Increasing Digital Activity and Billing in a Banking Entity using Machine Learning

Joselin Rosemary Diestra Ñañez [1]; Eduardo Carbajal López[2]
[1] Pontificia Universidad Católica del Perú, Perú, joselin.diestra@pucp.edu.pe
[2] Pontificia Universidad Católica del Perú, Perú, ecarbajal@pucp.pe

Abstract – This report describes a project in which Machine Learning technology is used to predict the propensity for e-commerce consumption of a banking entity's customers. The objective of the project is to increase two key business indicators: the percentage of clients who consume digitally in the month and billing, that is, the total amount consumed by clients; To this end, Machine Learning models were developed to predict which customers are the most prone to e-commerce consumption and which are the least prone, thus enabling the business to use this valuable information to redesign, improve and optimize its products. incentive strategies and launch campaigns to clients.

Keywords: Machine Learning, Banking, Supervised Models

I. INTRODUCTION

In recent years, the amount of digital data created or replicated in the world has grown by leaps and bounds. In 2010, there were only 2 zettabytes of data. By 2015, that figure had risen to 16. Growth will continue at an even faster rate, reaching 64 zettabytes by 2020. An average annual growth rate of 20% is expected for 2021-2025, reaching 181 zettabytes of information by 2025 (IDC, Seagate, Statista, 2021). This large amount of data, together with improvements in various technologies, such as the availability and accessibility of large storage and processing power at an affordable cost, has made tools related to generating value from data, such as data analytics and machine learning, very important (Colorado State University Global, 2021).

Data analytics and machine learning can be used in different types of industries (financial, retail, insurance, medical, marketing, etc.), if the company has a large amount of data available, then value can be generated through it. One of the main applications of these tools is in the commercial and marketing area, where it is used to make the right decisions to increase sales, lead conversion, among other KIPs that this area has. According to Christine Moorman, director of The CMO Survey, companies rely on marketing analytics 42% of the time to make decisions, which represents a growth of 12% in the last 5 years (BizEd, 2018). In this report, we will see a case of the application of data analytics and machine learning in the commercial and marketing area of a banking company. The objective of this project is to increase 2 main Key Performance Indicators (KPI): firstly, digital activity, which refers to the percentage of customers who make any type of purchase digitally, and secondly, billing, which refers to the total amount of purchases made by customers. These objectives will be achieved by using predictive models or algorithms to improve decision-making regarding the communication strategy and incentives for customers.

II. LITERATURE REVIEW

Machine Learning algorithms can be categorized into supervised, unsupervised, and reinforcement learning. In the banking sector, supervised learning is commonly used for tasks such as credit scoring and fraud detection. Unsupervised learning is employed for customer segmentation and anomaly detection. Reinforcement learning, on the other hand, has potential applications in algorithmic trading and risk management

Machine Learning has found numerous applications in the banking sector. For instance, it is used to detect fraudulent transactions by identifying anomalies in customer behavior as in [1], [2], [3] or [4]. Additionally, Machine Learning algorithms are employed to assess creditworthiness and predict loan defaults, enabling banks to make more informed lending decisions as in [5], [6] and [7]. Moreover, banks leverage Machine Learning to offer personalized financial products and services, enhancing customer satisfaction and loyalty as in [8] and [9]

The adoption of Machine Learning in banking offers significant benefits, including improved fraud detection rates, more accurate risk assessments, and enhanced customer satisfaction. However, challenges such as data quality, model interpretability, and regulatory compliance need to be addressed. Ensuring data privacy and security is also paramount, as banks handle sensitive customer information as described in [10]

In conclusion, Machine Learning has emerged as a powerful tool for transforming the banking industry. By enabling banks to analyze vast amounts of data and make data-driven decisions, Machine Learning has the potential to revolutionize various aspects of banking operations. Future research should focus on developing more explainable and transparent Machine Learning models, as well as exploring the ethical implications of AI in finance. Additionally, the integration of

Machine Learning with other emerging technologies, such as blockchain in [11] and the Internet of Things as in [12], presents exciting opportunities for the banking sector.

III. CURRENT SITUATION

The business had no model to help them make decisions, no way of identifying customers who were more or less likely to consume digitally, so they just ran information campaigns about the discounts or promotions they already had. In addition, the business unit had tried to launch general incentive campaigns for customers, for example, campaigns that incentivized them through a mileage draw, where if they consumed digitally a certain number of times or a certain amount, they would be entered into a draw to earn a certain number of miles, but these campaigns did not generate the expected profit; Until then, they had not been able to run campaigns with personalized incentives because the budget was limited and because the customer base was very large, they did not know how to prioritize them to give them a personalized incentive.

Given the situation mentioned in the previous point (As is), the company expected (To be) that the analysis project would help them to identify the customers most and least prone to digital consumption, so that a better incentive strategy could be designed. All this in order to achieve the company's main objective, which is to increase 2 key KPIs: the increase in customers' digital activity (percentage of customers who consume in the digital channel) and also the invoicing (total monetary sum of the consumptions made by customers). To do this, the company must encourage customers who do not usually buy online to do so now; it seeks to identify and encourage customers who have not bought online for some time, but who are currently inclined to do so, with the aim of increasing the digital activity of the customers and, as a consequence, also the billings; similarly, for customers who are active in the digital channel (they have recently consumed in this channel), the aim is to encourage them to consume more and thus generate more billings.

IV. DEVELOPED SOLUTION

Based on the survey of the project carried out, the solution was developed, which is made up of different stages: definition and generation of the Target Population, Target, Segmentation, and finally Model Development, these will be described in the following points.

A. Target Population

All clients of the bank that have at least 10% of the minimum monthly income as a line on their cards or as savings.

B. Target

The target was defined as a binary variable, in which success (1) represented that the customer had made at least one e-commerce purchase the following month (compared to the month of analysis) and failure (0) represented that the customer had NOT made at least one e-commerce purchase the following month.

C. Segmentation

By analyzing the client's digital consumption recency (number of months the client had not consumed digitally) and the target, using a decision tree, the optimal segmentation of the client population was determined.

As can be seen in Figure 1, the resulting segments were 3:

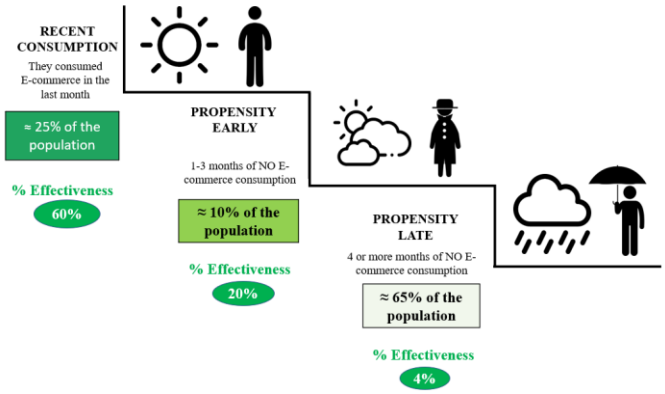

Fig. 1 Customer segmentation according to their digital consumption recentness

Recent Consumption: These are the customers who consumed e-commerce in the month of analysis (last month). They represent approximately 25% of the total target population and have an average effectiveness of 60%, which means that 60% of this population was a "success" in the Target variable, since they consumed on the digital channel the following month. "Effectiveness" in analytical projects represents the measurement of the percentage of successes (target=1) in the population. It is a widely used and important metric in the analysis of data prior to developing the model and also in measuring results.

Early Propensity: These are customers who have not made any e-commerce purchases for 1 to 3 months. They represent approximately 10% of the target population and have an average effectiveness of 20%, which means that 20% of this population is consumed on the digital channel the following month.

Late Propensity: These are customers who have not made any e-commerce purchases for 4 months or more. They represent approximately 65% of the total target population and have an average effectiveness of only 4%, which means that only 4%

of this population is consumed on the digital channel the following month. This group of customers is definitely the least likely to consume the following month. They will be the most difficult group of customers to convince to consume.

D. Model development

After having the Target Population, the target and the Segmentation of the population defined and generated, we proceed with the final stage of model development. In this case, as we have 3 segments, it was proposed to develop 3 models, one for each segment, in order to obtain greater precision in the prediction and implementation of the solution.

The development of the models begins with the search and engineering of variables, then the most relevant variables for the project are selected, then the preprocessing and/or treatment of the generated information is carried out, in order to be able to carry out the training stage of the models in the best way. All these stages are described in greater detail below:

Variable Search and Engineering: Variables are the key components to obtain a model that makes effective predictions. If you give the machine good information to learn from, regardless of the algorithm used, it will learn better and therefore be able to predict with greater certainty and effectiveness. For this reason, different sources of information were reviewed, extracting and/or generating different groups of variables: sociodemographic variables of the client, variables of the client's behavior in the Financial System, variables of the client's products, variables of the client's behavior in the different channels of the bank, campaign variables and lastly, and most importantly, variables of the client's transactions and consumption, both historical and specific. Thus, there were more than 400 variables that could add value to the prediction in the models.

Selection of Final Variables: For the selection of final variables, an XGBoost model was trained for each of the 3 segments using the previously extracted and/or generated variables. Of these variables, 95% were numerical and entered the model as such, and the remaining 5% were categorical variables, so before entering them into the model, the One Hot Encoding technique was used to generate numerical variables with them, that is, these categorical variables were transformed into binary numerical variables. Then, using the trained models, the 85 most relevant variables were selected according to their “gain” indicator, which refers to the “gain” or “benefit” provided by that variable to the prediction. That is, a variable that has a higher value of the “gain” indicator helps more in the prediction and therefore is a more important variable. Likewise, of the 85 variables previously chosen, only those with a correlation less than 0.65 were selected, thus concluding the selection of final variables.

Pre-processing of information: The information used in the models is controlled, that is, it has already passed quality filters, so that it can be easier to use and helps to optimize the time invested in this stage of model development. Thus, the pre-processing applied to the information in both the initial models (for variable selection) and the final models is specific and uncomplicated. The pre-processing applied to the information depends on the type of data in the variables.

Numeric variables: preprocessing consisted only of imputing nulls that numeric variables might have.

Categorical variables: pre-processing consisted of imputing nulls and applying One-Hot-Encoding to convert categorical variables to numerical variables. For variables with several categories, only the 5 most representative ones or those that covered 80% of the population were selected (whichever came first) and the remaining categories were placed in a new group called “Others”.

Model training: The model was trained using the hyperparameter tuning tool present in Amazon Web Services (AWS), this tool allows you to indicate a range of values for the hyperparameters of the model to be trained, in such a way that the same tool trains different models by changing the values of the hyperparameters in the indicated range and after "n" trained models that are indicated, it closes the training process and shows you the best model based on the evaluation indicator that you have indicated, in the case of models with binary target we usually use the AUC - ROC curve (AUC). It is worth mentioning that this tool must be entered into 2 databases, one which is the data with which it will be trained and another which is the testing base, which is data different from the training but which belongs to the same observation periods. AWS does not have this hyperparameter tuning tool suitable for any algorithm, it only has it enabled for a small group of algorithms, among which XGBoost stands out for models with binary target, which is why the final training was performed using the XGBoost algorithm together with the hyperparameter tuning tool. In order to obtain the final AUCs for the models, they were tested with a new test data that was now from different months than those used in the aforementioned data (2 months); as seen in Table 1, the AUCs per month have close values, which provides reliability in the stability of the model throughout the different months and also provides reliability to the final AUC of the model (the average); it is observed that the Recent Consumption and Late Propensity segments have an AUC greater than or equal to 80, which is very good, since it means that they predict, respectively, with an 86% and 80% probability or certainty of making a correct prediction; On the other hand, Early Propensity presents an AUC of 71%, which is considered an acceptable value for a prediction model, looking at the 3 models together as a single solution, the weighted AUC of the final developed project was 80%.

TABLE I
AUC OF THE FINAL MODELS

	Consumo Reciente	Propensión Temprana	Propensión Tardía
AUC mes 1	87%	71%	79%
AUC mes 2	85%	71%	80%
AUC del modelo	86%	71%	80%
% de Población	25%	10%	65%
AUC ponderado general de la solución	**80%**		

V. RESULTS ANALYSIS

The analysis of results is made up of two very important tasks. First, we have the analysis of the final performance of the trained models, for which business indicators are used and it is carried out in the most recent month available; second, there is the analysis of the importance of variables, which will help to know the variables that have contributed the most to the models and therefore the most important in the prediction.

A. Analysis of model performance (Testing)

The analysis of the model's performance includes the determination of business indicators that make it possible to verify that the model manages to adapt and function properly in the client's flow. The flow of the campaign process focused on consumption is defined on a monthly basis, that is why the model was trained with this temporality and that is why the analysis of the performance of the final business models must be carried out and presented per month, to To carry out these analyses, the most recent month available in the development was used.

The results of the Recent Consumption model are seen in the diagram of Figure 2, in the bar graph we can see the effectiveness by decile of the population of the month, the deciles are generated based on the prediction of the model, the higher the value in prediction, the more likely the customer is to consume digitally, that is why in order to analyze the performance of the ordering generated by the model, the deciles are generated and for each one the effectiveness is analyzed (% of "successes", i.e. target = 1 that have been identified in the decile). The graph shows that the model is ordering very well since the effectiveness makes a "staircase" by decile going from the highest effectiveness in Decile 1 to the lowest effectiveness in Decile 10; This segment has an average effectiveness of 78.7% (dotted lines) and with the model we achieved that the first 6 Deciles exceed this average effectiveness and therefore provide the business with a greater probability of finding the target customers; Likewise, it is important to highlight that in the first 5 GE we can see that the effectiveness is above 90% and if we focus on the first 2 GE we can see that its effectiveness is practically 100%, which is super incredible and valuable since with the model we have managed to identify a set of clients (Decile 1 and Decile 2) in which practically all of them will consume e-commerce. Regarding billing, it was also analyzed and has a very similar behavior to effectiveness, it is also going from the highest % Billing (with respect to the total) in Decile 1 to the lowest % Billing in Decile 10 and with only the first 3 Deciles we managed to identify 55% of the total billing. In conclusion, the model presents very good and positive results for business, so it will definitely help with your campaign strategy.

Fig. 2 Model performance results for Recent Consumption

The results of the Early Propensity model are seen in the graph in Figure 3, it can be seen that the model is ordering very well since the effectiveness makes a "staircase" by Decile going from the highest effectiveness in Decile 1 to the lowest effectiveness. in Decile 10; This segment has an average effectiveness of 31.8% (dotted lines) and with the model we achieve that the first 5 Deciles exceed this average effectiveness and therefore provide the business with a greater probability of finding the target customers; Likewise, it is important to highlight that, in the first Decile we have twice the average effectiveness, this means that with the help of the model we have managed to identify a group of clients (Decile 1), in which we have twice the probability of find target customers. Regarding billing, it was also analyzed and has a very similar behavior to effectiveness, it is also going from the highest % Billing (with respect to the total) in Decile 1 to the lowest % Billing in Decile 10 and with only the first 3 Deciles we managed to identify 57% of the total billing. In conclusion, this model also presents very good and positive results for business, so it will definitely help in your campaign strategy.

Fig. 3 Model performance results for Early Propensity

The results of the Late Propensity model are seen in the graph in Figure 4, it can be seen that the model is ordering very well since the effectiveness makes a "staircase" by Decile going from the highest effectiveness in Decile 1 to the lowest effectiveness. in Decile 10; This segment has an average effectiveness of 6.3% and with the model we achieve that the

first 3 Deciles exceed this average effectiveness and therefore provide the business with a greater probability of finding the target customers; Likewise, it is important to highlight that in the first Decile we have 3.7 times the average effectiveness and in Decile 2 we have 2.1 times the average effectiveness, this means that with the help of the model we have managed to identify some groups of clients, in those of us who are almost four times and twice as likely to find the target customers, respectively. Regarding billing, it was also analyzed and has a very similar behavior to effectiveness, it is also going from the highest % Billing (with respect to the total) in Decile 1 to the lowest % Billing in Decile 10 and with only the first 3 GE managed to identify 72% of the total billing. In conclusion, this model also presents very good and positive results for business, so it will definitely help in your campaign strategy.

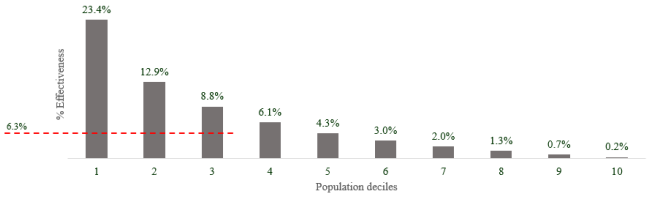

Fig. 4 Model performance results for Late Propensity

B. Importance of variables

The most important variables for each segment are shown in order in Figure 5, the bars refer to the importance (gain) of each variable shown, the larger the bar, the more important the variable is in the model's prediction for that segment. In general, we can observe that among the most important variables for the 3 segments, the variables related to consumption stand out, this was something that I expected to happen, I had the hypothesis that the historical behavior of customer consumption was what It would mainly define your future consumption, which is why the effort was made to design and create the different consumption variables and, as expected, all the effort was worth it.
It is worth mentioning that for each segment the same consumption variables do not stand out and not only these are the most important, which is why below we will carry out a more detailed analysis by segment:

Recent Consumption: in this segment, it is observed that practically all the important variables analyzed are variables related to the quantity and amounts of consumption, specifically e-commerce type consumption and consumption in the entertainment category stand out, the greater the quantity or amount Of these historical consumptions, the customer is more likely to consume e-commerce; In addition, it is observed that the variable of Possession of a Classic Visa Card also stands out (it takes the value "1" when the client has a classic Visa Card and "0" when he does not have that type of Card), clients who do not have a Viso Classic card are more likely to consume e-commerce.

Early Propensity: in this segment, with regard to consumption, it is observed that e-commerce consumption mainly stands out and also non-e-commerce consumption, the more historical consumption of these types, the more likely customers are to consume e-commerce. -commerce. Furthermore, both the recency and frequency of transactions in the last 15 days stand out. The more recently or more frequently the customer has made transactions in the last 15 days, the more likely they are to consume e-commerce. Likewise, we have a very interesting variable related to the campaigns in which the client has participated in the last 6 months, the greater the number of campaigns in which a client has entered in any of the last 6 months, the more likely it is to consume e-commerce. Finally, we also highlight the second most important variable in this segment, which is the number of months without e-commerce consumption, which indicates that the smaller the number of months without e-commerce consumption, that is, the more recently there the more e-commerce the customer consumes, the more likely they are to consume e-commerce in the following month.

Late Propensity: in this segment, as it is made up of clients who have not consumed e-commerce for 4 months or more, it is observed that the variable TOP1 (the most important) is no longer a variable related to consumption, in this case, The most important variable is the balance on the client's credit card. The more available balance the client has on their TC, the more likely they are to consume e-commerce. However, we highlight that among the most important variables shown there are consumption variables, specifically quantities and amounts of e-commerce, non-e-commerce consumption and revolving purchases (purchases without installments, these would be paid in full at closing). of the billing cycle), the more historical consumption of these types, the more likely customers are to consume e-commerce. Additionally, the recency of transactions in stores in the last 15 days stands out. The more recently the customer has made transactions in the bank's stores in the last 15 days, the more likely they are to consume e-commerce. Likewise, we have a very interesting variable related to the age of the customer's first card, the younger the age of the customer's first card, the more likely they are to consume e-commerce. Finally, we also highlight the second most important variable in this segment, which is the number of months without e-commerce consumption, the lower the number of months without e-commerce consumption, that is, the more recently you have consumed e-commerce. the more likely the customer is to consume e-commerce the following month.

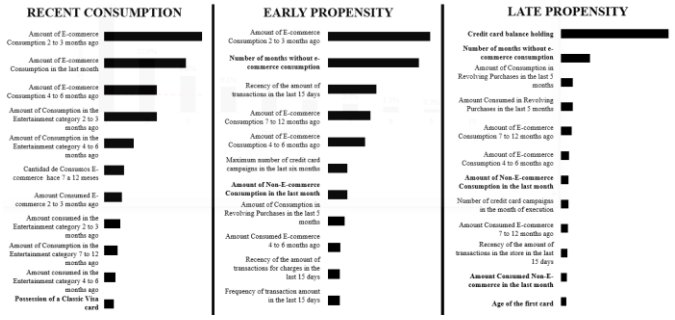

Fig. 5 Most important variables of the models

VI. ECONOMIC IMPACT / BENEFITS ACHIEVED

In order to measure the economic impact of the solution, an A/B testing was carried out for a campaign in which customers would be encouraged to consume digitally, the population was divided into 2 equal parts and 50% of the population, the personalized campaign was sent and communicated using customer prioritization with the deciles generated based on the output of the models and the remaining 50% were not communicated with this personalized campaign; In this way, after the duration of the campaign (1 month) the results in both groups could be compared, thus determining the impact and benefit that this solution provides.

Finally, when the results of the pilot vs control group were compared, very good results were obtained, the pilot group surpassed the control group in both Digital Activity and Billing, thus it was concluded that the use of the model in these campaigns would bring an approximate increase of 10% in digital activity and an increase of more than S/. 1MM turnover.

VII. CONCLUSIONS AND RECOMMENDATIONS

In conclusion, the developed project managed to meet the goals set by the business, thus helping to increase digital activity and billing. This project was fundamental for the business, since its goals were increased by management, but with a minimal budget increase, so advanced analytics turned out to be and continues to be its main ally for meeting its objectives and improving its strategies and competitiveness.

Regarding recommendations, there are points of improvement that can be considered for future models and to be able to have a greater impact on business indicators, these are mentioned below:

In order to seek an improvement in the model, it would be recommended to explore new sources of information, a type of relevant information that I would like to highlight is geographic information, at that time it was not available, but in the future it may be easier to obtain Detailed geographic information about customers, such as the zip code of their residence, the zip code of their workplace, the distance they are from a shopping center, the number of shopping centers, supermarkets and/or retail stores. that they have around their home, the number of times they visit a certain center commercial, the areas where the client usually spends more time according to their GPS, etc., with this information finer and more detailed predictions could be achieved with which we could seek to increase the AUC of each model and with it the % effectiveness in the first deciles.

It would be very good to complement this project with an item recommender model, so that with this new project we know which clients are the most likely to consume e-commerce and with the item recommender we know in which items the client is most likely to consume according to their profile and thus be able to encourage them in the areas in which they are most prone to consumption. In addition, with this you can also set up a campaign to seek to expand customers' digital consumption items, encouraging the most likely customers to consume digitally in new items.

REFERENCES

[1] M. Â. L. Moreira *et al.*, «Exploratory analysis and implementation of machine learning techniques for predictive assessment of fraud in banking systems», *Procedia Comput. Sci.*, vol. 214, pp. 117-124, ene. 2022, doi: 10.1016/j.procs.2022.11.156.

[2] A. Patil, S. Mahajan, J. Menpara, S. Wagle, P. Pareek, y K. Kotecha, «Enhancing fraud detection in banking by integration of graph databases with machine learning», *MethodsX*, vol. 12, p. 102683, jun. 2024, doi: 10.1016/j.mex.2024.102683.

[3] M. N. Alatawi, «Detection of fraud in IoT based credit card collected dataset using machine learning», *Mach. Learn. Appl.*, vol. 19, p. 100603, mar. 2025, doi: 10.1016/j.mlwa.2024.100603.

[4] N. S. A. Polireddi, «An effective role of artificial intelligence and machine learning in banking sector», *Meas. Sens.*, vol. 33, p. 101135, jun. 2024, doi: 10.1016/j.measen.2024.101135.

[5] H. Li y W. Wu, «Loan default predictability with explainable machine learning», *Finance Res. Lett.*, vol. 60, p. 104867, feb. 2024, doi: 10.1016/j.frl.2023.104867.

[6] N. Uddin, Md. K. Uddin Ahamed, M. A. Uddin, Md. M. Islam, Md. A. Talukder, y S. Aryal, «An ensemble machine learning based bank loan approval predictions system with a smart application», *Int. J. Cogn. Comput. Eng.*, vol. 4, pp. 327-339, jun. 2023, doi: 10.1016/j.ijcce.2023.09.001.

[7] C. Bockel-Rickermann, S. Verboven, T. Verdonck, y W. Verbeke, «Can causal machine learning reveal individual bid responses of bank customers? — A study on mortgage loan applications in Belgium», *Decis. Support Syst.*, vol. 190, p. 114378, mar. 2025, doi: 10.1016/j.dss.2024.114378.

[8] P. P. Singh, F. I. Anik, R. Senapati, A. Sinha, N. Sakib, y E. Hossain, «Investigating customer churn in banking: a machine learning approach and visualization app for data science and management», *Data Sci. Manag.*, vol. 7, n.º 1, pp. 7-16, mar. 2024, doi: 10.1016/j.dsm.2023.09.002.

[9] F. Mi Alnaser, S. Rahi, M. Alghizzawi, y A. H. Ngah, «Does artificial intelligence (AI) boost digital banking user satisfaction? Integration of expectation confirmation model and antecedents of artificial intelligence enabled digital banking», *Heliyon*, vol. 9, n.º 8, p. e18930, ago. 2023, doi: 10.1016/j.heliyon.2023.e18930.

[10] S. Wang, M. Asif, M. F. Shahzad, y M. Ashfaq, «Data privacy and cybersecurity challenges in the digital transformation of the banking sector», *Comput. Secur.*, vol. 147, p. 104051, dic. 2024, doi: 10.1016/j.cose.2024.104051.

[11] H. O. Mbaidin, M. A. K. Alsmairat, y R. Al-Adaileh, «Blockchain adoption for sustainable development in developing countries: Challenges and opportunities in the banking sector», *Int. J. Inf. Manag. Data Insights*, vol. 3, n.º 2, p. 100199, nov. 2023, doi: 10.1016/j.jjimei.2023.100199.
[12] E. Fazel, M. Z. Nezhad, J. Rezazadeh, M. Moradi, y J. Ayoade, «IoT convergence with machine learning & blockchain: A review», *Internet Things*, vol. 26, p. 101187, jul. 2024, doi: 10.1016/j.iot.2024.101187.
[13] UNCTAD. (15 de marzo del 2021). How COVID-19 triggered the digital and e-commerce turning point. https://unctad.org/news/how-covid-19-triggered-digital-and-e-commerce-turning-point
[14] Vlačić, B., Corbo, L., e Silva, S. C., & Dabić, M. (2021). The evolving role of artificial intelligence in marketing: A review and research agenda. Journal of Business Research, 128, 187–203.
[15] De Mauro, A., Sestino, A. & Bacconi, A. Machine learning and artificial intelligence use in marketing: a general taxonomy. Ital. J. Mark. 2022, 439–457 (2022). https://doi.org/10.1007/s43039-022-00057-w
[16] IBM. (s.f.). What is machine learning?. https://www.ibm.com/topics/machine-learning
[17] BizEd. (2018). Marketers: No on Politics, Slow on Tech (p. 11).
[18] Mena, M. (2021, 22 de octubre). El Big Bang del Big Data. Statista. https://es.statista.com/grafico/26031/volumen-estimado-de-datos-digitales-creados-o-replicados-en-todo-el-mundo/
[19] Colorado State University Global. (2021, 6 de julio). Why is Machine Learning Important?. https://csuglobal.edu/blog/why-machine-learning-important
[20] SAS. (s.f.). Marketing Analytics. https://www.sas.com/pt_br/insights/marketing/marketing-analytics.html
[21] Jo, T. (2021). Machine Learning Foundations: Supervised, Unsupervised, and Advanced Learning (p. 11). Springer. DOI:10.1007/978-3-030-65900-4
https://books.google.com.pe/books?id=0egdEAAAQBAJ&printsec=frontcover&dq=supervised+learning&hl=es&sa=X&redir_esc=y#v=onepage&q&f=false
[22] Brownlee, J. (2020, 15 de Agosto). Parametric and Nonparametric Machine Learning Algorithms. https://machinelearningmastery.com/parametric-and-nonparametric-machine-learning-algorithms/
[23] IBM. (s.f.). Decision Trees. https://www.ibm.com/topics/decision-trees
[24] Scikit-learn. (s.f.). 1.10. Decisión Trees. https://scikitlearn.org/stable/modules/tree.html
[25] Tyagi, N. (2020, 24 de Marzo). Understanding the Gini Index and Information Gain in Decision Trees. https://medium.com/analytics-steps/understanding-the-gini-index-and-information-gain-in-decision-trees-ab4720518ba8
[26] C3.ai. (s.f.). Gradient-Boosted Decision Trees (GBDT). https://c3.ai/glossary/data-science/gradient-boosted-decision-trees-gbdt/#:~:text=Gradient%2Dboosted%20decision%20trees%20are%20a%20popular%20method%20for%20solving,to%20a%20sufficiently%20optimal%20solution.
[27] NVIDIA. (s.f.). XGBoost. https://www.nvidia.com/en-us/glossary/data-science/xgboost/
[28] Geeksforgeeks. (2023, 1 de Febrero). ML | One Hot Encoding to treat Categorical data parameters. https://www.geeksforgeeks.org/ml-one-hot-encoding-of-datasets-in-python/
[29] Trotta, F. (2022). How and Why Performing One-Hot Encoding in Your Data Science Project. https://towardsdatascience.com/how-and-why-performing-one-hot-encoding-in-your-data-science-project-a1500ec72d85