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.

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.

Now, the bank is looking for your help in identifying customers that could show higher intent towards a recommended credit card, given:

- Customer details (gender, age, region etc.)
- Details of his/her relationship with the bank (Channel_Code,Vintage, 'Avg_Asset_Value etc.)

Data Dictionary

Train Data

Variable	Definition
ID	Unique Identifier for a row
Gender	Gender of the Customer
Age	Age of the Customer (in Years)
Region_Code	Code of the Region for the customers
Occupation	Occupation Type for the customer
Channel_Code	Acquisition Channel Code for the Customer (Encoded)
Vintage	Vintage for the Customer (In Months)
Credit_Product	If the Customer has any active credit product (Home loan, Personal loan, Credit Card etc.)
Avg_Account_Balance	Average Account Balance for the Customer in last 12 Months
Is_Active	If the Customer is Active in last 3 Months
Is_Lead(Target)	If the Customer is interested for the Credit Card 0 : Customer is not interested 1 : Customer is interested

Test Data

Variable	Definition
ID	Unique Identifier for a row
Gender	Gender of the Customer
Age	Age of the Customer (in Years)
Region_Code	Code of the Region for the customers
Occupation	Occupation Type for the customer
Channel_Code	Acquisition Channel Code for the Customer (Encoded)
Vintage	Vintage for the Customer (In Months)
Credit_Product	If the Customer has any active credit product (Home loan, Personal loan, Credit Card etc.)
Avg_Account_Balance	Average Account Balance for the Customer in last 12 Months
Is_Active	If the Customer is Active in last 3 Months

Evaluation

The evaluation metric for this competition is **roc_auc_score** across all entries in the test set.

Objective: To study selling of credit card to the customers

Target attribute: Is_Lead , It's a Binary classification problem since the target attribute has only two class.

The problem is approached is as follows

The problem is approached with checking the uniqueness, counts and value_counts of the attribute. And it was found that ID is 100% unique which does not show any significant in deriving the target attributes. Removed Id columns

Checking the patterns in the data.

1. For checking the patterns customised functions are developed is named as **Symbol** in code. There is possibility that entry of the data may be fault or mismatch in the data

pertaining to the particular attributes. But the attributes are clear when it checking with the pattern.

2. Separated the attributes by numeric and categorical attributes for visualisation. It was found that the **distribution of the numeric attribute Avg_Account_Balance is skewed right side**. Due to the **presence** of **outliers** from the customised function (**Univariate stats**). Thus applied **outlier clipping** and considering the distribution between 1 to 99% of the data.
3. Again checking the distribution and shows little **less skewed**. For this attribute transformation such as log transformation is applied. And checking the distribution was fairly found normally distributed.
4. Checking the Percentages of **null vales** and it was found that only in the credit_product null value is observed with 11% and later that part null value is imputed using **simple imputer function** with **most_frequent** observed value.
5. **Checking the target distribution: data is not imbalanced :**
   Label 0 : 76.27 Label 1: 23.72
6. **S**ince **region** and **channel not give any significant with respect to dummification or one hot encoding**. Need to be handled different way either by **categorical embedding or target encoding.**
7. Problem with the target encoding is entirely depend on the global mean and its weightage. Instead of target encoding, it is better to represent the each level of region_code and channel_code n-dimensional vectors using categorical embedding

Train test splitting:

1. Taking **region_code**, **channel_code**, **other independent numeric** and **categorical attributes separately** and **splitted** into train as 0.8 and validation set as 0.2. Since the **embedding** is need to be done for **region_code** and **channel_code** separately.
2. Checking the **null values** for each of the columns and is found that Credit_product is found to be 11%. And it is imputed with **simple_imputer** with a strategy **Most_frequent.**
3. **Before applying categorical embedding for Region and channel code**. That needs to be **converted** to **numeric attributes** and is achieved through **Label encoder** function.
4. **Other independent numerical and categorical attributes** are handled by applying **standard scalar** and **one hot encoder** and finally concatenating the independent attributes.

Categorical Embedding Through Functional API

1. For **Region and Channel embedding vectors are generated using Embedding layers** and these two layers are concatenated with other independent attributed and this is given to **Neural network of** one **hidden layer with a 8 neurons** with and activation function **relu** and **output** layer with a **sigmoid** activation function.
2. Model is then fitted for train data sets and **output of the embedded layer** is collected and **attached to the original train and validation data set using merging operations**.

Deriving Non Linear feature(Auto encoder and Decoder)

- Other Nonlinear features of 1**0 attributes** are derived using **auto encoder and decoder part** and fitted and transformed for both train and validation data sets. **These nonlinear features are derived and attached to the final train and validation data sets**.

Test Data sets

- All the processing and transformation is applied to test data as it is applied for train data sets

Model Building

1. Base model logistic regression is built with ROC_AUC_SCORE =0.55
2. Base model logistic regression with liblinear solver is applied and found is built with ROC_AUC_SCORE =0.668
3. Base model Ridge regression with class weight balanced is applied and found ROC_AUC_SCORE =0.667
4. Base model Decision Tree classifier with class weight balanced is applied and found ROC_AUC_SCORE =0.667

Grid Search

1. **Grid search with hyperparameter for Decision Tree classifier with class weight balanced is applied and found ROC_AUC_SCORE =0.707**
2. **Grid search with hyperparameter for Random Forest classifier with class weight balanced is applied and found ROC_AUC_SCORE =0.77**

Auto ML Model:

predicted_probablity_with_auto_ML = 0.872

Conclusion:

Finally, I Choose a model since getting 0.872 ROC_AUC_SCORE