International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 15(2023) pp. 583-594
© MIR Labs, www.mirlabs.net/ijcisim/index.html

Submitted: 1 Jan, 2023; Accepted: 1 Jan, 2023; Publish: 26 Jul, 2023

Customer Success Analysis and Modeling in Digital Marketing

Inês César[1,2], **Ivo Pereira**[1,3,4], **Ana Madureira**[2,4,5], **Duarte Coelho**[1,4], **Miguel Â. Rebelo**[1,6] **and Daniel A. de Oliveira**[1]

[1]E-goi, Av. Menéres, 840, 4450-190 Matosinhos, Portugal
{*icesar, ipereira, dcoelho, mrebelo, dalves*}*@e-goi.com*

[2]ISEP, Polytechnic of Porto, 4249-015 Porto, Portugal
{*1171426, amd*}*@isep.ipp.pt*

[3]Universidade Fernando Pessoa, Praça 9 de Abril, 349, 4249-004 Porto, Portugal
ivopereira@ufp.edu.pt

[4]ISRC - Interdisciplinary Studies Research Center, ISEP, 4249-015 Porto, Portugal

[5]INOV - Institute for Systems and Computer Engineering, Technology and Science
1000-029 Lisboa, Portugal

[6]i3s, Rua Alfredo Allen, 208, 4200-135, Porto, Portugal

Abstract: **Digital Marketing sets a sequence of strategies responsible for maximizing the interaction between companies and their target audience. One of them, known as Customer Success, establishes long-term techniques capable of projecting the sustainable value of a given customer to a company, monitoring the indexers that translate its activities. Therefore, this paper intends to address the need to develop an innovative tool that allows the creation of a temporal knowledge base composed of the behavioral evolution of customers. The CRISP-DM model benefits the processing and modeling of data capable of generating knowledge through the application and combination of the results obtained by machine learning algorithms specialized in time series. Time Series K-Means allows the clustering and differentiation of consumers characterized by their similar habits. Through the formulation of profiles, it is possible to apply forecasting methods that predict the following trends. The proposed solution provides the understanding of time series that profile the flow of customer activity and the use of the evidenced dynamics for the future prediction of these behaviors.**

Keywords: Digital Marketing, CRM, Time Series, Machine Learning

I. Introduction

Digital Marketing focuses its efforts on arranging mechanisms aimed at business success [11, 27]. Its adoption increases proximity to various customers, achieving in-depth knowledge about their daily choices among different types of products and services they are depending on [21, 32]. A company that contextualizes all of its elements that it is a priority to constantly seek new sources of information about its customers enhances its action and prestige, specially in B2B and B2C markets. In order to build an embedded structural base with insights into deep and unlabeled customer activity patterns, companies like E-goi that customize multichannel solutions for every business routine, fully leverage the stored data to solidify their competitive position [21, 2, 4, 27]. This process can be granted with a dynamic application of innovative technologies within the field of Machine Learning [8, 9, 10, 29, 28].

This case study considers several attributes specialized in measuring the presence or absence of customer activity in the functionalities provided by the company's application. Developing the solution plan with manipulation of the phases that make up models such as or based on CRISP-DM [23, 16] allows for the generation of intermediate steps customized to meet specific objectives [12, 14]. The beginning of the process reconciles the primordial validation of typology and data integrity, speeding up the search for discrepancies caused by streaming data capture. For the treatment of the anomalies found, data manipulation techniques [12, 14] are applied, where an analysis is performed to the methods applicable to the lack of information, with replacement or prediction by regression, or the existence of irregular information, by detecting outliers. The stability and consistency in the extracted set of observations expedite the time series framework without compromising temporal dependence [19, 24]. However, the size and quantity of time series per client generate a problem concerning the complexity of the entire solution, requiring a deeper exploratory update.

The subsequent requirements pursue some arrangements which are required for the creation of just one set of temporally equivalent observations for each client [25]. Thus, the generation of representative groups of similar behavioral patterns is achieved by the aggregation of the [34, 35, 5] information. Considering the aspects that characterize the performance of an algorithm such as Time Series K-Means, it is admitted in accordance with parameters that maximize the performance and the reliability of the segments obtained. These, in particular, respect the specificity in the definition of a metric capable of analyzing the similarity between two samples [3]. These parameters are selected using different supporting methods that allow the decisions made and, consequently, the results generated to be substantiated. The importance of controlling all the steps makes it possible to use the centroids of each admitted group as a solidified time series to be predicted using tools such as Facebook Prophet.

The organization of this paper is as follows: section II marks some background knowledge about related work developed, section III describes the problem domain through artifacts, and in section IV the data characteristics are enumerated. In section V, the methodology is explained and in section VI the results are presented, followed by a discussion. Finally, in section VII, the conclusions.

II. Related Work

It is with great prospection that a particular organization seeks to maximize the compatibility between the services they provide and the trends most requested by customers of the target audience they explore. Nowadays, following the consumer and the traces of his presence in the market represents one of the most relevant and conditioning responsibilities for true commercial success. It is, therefore, required that the formulation of promotional actions responds positively to the needs of the various customers. This process, composed of an analysis of several registers that represent the level of interaction, raises its complexity in the monitoring of all factors with significant influence on the relationship with the consumer in the long term. Considering not only its preservation but also its understanding and provisioning, the application of machine learning in digital marketing has revealed a long path enriched by the contributions exercised in different and challenging application areas [6].

Tekin et al. [33] detail several possible approaches to applying data mining that, through the documented concepts, provides the necessary steps for studying and evaluating the impact of cybernauts' digital marketing. His work effects as a standard benefit the volume of data produced by each user and its great proportions are evidenced in the efficiencies obtained by the results of the scientific contributions with which he exemplifies his arguments. The supervision of entities and the continuous tracking of their values, such as demand and supply between consumer and company, ensure the extraction of performance metrics and customer value for segmentation of priorities and updates of the implemented strategy. The areas of use of this knowledge enrich health and economics in addition to science.

Nethravathi et al. [26] present findings on how consumer behaviors are energized by knowledge of unique characteristics such as their hobbies. Focused on enriching the business with the application of business intelligence, they use the information regarding the pattern of purchases made and correlate it with the different hobbies surveyed. Through the development of a genetic algorithm, they give rise to populations as the new knowledge base to be analyzed to predict the behaviors of new customers. The promising results substantiate the importance of detail in studying the behavioral patterns of the target audience's activities.

Shah et al. [31] analyze how the challenges present in stock market prognosis and classification present and how to study and predict techniques encourage the use of various analytical methods where statistical and machine learning models gain popularity. Added to these, the authors raise in detail the studies where these techniques are positively supporting a viable alternative for application to the stock market. Still, the conclusions assume great uncertainty about the forecast's true behavior, considering the instability of the values is strongly related to today's political, economic, and social events.

Table 1 summarizes the study of some of the existing work related to the nature of the problem on which this paper is based. In this way, a set of characteristics is identified and analyzed that highlight the structural similarities in each case, accelerating the perception of the different alternatives and consequent reliability in the results obtained in the conclusions. The comparative is conducted having as foundations the strategic models on which they are based, the chosen approach and the relevant attributes of the available data, and the auxiliary decision methods that allow the analysis and validation of the machine learning models applied.

Maryani et al. [22] scientific contribution presents a solution for determining customer performance that benefits business improvement. With the adjustment and updating of time-based data to the problem, some still in physical format, a dataset was built with transactional records of customers simplified by the variables advocated by the RFM model. The adjustment of this information is justified by the search to segment, classify, and profile the different consumers in groups categorized by their interactions, forming guidelines for the entire corporation on maintaining the relationship with the customer. The segmentation is developed using the K-Means method and tested by the Davies Bouldien index to validate the performance of previously obtained results. The acquired data, composed of the groups to which they belong, are submitted to a tree model as a classification of the characteristics of each differentiated customer in each cluster. Consumer profiling occurs according to the determination of the patterns of each consumer: these are analyzed by the results previously obtained and labeled according to the Grid Hill theory.

Taking the case of the retail industry, Dogan et al. [13] define a study focused on the application of segmentation methods to deepen their understanding of the existing categorization process. The addition of the RFM model variables to the dataset considered and the application of segmentation methods such as Two-step and K-Means provide promising alternatives for updating marketing strategies with the target audience. While the former highlights discrepancies in customer classification according to the loyalty program previously adopted, the latter presents a new perspec-

Table 1: Customer Analysis' Related Work

Ref.	Strategy Models	Approach	Attributes	Auxiliary Decision Methods	ML Models
[22]	CRM	CRISP-DM	RFM	Davies Bouldin, Grid Hill Theory	K-Means, Decision Tree
[13]	Non Defined	CRISP-DM for Time Series	RFM	Elbow	K-Means
[17]	Non Defined	Customized	RFM	Elbow, Silhouette, Calinski-Harabasz, Davies Bouldin, Ratkowsky, Hubert, Ball-Hall, Krzanowski-Lai, R ratio	K-Means
[1]	Non Defined	Aggregated and Specialized Time Series	RFM	Silhouette, RSME, SMAPE	Time Series Clustering, ARIMA
[18]	CRM	CRISP-DM	RFM, ALC	Elbow	K-Means, Naive Bayes, Decision Tree
[36]	CRM	CRISP-DM	LRFMM	Non Defined	K-Means, DBSCAN, Hierarchical, IK-Means-+
[20]	Non Defined	CRISP-DM	Non Defined	Non Defined	K-Means, Decision Tree, Logistic Regression, Neural Networks

tive of the categorical organization where it emphasizes promotional actions and other strategic adjustments. Thus, Machine learning models enhance the need to keep this analytical process of their behaviors up to date.

Equally, Gustriansyah et al. [17] also thoroughly explore the effectiveness and performance of the method of differentiating patients from a given pharmacy by applying a series of validation indices to the process. The initial dataset, consisting of sales made over 12 months, was processed into variables with custom value ranges supported by the RFM model: recency represents the total number of days, frequency translates to the number of times that product was sold, and monetary the total amount spent on a single product by a customer. The versatility of this technique allowed the construction of a more intuitive dataset to segment customers and create optimizations to the establishment consistent with the annual dynamics. The use of the most suitable group number evaluation metrics, described in Table 1, allows for complementary success in finding similarities and differences between individuals. The conclusions obtained consider the stock management measures as one of the main results obtained for the service to the public according to the observed needs.

Focused on the identification of possible approaches for the interpretation of time series, Abbasimehr et al. [1] investigate new proposals where it is considered conditioning not only the variables used in the time series but also the context where the situation is inserted. To this end, several transaction observations from different areas are applied to a collection of machine learning models as reinforcement that the results obtained are different depending on the final objective. The dataset, composed of several variables, is classified using Laplace's method to select the most relevant characteristics to obtain results regarding consumer dynamics. This way, the methodology ensures the application of dedicated machine learning models in the segmentation and prediction of the time series representing the customers with the purpose of concluding improved performances caused by the precautions throughout the process.

Hartini et al. [18], from the practical case of the vast network of cosmetics businesses located in Indonesia, report the development of a consumer analytics solution to provide recommendations for the company. These, in turn, are derived from the results obtained through differentiating methods for profiling different individuals. However, and in similarity to other related works, despite referring to the adoption of the RFM model, it was carried out the conjugation with variables formatted according to the ACL (age, location, cell phone operator), believed to be a more complete analysis of the characteristics of the sample population. The data mining techniques used allowed the selection of the most complete sample portion for the recognition of the best number of groups to be defined with the elbow method. The results obtained are presented according to the combination of an unsupervised learning model with another supervised learning model, focusing on the intention of obtaining classification of the customers in each group according to the divergence of the ACL variables.

With the purpose of providing enrichment to the traditionally performed segmentation results, Zare et al. [36] propose a customized approach to the customer analytical process through behavioral variables related to activity and satisfaction with the service. This empowers the models with features that condition the results positively as well as negatively, which in the authors' view deserves special attention for the interpretation of segments where this information is not considered at all. Accordingly, they structure the research methodology in two phases: the preparation and transformation of data, where ideal parameters are generated and scaled to characterize a customer, and the classification of the total set of consumers in agreement with the most similar group they belong to and the evaluation of the observations evidenced by the models. Performance comparison is given by multiple testing with more than one machine learning model, enhanced by the previously stated guidelines for the final calculation of each individual's life cycle value. The results show strong benefits in using a more contextualized model in the particular situation, illustrating the simplicity in which the implemented changes show promising conclusions.

Meanwhile, Jamjoom [20] refers to the benefits of using knowledge extraction techniques to predict churn situations for customers who identify themselves as other companies. The compilation of information from insurance company databases demonstrates in which formats the predictive model design should be based for the application of methods and models that allow profiling, classification, and prediction of future consumer behavior. Using CRISP-DM, an initial study was conducted with decision trees for the classification and subsequent selection of the most relevant variables. Therefore, the comparison between the results obtained enunciates the distribution of the training set as one of the crucial factors, without affecting the effectiveness of the information obtained and implicit in the new changes made by the marketing department.

The related works previously compared and described are some of the subjects with relevance for the structuring of one or several action plans for further learning of the environment favorable to the prosperity of a business. Simultaneously, their components solidify the formula of the business continuity study based on the definitions and the most advantageous steps in adapting the reality. The similarity of the process motivated E Ernawati et al. [15] in building a model capable of analyzing and detecting the data patterns and key elements in the instruction about customers. Considering their characteristics, customs, and behaviors translated by the simplicity of the RFM model, the composition of the model intends to generalize all the possibilities of obtaining knowledge. In addition to exposing the styles and how they are found in a real context, there is the particularity of enumerating the number of projects found in the area, highlighting forecasting as the fewest. This fact allows greater potential for the development of new discoveries in the manipulation of these tools for the sake of mastering the near future.

III. Problem Domain

E-goi is a Portuguese company that provides automated multichannel marketing services through its digital platform. Positively impacting the success of entities inserted in B2B and B2C markets, it is responsible for maintaining and aggregating data that enables the continuity of strategic relationships with its customers.

Daily it registers a significantly high amount of information regarding its customers and the activity performed with the use of the platform's functionalities. The knowledge obtained allows the generation of the demographic, behavioral, and social history of each of the consumers who choose the platform as a complementary and customized tool to their own reality. This, in turn, presents a wide choice of lines of action functionally accessible to the user, enabling successful paths and resolutions in Fig. 1. Consequently, it becomes indispensable to analyze the information stored about customers' temporal choices and the dynamic flow of the platform. In order to highlight similarities and patterns between the performances of each, so far unknown, it is expected the steps towards the formulation of a model capable of differentiating and evaluating the factors conditioning the success of the users of the sample.

Given the considerable volume of elements that populate the company's databases, visual and manual diagnosis entails obstacles that easily mislead the employee responsible for customer service, with repercussions later on in the bond. The need to obtain an intelligent machine capable of managing and visualizing all traces of customers' temporal performance in a grouped manner highlights the reasons that triggered the proposed problem. Fig. 2 intends to present a domain model composed of the relevant entities to the situation and the connections established between them.

Table 2: Dataset characterization

Variable(s)	Type(s)	Description
Client Identifier	Numerical	Client Identification Number
Representative	Numerical, Categorical	Specific perspective inside the company
Contractual	Numerical, Categorical	Full available resources for the client
RFM	Numerical	Based on RFM model
Activities	Numerical	Presence or absence on resources usage

E-goi, which provides services to a Customer, is directly responsible for aggregating information about this agreement, generating a record of the Activity. The effect of the accumulation of activity records requires E-goi, through its Employees, to request a recurrent Analysis to understand the relationship formed.

IV. Dataset Analysis

Planning the solution requires prior structured knowledge of the dataset provided. For this particular case, there are more than 98 million lines between 2019 and 2022 that make up the company's database filled with numeric and categorical values. Differing only in content and size, they support continuing the practice of good customs in customer relations by allowing customers to be summarized by their activities. As much as possible, the building of the knowledge bases should agree with the streamlined marketing strategy as in the reproduction of procedures that reflect on customer interactions as positive or negative.

Table 2 provides a brief description of the categories in which the data are classified, explaining the type of information that can be consulted. Its configuration is mostly numeric, with only a few categorical elements, which allows the proper manipulation and processing of the variables without having to perform too many transformations on the data for testing. In addition to the identification items, it is possible to evidence the timestamp of the updates, the location of the client, and the activity performed in the use of the platform's resources. These properties are of great importance in assigning representative and contractual categories that are as similar as possible to the actual usage of the client in question. Complying with this conduct, the company's previous adoption of the RFM model provides a summary of these properties.

By means of a bar chart, Fig. 3 shows the distribution of customers by the different CRM categories. It adapts the understanding of one of the categorical variables where it presents the number of occurrences of each CRM state for each customer. These are obtained according to the number of significantly relevant actions for the company, intensifying the distinct advantage with the increase of hours dedicated to the success of each asset. This type of data usually requires modeling to initialize its transformation without losing information of great interest.

Considering Fig. 4, a set of variables is presented that daily update their value mirroring the presence or absence of consumers in different communication channels. The construction of the diagram refers to the singular study of each item, isolating the distribution range of the values it acquires and

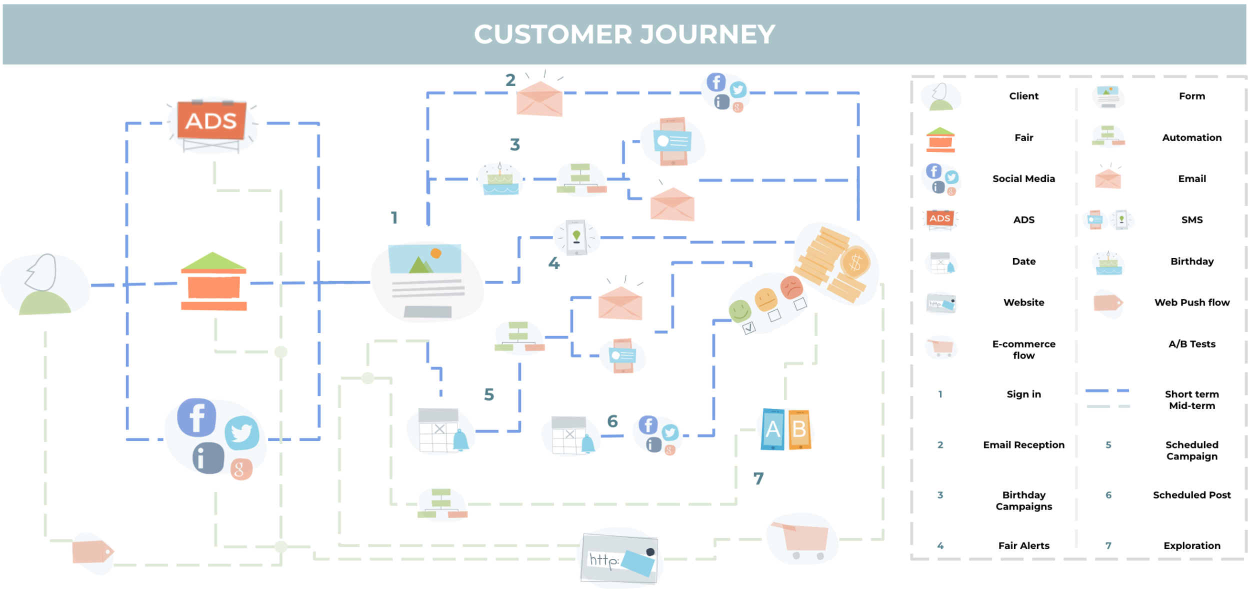

Fig. 1: Customer Journey

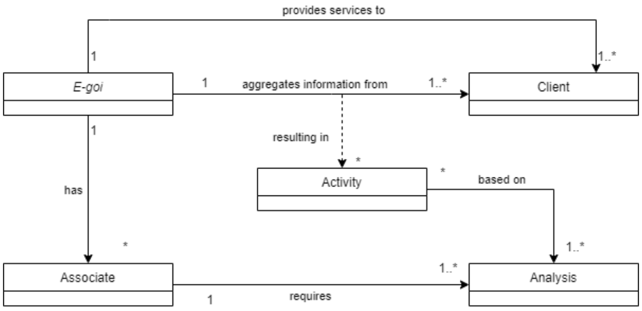

Fig. 2: Domain Model Diagram

Table 3: Milestones of CRISP-DM

Step	**Description**
Business Understanding	Acquisition of detailed knowledge about the business reality and the available resources
Data Understanding	Acquisition of detailed knowledge of the data's foundations, structure, and meaning
Data Preparation	Selection of the information based on the inclusion or exclusion criteria
Modeling	Element modeling for the application of techniques to model formulation
Evaluation	Validation between the results obtained and the objectives set for the formulation of conclusions
Deployment	Implementation of the assessed conclusions

resorting to statistics such as median and quartiles for the design of each boxplot. Thus, all values that are outside the outlined boxes are considered outliers, requiring mostly flexible and assertive modeling.

Fig. 5 states the distribution of numerical values obtained in a given variable responsible for capturing the use of a platform feature by the platform's direct stakeholders. However, the logic associated with this property is inverted, translating the number of days that have passed since the last use. That is, the time series is monitored by assigning the value 0 whenever the presence of the user when using the specific functionality is verified.

This characterization built a deeper understanding of the dynamics of the observations and represents one of the main steps to initialize the processing and modeling of the information. To recognize which can be transformed into time series, the alternative of working with more than 40 variables is discarded and tests are formulated to select the most important ones with feature engineering techniques [7].

V. Methodology

The proposed methodology for evaluating a client's success through data is built using the guidelines of the CRISP-DM model. Referred to as one of the most complete approaches, it is composed of steps independent of the reality of the problem and benefits from the comprehensive study of the observations obtained during the process. Its systematic application by the scientific community is supported by the success obtained by those who adopt it and by the flexibility inherent to its practice [30].

For a better understanding of what is expected, Table 3 describes the general ideas regarding each phase Next, it will be explained what was developed in each of the stages of the model for the solution substantiated by this document, supported by artifacts that represent the procedural actions.

A. *Business Understanding*

The first stage was carried out gradually with the initial integration with the platform and the monitoring of all the possibilities through the content made available and carried out by the company itself. The perception of the set of data acquired and stored allowed the meeting of the main goals to

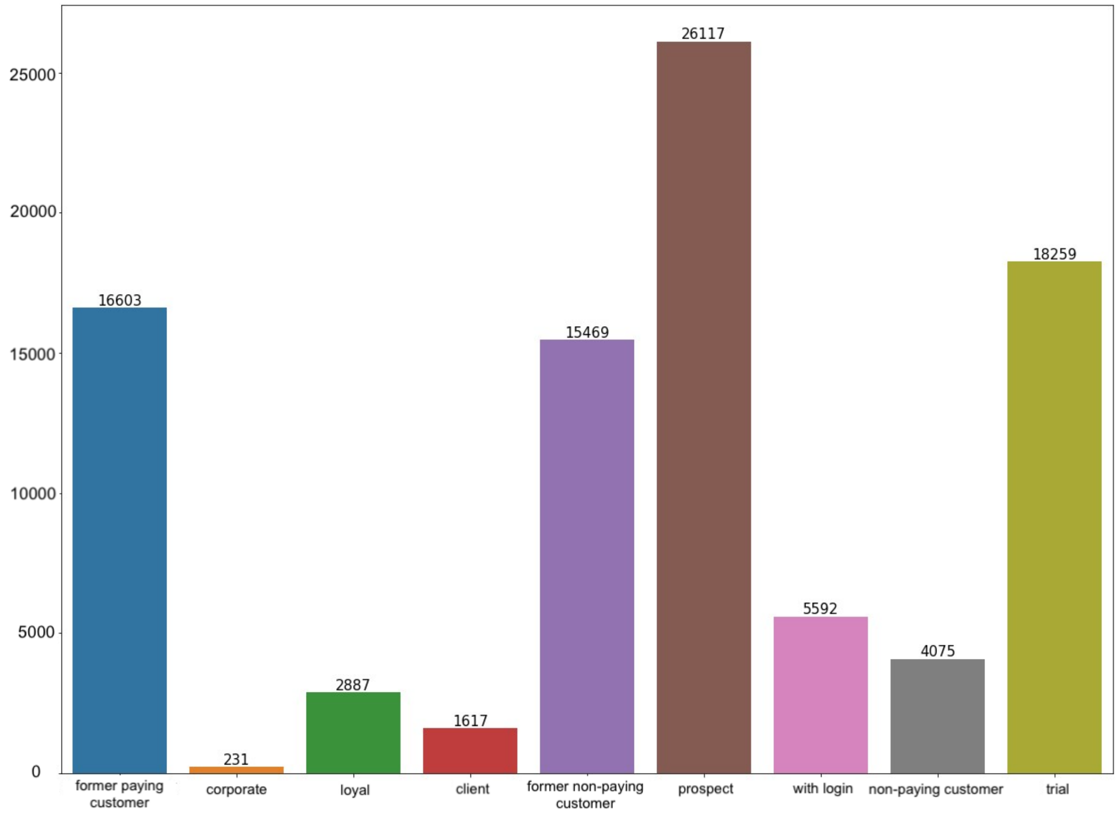

Fig. 3: Plot of the different CRM status

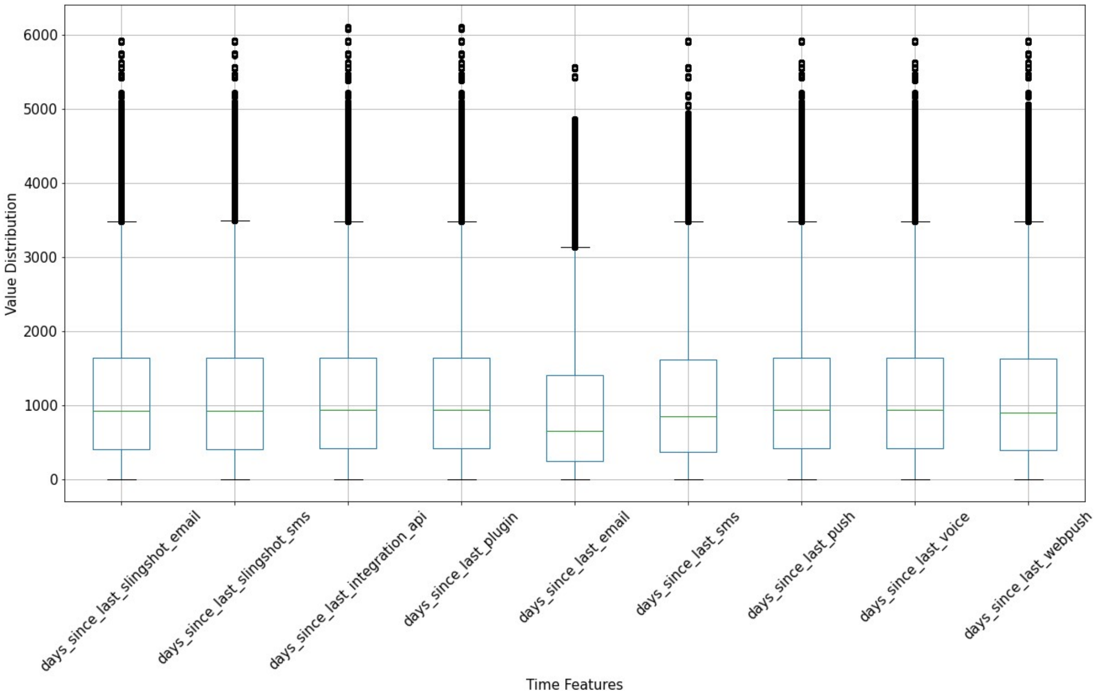

Fig. 4: Boxplot of time features

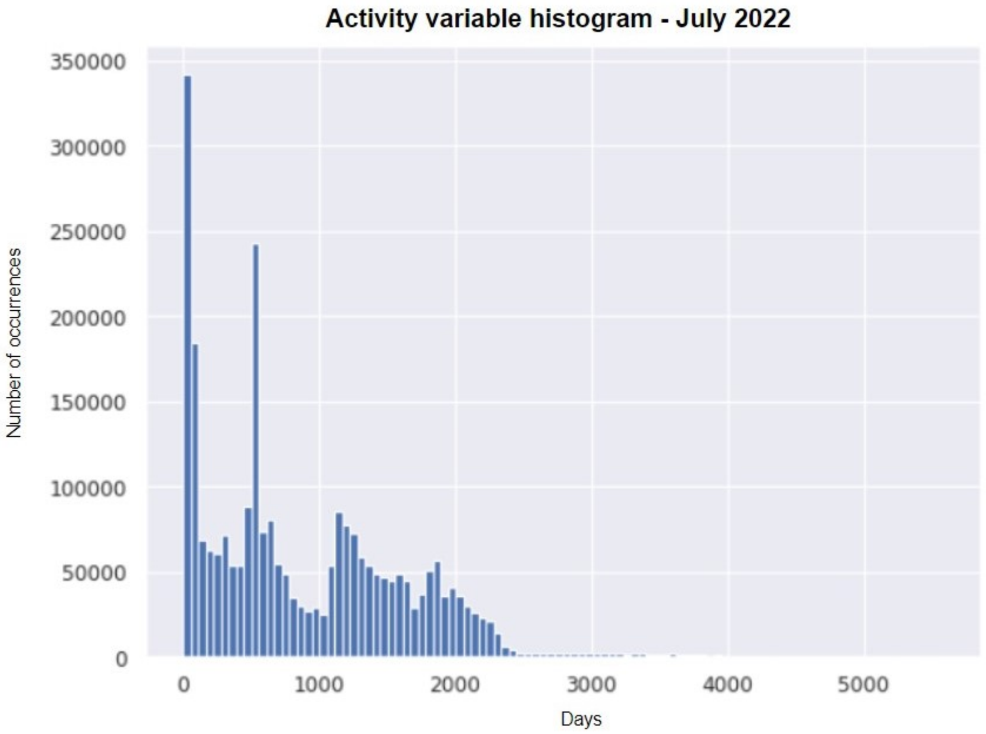

Fig. 5: Histogram of an activity variable

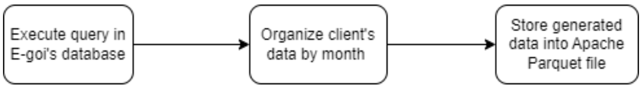
Execute query in E-goi's database
Organize client's data by month
Store generated data into Apache Parquet file

Fig. 6: Process Diagram of CRISP-DM phase *Data Understanding*

be achieved for the innovation of the business. The issues raised reveal a high priority in obtaining a means of in-depth analysis in the various time series of the client's activity.

B. Data Understanding

The understanding of the data is started by accessing information structures in the company's databases to extract sample portions for the first analysis. This step allows us to conclude the magnitude of the records entered daily and to choose the most appropriate type of storage files so that the time interval corresponding to one month would be correctly stored. In Apache Parquet format, the mapping of the values in each column is faithful to any dimension, speeding up the practice of analysis and modeling depending on the format of the information for a comprehensive knowledge base of inherited facts. Fig. 6 exposes the processes required for data extraction.

Fig. 7 was built in order to highlight the customization performed as intermediate processes integrated with the foundational steps of the CRISP-DM model, respectively presented in the image legend. In each category, a presentation of artifacts will be made where the tasks developed will be individually specified.

C. Data Preparation

The data preparation stage, as the CRISP-DM model indicates, allows for combination with the next stage for the purpose of thoroughly analyzing the properties of the variables under study. Due to that, several analyses are performed for completeness and the presence of anomalies with a view to grounding the data to apply to the chosen models. It is essential to optimize the gathered information to increase confidence in results and predict several failure points where data can be the main issue.

The exploratory analysis of the data is depicted in diagram 8 and shows the implicit dependence of deductive reasoning between previously obtained answers on the format and character of the data. It also makes perceptible the rules created for the formulation of a concise and feasible dataset.

D. Modeling

The modeling phase, combined with data preparation, leads to obtaining a broader knowledge base through the abstraction and flexibility of the process initiated. It is possible to achieve reliable results based on the details that make up the problem since they are all conditioning factors in the choice and performance of the algorithms.

In Fig. 9, it is documented how the formulation of the enhanced time series for each customer is done considering a large number of time series intended for each feature made available. The information transformation process compacts the user presence into percentage metrics equally spaced every 15 days for the development of a time series capable of representing a greater total range of behaviors performed as a consumer of the services made available. The normal-

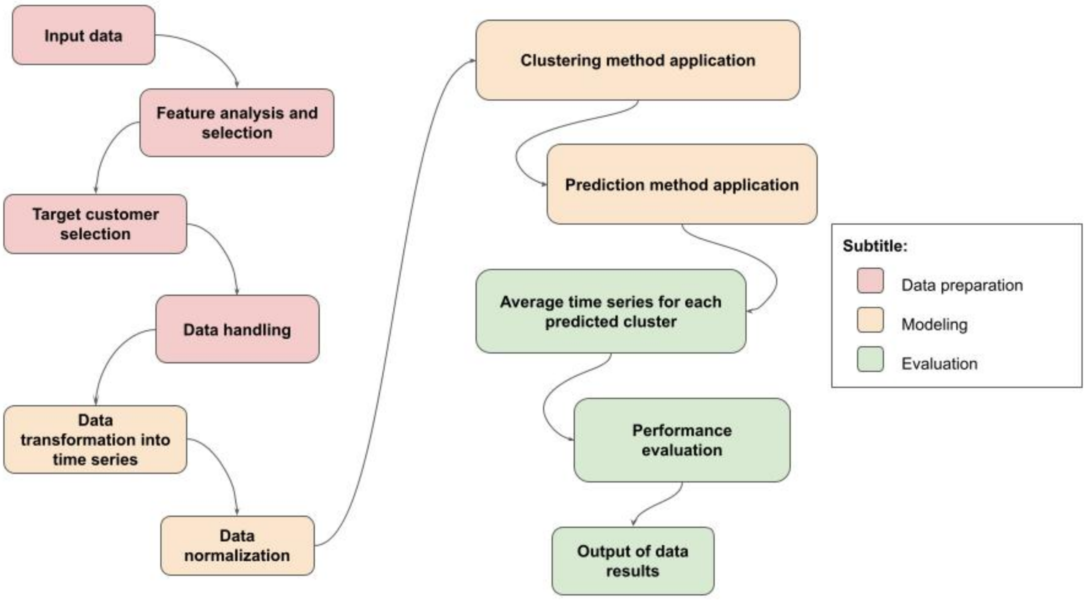
Input data
Feature analysis and selection
Target customer selection
Data handling
Data transformation into time series
Data normalization
Clustering method application
Prediction method application
Average time series for each predicted cluster
Performance evaluation
Output of data results
Subtitle:
Data preparation
Modeling
Evaluation

Fig. 7: Proposed methodology

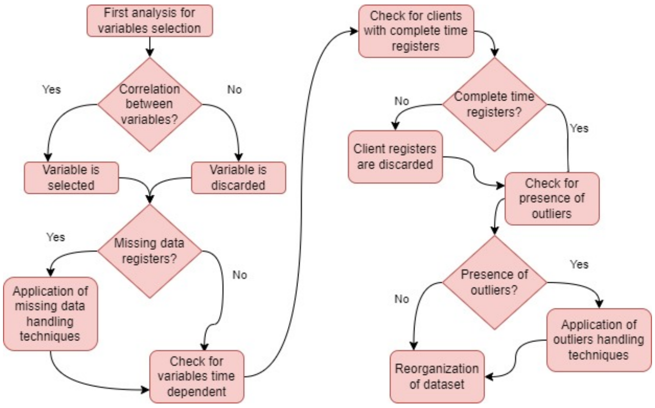
First analysis for variables selection
Correlation between variables?
Yes
No
Variable is selected
Variable is discarded
Missing data registers?
Yes
No
Application of missing data handling techniques
Check for variables time dependent
Check for clients with complete time registers
Complete time registers?
No
Yes
Client registers are discarded
Check for presence of outliers
Presence of outliers?
No
Yes
Application of outliers handling techniques
Reorganization of dataset

Fig. 8: Process Diagram of CRISP-DM phase *Data Preparation*

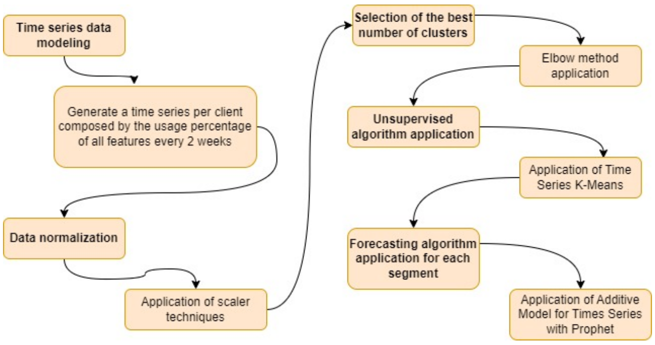
Time series data modeling
Generate a time series per client composed by the usage percentage of all features every 2 weeks
Data normalization
Application of scaler techniques
Selection of the best number of clusters
Elbow method application
Unsupervised algorithm application
Application of Time Series K-Means
Forecasting algorithm application for each segment
Application of Additive Model for Times Series with Prophet

Fig. 9: Process Diagram of CRISP-DM phase *Modeling*

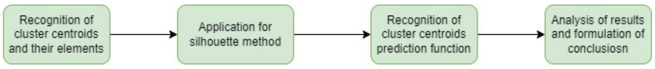
Recognition of cluster centroids and their elements
Application for silhouette method
Recognition of cluster centroids prediction function
Analysis of results and formulation of conclusiosn

Fig. 10: Process Diagram of CRISP-DM phase *Evaluation*

ization of the data, recommended in this type of problem, used standardized distribution techniques for the application of the Elbow method as one of the indexes supporting the decision of the number of segments to define. The centers marked in each group are obtained by applying an unsupervised learning model to obtain unknown details of the sample and, later, applied to a predictive method to generate time intervals promising future activities of the different sets.

E. Evaluation

The evaluation of the solution is the step responsible for the full validation of each developed portion that satisfies the requirements to evaluate the customers and define their current state of success and the range of values that define the possible future variations of a set of customers whose behavioral patterns are similar within at least 6 months. Thus, the recognition of the segments obtained and the validation of their similarities through the application of methods and the contextualization of graphical evidence represent some of the assessments recommended for the solution developed.

F. Deployment

The implementation aims to link all the previously identified phases so that the progress of learning and obtaining information is in line with the methodology. The presentation of these details is directly conducted to the employees

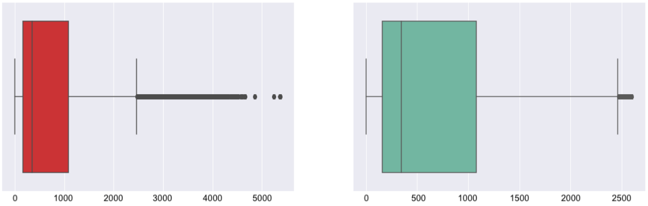

Fig. 11: Pair of boxplots of the winsorization effect on data outliers

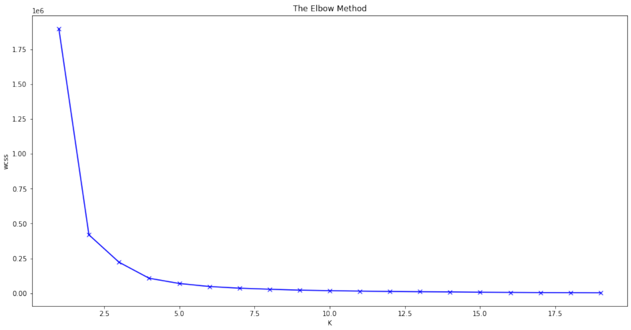

Fig. 12: Results of Elbow Method with WCSS

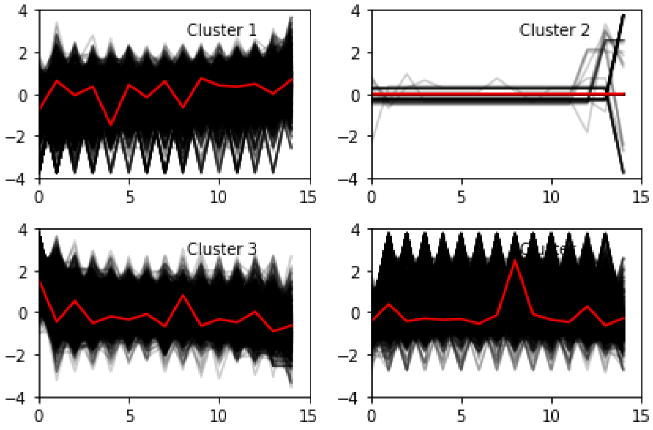

Fig. 13: Results of Time Series K-Means with Dynamic Time Warping

that monitor the assets every day and are available to answer questions or difficulties that may arise, and to those in higher hierarchies, responsible for the provisioning of the services developed by the company.

VI. Results and Discussion

The investigation of data that define different interaction flows needs to be done under special conditions. The first one refers to the fact that the data capture is done only once every day, promoting the increase of anomalous situations that compromise the entire solution. The winsorization, shown in Fig. 11, is an example of the data manipulation technique where it highlights the effect of a percentage cut equal to 1. The identification of the portion of data to be discarded follows according to the study of statistical operations performed that focus on the treatment of these discrepancies. Its development has an impact on the following operations because some of the values remain valuable.

Taking into account the integrity of the temporal component, there is a need to restore the missing data. To this end, two regression models compatible with the prediction of multiple variables without discarding the relationship established between the evolution of each one were developed. The motivation for using Machine Learning models is later compared with the evaluation of their performance, by calculating the mean square error and its standard deviation. The coefficients do not coincide with optimal values, limiting their proper application and forcing the use of other techniques.

As a method of validating the best number of clusters, parameterized as a prerequisite for performing the segmentation, the Elbow method was used. The illustration of the obtained result is visible in Fig. 12, where the function defined by the line is constructed by summing the distances from a given point to the correspondents of the same cluster and, consequently, to the center of this segment. In other words, the greater the result of the sum, the worse the differentiation performance of the elements of the sample. Therefore, as its name indicates, it identifies the best value by defining a break in the line of its graph even before it stabilizes. The solution relied on the development of tests to the different values that K could acquire, and K = 4 was the number with the highest success.

The choice of four segments followed as the most convenient, represented by the effect of the Time Series K-Means algorithm in formulating the different benchmarks in Fig. 13 and the time series that justify the centers of each group. To substantiate this division, the distance metric was previously formulated, seeking to be validated by the results of the Silhouette coefficient that conveys the stability of each recommended group. The one with the greatest use in the results obtained was the DTW, a measure algorithm capable of calculating the shortest distance between two sets of registers while respecting the temporal dynamic of the observations.

Time series components that demonstrate the tendency, seasonality, and cycle standards are key factors for the following process of predicting behaviors followed by previous ones. One way of granting this type of prediction without getting to focus on decomposing and analyzing one by one, Facebook Prophet provides an innovative tool with reliable results in time series where seeking future values would be inconclusive or without essential elements to verify its trustworthiness. Figures 14, 15, 16 and 17 instantiate graphical elements capable of understanding the predisposition of the future engagement the first cluster clients may acquire. The black points represent the observations that make up the average time series for the specific segment, defining the slope of the equation of the temporal function described. United with the translucent blue section, it concludes a positive range of future activity pattern values, predicting success for this group of clients.

Regarding the segments predicted in Fig. 14 and Fig. 15,

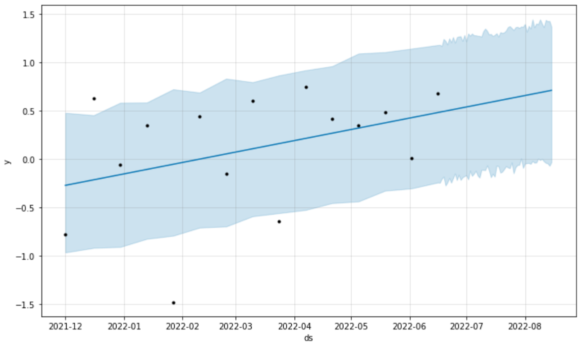

Fig. 14: Results of Prophet prediction for the first cluster

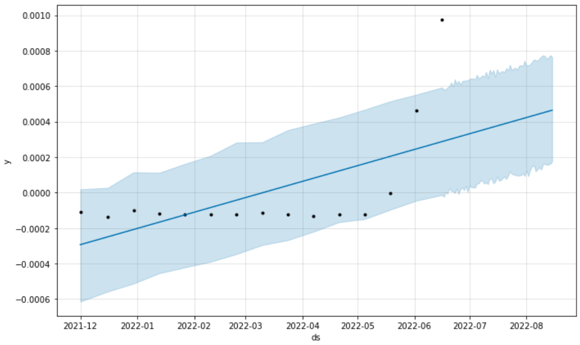

Fig. 15: Results of Prophet prediction for the second cluster

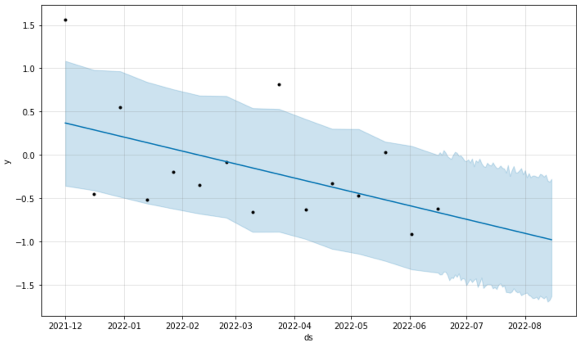

Fig. 16: Results of Prophet prediction for the third cluster

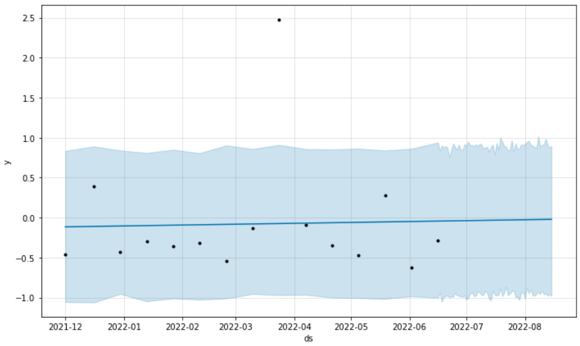

Fig. 17: Results of Prophet prediction for the fourth cluster

although with differences in the variation intervals, these are customer clusters that show positive behavioral patterns. The function drawn to help forecast the next 30 days presents a positive inclination, identifying both clusters as being representative of customers loyal to the use of the platform and customers whose interest in the functionalities made available created a gradually positive dynamic, respectively.

The graph in Fig. 16 shows disparities between the results observed so far, revealing a negative slope in the predictive function built by the time series points at the center of this segment. It is possible to draw conclusions about the lack of activity by this set of customers, giving rise to arguments that support the progressive abandonment of these subjects. This sort of visual content cooperates with the elaboration of strategies to reverse these situations and to recover the success of customers before they have even ceased to contract.

Finally, analyzing the results of the fourth segment in Fig. 17, we can consider that the slope of the function, although not very accentuated, is classified as positive. This type of temporal evaluation indicates the practice of constant behaviors, identifying consumers with reduced knowledge in the exploitation of the platform's utilities or that bases the exclusive use of a significant amount of functionalities. The formulation of marketing practices that operate in the evolution of this pattern should be instated considering the implementation of innovations that resemble customers to those present in clusters 1 and 2 than in cluster 3.

VII. Conclusions

The development of this project contemplates an analytical solution to the behavioral evolution of a customer to define its success. Incorporating a set of guidelines, the construction of innovative systems capable of serving the business world with knowledge regarding customer needs is enhanced.

Its development represents to be a beneficial addition in the strategy of defining and understanding the target audience. The set of assumptions favored the decision of the various steps that designed the implementation of different alternatives for grouping and predicting the activities of the group of clients under investigation. The dynamics of the variables that represent the activity of the customers of E-goi enabled the analysis of their evolution with the intention of demonstrating and predicting success in all cases. Time Series K-means, with better performance, display the distribution of more than a hundred thousand customers and effectively spectate their future behavior. The prediction of the profiled time series makes it possible to ponder the next strategic actions, proposing the maintenance or provisioning of actions.

The results that were obtained amplify the optimizations of this service, setting as future work the use of new technologies that promote flexibility and reliability in forecasting behaviors through success and failure observations, independently. Similarly, the exploration of new methods of segment validation to be defined, the application and testing of other segmentation and prediction methods, and consequent performance evaluation shape future work, leaving open several alternative paths. This makes it viable to pursue the search for new conditions to be added as determining factors to obtain conclusions as closely related to the type of information that is intended to be extracted and predicted.

References

[1] H. Abbasimehr and M. Shabani, ”Forecasting of Customer Behavior Using Time Series Analysis,” in *Data Science: From Research to Application*, pp. 188-201, 2020.

[2] E. Alpaydin, ”Introduction to machine learning,” MIT Press, 2020.

[3] A. Alqahtani, M. Ali, X. Xie, and M. W. Jones, ”Deep Time-Series Clustering: A Review,” *Electronics*, vol. 10, no. 23, 3001, 2021.

[4] C. Araújo, C. Soares, I. Pereira, D. Coelho, M.Â. Rebelo, and A. Madureira, ”A Novel Approach for Send Time Prediction on Email Marketing,” *Applied Sciences*, vol. 12, no. 16, 2022.

[5] P. E. Bhaskaran, M. Chennippan, and T. Subramaniam, ”Future prediction and estimation of faults occurrences in oil pipelines by using data clustering with time series forecasting,” *Journal of Loss Prevention in the Process Industries*, vol. 66, 104203, 2020.

[6] I. César, I. Pereira, A. Madureira, D. Coelho, M.A. Rebelo, and D.A. Oliveira, ”Analysing and Modeling Customer Success in Digital Marketing,” in *13th International Conference on Innovations in Bio-Inspired Computing and Applications (IBICA 2022)*, On the World Wide Web, December 15-16, 2022.

[7] D. Coelho, A. Madureira, I. Pereira, and R. Gonçalves, ”A Review on MOEA and Metaheuristics for Feature-Selection,” in *Innovations in Bio-Inspired Computing and Applications*, Lecture Notes in Networks and Systems, vol. 419, Springer, Cham, 2022.

[8] D. Coelho, A. Madureira, I. Pereira, and B. Cunha, ”A Machine Learning Approach to Contact Databases' Importation for Spam Prevention,” in *Hybrid Intelligent Systems*, A. Madureira, A. Abraham, N. Gandhi, and M. Varela (eds.), Advances in Intelligent Systems and Computing, vol. 923, Springer, Cham, 2020.

[9] D. Coelho, A. Madureira, I. Pereira, and R. Gonçalves, ”A Review on Dimensionality Reduction for Machine Learning,” in *Innovations in Bio-Inspired Computing and Applications*, A. Abraham, A. Bajaj, N. Gandhi, A. M. Madureira, and C. Kahraman (eds.), IBICA 2022, Lecture Notes in Networks and Systems, vol. 649, Springer, Cham, 2023.

[10] D. Coelho, A. Madureira, I. Pereira, and R. Gonçalves, ”Multi-Objective Evolutionary Algorithms and Metaheuristics for Feature Selection: a Review,” *International Journal of Computer Information Systems & Industrial Management Applications*, vol. 14, 2022.

[11] C. Dakouan, R. Benabdelouahed, and H. Anabir, ”Inbound Marketing vs. Outbound Marketing: Independent or Complementary Strategies,” *Expert Journal of Marketing*, vol. 7, no. 1, 2019.

[12] J. Densmore, ”Data Pipelines Pocket Reference,” O'Reilly Media, 2021.

[13] O. Dogan, E. Ayçin, and Z. A. Bulut, ”Customer segmentation by using RFM model and clustering methods: a case study in retail industry,” *International Journal of Contemporary Economics and Administrative Sciences*, vol. 8, no. 1, pp. 1-19, 2018.

[14] P. Duboue, ”The art of feature engineering: essentials for machine learning,” Cambridge University Press, 2020.

[15] E. Ernawati, S. Baharin, and F. Kasmin, ”A review of data mining methods in RFM-based customer segmentation,” *Journal of Physics: Conference Series*, vol. 1869, 2021.

[16] C. Schröer, F. Kruse, and J. M. Gómez, ”A systematic literature review on applying CRISP-DM process model,” *Procedia Computer Science*, vol. 181, pp. 526-534, 2021.

[17] R. Gustriansyah, N. Suhandi, and F. Antony, ”Clustering optimization in RFM analysis based on k-means,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 18, no. 1, pp. 470-477, 2020.

[18] S. Hartini, W. Gata, S. Kurniawan, H. Setiawan, and K. Novel, ”Cosmetics Customer Segmentation and Profile in Indonesia Using Clustering and Classification Algorithm,” *Journal of Physics: Conference Series*, vol. 1641, 012001, 2020.

[19] T. Hlupić, D. Oreščanin, and M. Baranović, ”A Novel Method for IPTV Customer Behavior Analysis Using Time Series,” *IEEE Access*, vol. 10, pp. 37003-37015, 2022.

[20] A. Jamjoom, ”The use of knowledge extraction in predicting customer churn in B2B,” *Journal of Big Data*, vol. 8, 110, 2021.

[21] E. Lindahl, ”A qualitative examination of lead scoring in B2B marketing automation, with a recommendation for its practice,” (Dissertation), KTH, Stockholm, Sweden, 2017.

[22] I. Maryani and D. Riana, ”Clustering and profiling of customers using RFM for customer relationship management recommendations,” in *2017 5th International Conference on Cyber and IT Service Management (CITSM)*, pp. 1-6, 2017.

[23] F. Martínez-Plumed, L. Contreras-Ochando, C. Ferri, J. Hernández-Orallo, M. Kull, N. Lachiche, M.J. Ramírez-Quintana, and P. Flach, ”CRISP-DM twenty years later: From data mining processes to data science trajectories,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 8, pp. 3048-3061, 2019.

[24] I. Met, A. Erkoç, and S. E. Seker, "Performance, Efficiency, and Target Setting for Bank Branches: Time Series With Automated Machine Learning," *IEEE Access*, vol. 11, pp. 1000-1010, 2022.

[25] T. C. Mills, "Applied Time Series Analysis," Academic Press, pp. 1-30, ISBN 9780128131176, 2019.

[26] P. Nethravathi, G. Bai, C. Spulbar, M. Suhan, R. Birau, T. Calugaru, I. Hawaldar, and A. Ejaz, "Business intelligence appraisal based on customer behaviour profile by using hobby based opinion mining in India: a case study," *Economic Research-Ekonomska Istraživanja*, vol. 33, pp. 1889-1908, 2020.

[27] M. Paulo, V. L. Miguéis, and I. Pereira, "Leveraging email marketing: Using the subject line to anticipate the open rate," *Expert Systems with Applications*, vol. 207, 117974, 2022.

[28] I. Pereira, A. Madureira, E. Costa e Silva, and A. Abraham, "A hybrid metaheuristics parameter tuning approach for scheduling through racing and case-based reasoning," *Applied Sciences*, vol. 11, no. 8, 3325, 2021.

[29] M.A. Rebelo, D. Coelho, F. Fernandes, and I. Pereira, "A New Cascade-hybrid Recommender System approach tailored for the Retail Market," *International Journal of Computer Information Systems & Industrial Management Applications*, vol. 14, 2022.

[30] C. Schröer, F. Kruse, and J. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," *Procedia Computer Science*, vol. 181, pp. 526-534, 2021.

[31] D. Shah, H. Isah, and F. Zulkernine, "Stock market analysis: A review and taxonomy of prediction techniques," *International Journal of Financial Studies*, vol. 7, pp. 1-22, 2019.

[32] A. T. Stephen, "The role of digital and social media marketing in consumer behavior," *Current Opinion in Psychology*, vol. 10, pp. 17-21, 2016.

[33] M. Tekin, M. Etlioğlu, Ö. Koyuncuoğlu, and E. Tekin, "Data Mining in Digital Marketing," in *Proceedings Of The International Symposium For Production Research 2018*, pp. 44-61, 2019.

[34] M. Wang, L. Wang, X. Xu, and Y. Qin, "Characteristics Analysis and Impact Cluster on Urban Rail Transit Perturbations: A Real Case in Beijing," in *2018 IEEE SmartWorld, Ubiquitous Intelligence and Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People and Smart City Innovation*, pp. 23-28, 2018.

[35] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," *Annals of Data Science*, vol. 2, no. 2, pp. 165-193, 2015.

[36] H. Zare and S. Emadi, "Determination of Customer Satisfaction using Improved K-means algorithm," *Soft Computing*, vol. 24, pp. 16947-16965, 2020.

Author Biographies

Inês César was born in Porto, Portugal, in 2000. She got hers BSc degree in Informatics Engineering by Institute of Engineering-Polytechnic of Porto in 2022. She is currently enrolled in MSc degree in Informatics Engineering by Institute of Engineering-Polytechnic of Porto and she is a Fellow Researcher at ISRC. Hers research interests involve artificial intelligence, machine learning and computational intelligence.

Ivo Pereira was born in Porto, Portugal, in 1984. He got his BSc degree in Informatics Engineering by Institute of Engineering-Polytechnic of Porto in 2006, MSc degree in Informatics Engineering by Institute of Engineering- Polytechnic of Porto in 2009 and PhD in Electronics and Computers Engineering by University of Trás-os-Montes and Alto Douro in 2014. He is currently an Assistant Professor at University Fernando Pessoa, Porto, Portugal, Researcher at ISRC and Consultant at E-goi. His research interests involve artificial intelligence, machine learning, optimization algorithms, intelligent marketing and computational intelligence.

Ana Madureira was born in Mozambique, in 1969. She got his BSc degree in Computer Engineering in 1993 from ISEP, master's degree in Electrical and Computers Engineering – Industrial Informatics, in 1996, from FEUP, and the PhD degree in Production and Systems, in 2003, from University of Minho, Portugal. Currently she is Coordinator Professor at the Institute of Engineering–Polytechnic of Porto and Director of the Interdisciplinary Studies Research Center (ISRC).

Duarte Coelho was born in Portugal, in 1995. He got his BSc degree in Computer Engineering in 2016 from ISEP, master's degree in Computational Systems, in 2018, from ISEP. He is currently a Data Scientist at E-goi, Matosinhos, Portugal, Researcher at. His research interests involve artificial intelligence, machine learning, optimization algorithms, intelligent marketing and computational intelligence.

Miguel Ângelo Rebelo was born in Espinho,Portugal, 1995. He got his First degree in Biochemistry, at the Faculty of Sciences, University of Porto (2016), Master's degree Molecular Biology and Biotechnology, at the School of Sciences, University of Minho (2018), Master's degree in Computational Statistics and Data Analysis, at the Faculty of Sciences, University of Porto (2022). His research interests involve population genetics, machine learning, and bioinformatics.

Daniel Alves de Oliveira was born in Santa Maria da Feira, Portugal, in 1982. He got is degree in Sports Management at University of Maia in 2006. He was a PhD student in Sports Science between 2012 and 2015 at University of Porto. He is currently Head of Innovation & Research at E-goi and also Manager of the Artificial Intelligence team. His research interests involve marketing, strategy, and management.