE->Tomek-Link sampling" on 20% of our initial dataset. We this partial dataset because of memory problem for cluster centroid sampling



```
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
```

```
Accuracy: 0.980385746976136
Precision: 1.0
Recall: 0.9605384311834086
```

```
Accuracy: 0.9801623083859333
Precision: 1.0
Recall: 0.9602072741493938
```

Figure 17: confusion matrix of XGBoost classification for Cluster Centroid under sampling, SMOTE over sampling and SMOTE -> TomekLink sampling methods on 20% of dataset.

Conclusion

Table 1,Table 2 and Table 3 summarize all results obtained for different sampling method.

We highlighted best result in each table. As we can see, on whole dataset, SMOTE and SMOTE-> Tomken have the best results on whole dataset and on partial dataset Cluster Centroid Under Sampling performs a little bit better than these two methods. However, when we used logistic classification SMOTE-> Tomken was better than Cluster Centroid method. So we cannot be sure which one is better for this dataset.

Table 1: result obtained by Logistic Classifier for different sampling method on whole dataset

	Accuracy	Precision	Recall
Random Under Sampling	0.5	0.5	0.56
Random Over Sampling	0.5	0.5	0.66
Tomken Link Under Sampling	0.96	0	0
SMOTE Over Sampling	0.51	0.51	0.66
SMOTE-> Tomken	0.51	0.51	0.67

Table 2: result obtained by XGBooster Classifier for different sampling method on whole dataset

	Accuracy	Precision	Recall
Random Under Sampling	0.59	0.6	0.57
Random Over Sampling	0.6	0.6	0.58
Tomken Link Under Sampling	0.96	0	0
SMOTE Over Sampling	0.98	1	0.96
SMOTE-> Tomken	0.98	1	0.96

Table 3: result obtained by XGBooster Classifier for different sampling method on 20% of dataset

	Accuracy	Precision	Recall
Cluster Centroid Under Sampling	1	1	1
SMOTE Over Sampling	0.98	1	0.96
SMOTE-> Tomken	0.98	1	0.96