CREDIT CARD LEAD PREDICTION

- First, I have loaded train, test and sample submission dataset to get a knowledge about them.
- Then for data understanding I checked for the shape of the dataset, info about the dataset, what’s the mean median mode value along with the min and max value of the numerical columns present in the dataset.
- Then I checked for the null values present in the dataset and found out that only one column Credit_Product is having null values.
- Then I checked whether there is any class imbalance present in the dependent variable or not.
- In EDA section I went for bar plots to check the majority class in all the categorical columns, and then checked for the lead count present in different section which can give us insights that which variables in which columns are important.
- From the Occupation column it was observed that the **Entrepreneur** section has a greater number of leads compared to the number of not leads.
- Then for numerical columns I checked for the distribution of the data using histogram and boxplot and found out that there is skewness present in the numerical columns.
- While checking for the region code section I found out that RG268, RG283, RG284 and RG254 are the top four regions in generating leads.
- In feature engineering as only Credit_Product column had null values and there was no clear info present regarding the null values, so I converted those null values into “No info”.
- In Age section I plotted two Kde plots taking one as lead=1 and another as lead=0. From there I observed that people between the age range of 35-65 are more eligible to take credit cards, followed by people between age range of 20-35 and 65& above. But in 20-35 age range it was observed that maximum number of not leads are also present. So, I gave High priority to 35-65(Adult) range followed by 65&above (Senior) and then 20-35(Young) During Label encoding.
- In the occupation section the Entrepreneur section has quite low count compared to other sections, so I combined them with Self_Employed section.
- Similarly for occupation column also High priority was given to Self_employed section followed by Salaried and Other.
- One hot encoding was applied to Gender, Channel_Code, Credit_Product, Is_Active.

- As there was skewness present in the average account balance column, so I applied log transform to it.
- For region code variable I applied label encoding in descending order of maximum to minimum lead generation by regions.
- The after all of this used standard scaler to scale all the values to a particular range where the values are centered around the mean with a unit standard deviation.
- For modelling I opted for Decision tree, Random Forest, XGBoost and light gradient boosting. As it was taking way too much time for hyperparameter tuning in notebook, did the tuning part for all the models in google colab and finally light gradient boosting with Kfold cross validation and hyperparameter tuning gave best roc_auc score.
- The ROC_AUC - CV Score I got from the model was 0.8717515332852127.