****

Subject:- CSE523 Machine Learning

**Weekly Report 6**

**Section-1**

Submitted to faculty: Prof. Mehul Raval

Date of Submission: 08-04-2023

**Student Details**

| Roll No. | Name of the Student | Name of the Program |
| --- | --- | --- |
| AU2040122 | Aditi Vasa | Btech CSE |
| AU2040002 | Shrey Somani | Btech CSE |
| AU2040196 | Vandan Shah | Btech CSE |
| AU2040048 | Ronit Shah | Btech CSE |

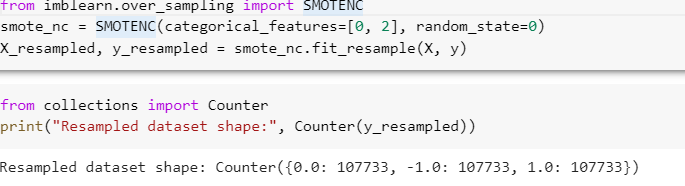
2022-2023 (Winter Semester)

For this week our goal was to do oversampling on our data. We have done this because there were some insufficient values in our data and to increase resolution and reduce noise.

Initially, in our project, we did data imputation but as we tried various prediction algorithms, we observed that our model was not learning the “large” class and in the confusion matrix, the whole column was turning out to be zero.

We realized that since the “fit” feature depends on multiple features, just doing data imputation on “weight” was not going to justify it.

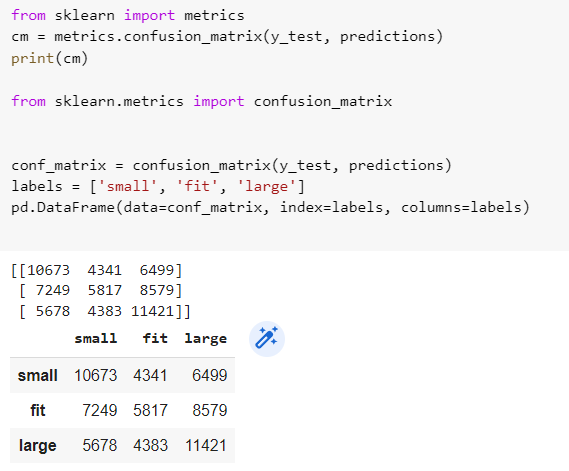
Henceforth, we decided to approach this problem using oversampling. Again in that we applied various approaches like random oversampling, ADASYM and finally used SMOTENC Oversampling.

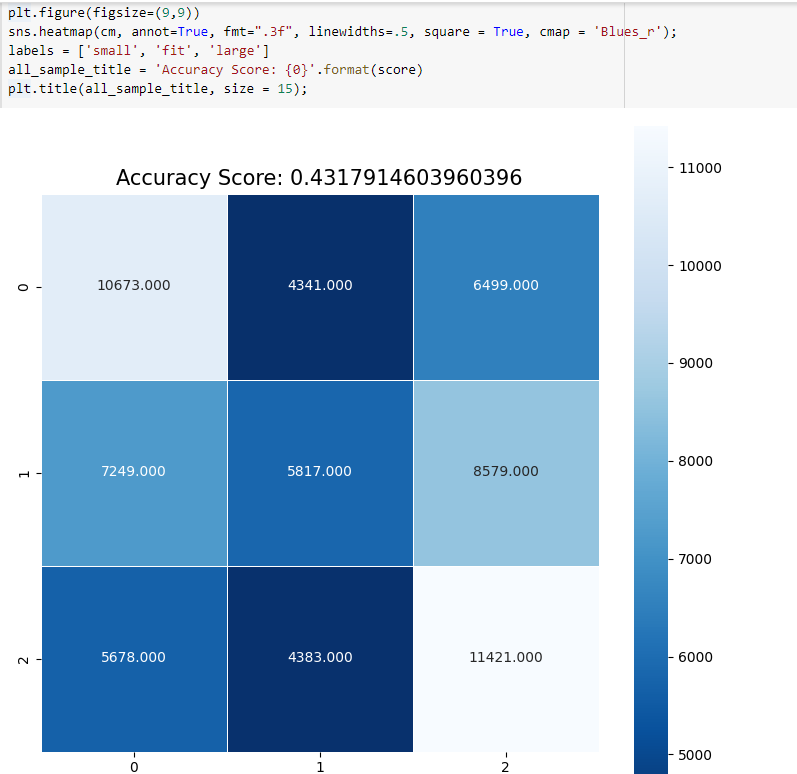


We used SMOTENC to generate synthetic data to oversample a minority target class in an imbalance dataset.

Oversampling is a technique used in data science to address the problem of imbalanced data. In an imbalanced dataset, one class of data may be significantly more prevalent than others, leading to biased machine learning models than favor the dominant class.

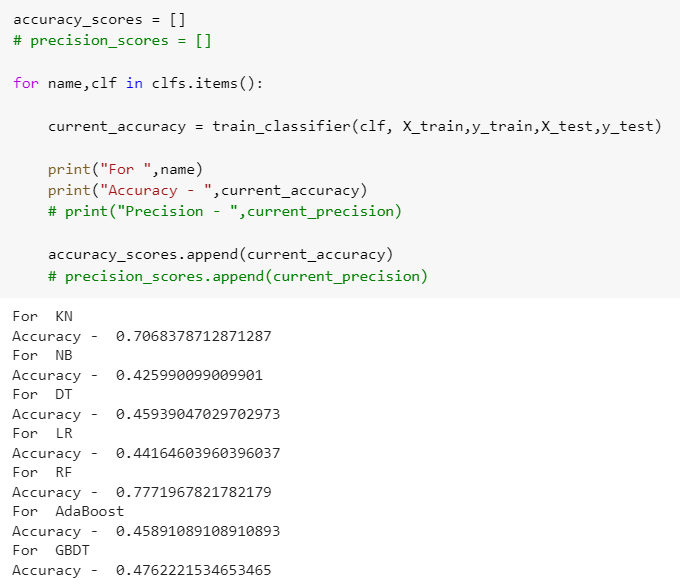
The data is finally normalized and confusion matrix turned out to be as we expected.





Lastly, we ran our model through different machine learning algorithms like K-Nearest Neighbors, Naive Bayes, Decision Trees, Logistic Regression, Random Forest, AdaBoost, and Gradient Boost algorithm and following are the accuracies we found.

Earlier we used the same algorithms on our model and accuracy was stagnant at 68%. However, we managed to increase the accuracy to 77% in Random Forest Classifier.



KN:- K-nearest neighbor

NB:- Naive Baye’s

DT:- Decision tree

LR:- Linear regression

RF:- Random Forest

GBDT:- Gradient Boosted Decision Tree

**TASK FOR UPCOMING WEEK**:

1. Improve the accuracy. The accuracy is about 70% for random forest, so we will work on further to improve the accuracy to 90%.
2. Implement ROC Curve and Hyperparameter Tuning.