**India ML Delinquency classification**

**Chosen Approach:**

1. Peformed EDA- 'number\_of\_borrowers' and 'co-borrower\_credit\_score' seems to be highly correlated, hence dropped 'number\_of\_borrowers'.
2. Removed outliers only considering using LOF with ‘knn’ algorithm and contamination = 0.001, this step increased the F1 score by 0.011
3. Constructed polynomial features of degree=2 for numerical features and One-Hot-Encoding for categorical features as the cardinality of categorical features was not high.
4. Trained a baseline Logistic Regression model, got poor F1 score = 0.1619.
5. Trained AdaboostClassifier and LGBClassifier with intense hyperparameter tuning, got a respectable F1 score = 0.5141. And chose LGBClassifier model as final mode.

**Failed trials:**

1. Since the data is highly imbalanced I experimented with different over-sampling techniques such as KMeansSMOTE, SVMSMOTE, ADASYN but every model was overfitting and it wasn’t improving my score in the leaderboard even though I was getting high accuracy and F1 score around 0.997
2. Tried SMOTENC for oversampling without converting categorical features to OHE for building a CATBoostClassifier. Even this model overfitted with high accuracy and F1 score around 0.997.
3. Had only considered OHE for categorical features with over-sampling got me high F1 score around 0.995. But considering polynomial features also with OHE features got me better F1 score.