# Report on Offline 2

Prepared by : Tanveer Rahman , (1905025), CSE, BUET

# Run the Notebook

### Assumptions

- The datasets are in the same directory as the notebook
- they are named as "churn.csv", "adult.data", "adult.test", "creditcard.csv"

### Procedure to Run the Code

There are Three Preprocessing cells in my code, each for the dataset.

- **First Dataset PreProcessing (Churn)**
- **Second DataSet PreProcessing (adult)**
- **Third Dataset Preprocessing (CreditCard)**

One just need to comment the two other cells in order to run the other.

For Example: Suppose I want to run the "Churn Dataset". For this :

- First Dataset PreProcessing (Churn) : This cell needs to be uncommented.
- Second DataSet PreProcessing (adult) and Third Dataset Preprocessing (CreditCard) : These cells need to be commented

# Some Preprocessing overview

### First Dataset

- In the TotalCharges column, some spaces make the datatype as object. So, after identifying the spaces and then filling up them with the mean of the column, the column is converted to float64 datatype. This helped to reduce a lot of columns when one hot encoding is done

### Second Dataset

- There are some columns of 'Object' datatype who has NaN values. So I have removed the rows whose columns have the NaN values. There is also option for finding the mode and then replacing the NaN with the mode value. Just need to uncomment.
- Here we have a file named 'adult.test' I have used this file as the test set
- From the file 'adult.data', I used it inorder to create the train and the validation set.

### Third Dataset

The third dataset is imbalanced, where the positive instances are scarce. And it also has huge amount to data. So the steps I have taken :

- Under Sampling : I have taken 20,000 of the negative data and all the positive data as my dataset and from them, I have created the test set, train set and validation set

- Assign weights : In order to reduce the biasness, I have introduced weights to the classes. So that the biasness is reduced and the model gives importance to the minority class
- Decrease the Learning rate: Reduced the learning rate in order to improve the stability of the gradient descent optimization.

# Feature Selection

Both the Correlation and infromation gain is used in order to select the features

> *First the correlations and information gain against the target is calculated Then the top features are selected combining both of them ( Sum and then sorting )*

# Performance Evaluation

## First dataset

| Metric | Accuracy | Sensitivity | Specificity | Precision | F1-Score | AURO( |
|---|---|---|---|---|---|---|
| Logistic Regression | 0.683468 | 0.851841 | 0.628366 | 0.432813 | 0.573223 | 0.7383 |
| Voting Ensemble | 0.710320 | 0.815341 | 0.675214 | 0.456280 | 0.585117 | 0.7452 |
| Stacking Ensemble | 0.707473 | 0.821023 | 0.669516 | 0.453689 | 0.584429 | 0.7452 |

- Voting Ensemble has the highest accuracy, means it correctly classifies a larger proportion of instances.
- Logistic Regression has the highest sensitivity, meaning it identifies a significant portion of actual positive cases. The Stacking ensemble does not fall far behind.
- Stacking ensemble has the higher specificity - identifying negative cases
- F1-Score of Stacking ensemble is also higher than LR models. Good balance between precision and recall
- AUROC means to distinguish between classes. Voting ensemble has the highest.
- AUPR means performance in predicting the positive class relative to all positive instances.

## Second dataset

| Metric | Accuracy | Sensitivity | Specificity | Precision | F1-Score | AURO( |
|---|---|---|---|---|---|---|
| Logistic Regression | 0.721817 | 0.805621 | 0.695668 | 0.462021 | 0.586538 | 0.7492 |
| Voting Ensemble | 0.735835 | 0.746757 | 0.732277 | 0.476133 | 0.581501 | 0.7395 |
| Stacking Ensemble | 0.734839 | 0.748108 | 0.730515 | 0.474949 | 0.581024 | 0.7393 |

- The accuracy is higher in stacking
- The mean LR models can predict the positves more (Sensitivity)
- Stacking is more effective in predicting the negatives than the LR models. (Specificity)
- Voting ensamble slightly outperforms both in precision ( Accuracy to positive predictions)
- Higher AUROC means it has a better ability to distinguish between classes.

## Third Dataset

| Metric | Accuracy | Sensitivity | Specificity | Precision | F1-Score | AUROC |
|---|---|---|---|---|---|---|
| Logistic Regression | 0.988260 | 0.550685 | 1.0 | 1.0 | 0.709596 | 0.7753 |
| Voting Ensemble | 0.973871 | 0.000000 | 1.0 | 0.0 | 0.000000 | 0.5000 |
| Stacking Ensemble | 0.980220 | 0.242991 | 1.0 | 1.0 | 0.390977 | 0.6214 |

- Since the dataset is imbalanced, The Voting ensemble fails other performance evaluations except good accuracy
- Logistic Regression has the highest sensitivity, F1-score, AUROC, AUPR. Since the data is imbalanced, Stacking assembly fails in this case. Because of the random sampling the training set may become more biased to the majority class and affect the performance metrics
- Specificity (1.0) means correctly predicts all the negative instances
- Precision (1.0) means when the predict true to a positive data, they are always correct.

# Observations

The Stacking ensemble sometimes, performs less that LR models. This may be caused because of the sampling of our data. Because of the sampling the data may overfit and caused some reduction in the performance.