**ZS Problem Solving Approach**

1. Appended the train and test data to create features and fill null values together
2. Creating new features & Filling up null values
3. Null values were filled up with mean values
4. Feature columns created and their significance:
   1. **“avg\_char\_wa\_msg”**: Avg. Characters used in each whatsapp message. Similarly for **“avg\_char\_twitter”** for each tweet, **“avg\_char\_fb”** for Facebook statuses, **”avg\_char\_email”** for each mail
   2. **“total\_charge”**: total charges across all social sites
   3. **“total\_posts”**: total posts made across all social sites
   4. **“%social\_sites”**: considering 5 social stuff (fb,whatsapp,twitter,snapchat & email) checking the % of social sites the user has been active on
5. Encoding the categorical variables and mapping “Yes”, “No” as 0 and 1, boolean values
6. **CatBoost** Algorithm for Modelling because of its performance. Its also robust and does not need extensive hyper-parameter tuning. Although I used default parameters only. Also since data is less Catboost has lesser chances of overfitting. Although I compared three models on the validation set, XGBoost and Random Forest out of which CatBoost gave me the best score on validation set. For validation part, I used stratified train\_test\_split because of the highly imbalanced dataset.