Data Glacier Group Project

YUHONG CHEN
HAOYUE CHANG
GIFTY OSEI

Group Name: DATA ALCHEMISTS

Team Members:

Name: Yuhong Chen
Email: ryechenn@gmail.com
Country: Canada
College/Company: McGill University

Name: Haoyue Chang
Email: kerrychy1215@gmail.com
Country: USA
College/Company: Northeastern University

Name: Gifty Osei
Email: gifty18osei@gmail.com
Country: USA
College/Company: Montana State University

Specialization: Data Analyst

CONTENT

- Problem Description
- Data Information
- Exploratory Data Analysis
- Modeling & Evaluation

Project Introduction

- We are determined to help XYZ Bank improve its cross-selling strategies and enhance customer engagement. The bank offers a wide array of financial products and services, including savings accounts, credit cards, mortgages, loans, and investment options. However, we've observed that many of our customers have limited product adoption and aren't fully utilizing the range of services available to them.
-

To tackle this challenge head-on, we plan to implement customer segmentation techniques to gain deeper insights into our customer base. By dividing the customers into distinct groups based on their demographics, financial behavior, and product usage patterns, we hope to identify specific customer segments that are more likely to use products and services. Armed with this valuable information, we aim to create personalized marketing strategies and tailored cross-selling initiatives to boost customer satisfaction and encourage higher product adoption.
As part of our data analysis team, the objective is to thoroughly analyze the extensive customer dataset provided by XYZ Bank and conduct a comprehensive customer segmentation analysis. The dataset includes detailed information about each customer, such as age, gender, income, transaction history, product holdings, and tenure with our bank.

Alidropship.com

Data Overview

fecha_dato The table is partitioned for this column
ncodpers Customer code
ind_empleado Employee index: A active, B ex employed, F filial, N not employee, P pasive
pais_residencia Customer's Country residence
sexo Customer's sex
age Age
fecha_alta The date in which the customer became as the first holder of a contract in the bank
ind_nuevo New customer Index. 1 if the customer registered in the last 6 months.
antiguedad Customer seniority (in months)
indrel 1 (First/Primary), 99 (Primary customer during the month but not at the end of the month)
ult_fec_cli_1t Last date as primary customer (if he isn't at the end of the month)
indrel_1mes Customer type at the beginning of the month ,1 (First/Primary customer), 2 (co-owner),P (Potential),3 (former primary), 4(former co-owner)
tiprel_1mes Customer relation type at the beginning of the month, A (active), I (inactive), P (former customer),R (Potential)
indresi Residence index (S (Yes) or N (No) if the residence country is the same than the bank country)
indext Foreigner index (S (Yes) or N (No) if the customer's birth country is different than the bank country)
conyuemp Spouse index. 1 if the customer is spouse of an employee
canal_entrada channel used by the customer to join
indfall Deceased index. N/S
tipodom Addres type. 1, primary address
cod_prov Province code (customer's address)
nomprov Province name
ind_actividad_cliente Activity index (1, active customer; 0, inactive customer)

renta Gross income of the household
segmento segmentation: 01 - VIP, 02 - Individuals 03 - college graduated
ind_ahor_fin_ult1 Saving Account
ind_aval_fin_ult1 Guarantees
ind_cco_fin_ult1 Current Accounts
ind_cder_fin_ult1 Derivada Account
ind_cno_fin_ult1 Payroll Account
ind_ctju_fin_ult1 Junior Account
ind_ctma_fin_ult1 Más particular Account
ind_ctop_fin_ult1 particular Account
ind_ctpp_fin_ult1 particular Plus Account
ind_deco_fin_ult1 Short-term deposits
ind_deme_fin_ult1 Medium-term deposits
ind_dela_fin_ult1 Long-term deposits
ind_ecue_fin_ult1 e-account
ind_fond_fin_ult1 Funds
ind_hip_fin_ult1 Mortgage
ind_plan_fin_ult1 Pensions
ind_pres_fin_ult1 Loans
ind_reca_fin_ult1 Taxes
ind_tjcr_fin_ult1 Credit Card
ind_valo_fin_ult1 Securities
ind_viv_fin_ult1 Home Account
ind_nomina_ult1 Payroll
ind_nom_pens_ult1 Pensions
ind_recibo_ult1 Direct Debit

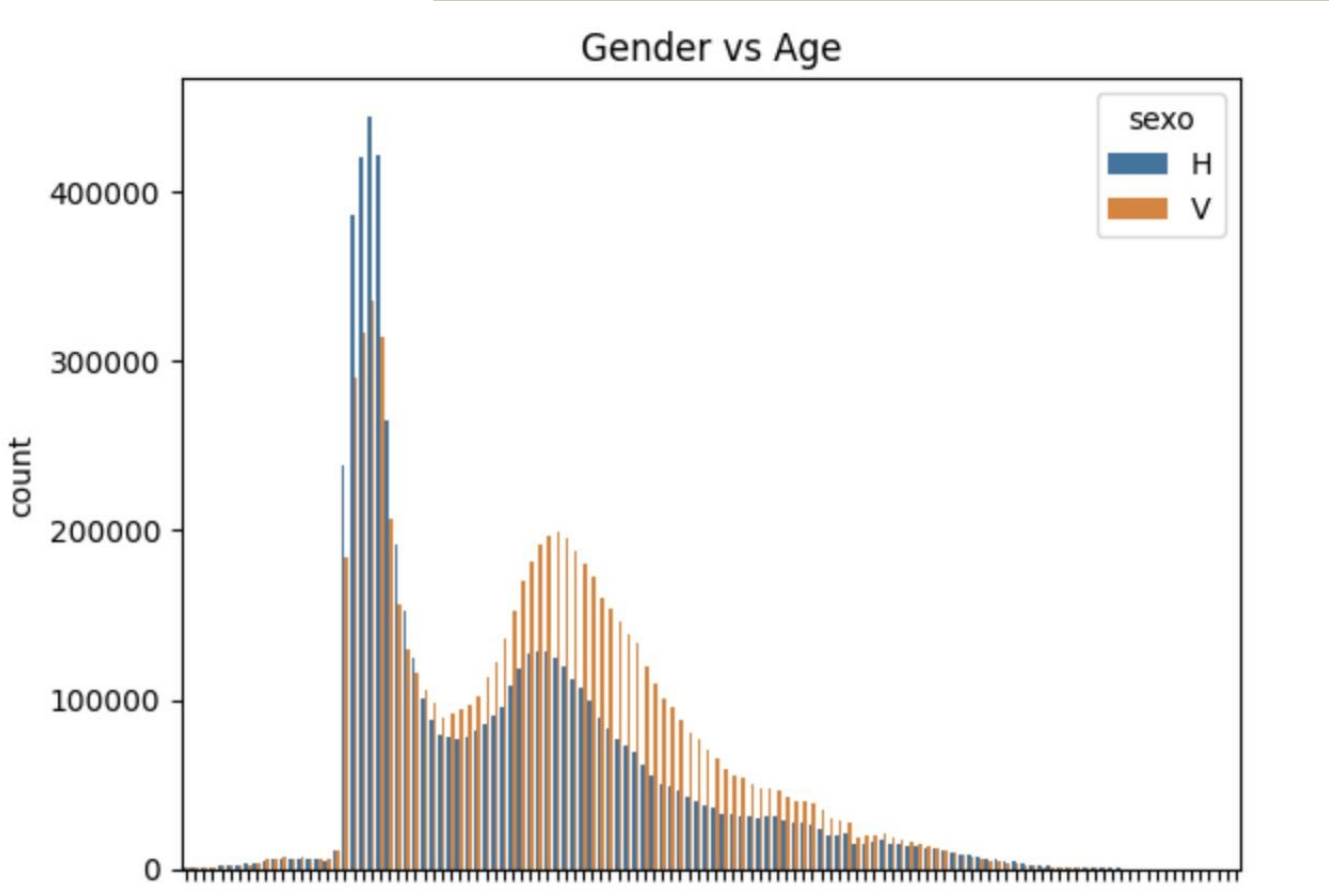
Descriptive Statistics

- Table 1 shows the summary statistics of some selected variables in our dataset.
- The selection was done at random.

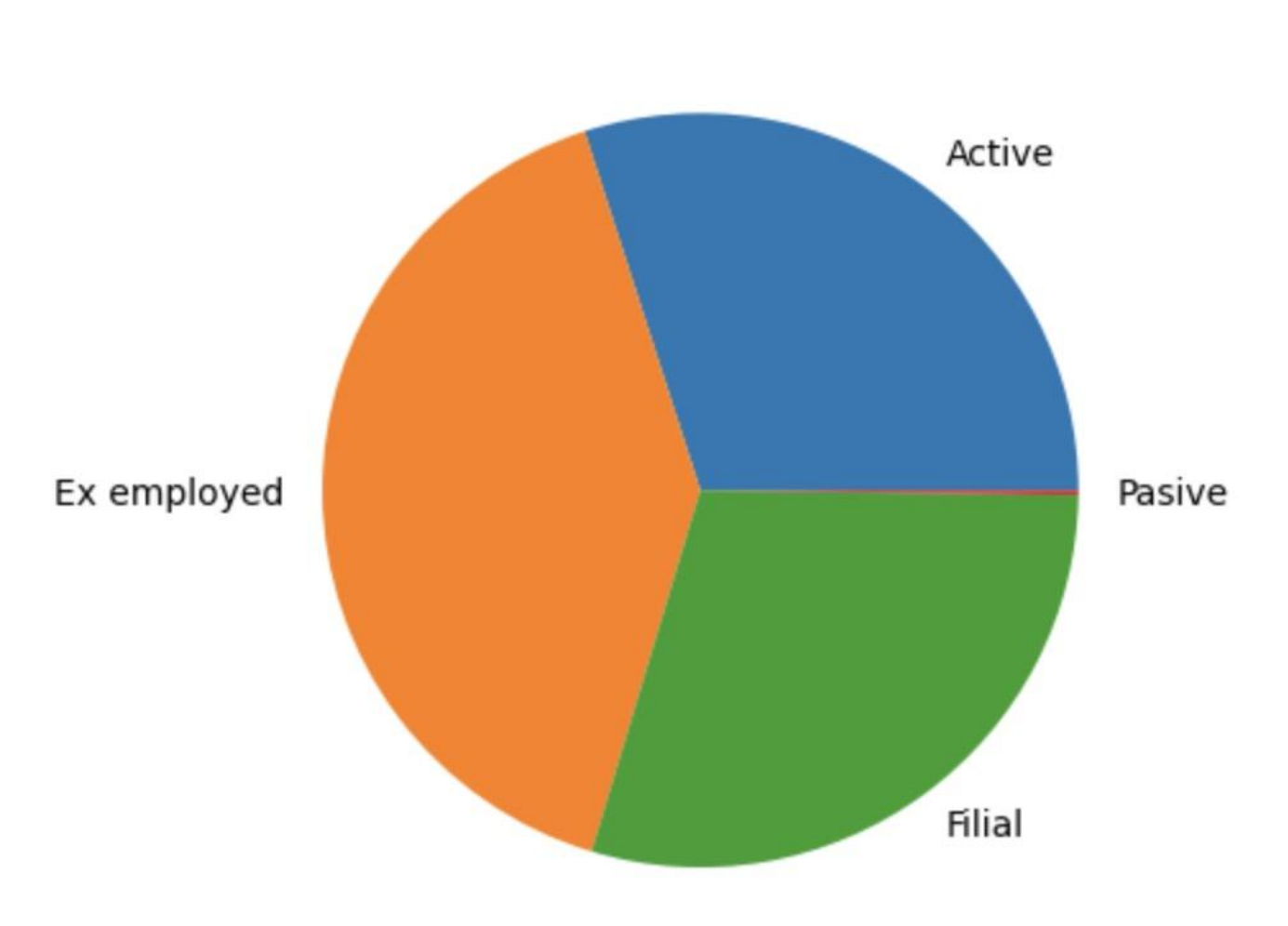
	ncodpers	ind_nuevo	indrel	tipodom	cod_prov	ind_actividad_cliente	renta
count	13647309.0	13647309.0	13647309.0	13647309.0	13647309.0	13647309.0	13647309.0
mean	834904.0	0.0	1.0	1.0	27.0	0.0	134254.0
std	431565.0	0.0	4.0	0.0	13.0	0.0	205659.0
min	15889.0	0.0	1.0	1.0	1.0	0.0	1203.0
25%	452813.0	0.0	1.0	1.0	15.0	0.0	76437.0
50%	931893.0	0.0	1.0	1.0	28.0	0.0	124680.0
75%	1199286.0	0.0	1.0	1.0	34.0	1.0	137452.0
max	1553689.0	1.0	99.0	1.0	52.0	1.0	28894396.0

EDA

- This chart plots the relationship between Gender and Age of customers.
- We could see the count of customers reaches highest point in younger age range(20) in 2 genders. Higher in H sexo type.
- And for another account which is second highest in middle age range(50) but is higher in another sexo type V.



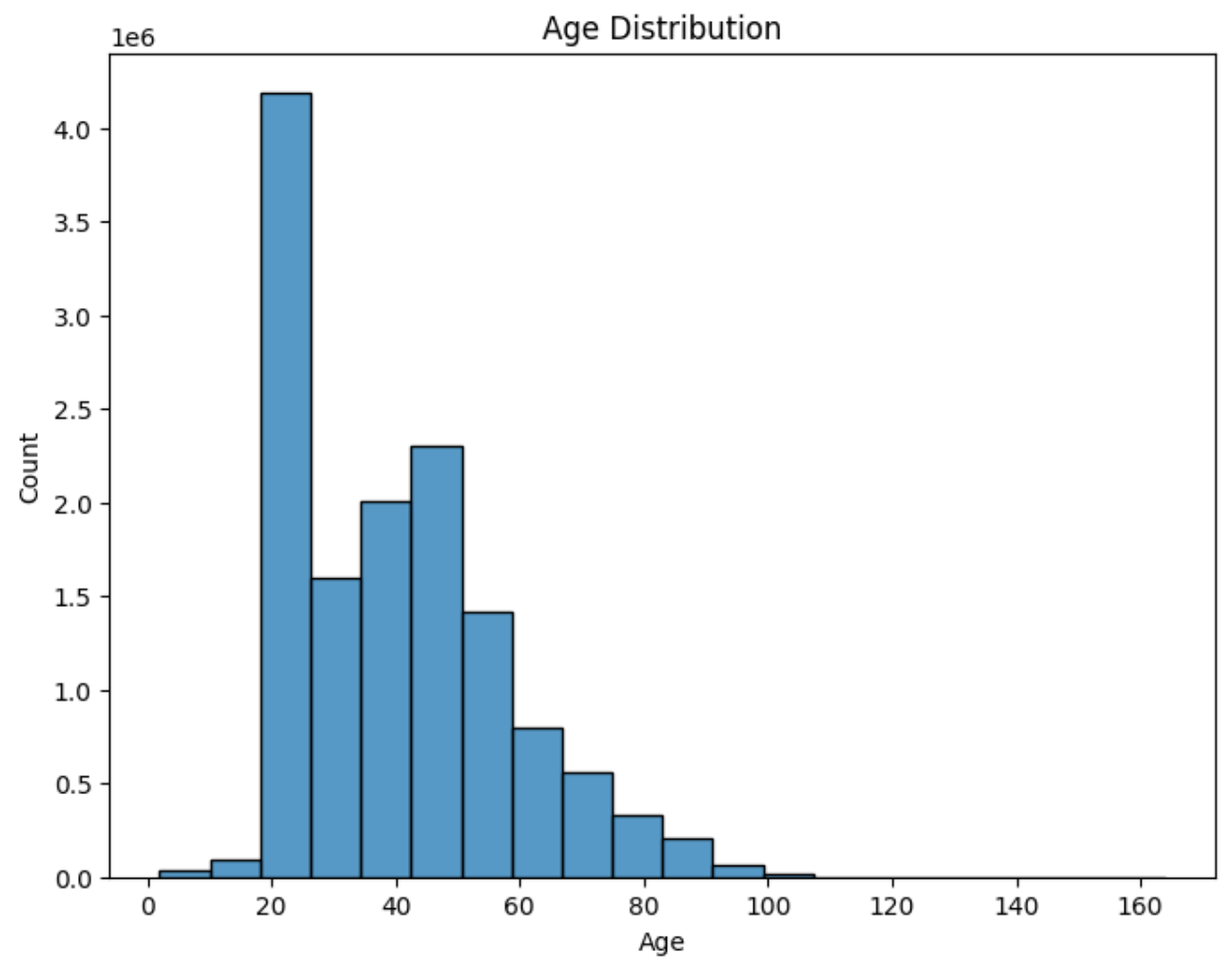
EDA



- The pie chart shows the percentage of employment index distribution.
- Unemployed ranks the 1st and then the Ex employed. Pasive is least here in the plot.
- So, we can see that the employment situation is not positive for the unemployment percentage and Ex employed count.

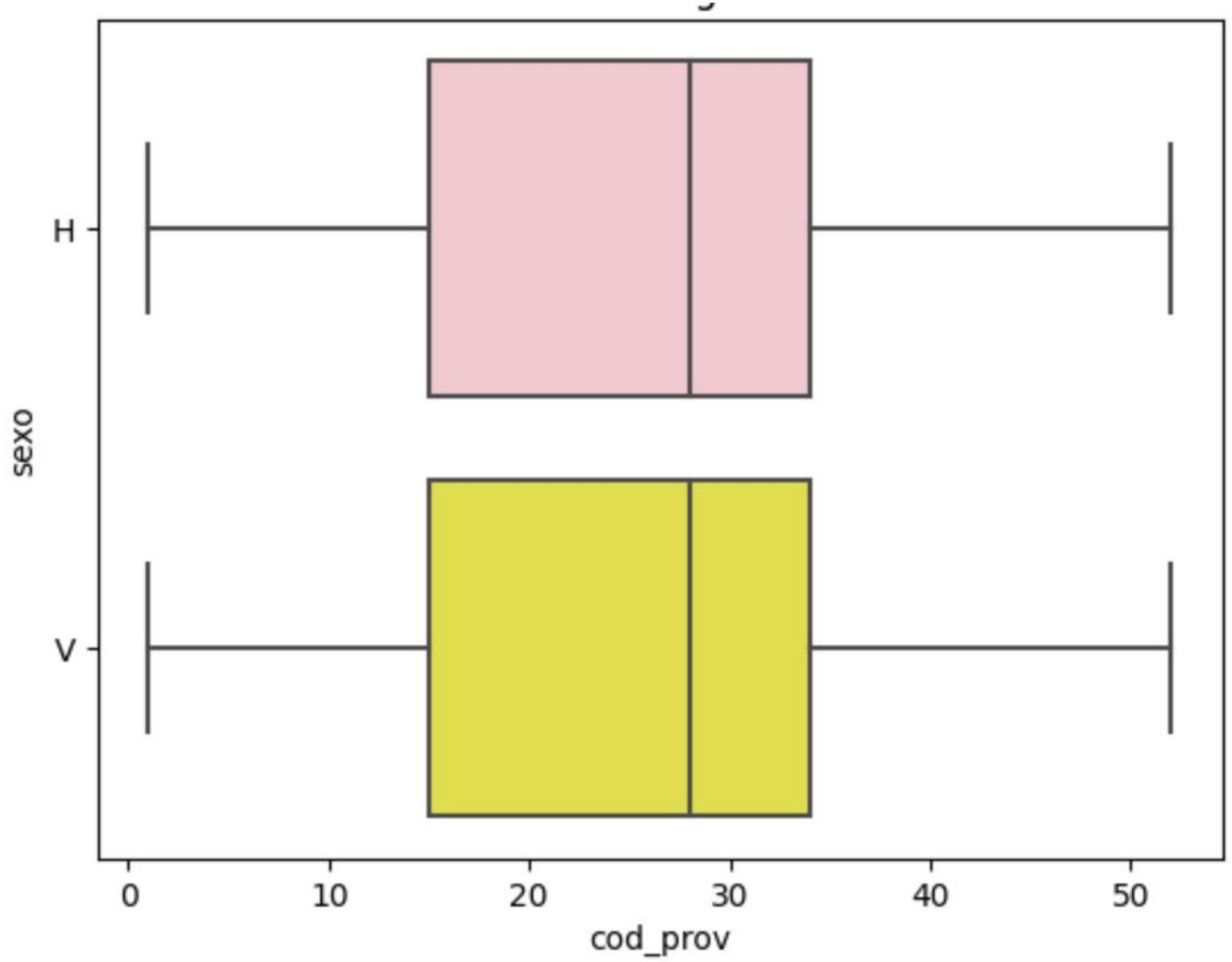
EDA

- We can see from the Age distribution plot that the majority of customers are between ages 20 and 50 meaning the products are patronized by the working force.



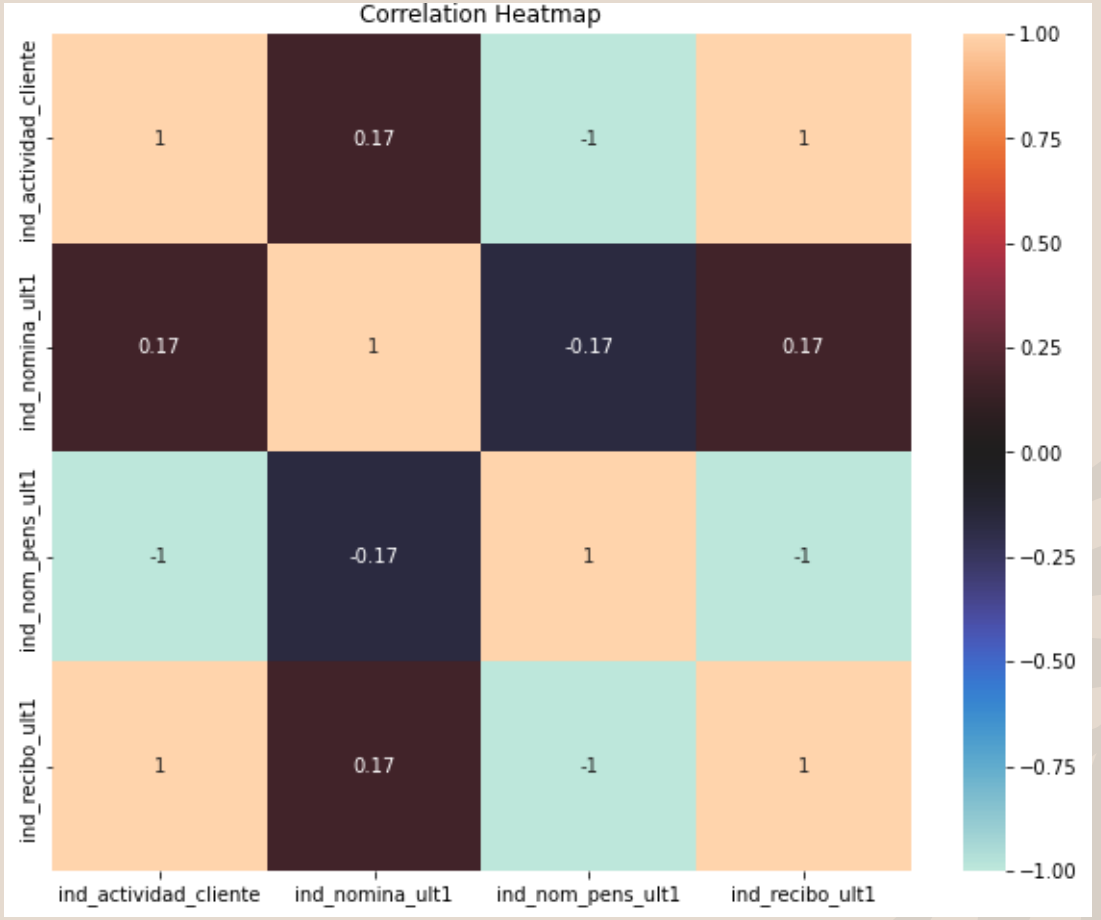
EDA

- We can see the Province code and gender distribution from this plot.
- For different gender, Province code (customer's address) are spreading nearly the same.



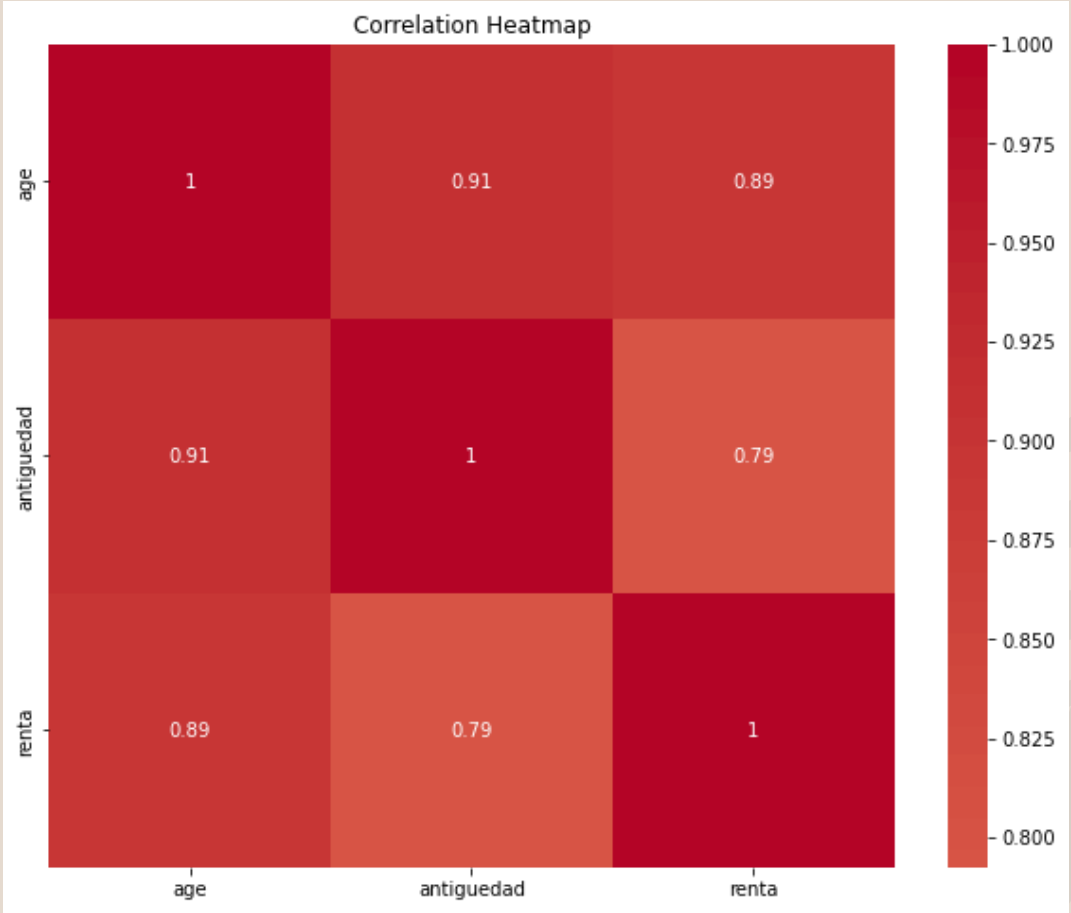
EDA

- The x-axis and y-axis display the variable of ind_actividad_cliente, ind_nomina_ult1, ind_nom_pens_ult1, and ind_recibo_ult1.
- "ind_actividad_cliente" and "ind_nomina_ult1":The correlation coefficient between "ind_actividad_cliente" and "ind_nomina_ult1" is close to zero. This suggests that there is little to no linear relationship between a customer's activity index and whether they received a payroll payment.
- "ind_actividad_cliente" and "ind_nom_pens_ult1":The correlation between "ind_actividad_cliente" and "ind_nom_pens_ult1" is -1, indicating a negative linear association between a customer's activity index and whether they received a pension payment.
- "ind_actividad_cliente" and "ind_recibo_ult1": The correlation between "ind_actividad_cliente" and "ind_recibo_ult1" (direct debit) indicates that there is a strong linear relationship between a customer's activity index and their participation in direct debit transactions.
- "ind_nomina_ult1" and "ind_nom_pens_ult1": Depending on the sample data, you might observe a correlation coefficient close to zero. This could indicate a possible connection between receiving payroll and receiving pension payments, though further analysis would be needed to establish causation.



EDA

- The x-axis and y-axis display the variable names age , antiguedad, and renta. A positive correlation between "age" and "antiguedad" because older customers tend to have longer relationships with the bank. There has a strong correlation between "age" and "renta". And there is a strong correlation between "antiguedad" and "renta" because the length of time a customer has been with the bank may necessarily be strongly related to their income.



Modeling- Linear on target variable 'ind_ahor_fin_ult1'

Linear model on target variable 'ind_ahor_fin_ult1'

Accuracy: The achieved accuracy of 1.00 (100%) might seem impressive at first glance. However, it's crucial to recognize that this high accuracy is largely due to the substantial class imbalance within the dataset, where the majority class is dominant. Relying solely on accuracy can be misleading in scenarios like this, and it should not be the sole determinant of model performance assessment.

Precision and Recall: The classification report underscores a significant contrast in the model's performance for the two classes:

The precision for class 0 is elevated (1.00), indicating that when the model predicts class 0, it's usually accurate. However, the precision for class 1 is exceedingly low (0.00), signifying the model's difficulty in making precise predictions for class 1.

The recall for class 0 is high (1.00), suggesting that the model captures most instances of class 0.

F1-Score: The F1-score, a harmonized measure of precision and recall, is substantial for class 0 (1.00), while being strikingly low for class 1 (0.00). This confirms that the model's performance is heavily skewed towards the majority class.

```
Accuracy: 1.00
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00   4085454
           1       0.00      0.00      0.00       419

    accuracy                           1.00   4085873
   macro avg       0.50      0.50      0.50   4085873
weighted avg       1.00      1.00      1.00   4085873
```

Logistic Regression

```
Logistic Regression:
Accuracy:  0.9998974515360609
Confusion Matrix:  [[4085454       0]
 [    419       0]]
```

```
Classification Report:                precision    recall  f1-score   support

           0       1.00      1.00      1.00   4085454
           1       0.00      0.00      0.00       419

    accuracy                           1.00   4085873
   macro avg       0.50      0.50      0.50   4085873
weighted avg       1.00      1.00      1.00   4085873


Random Forest:
Accuracy:  0.9998974515360609
Confusion Matrix:  [[4085454       0]
 [    419       0]]
```


```
Classification Report:                precision    recall  f1-score   support

           0       1.00      1.00      1.00   4085454
           1       0.00      0.00      0.00       419

    accuracy                           1.00   4085873
   macro avg       0.50      0.50      0.50   4085873
weighted avg       1.00      1.00      1.00   4085873
```

Linear Model for target variable "ind_cco_fin_ult1" by Yuhong

#Multiple regression

```
m2<-lm(formula=ind_cco_fin_ult1 ~ sexo+age+antiguedad,data=train)
summary(m2)
```

```
##
## Call:
## lm(formula = ind_cco_fin_ult1 ~ sexo + age + antiguedad, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8639 -0.5944  0.2640  0.3496  0.7131
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.541e-01  3.384e-04 2524.19  < 2e-16 ***
## sexoV       -1.482e-02  2.565e-04  -57.79  < 2e-16 ***
## age         -4.721e-03  7.434e-06 -635.01  < 2e-16 ***
## antiguedad  -2.654e-07  7.583e-08   -3.50 0.000466 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4678 on 13619501 degrees of freedom
##   (27804 observations deleted due to missingness)
## Multiple R-squared:  0.03008,    Adjusted R-squared:  0.03008
## F-statistic: 1.408e+05 on 3 and 13619501 DF,  p-value: < 2.2e-16
```

For model m2, with target variable current account, the terms of m2 model has viewed the AIC and the terms are the relatively independent terms that influence selling of current account significantly.

So, for cross selling strategy, company could use the terms of customers to improve the cross selling. Also find the same significant terms between different accounts or products, which could realize the cross-selling improvement. One product, Two product, or even 3 or 4, they have the same terms in models influencing significantly of the selling of them, then the business advisors could take more attention on these terms, like age, province or else to improve cross selling target.

Final Model Selection

For our final model selection, we finally chose a third model, named m2.
It is the model created for the current account and associated with the customer's gender, age and customer seniority.
According to other models (m1 to m4) comparison, this model has the most suitable AIC data, and more optimistic F-statistic data, and its standard deviation is also optimistic compared to other models, which belongs to the neutral category.
For model 1 and model 2 (logistic model), they are created for savings accounts, and their fitting results are near to 1, so we choose model 3.
So, in this case, the final selection and recommendation of customer business will be predicted based on this model.

According to the m2 model, we recommend business managers and business consultants to collect and analyze the characteristics of customers, obtain the most suitable model among different products (just like m2), and select repeated independent variables. Moreover, the overlapping independent variables of the company's existing products are used to classify customers. For different products, finding the same customer attributes in the process of modelling and confirm the plan for cross-selling products could increase the profit of cross-product sales.

THANK YOU !

Data Alchemist