Advances in Artificial Intelligence and Machine Learning; Research 5 (1) 3202-3215 Received 16-09-2024; Accepted 02-01-2025; Published 09-01-2025

Propensity Model Using Decision Trees (LightGBM) for the Management of the Effective Credit Product in a Financial Entity

Ulises Roman-Concha nromanc@unmsm.edu.pe
Professors at the Faculty of Systems Engineering and Computer Science, UNMSM Lima-Peru

Andrea López jenandreall@gmail.com
Graduated from the Professional School of Systems Engineering, UNMSM Lima-Peru

Kathy Ruiz-Carrasco c23370@utp.edu.pe
Professor at the Technological University of Peru. UTP Lima-Peru

Carlos Chavez-Herrera cchavezh@unmsm.edu.pe
Professors at the Faculty of Systems Engineering and Computer Science, UNMSM Lima-Peru

Dominga M. Cano dm.cano@unaj.edu.pe
Professor at the National University of Juliaca, Lima-Peru

José Piedra jpiedrai@unmsm.edu.pe
Professors at the Faculty of Systems Engineering and Computer Science, UNMSM Lima-Peru

Juan Carlos Woolcott jwoolcotth@unmsm.edu.pe
Professor at the Faculty of Chemistry and Chemical Engineering, UNMSM, Lima-Peru

Carlos Navarro cnavarrod@unmsm.edu.pe
Professors at the Faculty of Systems Engineering and Computer Science, UNMSM Lima-Peru

Corresponding Author: Ulises Roman-Concha

Abstract

The objective of this paper was to develop a propensity model based on decision trees (Light-Gbm) for the management of the Credit product in a financial institution. The CRIPS-DM methodology was used as a framework and Python/LightGBM was used for the development of the solution. As a result, it was possible to increase by 5% the effectiveness in credit for

Citation: Ulises Roman-Concha, et al. Propensity Model Using Decision Trees (LightGBM) for the Management of the Effective Credit Product in a Financial Entity. Advances in Artificial Intelligence and Machine Learning. 2025;5(1):184.

each month on a park of 200 thousand customers of the financial institution, which ensures the understanding of the applied model.

Keywords: Effective credit, Decision trees, Financial entity, LightGBM.

1. INTRODUCTION

Every day, large volumes of information are generated in financial entities, the same that require to create predictive models in order to offer better services on the different products that they offer to their clients, which are: Saving, Investment and Funding [1]. Within the main products of a financial entity are cash credits or loans also known in the researched entity as a parallel line. This loan is additional to their credit line, goes on a parallel way and it is not discounted from the credit card balance. Among other cash products is Maxicash, which are loans that are consumed from the credit card and is done in small amounts. In the case of the researched entity, the cash income in general corresponds to a total of 48% of the companies balance therefore, is a really important product and must be managed properly. The business looked for achieving the objective of adding the product based on increasing the response rate or conversion rate. To achieve the conversion increase, the area counted on a variable given by the risks area denominated "propensity". This variable helped the area to have a better reach of which clients were more prone to withdraw the cash credit along with a score of the client's behavior.

The propensity model mentioned with which the researched financial entity counted on was very limited, where there were only 3 propensity levels: High, Medium and Low and with that, decisions had to be taken and also had to send campaigns for clients based on emailing, mailing, SMS, push notifications, call center, etc. The achieved conversion in general reached 3% in a month.

The LightGBM model would help a financial institution by identifying and classifying customers more accurately in: (1) Retaining customers likely to leave the entity,(2) Fraud detection through analyzing patterns in transactions and suspicious activity in real time, (3) Cross-selling and Upselling opportunities by identifying customers with high probability of accepting additional products, (4) Classifying customers into specific segmentations, which facilitates personalization of services and improves the customer experience, (5) Predicting the risk of default on loans for each customer. LightGBM is fast, handles unbalanced data well and allows interpreting which factors influence each prediction, facilitating more effective decision making. [2–4].

To have a trustable method helps us to take the right decisions in order to have a better efficiency in the management of the cash product.

For those reasons, the propensity model, LightGBM was implemented for clients based on decision trees where differentiated probability scores by client were achieved in order to be more exact and focus on clients that are more prone to exit the campaign. This allowed them to be provided with a differentiated campaign and thus, obtain a better conversion. As for clients that had low probability, a strategy involving communication was proposed so conversion could be possible. This project helped significantly to reorganize the product, to get the probability by client more exactly and thus, remake a strategy of directed campaigns. It also allowed to get to know the clients and get their feedback on the different campaigns and also, achieve significant savings by not sending

communication campaigns to those clients that despite having lots of variables that classified them as an excellent scored client, was not going to do the conversion. In that case, the solution was to map those clients and give them different and unique communication [5].

Financial institutions face multiple challenges in managing customer relationships in an increasingly competitive and digitized environment. Among these challenges are customer retention, fraud detection, cross-selling optimization and credit risk assessment. These issues not only affect profitability, but also customer satisfaction and the financial security of the institution [6, 7].

To address these difficulties, propensity models have emerged as a powerful tool in predicting future customer behaviors and needs. In particular, decision tree-based models, such as LightGBM, offer notable advantages in accuracy, efficiency, and the ability to interpret factors that impact each prediction

2. MATERIALS AND METHODS

2.1 Decision Trees:

The Decision Tree algorithm, whose classification is owned by the supervised learning algorithms. As in other Machine Learning algorithms, automated learning must be achieved first, this is obtained by using the data from the Dataset and then it is applied to another database to get the desirable prediction. This is achieved after identifying the first case patterns. The objective of using a decision tree is to create a training model that can be used for predicting the class or value of the target variable by learning the simple inferred decision rules of past data (training data) [8, 9] as shown in FIGURE -1.

The advantages of the use of the decision tree are [10]:

1. There is a low number of indicators for the determination of consequences and probabilities.
2. There is a measure order of the indicators based on the importance of them. In this research, the decision tree method was used due to the number of variables that give an order in the measurement of the indicators and variables on themselves too.

2.2 Light Gbm Model

For [11], The LightGBM model is and open-source framework proposed originally by Microsoft. It is about an algorithm based on a decision tree that divides the parameters of the input layer on different parts and in that way, builds the relationship of mapping between the inputs and outputs. In FIGURE 2, is shown the main characteristics of the LightGBM model, that utilizes growth in the shape of leaf tree instead of the growth on levels to accelerate the training. Regarding the strategy of level growth, the tree structure grows level by level. As shown in the figure, assuming that the deepness of the tree is D, the number of nodes in the last level is $2 \wedge D$.

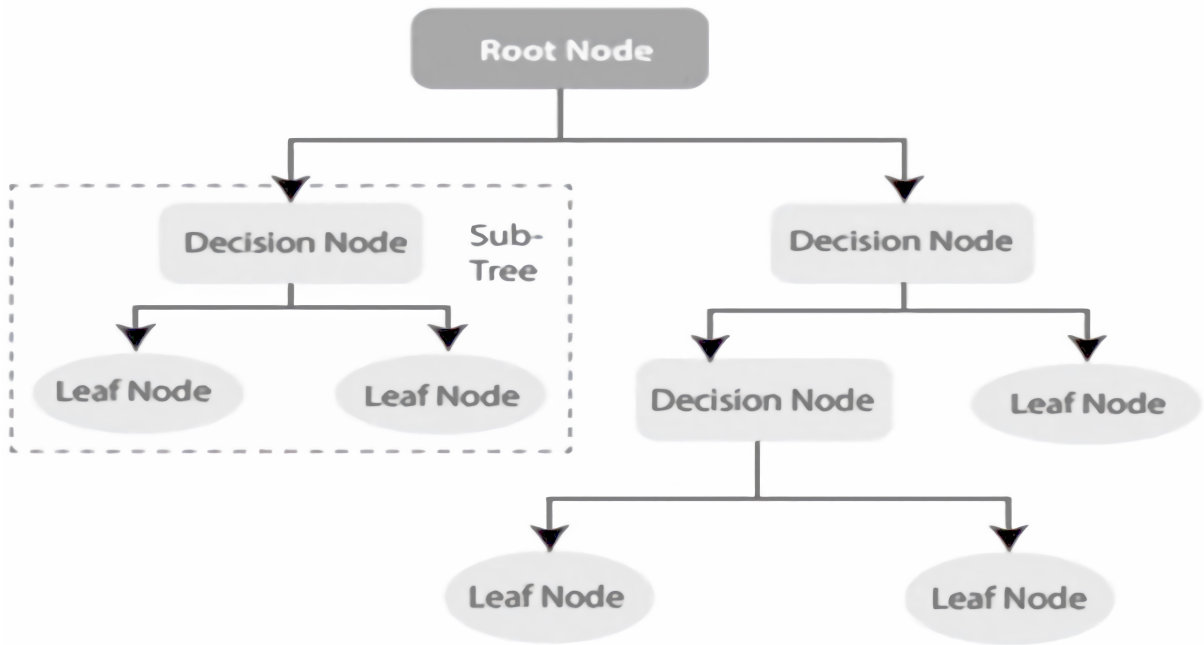

Figure 1: Decision Tree
Note. Extracted graph from [12]

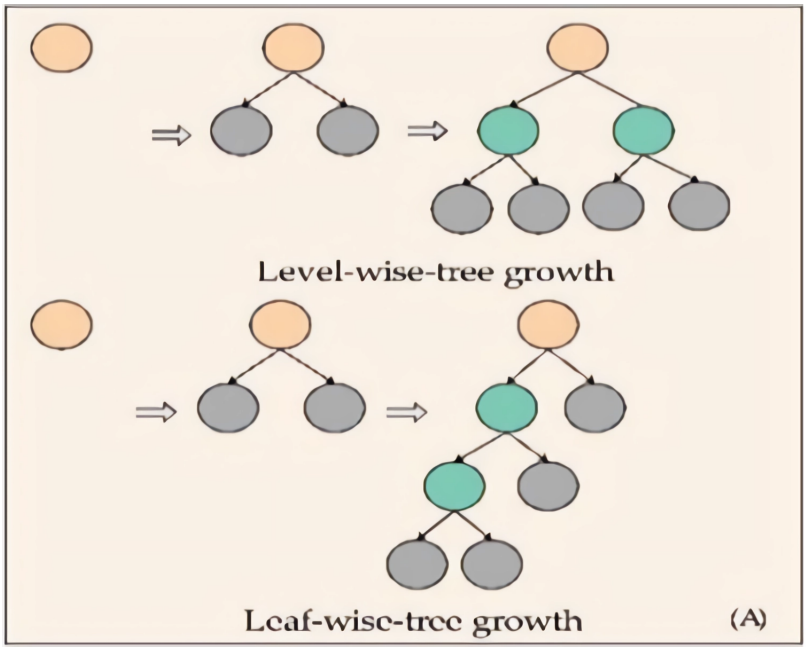

Figure 2: LightGBM model
Note. Diagram extracted from [11]

A growing leaves tree is different from a growing level tree because the first one allows to reduce considerably the number of nodes of the tree. This helps to accelerate the training process in case the dataset is large. So, it would not give the same result if the amount of data is low because the growing leaves tree algorithm has a tendency to adjust itself due to its onerous algorithm [11]. In this research, because it has a large amount of data it is useful to utilize the growing leaves tree method (LightGBM).

2.3 Propensity Model

The propensity model is a technique based on a set of approaches, which is utilized to predict the behavior of the targeted audience by doing an analysis previous to their past behaviors. In other words, they help to know the probability of a determined action to be done [13–15].

The origins come from the year 1983, however it was just recently when we could get to know the whole potential of this technique thanks to the randomized learning, which with the help of tools and equipment with experience in data science it has been possible to see a large growth and also, the generation and exploring of more advanced propensity levels have begun. This knowledge has become essential for building more accurate marketing campaigns in business or financial entities[12, 16].

The following advantages and disadvantages are presented in the use of a propensity model [17]:

Advantages of propensity models:

- By being used as models to predict behaviors based on factors influenced by it, benefits can be obtained from it and thus, guide selling capacities and efforts in order to make them more efficient, successful and reach its objectives.
- These models allow to optimize and promote the acquisition of new clients. This is able to determine and solve which clients can become into strongly loyal clients and thus, focus the commercial campaigns.
- Contributes to the decision in a Smart way to get to know those clients by having the information and knowledge that the competition would not have. It would be an added value.
- It is also employed in strategies to retain clients and prevent client attrition because, it allows to get to know clients that are more likely to leave.
- It is also utilized for improving satisfaction levels of new clients and increase their scores in case it is measured on a company.

Disadvantages of propensity levels:

- Since is being based on past data, the model cannot be exact in case incomplete information is used or it is not a representative sample of the general population.
- Suppositions can be presented if the context is not well comprehended.

The purpose of the research was to implement a propensity model that could predict the behavior of a client towards the given campaigns. This propensity should be differentiated by client unlike the last score where there were 3 general levels: High, medium and low.

2.4 CRISP-DM Methodology

According to the Documentation IBM [18], CRISP-DM, whose acronym means Cross-Industry Standard Process for Data, is a verified methodology or method that is utilize to guide and sequence the mining data works. Which include descriptions about the steps that involve the project, the assignments that are essential on each step in order to fulfill the model and an explanation of the relationships between those assignments.

CRISP-DM steps are detailed here:

- Comprehension of the problem: Phase that allows understanding and delimiting the issue.
- Comprehension of the data: Is the phase that comprehends the gathering, description and exploration of the data.
- Preparation of data: Data selection and cleansing.
- Modeling: The technique that is going to be used is chosen.
- Model testing: The calculated reliability is tested and the best model for the project is selected.
- Model implementation: Model deployment and use.

2.4.1 CRISP-DM application steps

It was taken as an implementation case a financial entity from PERU.

1. Business understanding phase:

The business presented a problem regarding the mistaken segmentation of clients by probability to direct the product campaigns Crédito Efectivo, maxicash and disef; representative products of the financial company in Peru and can be classified as a cash loan to clients with and active campaign. This problem happened because the utilized propensity model presented issues and was not precise. The biggest product and most representative because the credit amounts were greater is Crédito Efectivo, in which the execution model is going to be based on [19].

Each database of the different products counted with approximately 300 thousand clients who had an active campaign, these clients should have been evaluated previously and after the corresponding filters were done, only those who had a good credit score where kept. Also, the cash area should have assigned an interest rate according to the characteristics of each one. The division of the cash campaigns is shown in FIGURE 3.

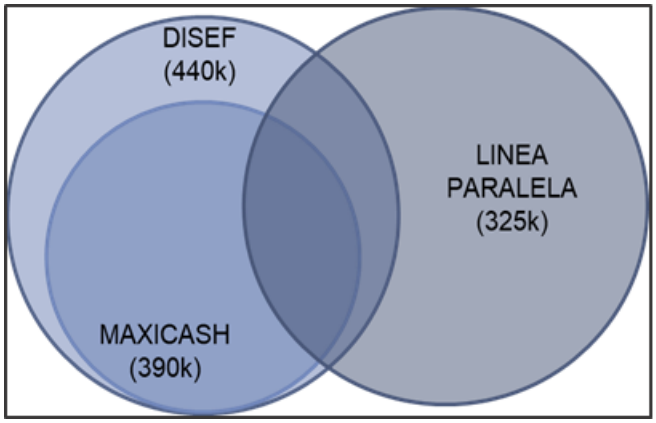

Figure 3: Cash producto types.

2. Data understanding phase.

The data of clients and other information sources to make possible this analysis was found stored in the data base of Oracle as shown in FIGURE -4.

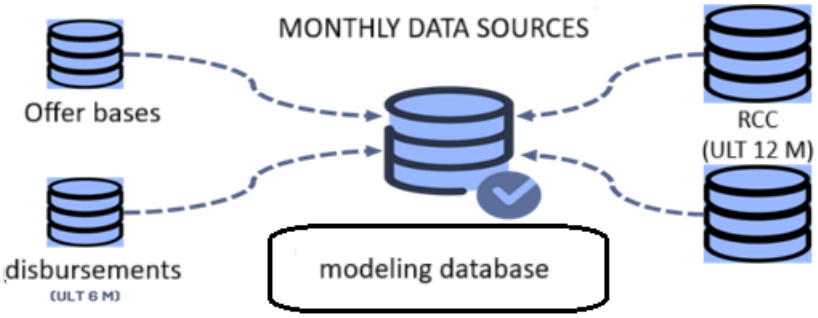

Figure 4: Data sources

The first and principal is the database of clients that the risks team generated (FIGURE -4). For the definition of this database, the database that contained the full client base was utilized and after doing the corresponding filters a Crédito Efectivo campaign database was generated. This database was a percentage of the total database of the base from the financial entity.

The second source is the RCC (Central Credit Report) which is a consolidated record of the different credits of the different financial entities. In this case a 12-moth antiquity was used. The third information source and data were the refunds, to build the model the refund flag was used on the last 6 months. This indicated that clients had taken the product of Crédito Efectivo on the last 6 months. At the same time, there were charts that complemented the refunds chart as shown on Table 1.

Finally, consolidated transactional databases were used, these databases had clients that went to the stores (Canal Retail) or clients that had used the card on different establishments. Withs these charts active clients with active campaigns could be recognize.

Table 1: Product charts of Crédito Efectivo.

Base de Oferta Mensual	RCC	Transaccional	Desembolsos
riesgos.asignar_TASA_aaaamm	esv_sbs.rcc	comercial_seg.tee_trn_retail	mcarbonero.solicitudes_lp_18_19
	esv_sbs.rcc_detalle	esv_gob.tee_trn_retail	esv_puc.puc_parametros
		esv_spsa.productos	esv_puc.sac_sucursal
			esv_finan.finan_lp_tipo_desembolso
			esv_finan.fpmvp_saldos
			esv_finan.finan_lp_tipo_tasa
			esv_finan.finan_tipo_vendedor
			esv_puc-puc_parametros
			esv_finan.finan_lp_solicitude_lp

The following charts are presented as for each source that was used for the construction of the model on Table 1.

3. Data preparation phase:

In the data preparation cross information was done and the query was executed to have the information by client of the RCC charts in order to continue with model execution.

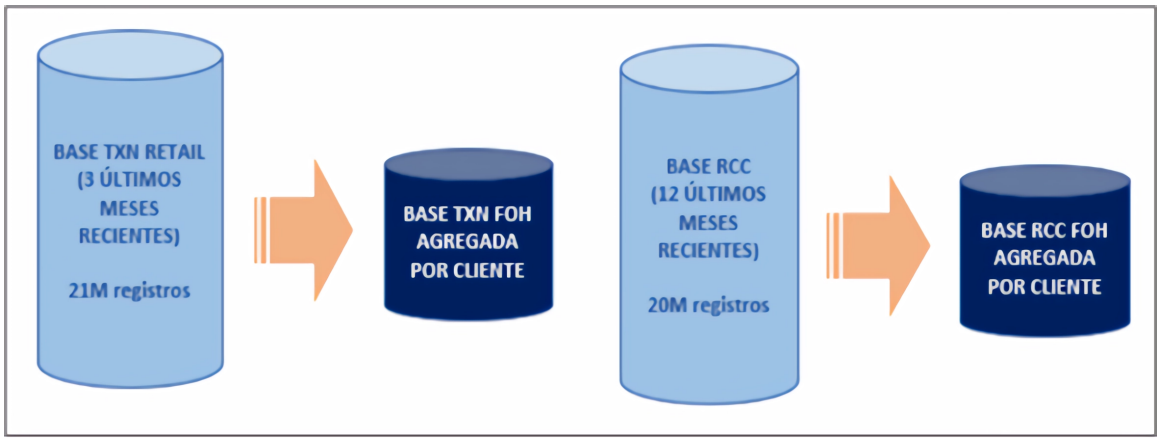

Figure 5: Data transformation

This process of the RCC data preparation is based on a series of executions of queries to obtain the credit information by client. This is because the RCC information chart keeps the information of the document number, while the RCC table detailed keeps the information of the ledger accounts that indicate to which financial product belongs to. In the research it was interesting to know the balances regarding the products of effective credits in other financial entities. That is why a preparation query is done, unifying both chats by the debtor code. The procedure is done in a SQL data server as shown in FIGURE -5.

The main database of the effective credit campaign consists of 106 variables, which are normalized and balanced, of which the fields of table RCC (type-document-identity, identity_document, debtor_code_sbs, first_name, last_name, year, month) and table RCC_DETALLE (debtor_code_sbs, first_name, last_name, year, month) are considered, identity_document, debtor_sbs_code, first_name, father_surname, year, month) and Table RCC_DETALLE (debtor_sbs_code, company_code, year, month, account_account, credit_sbs_type, account_account, credit_sbs_type, amount).

The fields of the RCC table, the first one was used to have the information of the client's document number and thus cross it with the data of the campaign and the second table RCC_DETALLE is where all the information of the credits that the clients had is stored, these tables were crossed by the field codigo_deudor_sbs. In this last table there was the accounting account field and type_credit_sbs, with which the different credit products that the client could have were obtained, among loans, mortgage loans, etc. The python code is shown in FIGURE-6.

```
--tiempo: 24.39 segundos
--drop table jlopez.base_lp_sbs_202010_2;
create table jlopez.base_lp_sbs_202010_2 as
select a.*,b.codigo_deudor_sbs,b.periodo_rcc,b.nombre_rcc,b.ape_paterno_rcc
      from jlopez.base_lp_sbs_202010  a
left join ( select distinct  tipo_documento_identidad,lpad(documento_identidad,12,'0')num_documento ,codigo_deudor_sbs ,primer_nombre nombre_rcc,
            apellido_paterno_razon_social ape_paterno_rcc, max(anio||lpad(mes,2,'0')) periodo_rcc
            from ESV_SBS.Rcc
            where tipo_persona='1' -- 1: persona natural , 2: persona juridica
            and anio=2020 and mes=08
            group by tipo_documento_identidad,documento_identidad,codigo_deudor_sbs,primer_nombre,apellido_paterno_razon_social
          )  b on a.tip_documento=b.tipo_documento_identidad and a.num_documento = b.num_documento;
-- select count(1),count(distinct(num_documento)) as cuenta from jlopez.base_lp_sbs_202010_2;
```

Figure 6: Query raw information

4. Modeling phase:

FIGURE -7 shows the facts and dimensions model is presented. In this model it is visualized the main database, which forms the campaigns database and relates with the other databases that have important information for knowing the significant variables that will help the prediction.

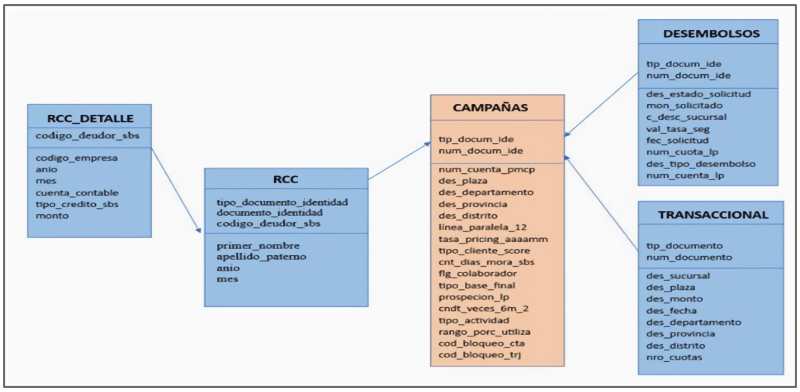

Figure 7: Facts model and dimentions

Additionally, in the construction of the model, the following components were used: Anaconda Navigator, Jupyter and Python with LightGBM model.

During the construction of the model, the training model was built and after the model was trained, the execution to get the desirable results proceeded.

5. Testing phase

Table 2: Results prediction

	No Toma	Toma	Total	
No Toma	911,936	328,886	1,240,822	73.5%
Toma	28,217	46,762	74,979	62.4%
Real	1,041,480	375,649	1,417,129	26.5%
	87.6%	12.4%	5.3%	

The testing results are shown in the Table 2 . This test was done between April 23th and July 23th. The result of the variables was:

- Exactitud(accuracy) = 73.5%
- Precision = 12%
- Recall (sensibilidad) = 62%
- F1 score = 21%
- AUC = 0.7734

6. Deployment phase

The additional placement that brought by the monthly deployment was 34 million soles on average.

The additional amount by Crédito Efectivo's product meant an increase on the product as shown in Table 3. For this reason, by having a database to exploit and having the necessary resources for classifying clients, an effective communication campaign was sent in order to get to the aimed public.

Table 3: Additional colocation calculus

	Promedio anterior modelo	Promedio posterior modelo
No desembolso	292,219	299,227
Desembolso	9,236	13,273
Total	301,455	312,400
%Tr	3.1%	4.2%

Promedio Linea	9,200
Base	312,400

Diferencia %Tr			Colocación adicion:
1.2%	3,702.08	9,200.00	34,059,168.59
	%Tr x Ticket	Ticket	

3. RESULTS

After the application of the new propensity model, the following results are presented. A control group was included, which was a percentage of the total amount of clients with Crédito Efectivo's campaign active, this was done to secure the correct execution and thus, the correct functionality of the new model. Also, to compare the results in a way that could be the correct execution of the model could be verified and the current results could be compared with the results from the last model. This is detailed bellow.

Table 4: Response rates in the testing period.

flag_y	Periodo				
	201904	201905	201906	201907	Total
No desembolso	297,668	298,548	296,517	300,574	1,193,306
Desembolso	13,187	14,076	13,954	15,076	56,293
Total	310,855	312,624	310,471	315,650	1,249,599
%Tr	4.2%	4.5%	4.5%	4.8%	4.5%

As seen in Table 4, the conversion rate increased on 4.2% on average when before was 3.1%, therefore, the additional colocation that the increase in the conversion can be calculated, which was 34 million soles on average.

This additional amount by Crédito Efectivo's product meant an increase on the product as shown in Table 3. For this reason, by having a database to exploit and having the essential resources for client classification, an efficient communication campaign was sent to reach the aimed public.

These campaigns were divided in text messages, emailing, IVR, shipping by receipts, radio, mailings and was kept in a Call Center area to be able to get in contact with the clients and offer the credit. With the application of the new model, it was more feasible to direct Crédito Efectivo's product campaigns and focus on those clients that had a high probability to withdraw the researched product and give them as clients the best options to get the credit and thus, fulfill the objectives already proposed. Table 5 shows the precision results of the models.

Table 5: Classification model table

Autor	Año	Método	Accuracy	Aplicación	Tipo de Data
Gross [14]	2018	XGBoost	79%	Clasificación	Valores
Zarabia [19]	2020	Árboles de Decisión	71.4%	Clasificación	Valores
Céspedes [15]	2022	Regresión Logística	71.1%	Clasificación	Valores

4. DISCUSIONS

The implementation of the model and the probability obtained by the client is a tool that allowed to enhance the realization of a correct analysis of the clients' behavior, this led to increase the placement and the conversion rate of the product. As mentioned before, a previous variable that tried to achieve the same regarding being able to get to know with it the behavior of the client existed, but it was a general segmentation, in that way the actual model does not underestimate the previous model and could not be used together for a different analysis and a support on the decision taking of the researched financial entity. The authors' opinion on this research this is an important aspect and characteristic in the propensity models that predict the client's behavior towards the product.

Among the researches that have a similar objective, all of them look for knowing the probability that clients can access to purchase a financial product between the most known are credit cards, Extralínea or financial necessities. It was noticed that the most used models are: Random Forest, Logistic Regression, Neural Networks, Support Vector Machine and Boosting.

In the different researches that were focused on financial entities their products can be checked and they looked for a prediction in which the clients that purchased the financial product could get to be known as explained previously credit products, credit cards, etc, that are based on the execution of classification and supervised analysis models.

5. CONCLUSIONS

The explained model was established following every requirement proposed by the company and achieving the main objective. The amount represented a monthly incremental placement of 30 million for the company by month and achieved the goal that was proposed with the incentive of Crédito Efectivo's product.

Many actions were done for sending the directed campaigns based on the analysis. It was established as a result that those clients that had a propensity above 80% were considered the most important database, in which the directed campaigns as SMS and social media were focused. These clients responded very well and the estimated conversion rate was achieved and with it, the colocations needed.

The correct cleansings and treatments to the different databases helped to achieve the objective and the realization of the diverse compared tests.

The results of both models, the old one provided by the risks area and the new one, were the ones needed to identify which suggested model was more exact and had many benefits.

The immersion to the data analysis in other process in the financial entity catch the attention of the other areas. Which thanks to this project it was concluded that data is a required information source that must be exploited in the correct way and with Machine Learning models mean an added value for the company.

6. RECOMMENDATIONS

The propensity model can be extrapolated using machine learning techniques for situations with a larger number of customers.

References

[1] https://www.mapfre.com/actualidad/economia/educacion-financiera-diferencias-entre-productos-financieros/

[2] Xiao Y, Zhao Q, Zhou W. Machine Learning Techniques for Customer Retention in the Financial Industry. J Financ Anal. 2020;12:45-57.

[3] Zhou P, Cao Z. Application of Machine Learning in Fraud Detection for Financial Transactions. Fin Technol Rev. 2019;7:87-98.

[4] Chen X, Li W, Wang S. Using Propensity Models to Improve Cross-Selling Strategies in Banking. J Mark Anal. 2021;9:145-159.

[5] Chakrabart R, Roy SG. A Novel Graph Clustering Algorithm Based on Discrete-Time Quantum Random Walk. In Quantum inspired computational intelligence. Morgan Kaufmann. 2017:361-389.

[6] Li H, Sun Y. Customer Segmentation for Personalization in Financial Services: A Machine Learning Approach. Int J Banking Fin. 2021;15:112-130.

[7] Wang L, Zhang Q, Liu J. Predicting Credit Default With Decision Trees: A Case Study Using LightGBM. Financ Risk Manag J. 2020;18:77-93.

[8] https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html.

[9] Singh Chauhan N. Decision Tree Algorithm, Explained. KDnuggets. 2022.

[10] Novoa-Hernández CV, SamaniegoMena NP. ÁRboles de Decision Para la Evaluación Del Riesgo Biológico de Procesos. 2018. Available from: https://www.redalyc.org/journal/5826/582661251001/582661251001.pdf.

[11] Gan M, Pan S, Chen Y, Cheng C, Pan H, et al. Application of the Machine Learning Lightgbm Model to the Prediction of the Water Levels of the Lower Columbia River. JMSE. 2021;9.

[12] IH Sarker. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN computer science. 2021;2:160.

[13] https://www.altexsoft.com/blog/propensity-model/.

[14] AG Sánchez. Elaboración de Un Modelo de Propensión Al Crédito de Consumo Capaz de Discriminar a Nivel de Individuo Utilizando Solamente Información Financiera de Carácter Pública. 2018. Available from: https://repositorio.uc.cl/handle/11534/21891.

[15] DM Céspedes Malpartida. Modelo de Propensión a la Adquisición de Un Producto Activo O Pasivo en Una Entidad Financiera Peruana. 2022. Available from: https://cybertesis.unmsm.edu.pe/handle/20.500.12672/18655.

[16] Ibarra-Vazquez G, Ramírez-Montoya MS, Buenestado-Fernández M, Olague G. Predicting Open Education Competency Level: A Machine Learning Approach. Heliyon. 2023;9.

[17] https://www.expressanalytics.com/blog/propensity-modeling-to-predict-customer-behavior-using-machine-learning/.

[18] https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

[19] Zarabia Yupanqui C. Identificación de la Propensión a la Adquisición de Un Sub-producto de Una Tarjeta de Crédito en Una Entidad Bancaria. 2020. Available from: https://repositorio.lamolina.edu.pe/handle/20.500.12996/4834.