Graphical Abstract

Detecting customers with inactive credit cards using a KNN machine learning model and descriptive statistics

José Eduardo Solís García, Dr. Álvaro Eduardo Cordero Franco

Highlights

Detecting customers with inactive credit cards using a KNN machine learning model and descriptive statistics

José Eduardo Solís García, Dr. Álvaro Eduardo Cordero Franco

- We analyzed various financial and demographic variables to create a machine learning model using the KNN method with the goal of predicting customer inactivity.
- A visual analysis (descriptive statistics) was conducted to ensure that the obtained data is easily understandable for any individual.

Detecting customers with inactive credit cards using a KNN machine learning model and descriptive statistics

José Eduardo Solís García[b,a], Dr. Álvaro Eduardo Cordero Franco[c,a]

[a] *–Universidad Autonoma de Nuevo Leon, Facultad de Ciencias Fisico Matematicas˝,–San Nicolas de los Garza˝, –66451˝, –Nuevo Leon˝, –Mexico.˝*
[b] *eduardo.solisgre@uanl.edu.mx*
[c] *alvaro.corderofr@uanl.edu.mx*

Abstract

This study focuses on predicting credit card inactivity to help businesses optimize costs. The research employed the K-Nearest Neighbors (KNN) machine learning algorithm, analyzing various demographic and credit-related factors.

Using tools such as GCP, Python, and Tableau, the study revealed key findings: credit card inactivity is influenced by factors such as registration location, cardholder age, and regional demographics.

Understanding these factors enables businesses to take targeted measures to prevent customer disengagement and associated financial losses.

Keywords: TDC, Data Analysis, KNN, Machine Learning

1. Introduction

The inactivity of customers in using their credit card can pose a challenge for card-issuing companies, as customer inactivity can directly impact their revenue and profitability **4**.

Banks are not inclined to have inactive credit cards for several reasons, such as:

- Operational costs: Maintaining an active credit card involves operational costs for the bank, including sending statements, issuing new cards, and managing the account. If a card remains inactive, the bank incurs these costs without generating associated income**2**.

- Credit risk: Inactive credit cards pose a potential risk to the bank, as the cardholder could use it at any time and accumulate debt. If the customer has no intention of using the card and is not generating income to pay it off, the bank assumes the risk of default or non-payment**2**.

- Idle capital: Inactive credit cards tie up the bank's capital that could be used more productively in other financial operations. By freeing up that capital, the bank can allocate it to loans or investments that generate interest or returns**2**.

- Rewards programs: Many credit cards offer rewards programs, such as points or miles, that accumulate with active card usage. If a card remains inactive, the bank does not reap additional benefits from these programs, reducing its ability to retain customers and generate higher income**2**.

To study this conduct, it is important to consider different variables. Typically, when conducting this type of analysis, the following variables are crucial to examine:

Inactivity period, usage frequency, transaction amount, demographic data (branch location, age, gender, etc.), and credit limit**1**.

In this research we will delve deeper into some of these variables to better understand the behavior of customer inactivity when using their credit card**3**.

2. Context

In the past decade, Banco Financiero has experienced significant growth in its customer base, reaching an impressive figure of 200,000 registered clients. However, a persistent concern has been the increasing number of inactive clients in the system. Despite efforts and strategies implemented to reactivate these accounts, results have been elusive to date.

This bank considers an inactive customer if the person does not use his or her credit card for three months.

This issue not only affects the financial health of the bank but also raises questions about the effectiveness of current customer retention strategies and the ability to anticipate account inactivity from their inception.

Identifying customers prone to becoming inactive has proven to be a considerable challenge. Although various tactics and reactivation campaigns have been applied, the lack of success suggests the need for a more predictive and proactive approach. Sole reliance on traditional methods and marketing strategies has not fully addressed the problem.

3. Justification:

To predict customer inactivity, we will rely on the supervised machine learning model KNN (K-Nearest Neighbors).

K-Nearest Neighbors is a supervised learning algorithm. Where the data is 'trained' with data points corresponding to their classification. To predict the class of a given data point, it takes into account the classes of the 'K' nearest data points and chooses the class in which the majority of the 'K' nearest data points belong to as the predicted class.

Assume we have a dataset with labeled points in an n-dimensional space. Each data point has a label indicating its class. Suppose we also have a new, unlabeled point that we want to classify.

Calculate the distance between the new point and all points in the training dataset. The most commonly used distance metric is the Euclidean distance in an n-dimensional space, defined as:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{1}$$

Select the k nearest points to the new point, where k is a predefined parameter in KNN. For classification, count the number of points in each class among the k nearest neighbors. The new point is assigned to the class with the majority of votes. For regression, the predicted value can be the average of the values of the k nearest neighbors.

The choice of the value of k is crucial in KNN. A smaller k can make the model sensitive to noise, while a larger k can overly smooth the decision boundary.

4. Overall Objetive

The proposal to implement a mathematical algorithm arises as an innovative response to address customer inactivity from its outset. This approach

involves using historical data and specific parameters to more accurately predict which customers have a higher probability of becoming inactive in the future.

The algorithm could be based on a combination of variables such as spending patterns, transaction frequency, online service interactions, and payment behavior. The collection and analysis of this data over time will enable the identification of trends and early signals of customers prone to becoming inactive.

The proposed algorithm will leverage the K-Nearest Neighbors (KNN) machine learning model. This model, known for its simplicity and effectiveness, will enhance the predictive capabilities of the algorithm by considering the proximity of customers in feature space. The utilization of KNN aligns with the bank's commitment to advanced technological solutions for addressing complex challenges.

5. Specific Objective

- **Proactive Prediction of Customer Inactivity:** evelop a cutting-edge predictive system using machine learning models, specifically the K-Nearest Neighbors (KNN) model, to proactively identify customers prone to becoming inactive from the moment of their onboarding.
- **In-Depth Understanding of Inactivity Variables:** Conduct a comprehensive analysis of historical and behavioral variables to identify crucial factors contributing to customer inactivity.

6. Problem Statement

Within the framework of the financial services optimization project, the goal is to anticipate and understand customer behavior when applying for a credit card at any branch in Mexico. The main objective is to discern whether a customer will exhibit inactivity or demonstrate significant recurrence in their interaction with the offered financial services.

To carry out this analysis, we will focus on the comprehensive exploration of the database that records the behavior of various customers over the last 12 months. This comprehensive study will incorporate multiple variables, including demographic aspects such as age, gender, region, among others. Relevant factors such as credit limit, the type of credit card applied for, as

well as details about the onboarding process, such as the registration method and dates of incorporation, will also be taken into account.

The methodology employed will rely on descriptive statistical techniques to conduct a detailed analysis of these variables. The goal is to identify patterns and relationships that allow us to understand the crucial points leading to customer inactivity. This statistical approach will not only help understand trends but also provide a clear and easily interpretable insight for all levels within the company.

Data visualization will play a crucial role in presenting the results. Intuitive graphs and tables will be used to highlight the most relevant findings. This approach will facilitate the communication of complex information in an accessible manner for all members of the organization, thus promoting widespread understanding.

In summary, this project seeks to provide the company with an effective tool to predict and understand customer behavior regarding credit card applications. The combination of demographic variables, credit limits, and other key factors, along with rigorous statistical analysis, will identify the inflection points leading to customer inactivity. The visual presentation of results will ensure that they are accessible and understandable for all departments in the company, thereby facilitating informed and strategic decision-making.

7. Hypothesis

We posit that utilizing the K-Nearest Neighbors (KNN) model will enable successful prediction of whether a customer, upon receiving their TDC, will become inactive or not. Additionally, through descriptive statistics, we aim to identify key factors that contribute to customer inactivity.

8. Methodology

To begin the analysis, the first step is to create the database of inactive users, and for this, the GCP tool was used. GCP (Google Cloud Platform) is a cloud services platform offered by Google. It provides a wide range of tools and services for storage, computing, data analysis, artificial intelligence, application development, and more. To manage this tool, users must have programming skills in SQL**7**,**8**.

The variables we considered were the following:

- Months on the portfolio

- Enrollment day (Monday, Tuesday, Wednesday, ...)
- Enrollment month (January, February, March, ...)
- Enrollment year
- Product (type of credit card)
- Enrollment center
- Origination type (Modulo o emboce)
- Credit limit
- Age
- First Purchase Amount
- Transactions during the month
- Purchases during the month
- External credit limit

The data was randomly obtained from clients who enrolled in different years, along with their behavior in the subsequent months after enrollment.

9. Results

classification report				
	precision	recall	f1-score	support
Activo	0.95	0.90	0.92	39,175
Inactivo	0.87	0.94	0.90	28,825
accuracy			0.92	68,000
macro avg	0.91	0.92	0.91	68,000
weighted avg	0.92	0.92	0.92	68,000

Figure 1: Classification Report table displaying the effectiveness of the KNN machine learning model.

Confussion Matrix						
	Active	Inactive			Active	Inactive
Active	35,246	3,929		Active	89.97%	10.03%
Inactive	1,822	27,003		Inactive	6.32%	93.68%

Figure 2: CThe confusion matrix of the KNN machine learning model employed.

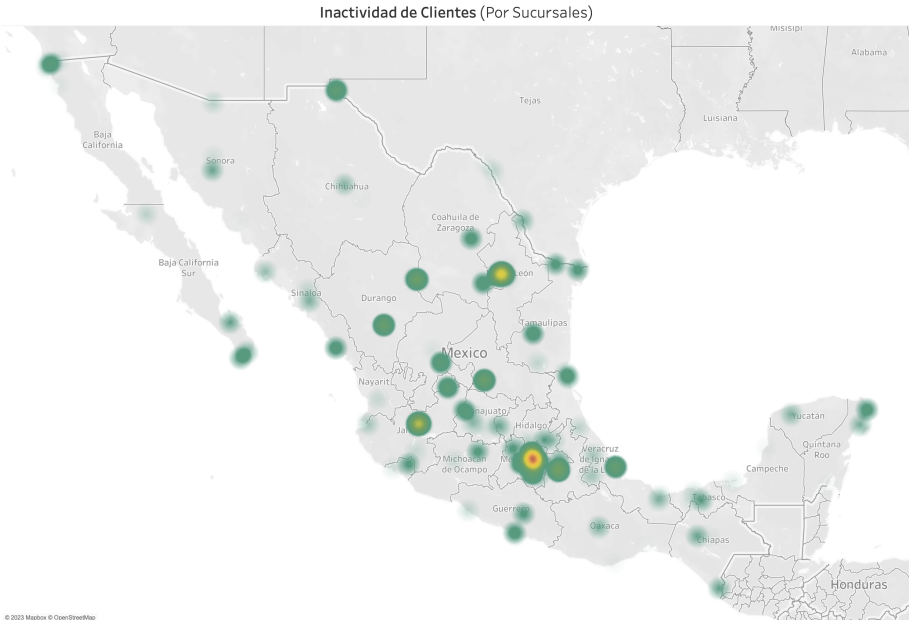

Figure 3: Graph shows the density of inactive customers per enrollment center in Mexico.

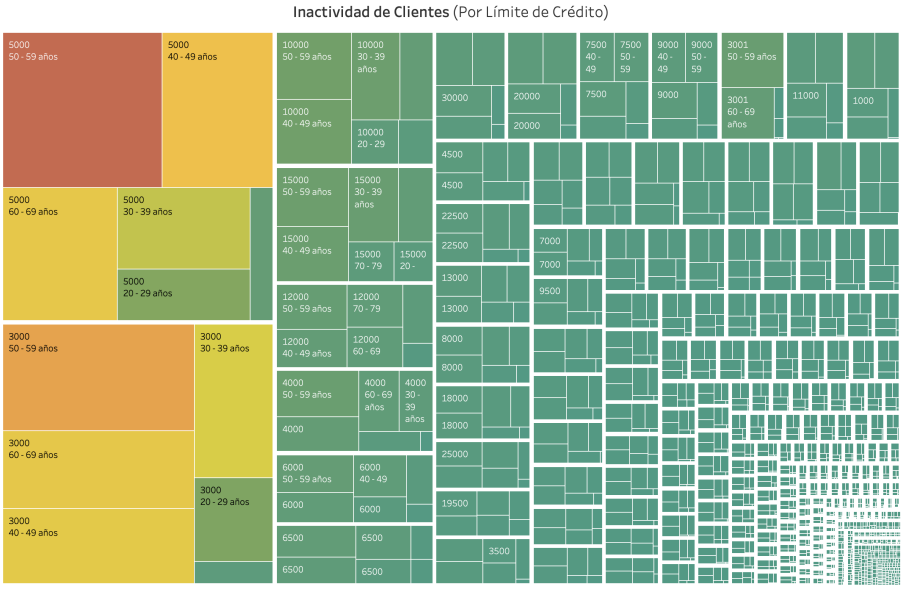

Figure 4: Graph that shows the credit limits most prone to inactivity categorized by age range.

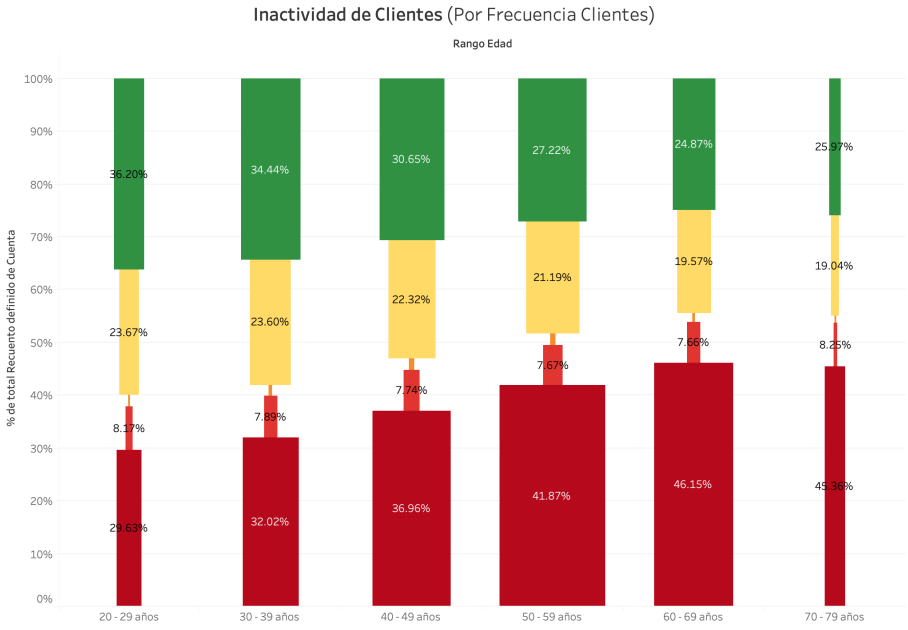

Figure 5: Graph that shows various customer recurrence types, where the thickness represents the quantity of customers in each recurrence type, segmented by age range.

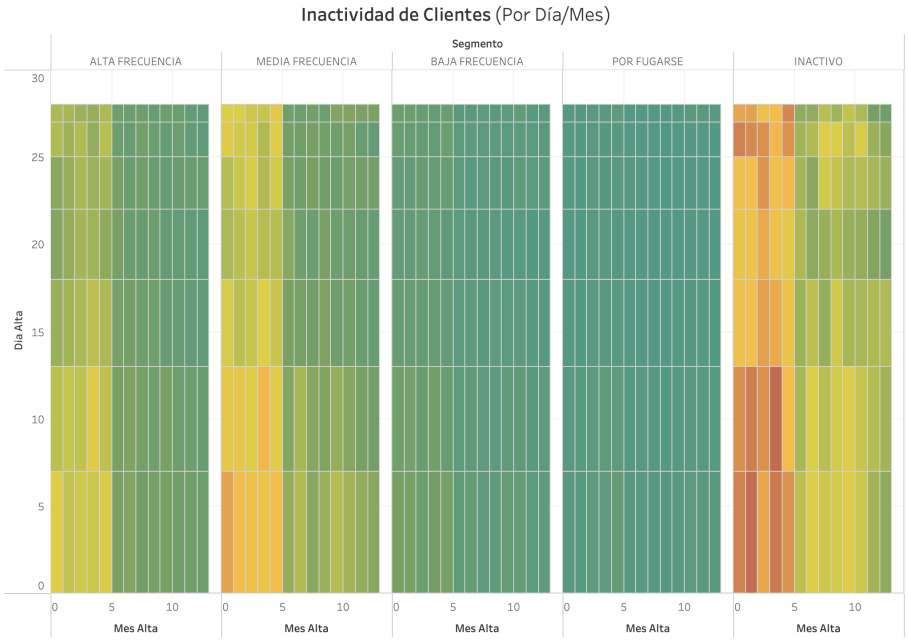

Figure 6: Graph that shows the number of inactive customers based on the day and month they registered, where red represents the highest number of customers and green the lowest.

10. Discussion

In the first table, a classification report is presented, evaluating the performance of a KNN machine learning model in a binary classification problem. Precision measures the proportion of instances classified as positive that are truly positive. A high value indicates that the model has few false positives. In this particular case, the values obtained are high, indicating a high precision value.

On the other hand, recall measures the proportion of positive instances that were correctly identified by the model. A high value indicates that the

model has few false negatives. The F1-Score is a metric that combines precision and recall into a single number. It is useful when there is an imbalance between classes, and in this case, both metrics show high values.

The confusion matrix provides information on how a classification model is making predictions in different classes. In this case, the confusion matrix is represented in Figure 2.

Active (Real) vs. Active (Predicted): 89.97pp of instances that are truly "Active" were correctly predicted as "Active" by the model, while 10.03pp were incorrectly classified as "Inactive" (false negatives).

Inactive (Real) vs. Active (Predicted): 6.32pp of instances that are truly "Inactive" were incorrectly classified as "Active" (false positives), and 93.68pp were correctly predicted as "Inactive."

These percentages are percentages of the total instances in each category. In summary, the model appears to have a strong performance in classifying both classes, with high precision for both, especially for the "Inactive" class.

Figure 3 displays the density of inactive customers per branch across the country. It is crucial to highlight that three cities stand out prominently (Guadalajara, Mexico City, and Monterrey), which are the primary cities processing a higher number of credit cards. Therefore, it is logical that these cities have a higher concentration of inactive customers.

Figure 4 presents a block diagram where each block represents the credit limits within specific age ranges. The size of each block corresponds to the number of inactive customers associated with that particular combination of credit limit and age range. Larger blocks indicate a higher number of inactive customers in that specific credit limit and age category.

In Figure 5, the number of inactive customers is presented, categorized by age range. Different blocks are displayed, where the size of each block represents the quantity of inactive customers. Furthermore, the color of each block is divided into five segments: high-frequency customers, medium-frequency customers, low-frequency customers, at risk of leaving, and inactive customers.

Finally, in the last figure, the quantity of customers is depicted by segments. The color red represents a high number of customers, while green indicates fewer customers. Each segment consists of a grid of 7 rows and 12 columns. Vertically, the first row at the top represents Monday, and the bottom row represents Sunday. Horizontally, the 12 columns represent each month, with the leftmost column representing January and the rightmost column representing December.

11. Conclusion

Given the visualized data, we can conclude the following

- Thanks to the results obtained from the classification report and the confusion matrix, we can assert that our machine learning model, with the provided parameters

- Examining the map of inactive customer density by branch (Figure 3), there are noteworthy details that draw attention. In branches bordering with United States states, a certain intensity of green color is noticeable, especially in the states of Baja California and Chihuahua. In border areas with high population mobility, such as the presence of migrant workers, credit card usage practices may be influenced by factors like temporary residence and employment circumstances.

  We also observe green points in southern states, particularly in Guerrero, Oaxaca, Chiapas, and Veracruz. This can be correlated with the socioeconomic level of these states compared to others in the country. They may be affected by specific economic conditions in the region, such as lower employment rates or lower incomes, which could influence their ability to maintain or use credit cards. This could be a significant factor contributing to customer inactivity.

- In Figure 4, we can observe that credit lines ranging between 3,000 and 5,000 Mexican pesos prominently occupy a place among inactive customers. Likewise, the age group that stands out the most is that of 'mature' individuals, aged between 40 and 60 years. We chose to analyze the credit line because the assigned value is closely linked to the client's socioeconomic situation. It is interesting to note that mature individuals are inactive customers, but this group shares the characteristic of having low incomes and, due to their age, typically holding stable employment. In contrast to younger users who may have lower credit lines, but given their short age, their economic level may improve over the years. Therefore, granting credit cards to mature users with low incomes makes them highly prone to falling into inactivity.

- We can confirm our previous point with Figure 5, as we observe that the largest blocks of inactivity fall within the age range of 40 to 69 years. In contrast, individuals between the ages of 20 and 39 have a

significant percentage of high frequency. Additionally, within the 40 to 49 age range, nearly half of the customers exhibit both high frequency and inactivity. In other words, this age group tends to approach the extremes in terms of customer behavior.

- In Figure 6, let's focus on the segment of inactive customers and observe how the customers registering on Saturdays and Sundays prominently appear in red. What catches the eye is that this behavior is noticeable only between the months of January and May, and subsequently, the number of inactivity decreases. The segments of low frequency and at risk of leaving do not exhibit interesting patterns, but the high and medium-frequency segments do. There is a behavior very similar to that of the inactive group. Thanks to this graph, we can realize that the segment with the highest number of customers in the portfolio is the medium-frequency segment.

In summary, the in-depth data analysis reveals valuable patterns, especially in identifying inactive customers. This knowledge provides a solid foundation for strategic decision-making. The ability to predict customer inactivity, as demonstrated in our visualizations, offers a powerful advantage. This information will enable the marketing team to develop personalized and targeted campaigns to retain these customers, preventing their loss and maximizing overall business performance.

References

[1]Mohammed J. Zaki and Wagner Meira, Jr. *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*, 2nd Edition. Cambridge University Press, March 2020. ISBN: 978-1108473989.

[2]Conover, W.J. (1999). *Practical Nonparametric Statistics*, 3rd Edition. Chapter 3: Categorical Attributes. New York: Wiley. ISBN: 0471160687.

[3]Haider, M. (2015). *Getting Started with Data Science: Making Sense of Data with Analytics*. Pearson Education.

[4]Olliffe, I. (2014). *Principal Component Analysis.* In Wiley Stats Ref: Statistics Reference Online. John Wiley and Sons, Ltd.

[5]Ren, C., Liu, Y. (2019). *Data Analysis using Tableau.* In 2019 4th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), pp. 114-118. IEEE.

[6]Gavrilov, M., Munzer, T. (2020). *SQL Pocket Guide: A Guide to SQL Usage.* O'Reilly Media.
[7]Borne, K. D. (2016). *Google Cloud Platform Essentials.* O'Reilly Media.
[8]Ye, Y.,Li, D. (2018). *Data Visualization with Tableau: Designing, Developing, and Delivering Accurate Visualizations.* Packt Publishing Ltd.

12. Acknowledgments

We would like to express our gratitude to Bank "F" for granting permission to manipulate their data for analysis. Special thanks to Lic. Francisco Pérez Morquecho and Ing. Abraham Mansur for their assistance in gathering the necessary data. Finally, we appreciate the guidance provided by Dr. Álvaro Eduardo Cordero Franco throughout the project.