**Problem Statement - Credit Card Lead Prediction**

**Happy Customer Bank is a mid-sized private bank that deals in all kinds of banking products, like Savings accounts, Current accounts, investment products, credit products, among other offerings. Now, the bank is looking for your help in identifying customers that could show higher intent towards a recommended credit card.**

**Approach -**

**The bank also cross-sells products to its existing customers and to do so they use different kinds of communication like tele-calling, e-mails, recommendations on net banking, mobile banking, etc.**

**In this case, the Happy Customer Bank wants to cross sell its credit cards to its existing customers. The bank has identified a set of customers that are eligible for taking these credit cards**

**Step 1 - Importing the Relevant Libraries & Load Dataset**

* **Import all the necessary libraries like sklearn, seaborb, matplotlip and all the relevant libraries.**
* **Load the dataset and check shape & size of the data.**

**Step 2: Data Cleaning & Feature Engineering**

* **Remove the ID column from the dataset. And check for the numeric and categorical features present in dataset.**
* **After finding out the categorical features implement Label Encoding to convert categorical data to numeric data. Label encoding is used because we don’t want to increase the size of the columns.**
* **Feature Region\_code has max numbers of sub categories if we apply One Hot Encoding, Dataset will suffer from curse of the dimensionality it will lead to poor performance of the model.**
* **The feature Avg\_Account\_Balanced has large numbers of outliers, data is rightly skew respectively.**
* **To remove the skewness log transform is used respectively**

**Step 3: Exploratory Data Analysis**

* **By using distplot from sklearn library we can plot the graph of data distribution.**
* **Data distribution shows that our data have right side skewness in Avg\_Amount\_Balanced feature and all other features are nearly normally distributed.**
* **The data is ranging differently for each & every column we need to scale down data to have standard distribution of data.**
* **At Initial approach with random forest and XGBoost are performing well but the accuracy of the model on test data was not good as expected.**
* **Thats why kfold cross validation is used to train the model on clusters to get bet accuracy.**
* **Kfold cross validation is used to obtain best cluster of the data. Total 9 folds are tested.**

**Step 4: Model Building**

* **First we built XGBoost model for prediction, without hyperparameter, got roc\_auc\_score = 0.619735.**
* **Random Forest Classifier without hyperparameter tuning got roc\_auc\_score = 0.62669**
* **XGBoost with hyperparameter tuning we got roc\_auc\_score = 0.71792241**
* **CatBoost Classifier with hyperparameter tuning we got roc\_auc\_score = 0.7311792**
* **LGBM Classifier with hyperparameter tuning we got roc\_auc\_score = 0.7314092**
* **By comparing the scores we can conclude that, LGBM Classifier performs good.**

**Final Model Selection**

**Hence by comparing the scores and performance we finally select the LGBM Classifier as prediction**

**model.**