CREDIT CARD LEAD PREDICTION

This Document contains the approach envisioned and taken to build a Machine Learning Model to predict Credit card leads

Business Problem:

- 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 wants to identify customers that could show higher intent towards a recommended credit card, given:

Dataset Description and Analysis:

There are 3 files given to us

- Train.csv: We will use this file for training our model. It contains variables or features that we will input to our model, and the target variable that we want to predict.

- Test.csv: This will contain all variables in the training dataset except the target variable. We will use our trained model to predict the target for this dataset.

- SampleSubmission.csv: This gives us a required format for submitting our solutions to the AV solution checker.

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

The second file is test dataset using which we need to predict the leads.

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

ML Formulation:

We need to create a model (Binary-Class Classification) that identifies the prospective customers for Credit Card Business

Performance Metric:

The Evaluation Factor is roc_auc_score.

First Cut Approach

BELOW are the STEPS I would be performing to arrive at a solution for the given Problem:

1. Data Description and Hypothesis Generation

    Here I will understand the data in high level and ask a few questions on what variables can influence the target variable.

2. EDA
    a. Target Distribution (multi-class classification)
    b. Analyzing Categorical Variables (Uni-variate and Bi-variate analysis)
    c. Analyzing Numerical Variables (Uni-variate and Bi-variate analysis)
    d. Analyzing date Variables (Uni-variate and Bi-variate analysis)

    Here I would perform a detail level analysis against each variable and see how the target variable is correlated with each of them. Will also perform multi-variate analysis if need be.

3. Feature Engg.

    Here based upon the EDA Inferences I will go back and do some feature engineering to see what else can be added to improve the performance.

4. Splitting the data and making the model ready
    a. One-Hot-Encoding Categorical Variables
    b. Normalization of Numerical Variables

Here I would do the necessary encoding against each of the variable so that modeling will be easier and would have no impact of outliers.

5. Choosing the metric for Binary-Class Classification

   Here I will be using roc_auc_score as suggested

6. Model Building

7. Logistics Regression
   Here I will be using the Logistic Regression, and predicting the Leads

8. Decision Tree and Random Forest

   Here I will be using the Decision Tree and Random forest and see the performance

9. Final Conclusion/ comparisons of the Models

   Here I will be showcasing the performance of all the models