**DATA GLACIER VIRTUAL INTERNSHIP**

**CROSS SELLING RECOMMENDATION-GROUP PROJECT**

**WEEK 8: DELIVERABLES**

**GROUP NAME: HEGY**

**Team members:**

Name: B. Harika

Email: harikabreddy444@gmail.com

Country: India

College/ Company: Data Glacier

Specialization: Data Analyst

Name: Yusuf Yuhan

Email: yusufyuhan98.yy@gmail.com

Country: Srilanka

College/ Company: The Open University of Srilanka

Specialization: Data Analyst

Name: Ebaghae Imhanlahimi

Email: imhanlahimiw@gmail.com

Country: America

College/ Company: Data Glacier

Specialization: Data Analyst

Name: Gladys Kalas

Email: gladys@kalas.me

Country: USA

College/ Company: Data Glacier

Specialization: Data Analyst

**Problem Description**

The problem for XYZ Credit Union is that despite having a successful track record of selling banking products, they are not able to maximize their revenue potential from their existing customer base. The customers are not purchasing more than one product, indicating that the bank is not effectively cross-selling their other offerings to their existing customers. This is a significant challenge for the bank as they are missing out on potential revenue opportunities, which could impact their overall growth and profitability.

To address this problem, XYZ Credit Union has approached ABC Analytics to provide insights and recommendations on how to increase cross-selling of banking products. The objective of this engagement is to help the bank identify the factors that are hindering their cross-selling efforts and provide actionable recommendations that can help them increase their product penetration rates among their existing customer base.

**Data Understanding**

The dataset appears to contain information about bank customers and their product holdings. There are several columns related to customer demographics, including age, sex, and country of residence. The dataset also contains both numerical and categorical variables.

The 'ncodpers' column appears to be a unique identifier for each customer, while the 'fecha\_dato' column represents the date when the data was collected, and the 'fecha\_alta' column represents the date when the customer joined the bank. The 'ind\_nuevo' column indicates whether the customer is new to the bank or not, and the 'antiguedad' column represents the number of months since the customer joined the bank.

There are multiple columns related to the customer's product holdings, such as 'ind\_ahor\_fin\_ult1' for savings accounts and 'ind\_cco\_fin\_ult1' for current accounts. These columns appear to be binary, indicating whether the customer holds a particular product or not.

The dataset contains missing values, which need to be handled appropriately before further analysis. Exploring the relationship between customer demographics and their product holdings could be an interesting avenue for further analysis. It may be possible to identify patterns or trends that could inform marketing strategies or product development.

**Data Analys**

Column\_Names Data Types

fecha\_dato Data Object

ncodpers/ Customer\_code Int64

ind\_empleado/ Employee\_index Object

pais\_residencia/ Country Object

Sexo/ Gender Object

age/ Age Object

fecha\_alta/ Customer\_join\_date Object

ind\_nuevo/ Customer\_index Float64

antiguedad/ Customer\_senoirity Object

indrel/ primary\_customer Float64

ult\_fec\_cli\_1t/ Customer\_leave\_date Object

indrel\_1mes/ Customer\_type Object

tiprel\_1mes/ Customer\_relation Object

indresi/ Residence\_index Object

indext/ Foreign\_index Object

conyuemp/ Spouse\_index Object

canal\_entrada/ Channel Object

indfall/ Deceased\_index Object

tipodom/ Primary\_address Float64

cod\_prov/ Customer\_address Float64

nomprov/ province\_name Object

ind\_actividad\_cliente/ Activity\_index Float64

renta/ Gross\_income Float64

segmento/ Segmentation Object

ind\_ahor\_fin\_ult1/ Saving\_account Int64

ind\_aval\_fin\_ult1/ Guarantees Int64

ind\_cco\_fin\_ult1/ Current\_account Int64

ind\_cder\_fin\_ult1/ Derivative\_account Int64

ind\_cno\_fin\_ult1/ Payroll\_account Int64

ind\_ctju\_fin\_ult1/ Junior\_account Int64

ind\_ctma\_fin\_ult1/ More\_private\_account Int64

ind\_ctop\_fin\_ult1/ Private\_accoun t Int64

ind\_ctpp\_fin\_ult1/ Private\_plus\_account Int64

ind\_deco\_fin\_ult1/ Short\_term\_deposits Int64

ind\_deme\_fin\_ult1/ Medium\_term\_deposits Int64

ind\_dela\_fin\_ult1/ Long\_term\_deposits Int64

ind\_ecue\_fin\_ult1/ E\_account Int64

ind\_fond\_fin\_ult1/ Funds Int64

ind\_hip\_fin\_ult1/ Mortgage Int64

ind\_plan\_fin\_ult1/ Pensions Int64

ind\_pres\_fin\_ult1/ Loans Int64

ind\_reca\_fin\_ult1/ Taxes Int64

ind\_tjcr\_fin\_ult1/ Credit\_card Int64

ind\_dela\_fin\_ult1/ Securities Int64

ind\_dela\_fin\_ult1/ Home\_account Int64

ind\_dela\_fin\_ult1/ Payroll Float64

ind\_dela\_fin\_ult1/ Pensions\_2 Int64

ind\_dela\_fin\_ult1/ Direct\_debit Int64

**Problems and solutions for Data**

* The dataset contains missing and null values that need to be addressed. These missing values can affect the accuracy and reliability of the analysis, and therefore, they need to be handled appropriately.
* The column names in the dataset are not user-friendly and may not be easily understood by analysts. It is essential to rename and reformat these column names to make them more readable and understandable.
* There are a few missing values in the dataset, with approximately 27734 null values detected. The presence of these missing values can impact the quality of the analysis and make it difficult to draw accurate conclusions.
* The dataset's columns do not have appropriate data types, which can lead to errors in analysis and modeling. It is crucial to ensure that each column has the correct data type to facilitate accurate analysis and modeling

**Data Cleaning**

* To address the issue of missing values in the dataset, we have dropped several records containing missing values. This approach helps to ensure that the remaining data is reliable and can be analyzed accurately.
* We have translated and renamed the column names in the dataset to make them more readable and user-friendly. This approach makes it easier for analysts to understand the data and extract valuable insights from it.
* To handle missing values in the dataset, we have used various imputation techniques such as mean, mode, median, and zeroes. This approach helps to ensure that the missing values are filled in a way that does not significantly impact the overall distribution of the data.
* We have assigned variables such as gender, residence index, spouse index, customer relations, employee index, etc., to their respective categories. This approach helps to ensure that the data is organized and structured in a way that is meaningful and useful for analysis.
* Outliers in the dataset have been detected using various methods and treated accordingly. This approach helps to ensure that the data is clean and reliable, which is critical for accurate analysis and modeling.