GROUP NAME: DATA ALCHEMIST

TEAM MEMBERS:

Name: Yuhong (Rye) 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

Problem description

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.

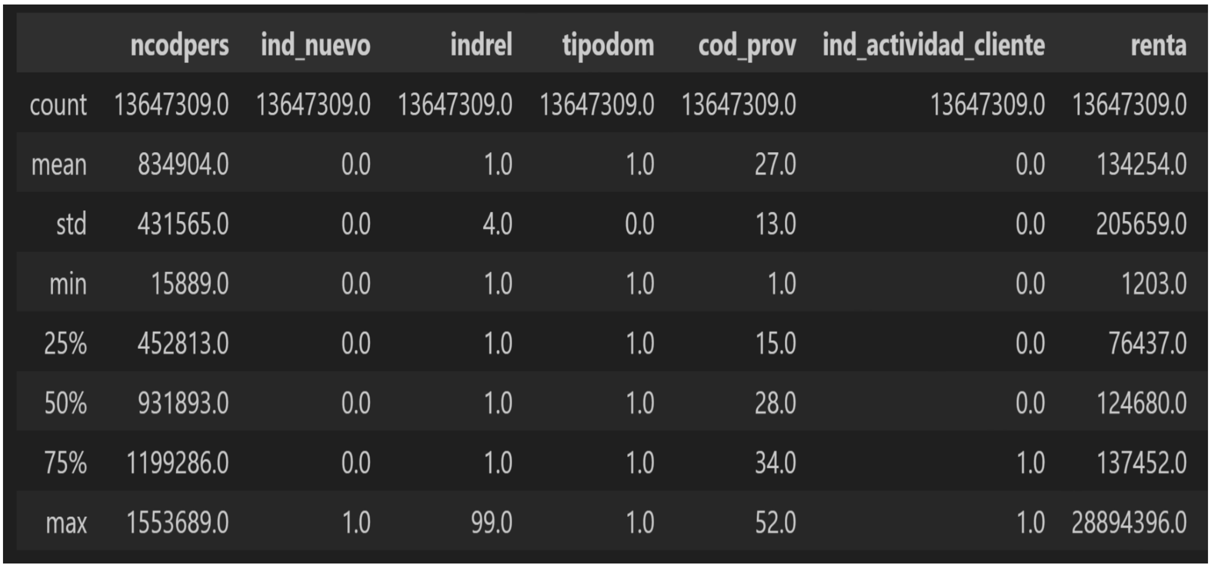
To tackle this challenge head-on, we plan to implement customer segmentation techniques to gain deeper insights into our customer base. By dividing our customers into distinct groups based on their demographics, financial behavior, and product usage patterns, we hope to identify specific customer segments 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.

Exploratory Data Analysis (EDA)

Descriptive Statistics

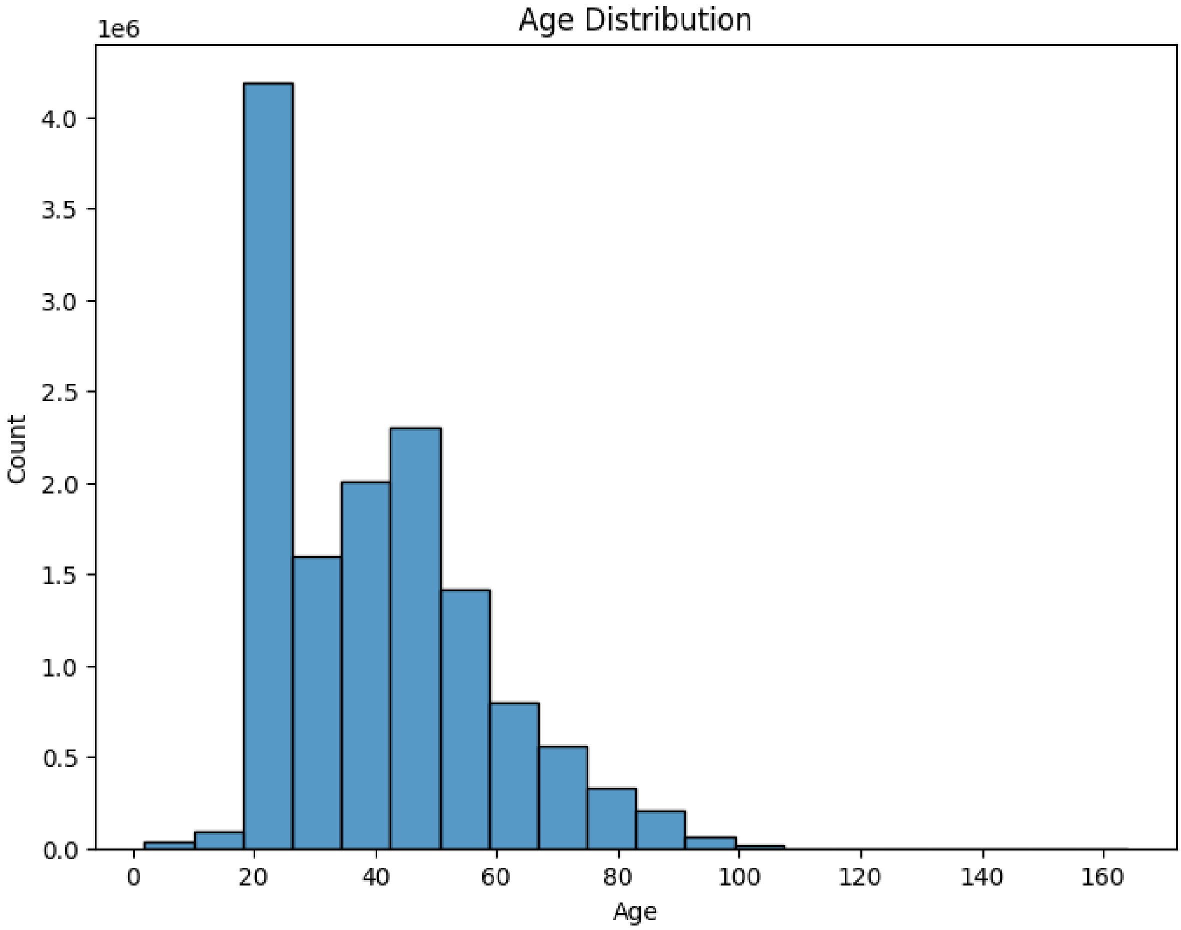
Before analyzing this data, we are interested in knowing how these variables collected behave by running a simple summary function to check the spread and dispersion rate of these variables. Below is a glimpse of the summary on some selected variables.



	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

Cod_prov has an average of 27.0 with a standard deviation of 13 and might be a variable to consider in this cross-selling recommendation system.

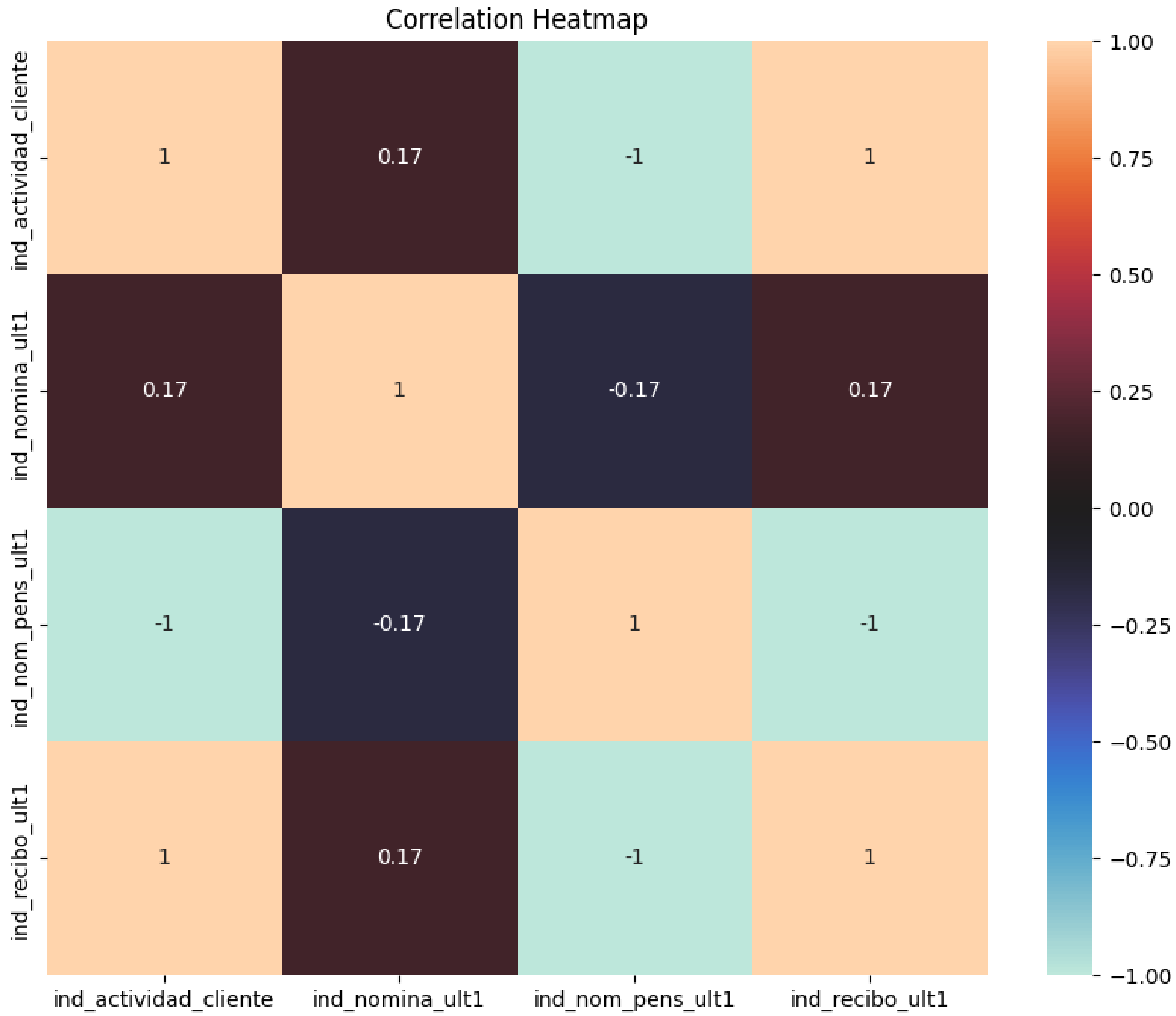
.ncodpers and renta variables might be significant in our model building



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

In the future, companies will want to make changes to their advertising to be able to retain most of their young customers because they tend to want to explore other options.

Correlation Analysis



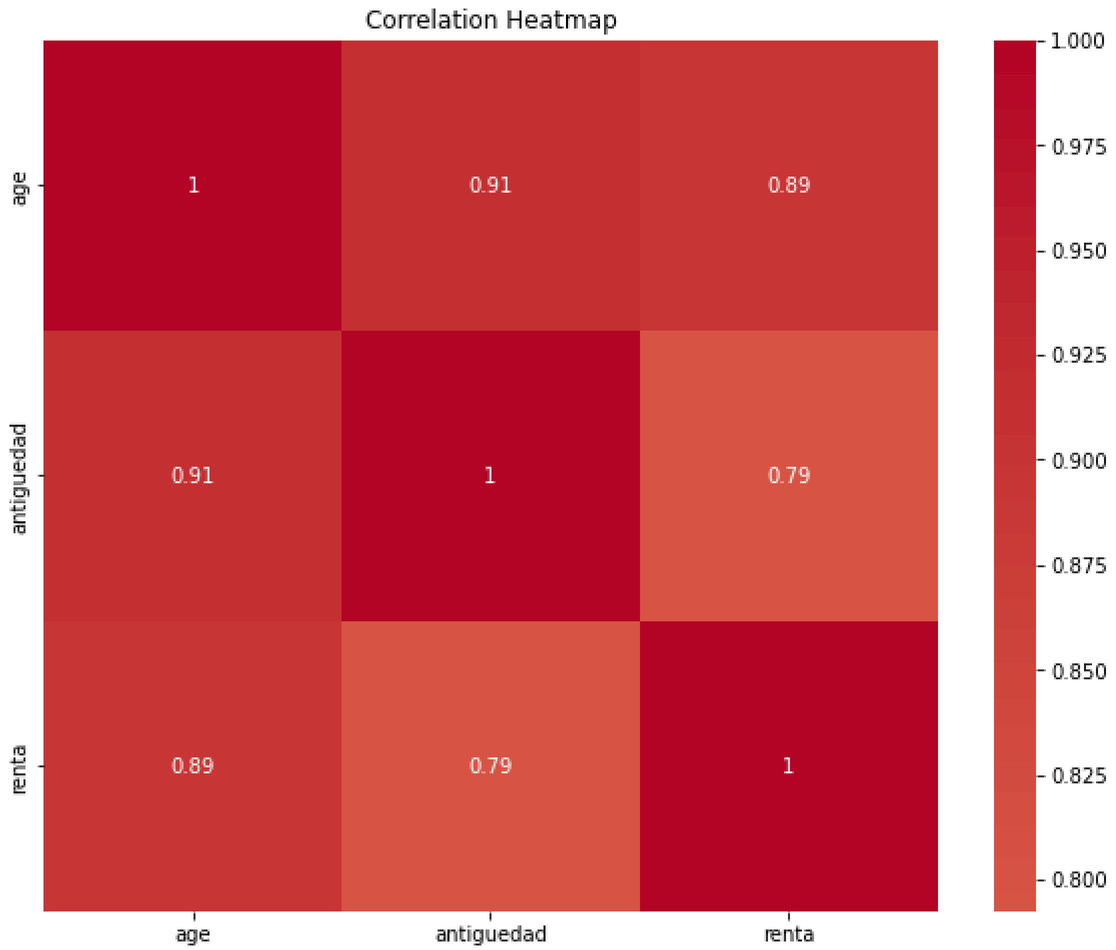
The x-axis and y-axis display the variable names ind_actividad_cliente, ind_nomina_ult1, ind_nom_pens_ult1, and ind_recibo_ult1.

"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.

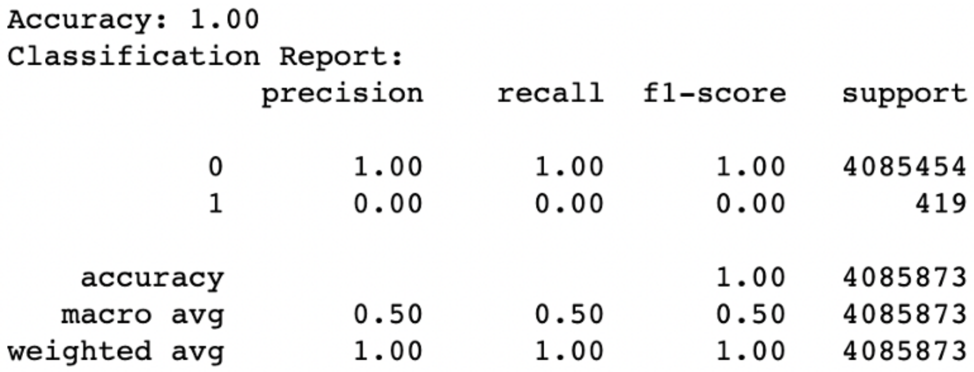
"ind_nomina_ult1" and "ind_nom_pens_ult1": Depending on the sample data, you might observe a correlation coefficient close to zero. This could indicate a possible connection between receiving payroll and receiving pension payments, though further analysis would be needed to establish causation.



The x-axis and y-axis display the variable names age , antiguedad, and renta. There is a positive correlation between "age" and "antiguedad" because older customers tend to have longer relationships with the bank. There is 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

1. **Modeling- Linear on target variable 'ind_ahor_fin_ult1’**



```
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
```

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.

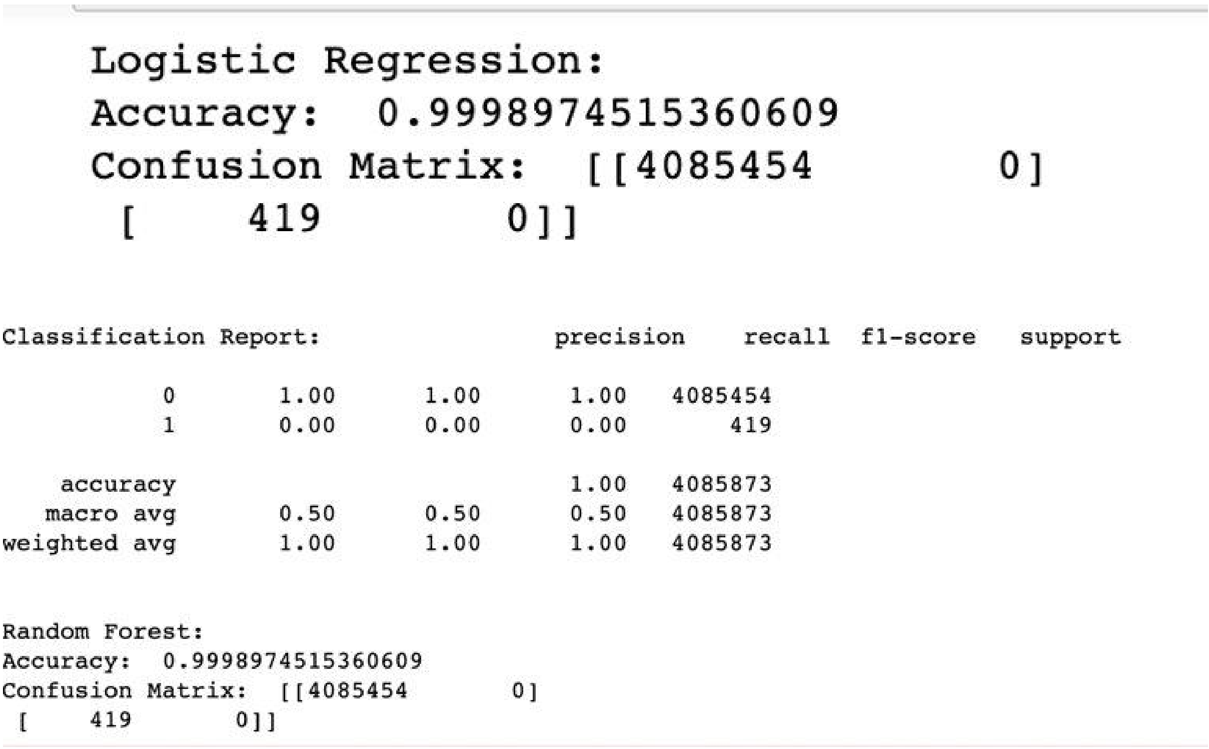
Precision and Recall: The classification report underscores a significant contrast in the model's performance for the two classes:

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 toward the majority class.

2. Logistic Model



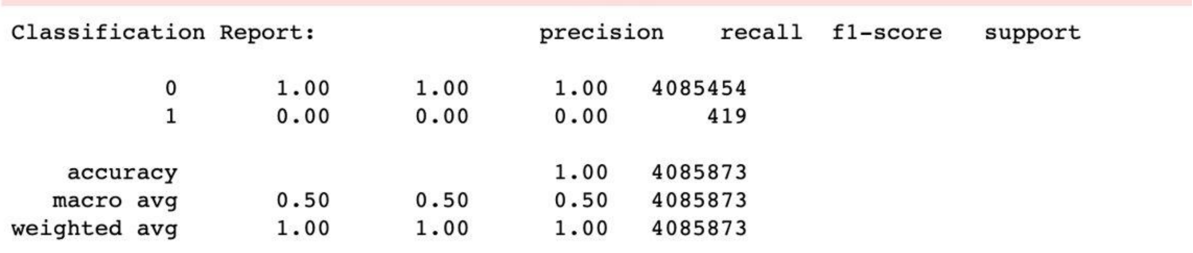
```
Logistic Regression:
Accuracy:  0.9998974515360609
Confusion Matrix:  [[4085454       0]
 [    419       0]]


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]]
```

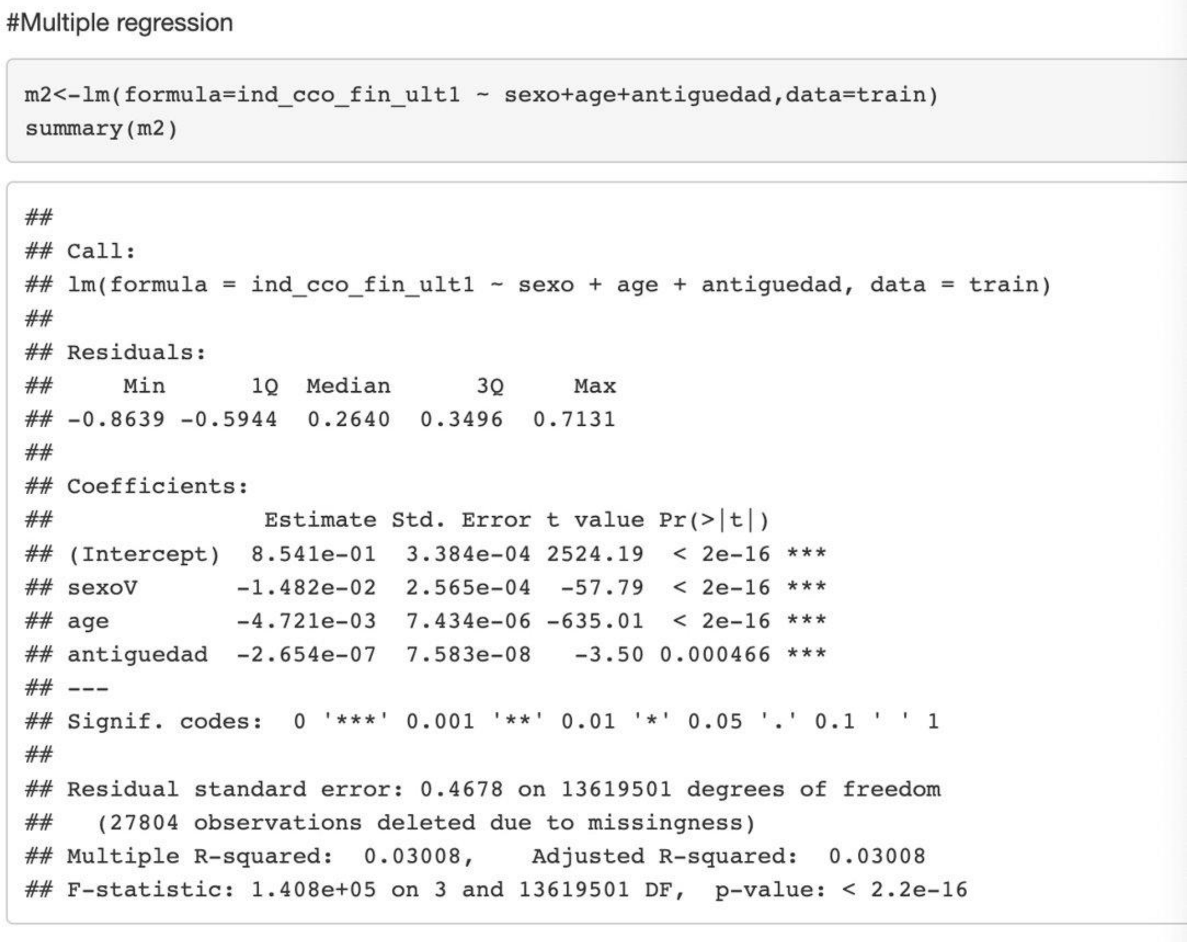


```
Classification Report:               precision    recall  f1-score   support

           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
```

3. m2 of 4 Linear Models (m1-m4) for target variable "ind_cco_fin_ult1"



#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
```

Model Comparing

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 seniority.

According to other model (m1 to m4) comparisons, 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 belong to the neutral category.

For models 1 and 2 (the logistic model), they are created for savings accounts, and their fitting results are near 1, so we choose model 3.

So in this case, the final selection and recommendation of customer businesses will be predicted based on this model.