**Analytics Vidhya JobAThon**

**Brief Approach:**

I have one hot encoding to encode the categorical variables. I have created a new feature for the missing column in Credit\_Product. I have used 15 sets of stratified K-Fold to split the data into training and validation set. I have used the stratified model (XgBoost, LightGBM and CatBoost) to get the probabilities of prediction.

**Feature Engineering/Preprocessing:**

I had decided to create a separate feature for missing data column in Credict\_Product because I found the missing rows in Credit\_Product has a different data distribution than yes/No values for Credit\_Product. Yes/No values has a more uniformly distributed to the target variable whereas the missing data predominantly related to the ‘yes’ value of the column.

Initially, I decided to use the count as a value for the region column but I found the AUC-ROC value to be lower than when I used the OHE of the region column. So I used OHE for the region column. For all other categorical columns. I used OHE to encode the data.

**Modeling:**

I used stacked XGBoost, CatBoost and LightGBM to calculate the probabilities for the prediction. I had used Logistic Regression, Random Forest Classifier, Artificial Neural Network, XGBoost, CatBoost and LightGBM but I found the stacked XGBoost, CatBoost and LightGBM to product the best AUC-ROC score.