ASSIGNMENT 2

Group no 57 - Akatsuki

Kakumanu Vamsee Krishna - 22111065
Sanket Sanjay Kale - 22111052
Pratik Mahipal Patil - 22111047
Shrawan Kumar - 22111080

1 Question 1 & 3

1) Methods that we would use to solve the problem:

1) Logistic Regression(OVA)
2) LWP
3) Decision trees
4) Neural networks

In the end we have chosen **Logistic Regression(OVA)** because it is more accurate than **LwP**, does less overfitting compared to **Decision Trees**, and takes less time and space than **Neural Networks**.

2) Logistic Regression(OVA)

Given data is pretty imbalanced, only 4-5 classes are accounting to majority of the data labels.
Because of this, Our Model was giving very poor Macro Precision.
It was giving around **65**% accuracy on Macro Precision for **Logistic Regression on Normal Weights (equal for all classes)**.

So,we tried using **Balanced Weights i.e., Giving more weight to Minority classes and lesser weight to Majority classes**.
It resulted in an increased **Macro Precision of 85% for Top 5 classes**, But it still gives **less accuracy of 55% for Macro for Top 1 class**.

To solve this we tried :

a)Upsampling minority classes:

We generated data for minority classes by using **sklearn library imblearn**.
We tried **SMOTE** which gave accuracy for Macro Precision for **Top 5** classes of **90**% and at **Top 1** around **67**%.

b)Hybrid sampling(Upsampling and downsampling both)

We generated data using **imblearn** library.
We used **SMOTETENN** function in **imblearn** for generating data.
It resulted in **Macro Precision of 87% for Top 5 classes and 59% for Top 1 class**.

Preprint. Under review.

c) Hyperparameters : C, iterations, tol(learning rate), solvers

We applied **Grid Search** for finding out the Optimal Parameter Values.
Final set of values of Hyperparameters are
C = 5, **iterations** = 5000, **tol** = 0.0001, **solver** = 'liblinear'

Gridsearch	C-value	Iterations	Learning Rate	solver	Macro@5 precision
1	0.1	1000	0.1	newton-cg	0.85
2	5	5000	0.0001	liblinear	0.90
3	10	5000	0.0001	lbfgs	0.88
4	100	5000	0.01	lbfgs	0.81
5	1000	10000	0.01	liblinear	0.80

2 Question 2

Advantages of Logistic Regression(OVA)

- Increased accuracy(macro precision)
- Reduced file size
- Reduced total time
- Reduced Model size
- Ease of coding and deployment

Disadvantages of Logistic Regression

- Overfitting
- Training time