Tanmoy Baidya
B.Tech,Part-IV
IIT(BHU) Varanasi
Metallurgical Engineering

Event: **Analiticity, Technex'18**
Problem: **Loan Defaulter Prediction**

Methodology:

1. Missing Values Imputing:

1.1 **VALUE , TJOB, RATIO, CL_COUNT** columns are imputed with their respective mean value.

1.2 **DUE_MORTGAGE, CLT** features are imputed with their respective median value.

1.3 **OCC** and **DCL** are imputed with value 0.

1.4 **REASON** is imputed with value 1.

2. Modelling

2.1 Logistic Regression: A simple linear logistic regression model is trained using standardised data which gave me Cross-Validation Score("Auc-Roc") of 0.7824 with Standard Deviation of 0.0462.

2.2 XGBoost Model: Tree-based Xgboost model is used. It gave me Cross-Validation Score("Auc-Roc") of 0.9364.

XGBoost Model Parameters:

```
" learning_rate":0.1,
 "n_estimators":1000,
 "max_depth":8,
 "min_child_weight":6,
 "gamma":0.1,
 "subsample":0.95,
 "colsample_bytree":0.95,
 "reg_alpha":2,
 "objective":'binary:logistic',
 "eval_metric": 'auc',
 "scale_pos_weight":1
```

2.3 Ensembling: To reduce bias and maintain homogeneity of the model 90% weightage of Xgboost and 10% weightage of linear model is taken.

3. Tools Used: Python 3.6(Anaconda Distribution), JupyterNotebook, Pandas, Numpy, Scikit-learn, Seaborn, Matplotlib, Xgboost.

Note: Please open my JupyterNotebook file for Data Exploration, Data Visualization and Data preprocessing part. Everything is explained there.