**Problem Statement:**

* Bank Customer Turnover

Description:

The data contains customer information of a leading bank from European region. Data is generated for a certain duration of time regarding the customer behavior. These are Some Details: CustomerId,Surname,CreditScore,Geography(Which Country of Customer ,Gender,Age,Tenure,Balance,Num\_of\_Products,HasCrCard,IsActive Member,Estimated\_Salary,Exited .To Identify This process they have given a Problem to Identify the Turnover of the Customer

It is a Classification Problem Where we have to predict Whether a Customer will left or Not. In a Classification Problem, We have to Predict Discrete values Based on a given Set of Independent Variable.

-----------------------------------------------------------------------------------------------

After Looking at the Problem Statement we will Now Move into Hypothesis Generation. It is the Process of Listing out all the Possible Factors that can Affect the Outcome.

**Hypothesis Generation :**

Below are Some of the factors which I think can affect Customer Turnover:

Credit Score : Customer have Low Credit Score There Has Chances to Customer Turnover

IsActiveMember : If it is Not the Active Member There has also Chances Num Of Products: Customer Having a More than One Account

HasCrCard : Customer having a card or not Getting its Benefit.

These are Some Factors Which I think can Affect the Target Variable.

-----------------------------------------------------------------------------------------------**Understanding the Data :**

Columns of Data

Bank\_data.columns

**Index(['RowNumber', 'CustomerId', 'Surname', 'CreditScore', 'Geography',**

**'Gender', 'Age', 'Tenure', 'Balance', 'NumOfProducts', 'HasCrCard',**

**'IsActiveMember', 'EstimatedSalary', 'Exited'],**

**dtype='object')**

**Give Below is the Description for Each Variable:**

|  |  |
| --- | --- |
| **Row Number** | **Serial Number** |
| CustomerId | Bank Customer Id |
| Surname | Surname of Customer |
| Credit Score | Integer Value |
| Geography | Country of Customer |
| Gender | Gender of Customer |
| Age | Age in Year |
| Tenure | A factor with 10 levels |
| Balance | Account Balance |
| Num Of Products | How Many Accounts |
| HasCrCard | Does the Customer Have a Credit Card through the Bank |
| IsActiveMember | Subjective |
| Estimated Salary | Estimated Field of Customer Salary |
| Exited | Did They Leave the Bank after all |

**# Print data types for each Variable Train. Types :**

RowNumber int64

CustomerId int64

Surname object-----

CreditScore int64 ----ordinal

Geography object-----

Gender object----

Age int64===

Tenure int64===

Balance float64==

NumOfProducts int64==

HasCrCard int64==

IsActiveMember int64==

EstimatedSalary float64==

Exited int64

dtype: object

We can see there are three Format of Data Types :

Object : Object Format means variable are Categorical .Categorical variables in our Dataset are: Surname,Geography ,Gender.

Int64: It represents the integer variables.Int variables in our Dataset are :RowNumber,CustomerId,CreditScore,Age,Tenure,NumofProducts,HasCrCard,IsActiveMember,Exited

Float64: It represents the variable which have some Decimal Values involved.They are also Numerical Variables.Numerical Variables in our Data sets are : Balance,EstimatedSalary

Let’s Look at the Shape of the dataset:-

Bank\_data.shape

(10000, 14)

We have 10000 rows and 14 columns in a Data sets.

-----------------------------------------------------------------------------------------------

Univariate Analysis :

We will First Look at the Target variable .As it is a Binary Categorical Variable .Let us Look at its Frequency table,Percentage Distribution and Bar Plot

Frequency Table of a variable will gives us the count of each Categorical in that variable .