# **JOB-A-THON - June 2022**

**POTENTIAL LEAD IDENTIFICATION**

**EDA and Feature Engineering -**

* Checked for the proportion of both the classes in the dataset. The data turned out to be imbalance- 95% data of class ‘0’ and only 5% of the data was of class ‘1’
* Checked the correlation between the ‘product\_purchased’ and the dependent variable by plotting count plots. No discernible dependence was visible. Therefore filled the null values with ‘0’ as the null values are distributed with the same proportion as the other product\_purchased values (1,2,3,4) among the two classes (of dependent\_variable)
* Checked the distribution of ‘campaign\_var\_1’ and campaign\_var\_2’ with respect to the two classes. There were 16 and 15 categories present respectively in each of the variables. Clubbed some of the categories within these variables based on their distribution w.r.t the dependent variable i.e. if any two categories have the almost same % of 1( in dependent\_variable) then those categories are clubbed as they exhibit the same variance.
* Analyzed the ‘signup\_date’ feature. Around 15k values were null but only a handful (25) of those null are present in the rows with buy=’1’. Therefore, made an assumption that these users have not signed in because if a user has to buy a product he/she has to sign in majority cases. Created a new variable ‘has\_signup’ which is 1 if signup\_date is present and 0 if it is null
* Created a column - ‘date-diff’ = ‘created\_date’- ‘signup\_date’ which indicates for how long the user has been using the website. Plotted a distplot to identify the trend in this column across both the classes. Found out that for class ‘1’ the value is more positive (date\_diff is large) and for class the values are more negative ( user has signed up after the created\_date)
* Checked the ‘user\_activity\_val’ columns against both the classes. Most of the categories in these features differ in terms of their distribution across the classes in dependent variable except for categories 2,3,4 in the ‘user\_activity\_val\_11’ feature which are quite similar w.r.t the dependent variable

**Data Preparation for Modeling -**

* Encoded the categorical variables
* Dropped ‘id’,’created\_at’ and ‘signup\_date’ column
* Checked for the correlation among the features by creating heatmap
* Performed train-test split - used stratification to ensure that the percentage of class-’1’ is same in both training and testing set

**Modeling -**

**Random Forest Classifier -**

* Implemented random forest classifier with grid search , the evaluation metrics results differ substantially for the train and the test data set
* Checked for the feature that was having the highest feature importance, the feature was date\_diff, removed this feature from modeling as it was causing overfitting
* Ran the model again after removing the ‘date-diff’ column and got the base f1 score of 68.7 ( after uploading the results on the website)
* Tried oversampling of the minority class but the result didn’t improved
* Tried the model run using the ‘class\_weight’ = ‘balanced’ parament , but the results got worse

**LightGBM -**

* Implemented LightGBM with grid search, the results depict that there was overfitting
* Tried dealing with class imbalance with oversampling and setting up ‘class\_weight’ parameter but the results didn’t improve.

**XGBoost -**

* Implemented the Xgboost algorithm with grid search but it didn’t yield better results

**Final Algorithm -**

*Finally used random forrest with the below parameters -*

* *Undersample the data such that sampling\_strategy=0.2*
* *N\_estimators : 300*
* *Max\_depth : 12*
* *Max\_features : log 2*

*This resulted in the score of 72.6 on the analytics vidhya site*