**Credit Card Lead Prediction**

**Introduction**

* 1. **Problem Statement**

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.)
  1. **Data**

Our Task is to Build a Classification model so as to predict for the new customers whether they will they will be lead or not based on the below variables.

**Figure 1.1 ( glimse of the 1- 6 columns)**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | | Gender | Age | | Region\_Code | Occupation | | | Channel\_Code | |
| NNVBBKZB | | Female | 73 | | RG268 | Other | | | X3 | |
| IDD62UNG | | Female | 30 | | RG277 | Salaried | | | X1 | |
| HD3DSEMC | | Female | 56 | | RG268 | Self\_Employed | | | X3 | |
| BF3NC7KV | | Male | 34 | | RG270 | Salaried | | | X1 | |
| TEASRWXV | | Female | 30 | | RG282 | Salaried | | | X1 | |
| ACUTYTWS | | Male | 56 | | RG261 | Self\_Employed | | | X1 | |
| ETQCZFEJ | | Male | 62 | | RG282 | Other | | | X3 | |
| JJNJUQMQ | | Female | 48 | | RG265 | Self\_Employed | | | X3 | |
| ZMQFYKCB | | Female | 40 | | RG283 | Self\_Employed | | | X2 | |
| NVKTFBA2 | | Female | 55 | | RG268 | Self\_Employed | | | X2 | |
| NVC424KZ | | Male | 53 | | RG254 | Self\_Employed | | | X3 | |
| GZ5TMYIR | | Male | 27 | | RG270 | Self\_Employed | | | X1 | |
| FCPEEIY3 | | Female | 27 | | RG277 | Salaried | | | X1 | |
| KCE7JSFN | | Male | 31 | | RG254 | Salaried | | | X1 | |
|  | |  |  | |  |  | | |  | |
| Vintage | Credit\_Product | | | Avg\_Account\_Balance | | | Is\_Active | Is\_Lead | |
| 43 | No | | | 1045696 | | | No | 0 | |
| 32 | No | | | 581988 | | | No | 0 | |
| 26 | No | | | 1484315 | | | Yes | 0 | |
| 19 | No | | | 470454 | | | No | 0 | |
| 33 | No | | | 886787 | | | No | 0 | |
| 32 | No | | | 544163 | | | Yes | 0 | |
| 20 |  | | | 1056750 | | | Yes | 1 | |
| 13 | No | | | 444724 | | | Yes | 0 | |
| 38 | No | | | 1274284 | | | No | 0 | |
| 49 | Yes | | | 2014239 | | | No | 0 | |
| 123 | No | | | 980664 | | | Yes | 0 | |
| 14 | Yes | | | 502787 | | | No | 0 | |
| 20 | No | | | 811591 | | | Yes | 0 | |
| 31 | Yes | | | 938754 | | | No | 0 | |

**Figure 1.2 ( glimse of the 7- 11 columns)**

**Predictor variable description**

|  |  |
| --- | --- |
| **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 |

**Methodology**

**Exploratory Data Analysis**

**In EDA we will look at Univariate, Bivariate, Multivariate Analysis**

**Univariate Analysis:**

Univariate analysis is the simplest form of data analysis where the data being analyzed contains only one variable. Since it's a single variable it doesn't deal with causes or relationships. The main purpose of univariate analysis is to describe the data and find patterns that exist within it.

**For Numerical Variables:**

We will look at min, max, quantiles,mean, median, sum, distribution plots, boxplots etc…

**For Categorical Variables:**

We will look value\_counts(), to check the variation in the data we can use pie plots, countplot, jointplot etc..

**Bivariate Analysis:**

Bivariate analysis is one of the simplest forms of quantitative (statistical) analysis. It involves the analysis of two variables (often denoted as X, Y), for the purpose of determining the empirical relationship between them. Bivariate analysis can be helpful in testing simple hypotheses of association.

In Bivariate analysis will look at the various variables and their relations we will use various plots such as , pair plot , density plots etc.. to understand relationship between variables .

We also look at the relation of variable with respect target variable as well in order to and try to understand their relation and draw various conclusions based on them

**Multivariate Analysis:**

There are multiple ways to look at multivariate analysis for this problem we will stick with correlation plot to understand effect of each predictor variable with other variables

**Missing Value Analysis**

Missing Value analysis is very important part of EDA plays a significant role in the data analysis

We will remove all the missing values from the dataset to a new dataset and we will treat it as test data to predict values of the missing target column.

**Outlier Analysis**

Outlier analysis is an important step in data preprocessing outliers can offset the algorithm There are a lot of ways to deal with outliers like use mean & 3 standard deviation method, or you can make outliers as null values and later impute them with mean / median

In this case we will use Quintile method

**Some of main findings of Exploratory Data Analysis are**

* **Based on data the target column ‘Is\_Lead’ is unbalanced to the ratio of (76:24)**
* **Credit\_Product has around missing values of approx 12% in both train and test**
* **Avg\_Account\_Balance,Vintage,Age variables are right skewed**
* **Mean avg account balance is is 12 % higher in customer converted as lead compaired to non-lead**
* **Occupation as entrepreneur is more likely to convert as lead than any category and occupation and salaried are less likely to convert as lead**
* **Male are more likely to convert as lead than female**
* **Cutomer who are associated with the bank for a longer period are likely to convert as lead**
* **Avg\_Account\_Balance has a lot of outliers**
* **Customers how are aged lessthan 35 has high mean avg balance but less likely to convert as lead based on data**
* **Vintage has a lot of outliers**
* **Credit**\_**Product as yes are less likely to convert as Lead**
* **Most of the missing values in Credit\_product in train are Leads**

**Some of the Approach that have worked while building the model are**

1. **Imputing the missing in Credit\_Product as missing rather than mode have drastically increased score.**
2. **Reducing the extreme outliers in Avg\_Account\_Balance resulted in increased roc\_auc\_score**
3. **Log transforming the Avg\_Account\_Balance variable**
4. **Log transforming the Age variable**
5. **Frequency encoding the Region\_Code and Channel\_Code has increased score**
6. **Bining the less frequent values in region code has increased the score**
7. **Transforming the vintage variable into no.of days**
8. **Out of fold predictions have worked well in increasing the score**
9. **GBDT worked well rather than the Logistic,Knn,Naivebayes algorithms**