CS771 Introduction to Machine Learning
Assignment 2

Bharat
22111074
bharat22@iitk.ac.in

Pradeep Chalotra
22111045
pchalotra22@iitk.ac.in

Pulkit Sharma
22111048
Pulkits22@iitk.ac.in

Sarthak Neema
22111079
sarthakn22@iitk.ac.in

Sumit Kumar Chaudhary
22111060
sumitkc22@iitk.ac.in

Problem 1:-

In this problem, we have used the Logistic Regression ML model as it has given us higher accuracy for the given training data as compared to the Decision tree ML model.

Processing performed on the feature:

1. An uneven dataset is presented in the training data. It's possible that models developed using such a dataset wouldn't be aware of the unusual class. Therefore, sampling has been done to balance the dataset so that the model will give the minority classes greater weight. To perform random over-sampling, we used RandomOverSampler, a Imblearn class has this function implemented. Selecting samples at random with replacement is intended to oversample the minority classes.

2. The training data which we had given is in the compressed matrix form so to feed it to the ML model, we have to convert it into feature vectors. For that we use .toarray() function which convert it into a dense ndarray.

Logistic Regression implementation:

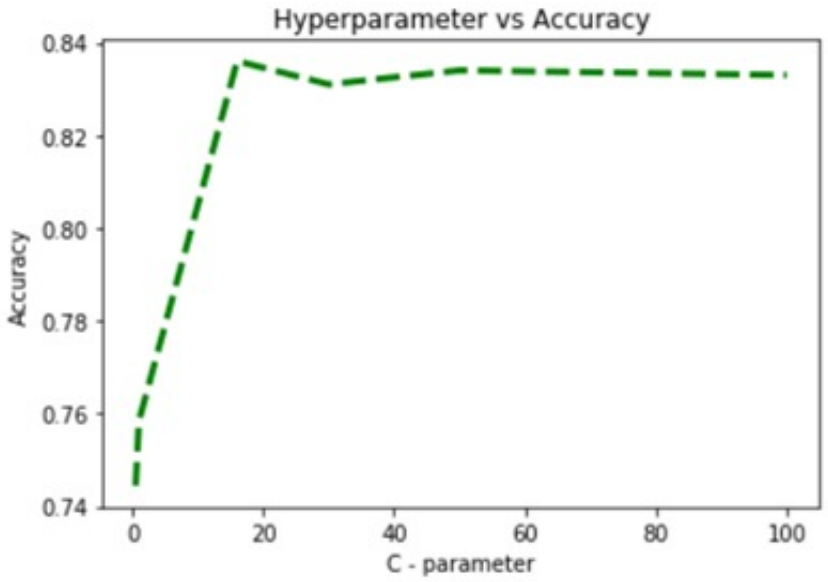
1. In the process of making the Logistic Regression ML model we have used the parameter 'multi_class' = 'ovr' (one vs rest) to implement it.

2. After that we have fit the model with the training data.

Hyperparameter Tuning:

1. We have used different set of values for 'C' hyperparameter 'C':[1,2,3,4,5,6,10,12,14,16,20,22,28,30,40,50] , after that we have applied Grid search CV on the trained model with cross validation (cv) = 5 which gives us the best 'C' value after computations.

   Reference: https://www.youtube.com/watch?v=n40hS9tQmcY&t=1042s



Problem 2:-

Predicting the top 5 classes of the error to which our data may belong is a condition of our problem statement. In order to do that, we are determining the likelihood of each class of error to which the provided data point may belong. From those classes, we will choose the top 5 most likely classes, which can be readily accomplished using logistic regression.

Advantages of Logistic Regression:

1. **Ease of Coding :**
   One of the simplest machine learning algorithms, logistic regression is straightforward to use and offers excellent training efficiency in our application. These factors also contribute to the fact that this technique doesn't need a lot of processing resources to train a model. The predicted parameters (trained weights) provide information on the relative weights of the various features. It also specifies if the relationship is positive or negative. In order to determine the link between the features, we canutilizee logistic regression.

2. **Model Size :**
   Logistic Regression is one of the easiest machine-learning algorithms to implement and provides great training efficiency. The Decision Tree model is considerably large as compared to the Logistic machine learning model.

3. **Training Time:**
   Due to the small model size, Logistic Regression model requires less time to train as compared to Decision Tree model which takes a bit more time to train.

4. **Less prone to over-fitting :**
   Although it can overfit in large dimensional datasets, logistic regression has a lower tendency to do so. To prevent over-fitting in these cases, regularisation (L1 and L2) approaches may be taken into consideration.

Disadvantages of Logistic Regression :

1. **Prediction:**
   The major limitation of Logistic Regression is the assumption of linearity between the

dependent variable and the independent variables. Non-linear problems can't be solved with logistic regression because it has a linear decision surface. Linearly separable data is rarely found in real-world scenarios.

2. **Overfitting :**
   If the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise, it may lead to overfitting.

3. **Dataset:**
   A large dataset and enough training examples are needed for all the categories that Logistic Regression needs to detect. Each training example must be separate from every other sample in the dataset.The model will attempt to prioritise those particular training instances if they are somehow connected. Therefore, matching data or repeated measurements shouldn't be used to generate the training data.