Analytic Vidya – Job A Thon May 2021

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

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

Solution Approach:

Exploratory Data Analysis:

1)Shape of Dataset

```
Train dataset   : (245725, 11)
Test dataset    : (105312, 10)
```

2) Datatype of dataset:

```
In [20]: train.dtypes

Out[20]: ID                    object
         Gender                object
         Age                    int64
         Region_Code           object
         Occupation            object
         Channel_Code          object
         Vintage                int64
         Credit_Product        object
         Avg_Account_Balance    int64
         Is_Active             object
         Is_Lead                int64
         dtype: object
```

3)Checking for Duplicate values in Test and Train dataset:

Duplicate value is not available in test as well train dataset.

```
In [26]: train.duplicated().sum()

Out[26]: 0
```

4)Checking for null values in the dataset:

Train Set:

Null values present in feature **Credit_Product.** which has **11.9%** [29325] of missing value.

Test Set:

Null values present in feature **Credit_Product.** which has **11.8%** [12522] of missing value**.**

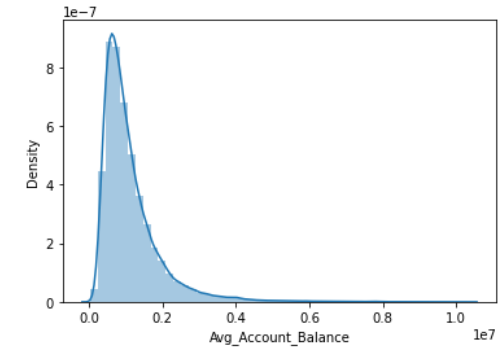
For handling the Null values, we were imputing the null values with the value **'Yes'** in both train and test data set.

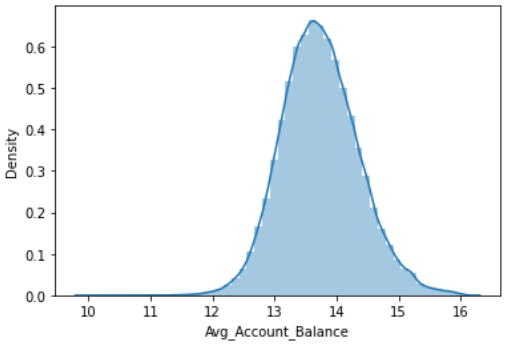
```
In [93]: train['Credit_Product']=train['Credit_Product'].fillna('Yes')

In [94]: test['Credit_Product']=test['Credit_Product'].fillna('Yes')
```

5) Handling Outliers:

We observed that feature [**Avg_Account_Balance**] is highly right Skewed. So we apply log transformation method to normalize the feature.



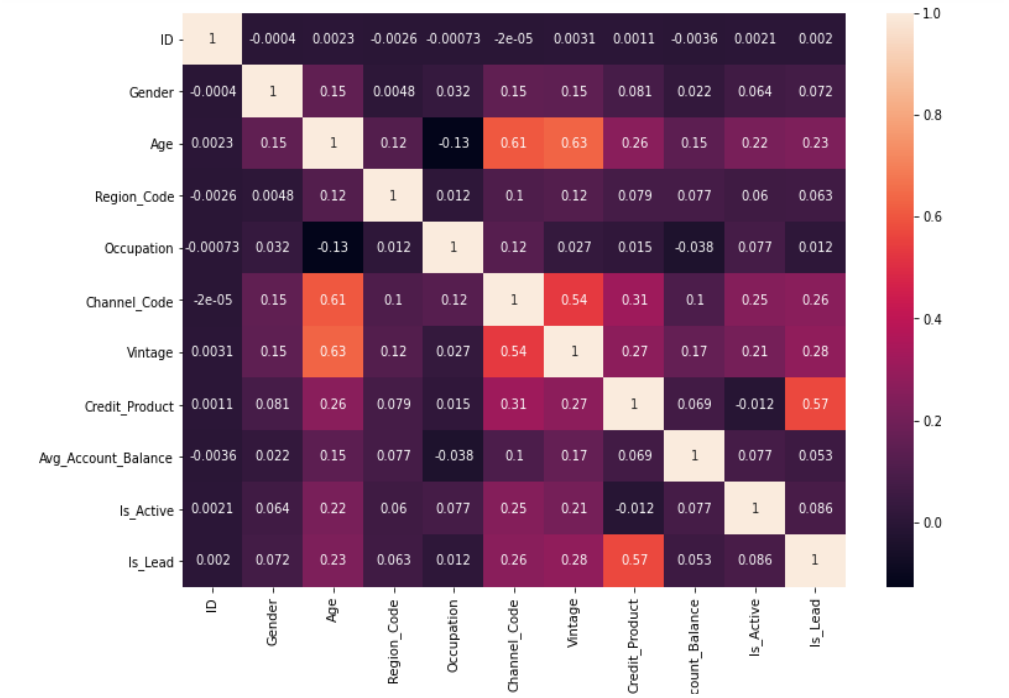


6) Encoding Categorical features:

We were using Label Encoder method to encode features having more than two categories.

7) Checking Multicollinearity:

Here we observed thers is no high corelation among the features



Note:

Our target variable in the dataset is highly imbalanced.

```
In [22]: train['Is_Lead'].value_counts()

Out[22]: 0    187437
         1     58288
         Name: Is_Lead, dtype: int64
```

Modelling Technique:

1) Logistic Regression :

Here we used Based model as Logistic Regression model.Which is a statistical model that in its basic form uses Logistic function, also known as sigmoid function to model the binary dependent variable.

Classification Report:

Test Set	Precision	Recall	F1-Score
0	0.77	0.94	0.85
1	0.31	0.12	0.18

Confusion Matrix:

True Positive	58037
True Negative	2225
False Positive	3732
False Negative	17096

Accuracy: 0.7431495868787766

Since , Our model is not performing well in False Negative/Positive rate and Accuracy rate is also not good.So we are going to test our model in other classification techniques and will select the model which performing well in our dataset and having high accuracy and AUC-ROC score.

2) Stacking Classifier with Ada Boosting:

As our model is a slow leaner ,so we were using boosting technique to convert our model to strong leaner.Here we were using ADA Boosting method as an estimator .

Base learner for our model will be :

1)Decision Tree clasifier

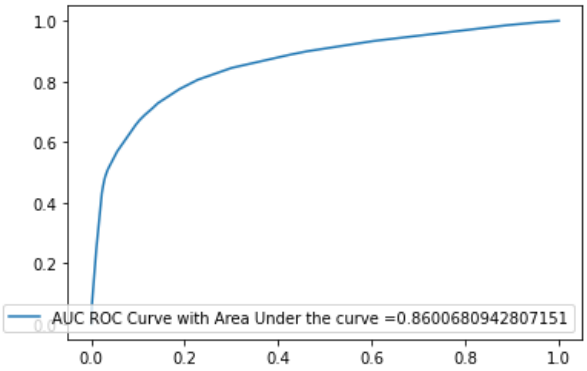
2) Random Forest Classifier

Final Estimator:

1)Ada Boosting Classifier

AUC-Score: 0.860068094

```
In [50]: generate_auc_roc_curve(stack_model_AdaBoost, X_test)
```



Using stacking classifier technique our model performed well and AUC-ROC score good compared to the logistic regression.We will compute further modelling with other techniques and choose the model with high accuracy / AUC-ROC score.

3) Extreme Boost Classifier (XG Boost Classifier):

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. XGBoost is fast. Really fast when compared to other implementations of gradient boosting.

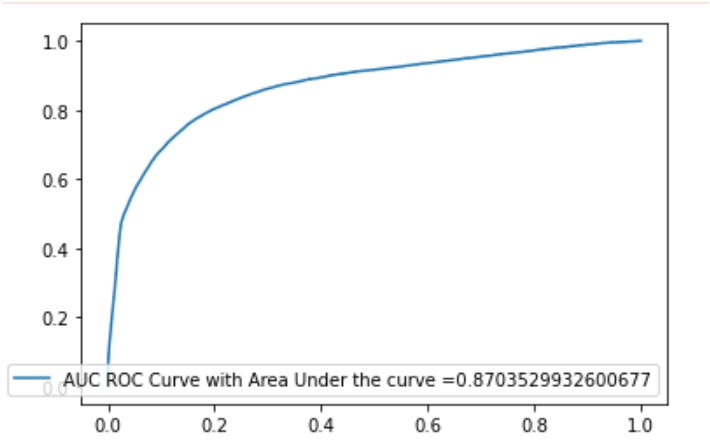
Classification Report:

Accuracy	0.85911
Precision	0.79194
Recall	0.55437
F1-Score	0.65219

Confusion Matrix:

True Positive	58955
True Negative	10711
False Positive	2814
False Negative	8610

ROC-AUC Score:



AUC-ROC Score: 0.8703

Model Comparision :

Scores Comparision Logistic Regression vs XGBoost technique:

Metrics	Logistic Regression	XG Boost
True Positive	58037	58955
True Negative	2225	10711
False Positive	3732	2814
False Negative	17096	8610
ROC-AUC Score	0.74	0.87

In XG boost, False Negative rate is drastically less than the logistic regression and accuracy score also high compared to logistic model.

Classification Report for Logistic Regression:

Test Set	Precision	Recall	F1-Score
0	0.77	0.94	0.85
1	0.31	0.12	0.18

Classification Report for XG Boost:

Test Set	Precision	Recall	F1-Score
0	0.87	0.95	0.91
1	0.79	0.55	0.65

In classification report Precision,Recall, F1 score is comparatively high than logistic model.

Conclusion :

From the above comparisons **Extreme Gradient boosting** perform well in our dataset.So in our model we had used XG Boosting model for predicting the target feature and the performance of the model is good in test data with an accuracy of **0.8703.**