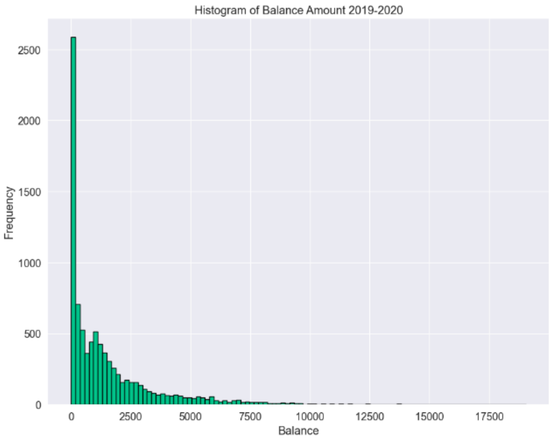
STORI DATA CHALLENGE CASE

LUIS BALDERAS

This analysis aims to find insights into the most common balance amount, predict fraud and find any pattern related of data between 2019-2020 about customers. The dataset presented has 22 columns with 8950 rows respectively.

1.1 Plot an histogram of the balance amount for all the customers.



A histogram was plotted to visualize the distribution of balance across all customers. The distribution is right-skewed, with most customers having low balances, and a small number holding very high balances.

bin with more than 2500 elements:

- frequency: 2588.0

- Range: 0.00 a 186.70$

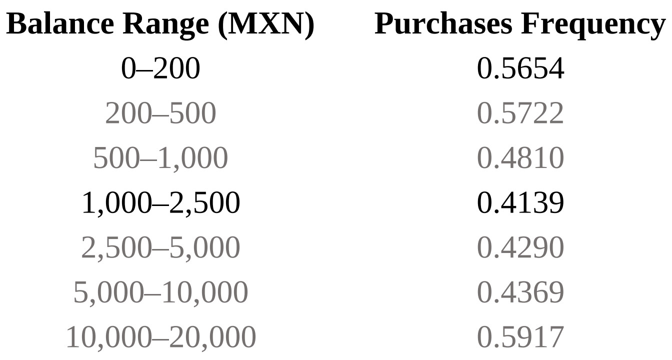
1.2 Report any structure you find and any hypotheses you have about that structure.

We observe a positive skew in the balance distribution, suggesting that a large group of customers keeps low or zero balances, while a minority maintains higher levels. This could indicate:

- *Limited credit usage among most users.*
- *A segmentation between non-active and active cardholders.*

*A potential hypothesis: customers **with higher balance frequency** might also exhibit **greater use of installment purchases.***

I propose this hypothesis based on the observation that accounts with balances between 0–200 MXN and 1,000–2,500 MXN represent the two most common balance ranges across all customers. Notably, accounts within the 0–200 MXN range stand out, as they also rank among the top three in terms of purchase frequency, suggesting active usage despite having low available balances.



Balance Range (MXN)	Purchases Frequency
0–200	0.5654
200–500	0.5722
500–1,000	0.4810
1,000–2,500	0.4139
2,500–5,000	0.4290
5,000–10,000	0.4369
10,000–20,000	0.5917

*Accounts with balances between 0–200 MXN are most likely to make purchases—but **¿do they prefer one-time payments or installment plans?***

Balance Range	One-off Purchase Frequency	Installment Purchase Frequency
0–200	0.1394	0.4408
200–500	0.2966	0.4082
500–1,000	0.2394	0.3533
1,000–2,500	0.1895	0.2985
2,500–5,000	0.2264	0.3025
5,000–10,000	0.239	0.3318
10,000–20,000	0.329	0.4898

This indicates that customers in this low-balance segment are likely to rely on installments plans to manage spending, potentially reflecting a preference for structured payments.

We can conclude that the high frequency of accounts with balances between 0–200 MXN is likely due to customers making high-value purchases through installment plans, which results in a low remaining balance in their accounts.

1.3 Report mean and median balance, grouped by year and month of activated_date.

Month Activated	Mean Balance (MXN)	Median Balance (MXN)
October 2019	2,482.23	1,524.41
November 2019	1,848.70	1,082.07
December 2019	2,018.79	1,162.59
January 2020	1,854.54	1,175.75
February 2020	1,747.35	994.84
March 2020	1,554.97	828.95
April 2020	1,483.18	910.14
May 2020	1,214.33	734.56
June 2020	940.00	472.79
July 2020	649.72	221.29

3.1 Build a predictive model for fraud.

To build a fraud prediction model, I began by confirming that the target variable fraud was binary, consisting of values 0 (no fraud) and 1 (fraud).Once I confirmed it was about a binary class then I analyzed the class distribution to verify if it is about a class imbalance problem and actually since only less than 1% of records were labeled as fraud (value : 1) (70 out of 8,950 records) I can confirm this is a class imbalance problem.

This is important to mention, because it is worth emphasizing that several contrast pattern-based classifiers do not have a good behavior with class imbalance problems due to some objects belonging to the minority class can be identified as noise.

Given this imbalance, standard classification models would likely be biased toward the majority class. Therefore, we explored and proposed several modeling strategies designed to address class imbalance, including:

- *RUSBoostClassifier*
- *XGBoost with scale_pos_weight adjustment*

It might be worth exploring other models such as PBC4CIP, a contrast-pattern-based classifier proposed by Octavio Loyola-González et al. (2017), specifically designed to address class imbalance problems and out-performing other models according to AUC and GM.

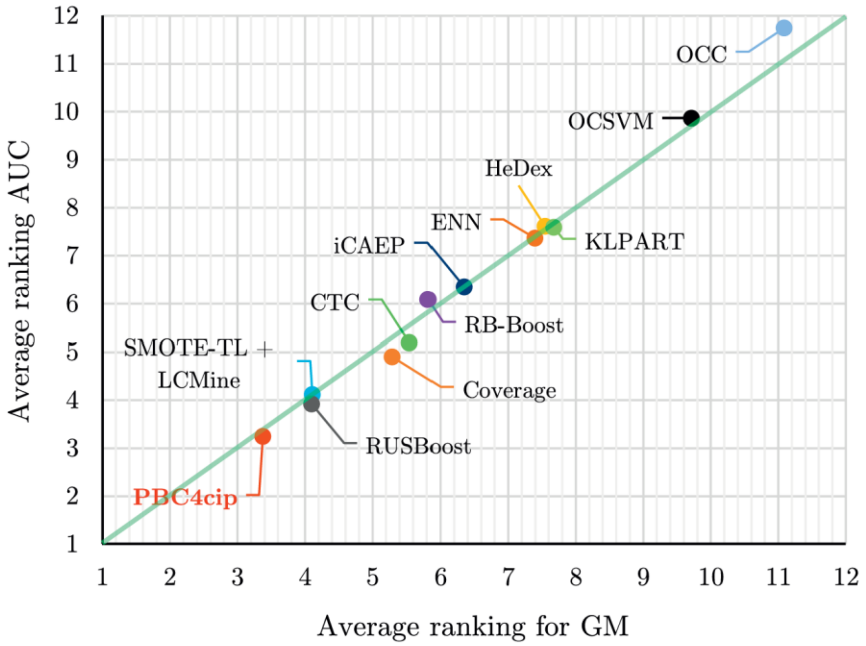
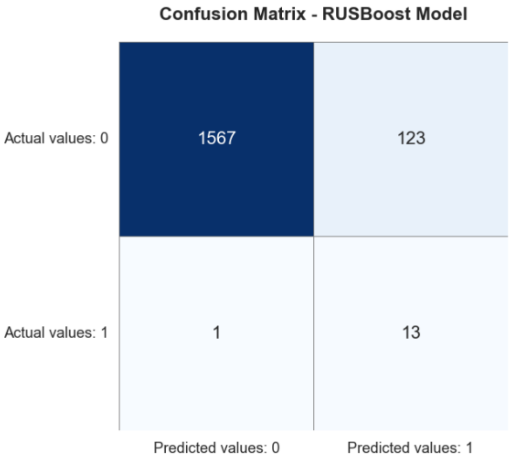


Fig. 2. Average ranking for the tested classifiers, according to AUC vs GM.

*Afte removing rows with missing values, stratified the train-test split to ensure that the proportion of classes in the target variable is preserved in both the training and test sets and trained both models (**RUSBoost** and **XGBoost**) these are the results presented as confusion matrix.*

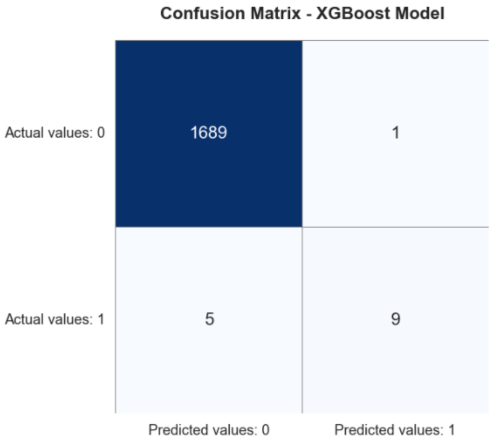
RUSBoost:



Confusion Matrix - RUSBoost Model

	Predicted values: 0	Predicted values: 1
Actual values: 0	1567	123
Actual values: 1	1	13

XGBoost:



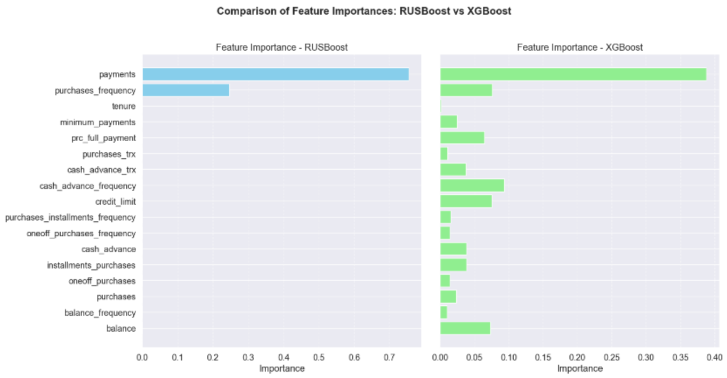
Confusion Matrix - XGBoost Model

	Predicted values: 0	Predicted values: 1
Actual values: 0	1689	1
Actual values: 1	5	9

RUSBoost demonstrated stronger performance in correctly identifying true positive fraud cases; however, it also produced a higher number of false positives. Despite this, I recommend the RUSBoost model, as it is generally more acceptable to incur false positives than to miss actual fraud cases (false negatives).

3.2 What explanatory variable was the most powerful predictor for fraud?

Payments variable was the most powerful predictor for fraud column target.


Comparison of Feature Importances: RUSBoost vs XGBoost
Feature Importance - RUSBoost
Feature Importance - XGBoost
payments
purchases_frequency
tenure
minimum_payments
prc_full_payment
purchases_trx
cash_advance_trx
cash_advance_frequency
credit_limit
purchases_installments_frequency
oneoff_purchases_frequency
cash_advance
installments_purchases
oneoff_purchases
purchases
balance_frequency
balance
Importance
Importance