Article

XAI for Churn Prediction in B2B Models: A Use Case in an Enterprise Software Company

Gabriel Marín Díaz *[iD], José Javier Galán and Ramón Alberto Carrasco [iD]

Faculty of Statistics, Complutense University, Puerta de Hierro, 28040 Madrid, Spain
* Correspondence: gmarin03@ucm.es

Abstract: The literature related to Artificial Intelligence (AI) models and customer churn prediction is extensive and rich in Business to Customer (B2C) environments; however, research in Business to Business (B2B) environments is not sufficiently addressed. Customer churn in the business environment and more so in a B2B context is critical, as the impact on turnover is generally greater than in B2C environments. On the other hand, the data used in the context of this paper point to the importance of the relationship between customer and brand through the Contact Center. Therefore, the recency, frequency, importance and duration (RFID) model used to obtain the customer's assessment from the point of view of their interactions with the Contact Center is a novelty and an additional source of information to traditional models based on purchase transactions, recency, frequency, and monetary (RFM). The objective of this work consists of the design of a methodological process that contributes to analyzing the explainability of AI algorithm predictions, Explainable Artificial Intelligence (XAI), for which we analyze the binary target variable abandonment in a B2B environment, considering the relationships that the partner (customer) has with the Contact Center, and focusing on a business software distribution company. The model can be generalized to any environment in which classification or regression algorithms are required.

Keywords: churn detection; XAI; interpretability; B2B; RFM; RFID

MSC: 68T20

1. Introduction

In a global market, where customers can change their preferences and buy from competitors, it is necessary to adopt strategies that encourage customer-brand engagement, for example, by proposing alternatives to strategically profitable customers, who could have a positive tendency to abandon the relationship with the brand, or by letting the less profitable ones go [1,2]. Through the RFID model [3] based on the recency, frequency, importance, and duration of interactions between the customer and Contact Center, a metric is proposed that makes it possible to determine the value of the customer from the perspective of after sales services, and therefore to design the most recommendable actions in order to build customer loyalty. However, these decisions, which could be left in the hands of black box algorithms, must be subject to interpretability to avoid discriminatory biases and to be able to make explainable decisions, thus constituting the cornerstone of this document.

Traditionally, customer churn studies are closely related to Business to Customer (B2C) environments. In fact, customer characteristics and behavior can vary considerably depending on whether the relationships are Business to Business (B2B) or B2C. [4]. Although companies that base their business model on relationships with other companies tend to have fewer customers, these customers make larger and more frequent purchases compared to their counterparts in a B2C environment [5], and their retention is seen as fundamental in the development of sustainable business relationships [6,7], hence the importance of

working on a definition of interpretable models related to customer churn in B2B environments. As we will see in the use case that we will present throughout the document, the characteristics that define the tendency of customer churn in B2B environments differ from the behavior in B2C environments.

In several preliminary studies, the data used for the development of predictive models are based on the RFM model [8], based on recency, frequency, and monetary parameters. This model is generally used to segment customers so that marketing, cross-selling, or up-selling actions can be developed according to the category to which the customer belongs. The metrics of this model are also often used for a customer churn analysis [9]. An extension of this model is the LRFM [10], based on the relationship time; recency, frequency, and monetary value of the set of transactions; and a new extension corresponds to the LRFMP model, where periodicity and seasonality in purchase interactions are included [11]. All these models are based on the processing of purchases made by the customer in a period.

A crucial aspect of the customer-centric philosophy is to consider that communications between the company and customer are bidirectional, and that the customer wants to be served in an integral, consistent way and through any channel. The importance of technology combined with strategy is fundamental, and systems based on customer relationship management (CRM) allow multichannel integration and therefore provide a deeper knowledge of the customer for better customer management [12]. Therefore, for any customer-centric strategy, the proper implementation of customer support processes integrated in the CRM and carried out by the Contact Center is essential [13].

Despite the proven and validated usefulness of the RFM model, this model does not consider the interactions that occur between the company and the customer after the sales process which, in many cases, are as important or more than the sale itself, to establish a true customer-centric strategy. Everyday examples in relation to these post-sales interactions can be delay in the delivery of an order, defective delivery, poor quality of the delivery service, etc. It is important to note that the literature is poor in relation to the interpretation of customer value or customer typification from the point of view of the interaction that the customer has with the Contact Center. In this sense, the RFID model applied to B2C and B2B environments is particularly relevant, and uses the criteria present in this model for the prediction and interpretability of customer churn.

In this paper, we will carry out a research process on the tendency to abandon the technological partners (customers) of a software company. The relationship between partner and software company is very close; it is usually cemented over years of work; the partner is responsible for implementing the software's solutions to end customers, and, therefore, needs qualified personnel and a high level of knowledge of the software to be implemented. Therefore, the relationship between the two technological partners is long-lasting, long-term, collaborative, and mutually supportive, and therefore the Contact Center takes on special relevance.

Given the importance of the Contact Center in the interaction between brand and customer, we have chosen to apply a working methodology that allows us to use Artificial Intelligence (AI) to approach interpretable predictive models [14] that help to understand the causes related to the abandonment rate of technology partners. In this sense, we have worked on correlating the customer classification obtained from the RFID model with the abandonment rate, with the aim of analyzing the possible causes that determine a business partner to leave the relationship with the software company. To do so, we have proposed a methodological process that helps to develop and apply the concepts of Interpretable Artificial Intelligence (XAI) [15], and based on these techniques, a working model has been implemented that includes a set of interpretability agnostic algorithms, independent of the selected machine learning (ML) model, in order to provide a methodological guide that allows the development of an explanatory model applicable to algorithms that are not very interpretable (black box).

The novelty of this work lies in the following factors, the first of which is to address the churn rate within a B2B environment where, both in the academic and industrial

fields, there are a limited number of studies that report a research process applied to the real world. Secondly, studies related to customer churn are very numerous in the B2C environment, where machine learning (ML) or deep learning (DL) procedures are applied with a significant degree of prediction, but without applying interpretability (XAI). Finally, we use a complementary model to the traditional RFM and LRFMP as a predictive analysis criterion, the RFID model, to respond to potential customer churn, as we have not found works that use the variables that typify customer service (RFID) in the prediction of abandonment.

In the rest of this paper we will develop and apply the XAI model, according to the following structure: in Section 2, we will review the current status of the use of XAI methodologies and their application scenarios, contrasting the GAP between the use of ML algorithms and the use of explainability in relation to the customer churn rate; in Section 3, we will address the methodological framework that we will use in the prediction and explainability of churn; in Section 4, we will detail the proposed model; in Section 5, we will implement the XAI model applied to the customer churn rate in a B2B model and within a business environment dedicated to the distribution and implementation of software licenses; and finally, in Sections 6 and 7 we will present the conclusions and future work.

2. Related Work

2.1. Related Work on Customer Churn B2C and B2B

Customer churn has been one of the main topics of attention for researchers and companies, with abundant literature in B2C environments (Figure 1), as the loss of a customer has a direct impact on the bottom line of any company, in addition to the loss of brand image, and since attracting a new customer is substantially more costly financially than retaining existing customers [16].
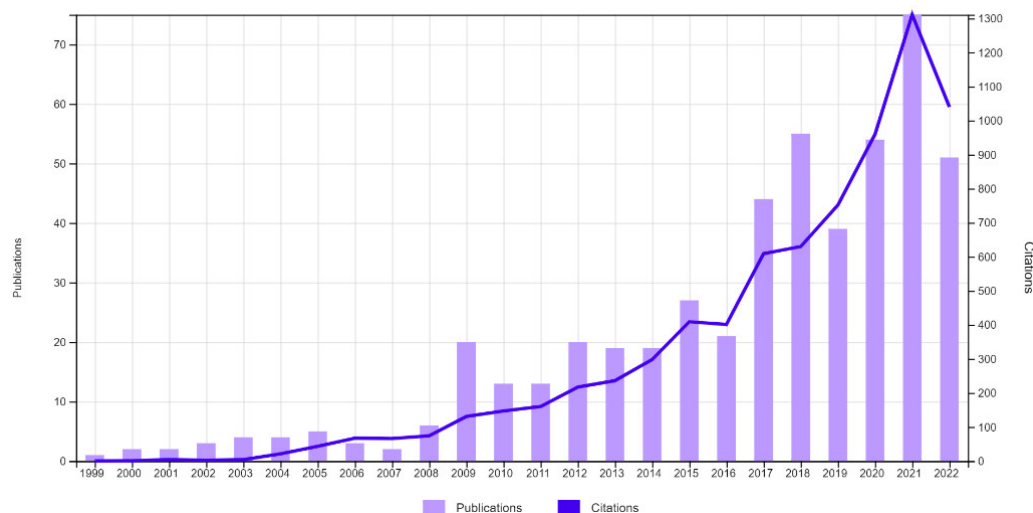

Figure 1. Publications (502) and citations, B2C. TS = (CUSTOMER CHURN NOT B2B) AND TS = ((MACHINE LEARNING) OR (DEEP LEARNING)).

The following graph shows the studies related to customer churn in B2C environments, until September 2022.

Research production in recent years has been mainly oriented towards the telecom, commerce, banking, and insurance sectors, Table 1.

Table 1. B2C publications by business sector.

Sector	Publications	%
Telecom	108	21.95%
Banking	67	13.62%
Commerce	121	24.59%
Insurance	48	9.76%
Others	148	30.08%

Some significant examples are related to churn in the telecommunication industry [17], in the banking sector [18], in the insurance sector [19], in the retail sector [20], and in Ecommerce [21].

B2B models have received less attention than B2C models, and there is a total of 17 articles published from 1999 to August 2022 (Figure 2, Table 2). The characteristics of the B2B business, with a lower impact in number of customers, but with much higher transactional values, make these models acquire special connotations, since the loss of any customer can have a very negative impact on turnover and brand image [22,23]. In addition, customer churn in B2B scenarios has been studied mainly from the perspective of resource allocation for business development, or in the analysis and prediction of current and future customer profitability [24].
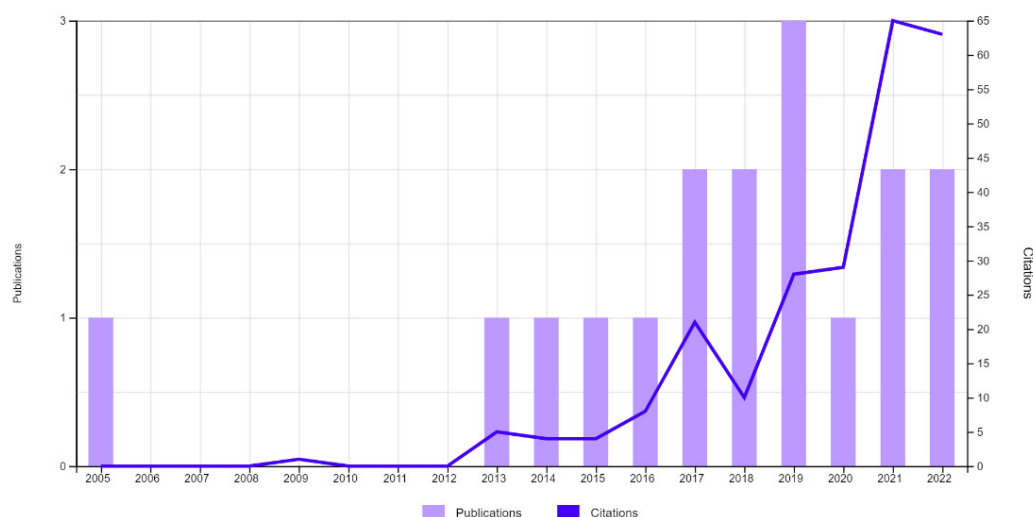
Figure 2. Publications and citations, B2B. TS = (CUSTOMER CHURN) AND TS = (B2B)) AND (PY = (1999–2022)).

Table 2. Publications B2B. TS = (CUSTOMER CHURN) AND TS = (B2B) AND (PY = (1999–2022)).

Ref.	Fundamentals	Datasets	Application
Figalist et al. 2019 [25]	Several modeling techniques	Supplier of SW products	Customer churn, challenge
Janssens et al. 2022 [26]	XGBoost	North American B2B beverage retailer	Maximum expected benefit from customer retention campaigns
Jahromi et al. 2014 [27]	Several modeling techniques	Australian online Fast Moving Consumer Goods (FMCG) retailer	Data mining and retention campaign modeling
Chen et al. 2015 [28]	LRFMP, AHP, SVM	Logistics company	Applicability of the LRFMP model to the B2B context in the logistics sector

Table 2. *Cont.*

Ref.	Fundamentals	Datasets	Application
Mirkovic et al. [9]	LRFMP, Logistic Regression, SVM, Random Forest	Eastern European seller and distributor of agricultural goods and equipment	Different churn definitions and variable window widths for feature extraction and a multislicing approach to dataset
Zhang et al. 2019 [29]	RFM, XGBoost, Random Forest, Others	Software maintenance service provider	Churn prediction to the context of software maintenance contract
Hopmann et al. 2005 [30]	Stochastic, and data mining method	Multisectoral	Contrasting the usefulness and quality of churn prediction
Sheikh et al. 2019 [31]	LRFMP, K-means	Fintech industry	Customer clustering method, helps to predict customer behavior
De Caigny et al. 2021 [32]	Uplift LLM	European software provider	Segmentation-based algorithm that combines predictive performance with interpretability
Barfar et al. 2017 [33]	Logistic Regression, Classification Trees	B2B service database	Quality of service vs. B2B churn
Gordini et al. 2017 [5]	SVM	eCommerce	Customers' churn prediction and marketing retention strategies
Lee et al. 2018 [34]	Analysis	Telecom Company	Probability of change with varying service conditions
Jamjoom 2021 [19]	Logistic Regression, RNN	Insurance	Loss of customers in insurance companies
Liu et al. 2016 [35]	Logistic Regression	Multisectoral	How B2B sales professionals deal with customer defection
D'Haen et al. 2013 [36]	Logistic Regression, Decission Trees, Bagging	Multisectoral	Investigate which data mining techniques worked best in predicting customer profitability
Schaeffer et al. 2020 [37]	SVM	Prepaid unitary services	Analysis of the probability of churn of prepaid customers
Gattermann-Itschert et al. 2021 [38]	Multislicing, Logistic Regression, SVM, Random Forests	Model for one of Europe's largest convenience wholesalers (company) to small convenience retail stores	Training on multiple time slices improves performance in churn prediction

Figure 2 below shows the publications per year related to customer churn in B2B environments, and Table 2 shows these papers.

The strategy followed in the work is related to the use of transactional data based on the RFM model [9,28,29,31]; through the relationship between supplier and customer over time in different phases, before, during, and after the purchase process, known in marketing as the Customer Journey Map [25]; or by collecting sales and interaction data [19,27,30,36–38]. The following study [26] uses the metric of the benefit implied by the correct classification of a customer and a cost associated with those who are incorrectly classified, and this other study uses the quality of service to determine the subscription rate [33]. As a general rule, it can be seen that all studies are based on a combination of interpretable (white box) and non-interpretable (black box) predictive models, but without using interpretability; however, only the following study proposes a customer segmentation that combines a predictive analysis with interpretability [32].

2.2. Related Work XAI

Decisions based on ML algorithms are having an increasingly significant social impact; however, most of these systems are based on black box algorithms, i.e., models whose rules are not understandable to humans [39].

AI research since its beginnings has been characterized by the development and implementation of predictive models. However, the first steps in interpretable models were taken in the 1970s and 1990s towards initiatives such as MYCIN [40], seeking an explanation in the diagnosis of infectious diseases; GUIDON [41], in the elaboration of computer-assisted learning; systems based on alternative lines of reasoning (TMS) and neural networks applied to the healthcare field were developed. Since 2010, the concern derived from bias in decision making has led to more focus on the development of Explainable Artificial Intelligence (XAI) models. Explainability requires interpretability, but explainability has to do with the need for the explanation to be deep enough to be audited [42].

According to Miller, "Interpretability is the degree to which a human can understand the cause of a decision." [43]. It is essential to understand why a given prediction was made by the model in question.

Features that should be incorporated in interpretable models [14] can be enumerated as follows: explanations should be contrasting [44], why a certain prediction was made rather than another. In addition, explanations are selected: we are interested in selecting the criteria that fit as most probable in the elaboration of the explanation. Explanations should be social, i.e., an explanation is linked to the explainer and the receiver of the explanation. Explanations focus on the abnormal [45], i.e., causes that are attributed with high potential but low probability. Explanations are true, so the event should be predicted with the highest possible probability. The explanations are consistent with previous beliefs: this is what is called confirmation bias, devaluing those explanations that do not agree with our beliefs [46].

The first formula to achieve interpretability is to use interpretable ML algorithms, including linear regression, logistic regression, decision trees, RuleFit, and Naive Bayes [14], thus deducing correlations between features that allow defining and interpreting the model at a global level [47].

Another option is to extract knowledge from a black box model by approximating it to interpretable models [48,49].

Finally, we have agnostic methods, whose implementation does not depend on the ML model used [50]. A review of agnostic models according to their global/local character is presented in Figure 3.

Figure 3. Global and local model agnostic methods. Source [14].

The current trend is to focus on model-independent interpretation tools [14,50,51]. The following is a list of studies related to interpretability methods applied to black box ML models (Table 3).

Table 3. ML interpretability studies.

Models	Author	Description
Lime Eli5 InterpretML AIX360 Skater	Ribeiro 2016 et al. [52]	"Why Should I Trust You?" Explaining the Predictions of Any Classifier.
PDPbox InterpretML Skater	Friedman, 2001 [53]	Greedy function approximation: A gradient boosting machine.
Shap Alibi AIX360 InterpretML	Lundberg et al., 2017 [54]	A unified approach to interpreting model predictions.
Eli5	Altmann et al., 2010 [55]	Permutation importance: A corrected feature importance measure.
PyCEbox	Goldstein et al., 2015 [56]	Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation.
DLIME	Zafar et al., 2019 [57]	DLIME: A Deterministic Local Interpretable Model-Agnostic Explanations Approach for Computer-Aided Diagnosis Systems.
AIX360	Gurumoorthy et al., n.d. [58]	Efficient Data Representation by Selecting Prototypes with Importance Weights.

In the following studies in Table 4, the interpretability models applied to the churn rate are explored in more detail:

Table 4. XAI, (TS = (CUSTOMER CHURN) AND TS = (XAI)) AND (PY = (1999–2022)).

Ref.	Title	Description
Leung et al., 2021 [59]	Explainable Artificial Intelligence for Data Science on Customer Churn	Explainable artificial intelligence (XAI) solution is presented to explain a Random Forest-based predictive model of customer churn.
Na et al., 2020 [60]	A Securities Company's Customer Churn Prediction Model and Causal Inference with SHAP Value	Presents the case for the development of a predictive model for financial churn, compares and analyses a total of six churn models, and infers the cause of churn by classifying and analyzing SHAP value data.
Ullah et al., 2022 [61]	Explaining deep learning models for tabular data using layer-wise relevance propagation	Layered Relevance Propagation (LRP), an established explanatory technique developed for deep models in computer vision, is applied in this study using a deep neural network (1D-CNN) for the use cases of credit card fraud detection and telecom customer churn prediction.

As can be seen, interpretability applied to customer churn prediction is a technique that is in the process of research and practical application, especially in B2B models. In this paper, we provide a set of interpretability techniques applied to real data, corresponding to a management software manufacturing company, using the RFID model by aggregating the interactions between customer and supplier in a predetermined period.

3. Methodology

To achieve our goals, we propose a methodology based on knowledge discovery databases (KDD) and the cross-industry standard process for data mining (CRISP-DM) [62]. Figure 4 shows the stages and the models used in each of them.
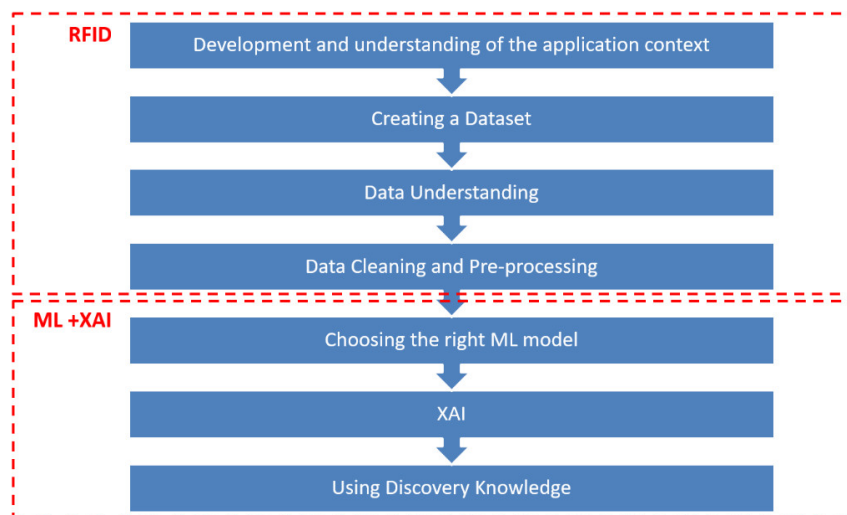

Figure 4. Proposed methodology.

3.1. RFID Model

The RFID model is based on the parameters of recency, frequency, importance, and duration of interactions between the customer and Contact Center during a defined period of time [3]. This model helps us to determine the value of the customer from the point of view of their interactions with the Contact Center, as well as providing us with a segmentation and a strategy of actions to be carried out for each group of customers.

From the ticket information stored in a conventional operational CRM, the model obtains two types of recommendations for customers based on the history of their interactions with the customer service: individualized and grouped. The model is parameterized with the information provided by customer service experts. These same users are also in charge of determining and implementing the final strategies for the treatment of marketing campaigns and/or interaction with customers (Figure 5).
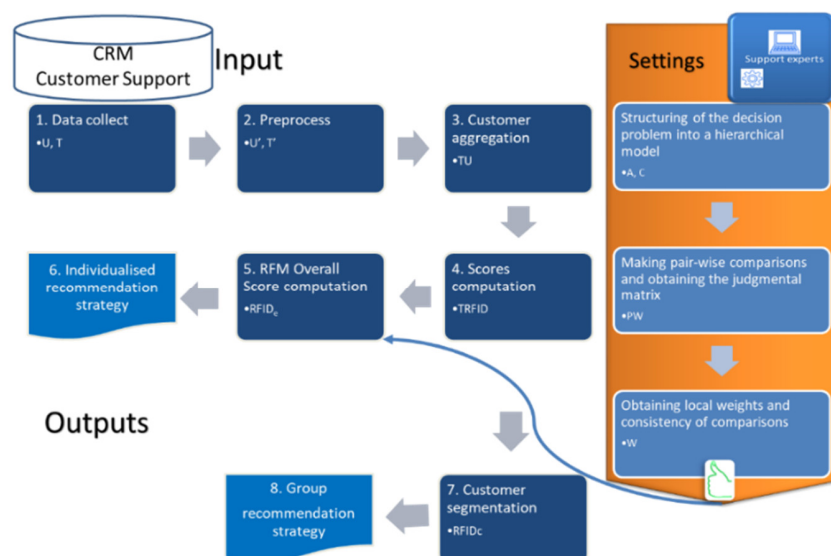

Figure 5. Stages of the RFID model. Source: [3].

The process is detailed below:

1. Obtaining data from the CRM, which correspond to the set of tickets opened by each customer.
2. Pre-processing of the information, the period of analysis is defined, and an initial exploratory data analysis (EDA) is addressed.
3. Information aggregation process: for each customer, and for the period considered, the values of the recency, frequency, duration, and importance of the interactions are obtained. A process of information aggregation is carried out, so that an aggregate value is obtained for each customer for each of the characteristics that make up the RFID model.
4. Application of the 2-tuple model [63] on the data obtained in the previous step, the aim of which is to bring all the information into the same working domain. The 2-tuple model allows working with heterogeneous information, unifying this information in linguistic evaluations, expressed in a basic set of S linguistic terms. In this way, all the heterogeneous information based on numerical, interval, or linguistic ranges can be unified in a fuzzy set, through an aggregation process (Figure 6).
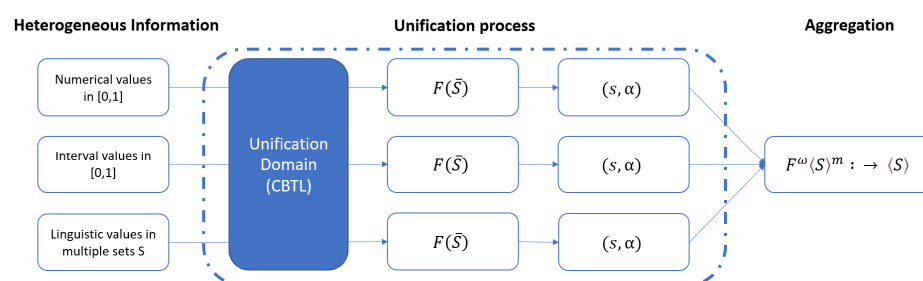

Figure 6. Heterogeneous information unification model in 2-tuple linguistic assessments [64].

5. Obtaining the global valuation of each client, by applying the AHP model [65] to each of the features that make up the RFID model (Table 5). In our model, we will use the AHP method to establish the weights of each of the criteria that will determine the total score of each customer, after the aggregation and ranking process using the RFID 2-tuple model.

Table 5. Saaty's scale [65].

Degree of Importance	Definition	Description
1	Equal importance	The comparative weighting of the criteria i and j is the same.
3	Moderate importance	The weighting of the criteria compared is moderately higher for the criterion i over the criterion j.
5	Strong importance	The weighting of the criteria compared is strongly higher for the criterion i over the criterion j.
7	Very strong importance	The weighting of the criteria compared is very strong for the criterion i over the criterion j.
9	Extreme importance	The weighting of the criteria compared is extremely strong for the criterion i over the criterion j.
2, 4, 6, 8	Intermediate values	Intermediate weighting criteria.
Reciprocals		If criterion i compared to criterion j is associated with one of the preceding numbers, then j has a reciprocal when compared to i.

The vector of weights for each of the criteria, w, is constructed using the eigenvector method through the following equation:

$$PW_i = \sum_{j=1}^{n} pw_{ij}w_j = \lambda_{max}\, w_i \qquad (1)$$

where λ_{max} is the maximum eigenvalue of PW_i and w_i is the normalized eigenvector associated with the principal eigenvalue of PW_i. This approach provides the best priority weightings for each criterion or sub-criterion (Figure 7).
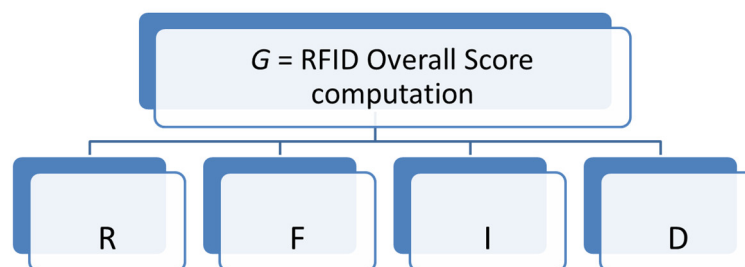

Figure 7. AHP hierarchy. Source [3].

A review of the *AHP* method and its applications can be found in the following references [66,67].

6. Establishment of an individualized recommendation strategy.
7. Customer clustering, according to the k-means model [68].
8. Obtaining a recommendation strategy by groups.

In our study, we will apply a set of ML algorithms and try to analyze the interpretability of the algorithm as a higher accuracy and higher ROC AUC curve score.

In our study, we will transform the data into a numerical domain, integrating the variable abandonment, and then develop the predictive model and analyze its interpretability.

3.2. XAI

From the interpretability point of view, some authors distinguish two types of models [69]: white box models, that allow one to establish correspondence between input and output; and black box models, in which the rules on which they base their decision making must be interpretable [69]. In this other study, the interpretability of white box models is questioned [70]. An inverse correspondence between interpretability and accuracy can be seen in Figure 8.
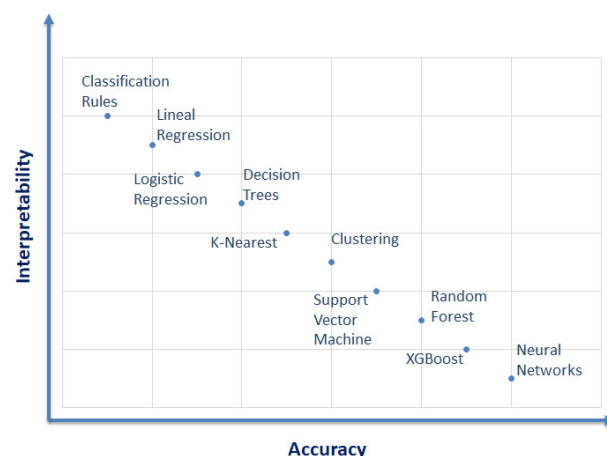

Figure 8. Explainability vs. accuracy. Source [71].

Interpretation methods for machine learning can be classified according to several criteria [70].

- ¿Intrinsic or post hoc? Interpretability is either inherent to the learning model (intrinsic), or it allows for analysis after model training (post hoc).
- ¿Specific or agnostic? Interpretability is achieved in a specific way by applying specific models for the object of study. Agnostic models are independent of the ML algorithm.
- ¿Local or global? It is necessary to respond to individual or global predictions. Global methods describe the average behavior of the ML model, and they are very useful when you want to analyze the overall mechanism of the data. Local methods, however, explain individual predictions.

The main XAI techniques are shown in Table 6. These techniques will be applied to the RFID dataset to analyze the interpretability of the ML algorithms used for the detection of technology partner abandonment in the proposed business case.

Table 6. XAI techniques applied to the case study.

Explanation Method	Scope	Description	Result
Partial Dependence Plot [53]	Global	It analyzes the partial dependence of one or more variables on a third variable.	Feature Summary
Individual Condition Expectation [56]	Global/Local	Visualizes the prediction dependence of a feature individually, and the result is a graph for each instance.	Feature Summary
Feature Importance [55]	Global/Local	Assess the importance of a given feature by calculating the increase in prediction error after making a permutation of it.	Feature Summary
Local Surrogate Model [52]	Local	LIME acts by checking what happens to the predictions when variations in the input data are introduced.	Surrogate Interpretable Model
Shapley Values [54]	Local	SHAP values attempt to explain the output of a function f as a sum of the effects ϕ_i of each conditionally entered characteristic.	Feature Summary

In our study, we will apply a set of ML algorithms and try to analyze the interpretability of the algorithm with higher accuracy and a higher ROC AUC curve score.

4. Application of XAI to Customer Churn Rate

One of the strengths of the Contact Center is to try to maximize customer satisfaction, and an important variable in this regard is the degree of satisfaction of the Contact Center staff [72].

In this methodological guide, we will approach a working procedure whose objective is to analyze the binary target variable abandonment, based on the values obtained by the RFID model, and in future works the proposed model will be extended to Contact Center staff turnover.

Following KDD and CRISP methodology [62], Figure 4:

- In Section 4.1, we will review the problem domain, create a target dataset based on the RFID model to which we will add the customer cancellation request variable (abandonment), then pre-process and transform the data to a numeric domain.

- In Section 4.2, a set of pre-model techniques will be applied to obtain the first knowledge provided by the dataset. Subsequently, the ML algorithms detailed in Table 7 will be applied to obtain the optimal algorithm for the case study.
- In Section 4.3, we will interpret the results and the convenience or not of using interpretability, in which case we will apply the global and local agnostic algorithms seen in Section 3 of this paper; finally, conclusions will be drawn.

Table 7. ML algorithms applied to the model.

Algorithm
Logistic Regression
Random Forest
Support Vector Machine
K-nearest neighbors
Decision Tree Classifier
Gaussian NB
XGboost

4.1. Data Collection

The data are collected from the CRM, and we will execute the process described in Section 3.1. In this case, we are interested in the value of recency, frequency, importance, and duration in numerical format. We will carry out the aggregation process and from there we will store this information to determine the binary target variable abandonment in relation to the rest of the criteria.

The value associated with the type of incident in the CRM indicates whether or not it is a request for cancellation by the customer, the value in the CRM of attribute *Type = "Cancellation Request"*.

Given $T = \{(u_1, r_1, f_1, i_1, d_1, t_1), \ldots, (u_{\#T}, r_{\#T}, f_{\#T}, i_{\#T}, d_{\#T}, t_{\#T})\}$,

- r_e: corresponds to the number of days since the last service request by the customer u_e (using the end date of the analysis period as a reference). Therefore, $r_e = diffdays(t_2 - max(ticket\_date_i))$, where $diffdays$ is a function that returns the difference in days between two dates, and max is a function that returns the last date of the different incoming dates.
- f_e: is the number of times the customer has made a service request, i.e., with different ticket codes, $ticket\_id_i$.
- i_e: is the average importance. This is a linguistic variable that must be transformed into a numerical variable, $i_e = \bar{x}^e[ticket\_importance_i]$.
- d_e: contains the total duration in days of all the customer's tickets. Therefore, $d_e = \sum_i (diffdays (status\_date_i - ticket\_date_i)$.
- t_e: contains the value of the service type, in this case the customer's service cancellation request.

The next step will be to perform a data cleaning, that is, we will check if the information collected requires some kind of debugging, for example, outliers. It usually happens that incidents can be opened without a specific customer, with these being imputed to generic customers, growing in number above the average, and thus distorting the information collected and therefore the analysis.

Once the data cleaning is done, a normalization process is carried out. The machine learning algorithms work best when the numerical input variables fall within a similar scale. In this case, we will normalize in the range (0,5).

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{2}$$

4.2. Customer Churn Prediction

Once the above steps are completed, some of the pre-model, data visualization, and exploration techniques are applied to explore, interpret, and gain initial insight into the dataset and thus predict churn or non-churn. The application of these techniques will help to identify the key features of the model and, being model-independent, they are applicable to any dataset and prior to any initial selection of the chosen ML model.

The first technique used is the univariate analysis, through histograms. Secondly, a multivariate analysis allows us to establish a correlation map between variables [73], as well as the distribution of outcomes, and thus obtain an initial data analysis.

Once the first approximation and evaluation of the dataset has been made, we can divide it into training and test, considering that the variable x corresponds to the RFID criteria (recency, frequency, importance, and duration) and y is the variable to be predicted, i.e., customer abandonment data (yes/no).

An analysis will be carried out using the algorithms shown in Table 7 to determine which of them best fits the predictive model.

Each of the models described in Table 7 is evaluated through a cross-validation process (K-fold), and the receiver operating characteristic (ROC) and area under the curve (AUC) curves are analyzed, and the accuracy mean [74,75]. The higher the area under the curve, AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1. The ROC curve can be seen in Figure 9, where on the y-axis we have the true positive ratio (TPR), and on the x-axis we find the false positive ratio (FPR). Accuracy can be obtained as the result of the quotient of the sum of correct predictions by the total number of predictions.
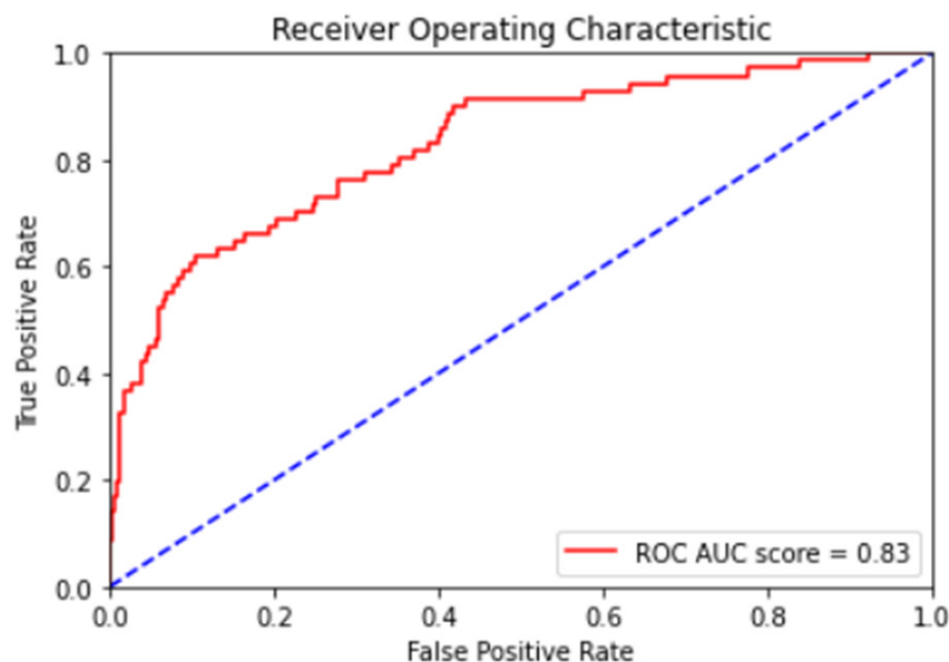

Figure 9. Example of ROC/AUC curve applied to customer abandonment.

Other algorithms could also have been used in the predictive process, such as deep neural networks [76], CatBoost, or LightGBM [77,78]. The objective is not so much to seek accuracy as to generate an ML model-independent explainability methodology.

Once the different machine learning models have been tested, we will discuss the explainability of each one of them versus the predictive capacity. We will keep the model that best meets the predictive expectation, and we will use interpretability in case it is necessary.

4.3. ML Interpretability Analysis

Next, this section will analyze the interpretability of the ML models described in Section 4.2. The methodology designed in this paper is extensible to any case in which we must predict a variable (classification or regression) based on the rest of the characteristics (Figure 10).
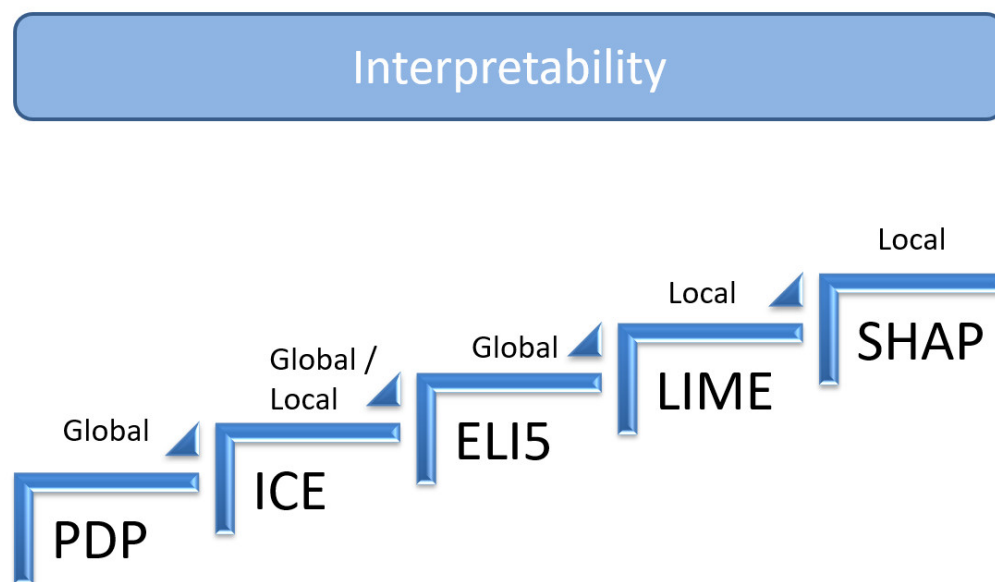

Figure 10. Algorithms used in the ML interpretability process.

We will start by applying the partial dependency plot (PDP), which shows the effect that one or two features have on the prediction result of an ML model [53]. The diagram allows us to work with univariate and bivariate graphs, and it will help us to determine the correlation between variables.

The PDP is an average of the lines of an ICE diagram; in the next step, we work with the individual conditional expectation curves (ICE) model, offers a local expectation, focusing on individual data instances [56].

Next, the ELI5 model is used to measure the importance of the features; it helps us to see when our model can respond to counterintuitive results. ELI5 allows us to fit the model using the XGboost library, and then analyze the importance of each feature within the applied model [55].

In the next step, the LIME model is based on approximating the black box model through explainable models (linear regression, decision tree), to make its predictions interpretable [52].

Finally, we will apply SHAP, which allows us to know which characteristics were the most influential for the model to make the correct decision to predict whether the customer was rated with a low or high possibility of abandonment [54].

In addition to the methods indicated above, the reliability of this study could be complemented with other methods of measurement by contrast, such as the Gini index, analysis of variance, Chi-squared test, regression t-test, and variance test.

All these evaluations will give us a global vision of the selected model and will explain which characteristics are determinant in customer abandonment, and thus guide the necessary compensatory actions to mitigate it.

5. Proposed Model Applied in an Enterprise Software Company

In this section, we present an example of the application of the methodological guide developed in Sections 3 and 4 of this paper. We will try to predict whether a partner (customer) abandons the relationship with the software company, based on the valuation

of its relationship with the Contact Center. For a total of 200,615 partners, 198,493 remain after 3 years (2018–2020) and 2122 have left the partner relationship.

In the clustering process (k-means) applied to the above dataset, five clusters of partners are anticipated, Table 8, with the following drop-out rate represented in Figure 11 below.

Table 8. Results of the k-means algorithm expressed in the 2-tuple model.

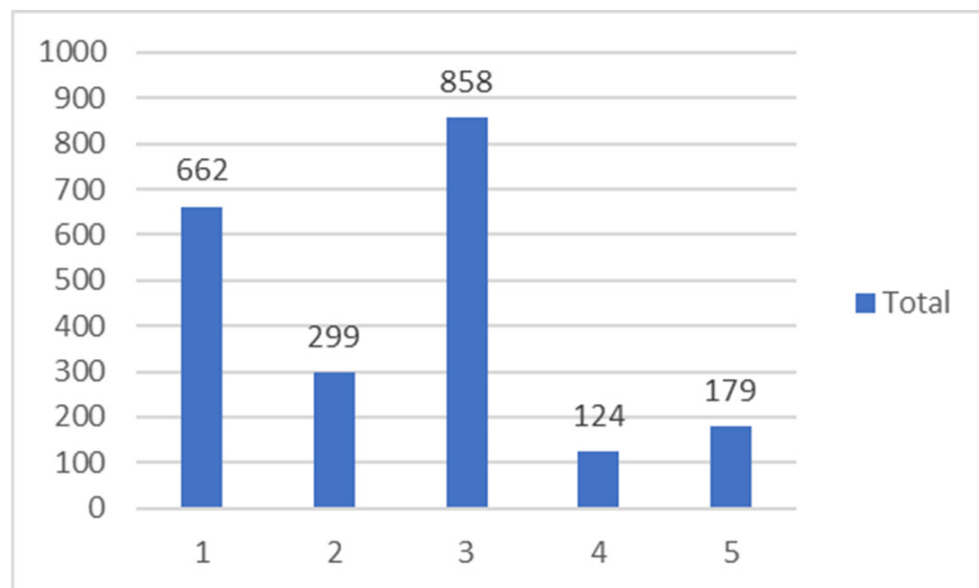| Cluster	R	F	I	D
c	vc1	vc2	vc3	vc4
1	(L, −0.041)	(L, 0.057)	(M, 0.092)	(L, 0.077)
2	(H, −0.008)	(H, 0.047)	(M, −0.006)	(H, 0.074)
3	(L, 0.007)	(L, 0.109)	(M, −0.103)	(H, 0.090)
4	(H, −0.091)	(H, −0.032)	(M, 0.020)	(L, 0.077)
5	(H, −0.050)	(L, −0.042)	(M, 0.019)	(M, −0.090)

Figure 11. Partners per cluster that have churn = "yes" (Type = 1).

Following KDD and CRISP-DM methodology [62], Figure 4, and once the problem domain has been revised, we add the cancellation request (churn) variable, Type = "Cancellation Request", Figure 12, to the RFID dataset, thus obtaining the set *RFIDT*. Next, the data are cleaned and transformed to a numeric domain between 0 and 5, using the Python function *MinMaxScaler*; next, since the data are unbalanced, we will adjust the datasets to avoid this problem. Then, we will use different ML classification algorithms (Table 7) to analyze the relationship between accuracy and interpretability, applying interpretability in the case of higher accuracy and low explanation. Finally, we will apply the global and local agnostic algorithms described in Sections 3 and 4 of this paper, and obtain and analyze the conclusions.

```
newdata['Type'].unique()

array(['Problem', 'Question', 'Subscription Management',
       'Feature Request', nan, 'Cancellation Request',
       'Professional Services', 'Proactive', 'Developer Support'],
      dtype=object)
```

Figure 12. Type's column value.

5.1. Data Acquisition, Processing, and Transformation

The data are collected from the CRM platform, once processed, and transformed, we have the following description of the model (Table 9).

Table 9. RFIDT description.

	Recency	Frequency	Importance	Duration	Type
count	200,615	200,615	200,615	200,615	200,615
mean	416.21	3.16	0.51	10.85	0.02
std	276.97	3.04	0.05	25.70	0.10
min	0.00	1.00	0.25	0.00	0.00
25%	178.00	1.00	0.50	0.00	0.00
50%	378.00	2.00	0.50	0.00	0.00
75%	621.00	4.00	0.50	3.00	0.00
max	1428.00	15.00	1.00	137.00	1.00

As part of the process described in Sections 3 and 4, we will apply some of the pre-model, data visualization, and exploration techniques necessary to explore, interpret, and gain initial knowledge of the dataset. They help us to identify the key features of the model and, being model-independent, they are applicable to any dataset and prior to any initial ML model selection.
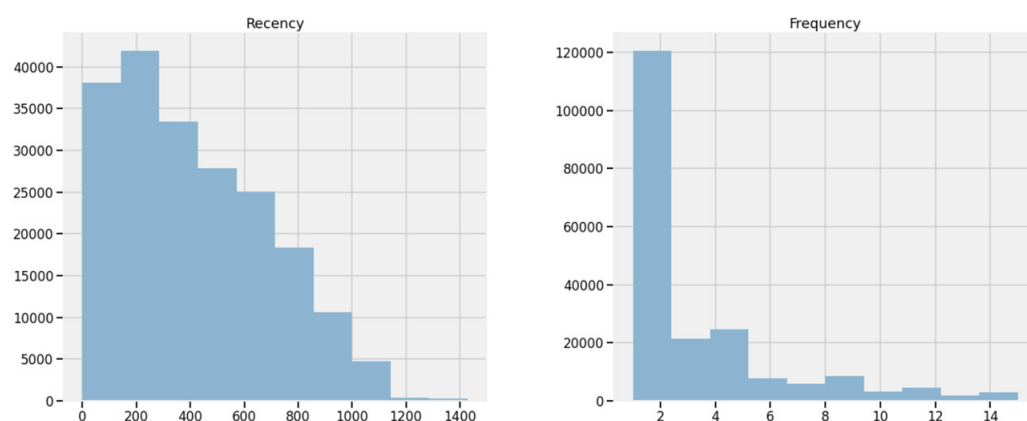
The following is a univariate analysis (Figure 13).

Figure 13. Recency and frequency of transactions by partner.

For both cases, the number of partners is represented on the y-axis, and on the x-axis, there is the time interval in days for recency and the number of interactions in the case of frequency.

In addition, for duration and importance we obtain (Figure 14):

There is a correlation between recency and frequency, but there is hardly any correlation between criteria and abandonment. Some other metrics can help to measure the nonlinear relationship of the characteristics, such as distance correlation, mutual information, and maximum information coefficient. For the case study, we will use Pearson's correlation.

Next, in the Table 10, we perform a transformation to a range [0,5] through of the function *MinMaxScaler*.
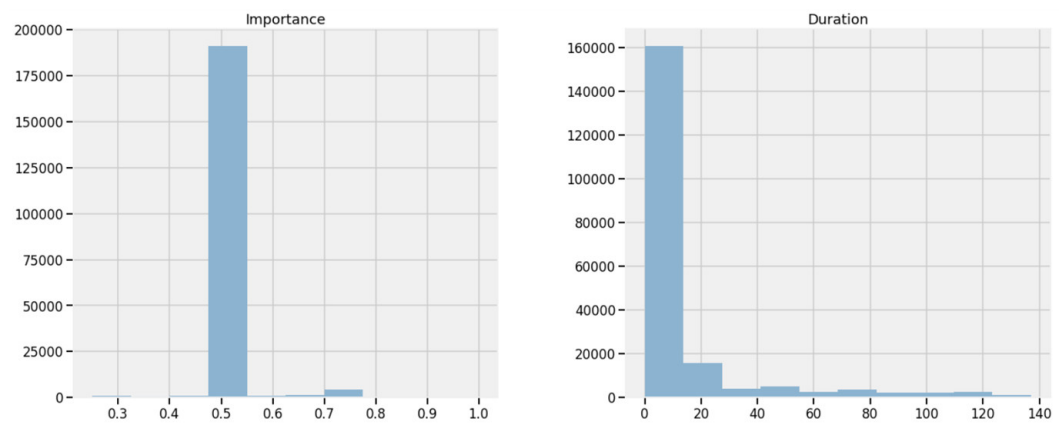
Figure 14. Importance and duration of transactions by partner.

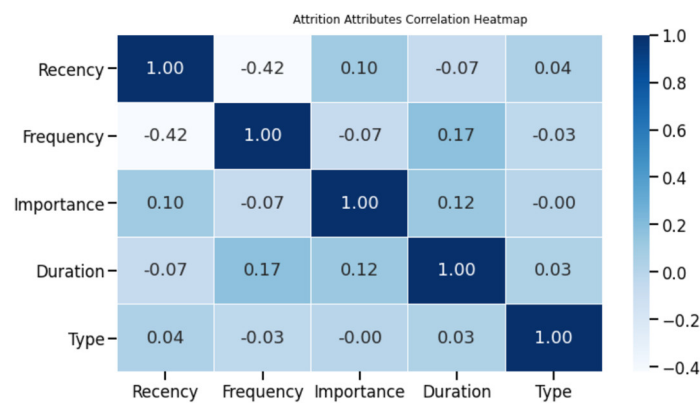The following step displays the correlation matrix (Figure 15).

Figure 15. Correlation matrix.

Table 10. RFIDT, MinMaxScaler (0,5).

	Recency	Frequency	Importance	Duration	Type
count	200,615	200,615	200,615	200,615	200,615
mean	1457	0.77	1.70	0.39	0.011
std	0.97	1.08	0.30	0.93	0.10
min	0.00	0.00	0.00	0.00	0.00
25%	0.62	0.00	1.67	0.00	0.00
50%	1.32	0.36	1.67	0.00	0.00
75%	2.17	1.07	1.67	0.11	0.00
max	5.00	5.00	5.00	5.00	1.00

5.2. ML Algorithms Evaluation

Next, to analyze whether to apply the set of interpretability algorithms described in Section 3, a set of pre-model techniques will be applied to obtain the first knowledge of the dataset. Subsequently, each of the models described in Table 7 is evaluated through a cross-validation process (K-fold), and the receiver operating characteristic (ROC) and area under the curve (AUC), and accuracy mean are analyzed, to obtain the optimal algorithm for the case study.

The results obtained can be analyzed in Table 11,

Table 11. Evaluation of applied predictive algorithms.

	Algorithm	ROC AUC Mean	ROC AUC STD	Accuracy Mean	Accuracy STD
6	XGboost	70.40	1.24	98.75	0.08
5	Gaussian NB	67.25	1.78	98.83	0.09
0	Logistic Regression	67.14	1.76	61.02	0.44
1	Random Forest	60.40	1.34	98.80	0.08
3	KNN	55.27	0.85	98.93	0.07
4	Decision Tree Classifier	52.32	1.41	98.43	0.10
2	SVM	52.16	4.77	98.94	0.07

The model selected according to the procedure described is XGboost, a black box model, and responds to the need to use interpretability.

XGBoost is used in supervised learning problems, and the objective is to predict a target variable y_i from a set of variables x_i. A common example of supervised learning is linear regression, where the prediction of a variable y_i is obtained as, $y_i = \sum_k (\beta_k x_{ik})$, and the characteristics making up the input are weighted by weights β_k.

When we talk about training a model, we are talking about adjusting parameters β, for which we need to define the objective function that best fits the training data x_i, and produce as a response the best fitted value to y_i. A notable feature of the objective functions is that they consist of two parts, the training loss, and the regularization term:

$$Goal(\beta) = L(\beta) + \Omega(\beta) \tag{3}$$

where L is the training loss function and Ω is the degree of complexity of explainability of the model. In this case, the model is defined as follows [79]:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \ f_k \in F \tag{4}$$

where K is the total number of trees and f_k is a function in the function space F.

The objective is to mix classification trees to measure which of the combinations is the best for our model.

5.3. Data Unbalancing

Our churn class has very few samples in relation to the majority class (no churn = partner). This causes an imbalance of data, and therefore the training of the model will be deficient, responding in an unbalanced way to the detection of the dropout pattern to be predicted.

In the first analysis performed using XGBoost, the model gave us the following results (Figure 16).

Figure 16. Evaluation of XGBoost predictions, unbalanced data.

As can be seen, the model presents an extraordinary result, with almost 99% prediction, but based on an accuracy of 100% in the majority class (no churn = 0) and 0% in the minority class (churn = 1). It is therefore essential to perform a data unbalancing process, and we must try to increase the degree of prediction of the minority class.

To deal with the possible problem of data imbalance in the dropout class, we have resorted to modifying the XGBoost training algorithm by introducing a value to the hyperparameter *scale_pos_weight*, which is designed to adjust the behavior of the algorithm in unbalanced classification problems. A suitable value for this parameter is found in estimating a correction corresponding to the inverse of the class distribution. For example, in a dataset where the ratio between the minority and majority class is 1 to 100, it is correct to apply a value of *scale_pos_weight* = 100 [80].

In addition, *scale_pos_weight* has been combined with the Smote-Tomek process [81], which consists of simultaneously applying a subsampling and oversampling algorithm to the dataset. This obtains in one of the different model trainings the following best result (Figures 17 and 18).

Precision:	0.7628			
	Precision	recall	f1-score	support
Churn (No)	0.99	0.76	0.86	59548
Churn (Yes)	0.03	0.57	0.05	637
accuracy			0.76	60185
macro avg	0.51	0.67	0.46	60185
weighted avg	0.98	0.76	0.86	60185

Confusion Matrix
[[45549 13999]
 [274   363]]

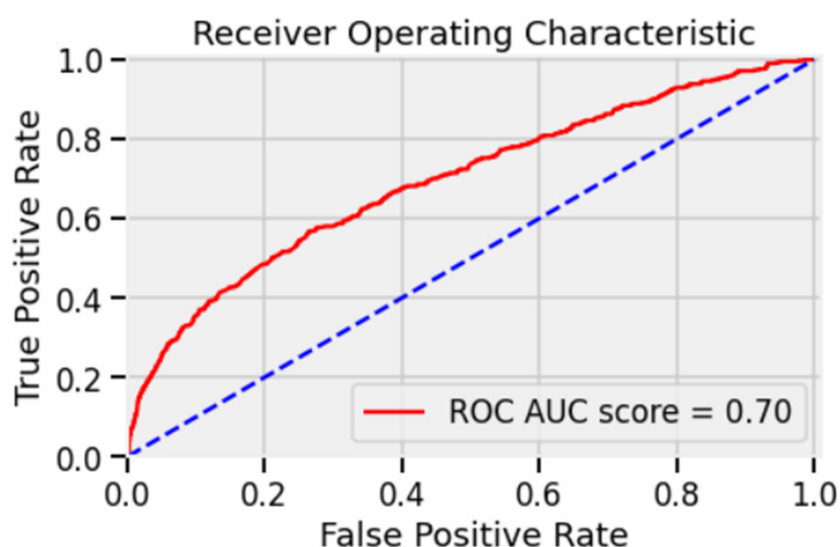**Figure 17.** Evaluation XGBoost predictions, balanced data.

Figure 18. ROC/AUC curve of balanced model XGBoost vs random classifier (blue line).

In exchange for losing precision in the majority class, we gain it in the minority class.

The next step, because it is *XGBoost* of a black box algorithm, consists of developing the interpretability process described in Sections 3 and 4 of this paper.

5.4. Interpretability Techniques Application

Next, the interpretability model detailed in Sections 3 and 4.3 of this paper is applied, and a study of the results obtained will be carried out.

5.4.1. Partial Dependency Diagram (PDP)

When we consider more than a certain number of variables, it is necessary to analyze the partial dependence of one or two variables in relation to the prediction of the response variable. Through the PDP diagram, we can perform this type of analysis, and the shaded area represents the confidence interval [53]. As can be seen in the graphs, normalization has been carried out between 0 and 5.

The diagrams in Figures 19–21 show the influence of recency, frequency, and duration on the prediction of abandonment, and the diagram in Figure 22 shows the degree of correlation between recency and frequency. In the graphs presented, importance has not been considered, since the value is biased towards the mean value (M).
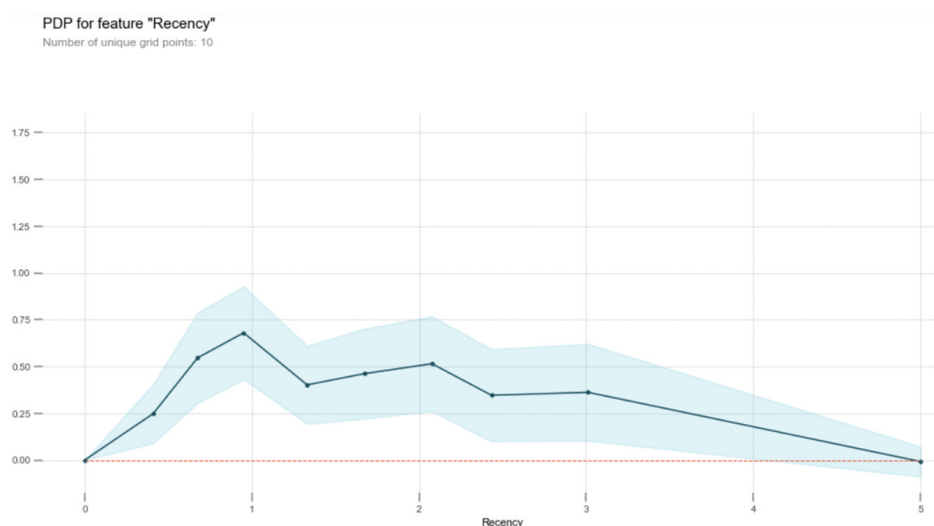

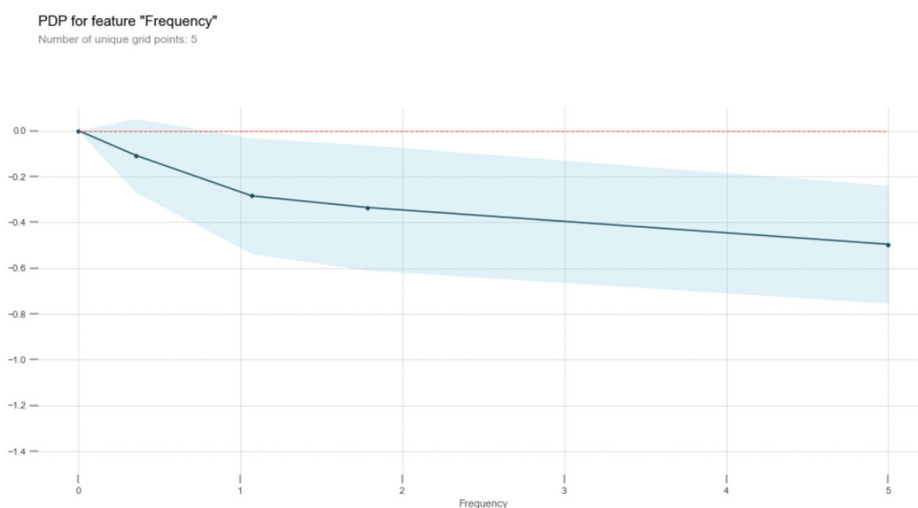
Figure 19. PDP partial dependence diagram (recency).

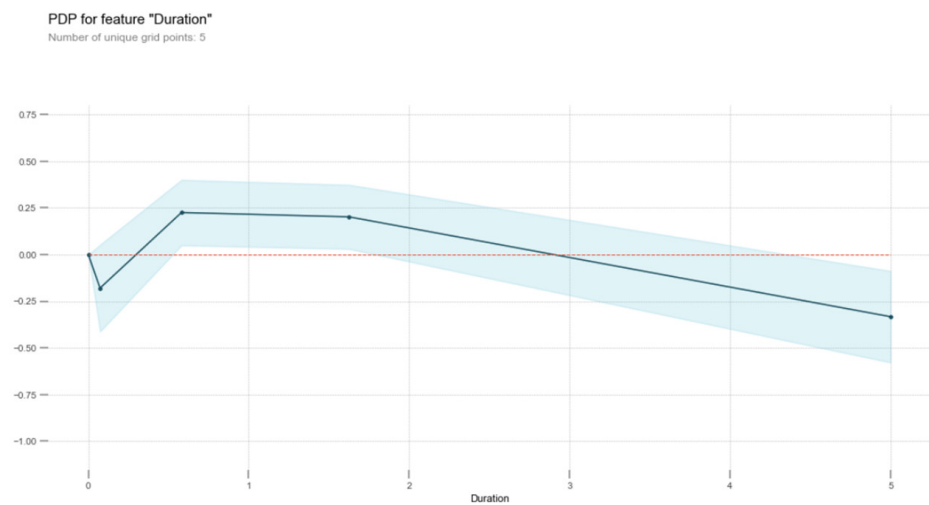
Figure 20. PDP partial dependence diagram (frequency).

PDP for feature "Duration"
Number of unique grid points: 5

Figure 21. Partial dependency diagram PDP (duration).

PDP interact for "Recency" and "Frequency"
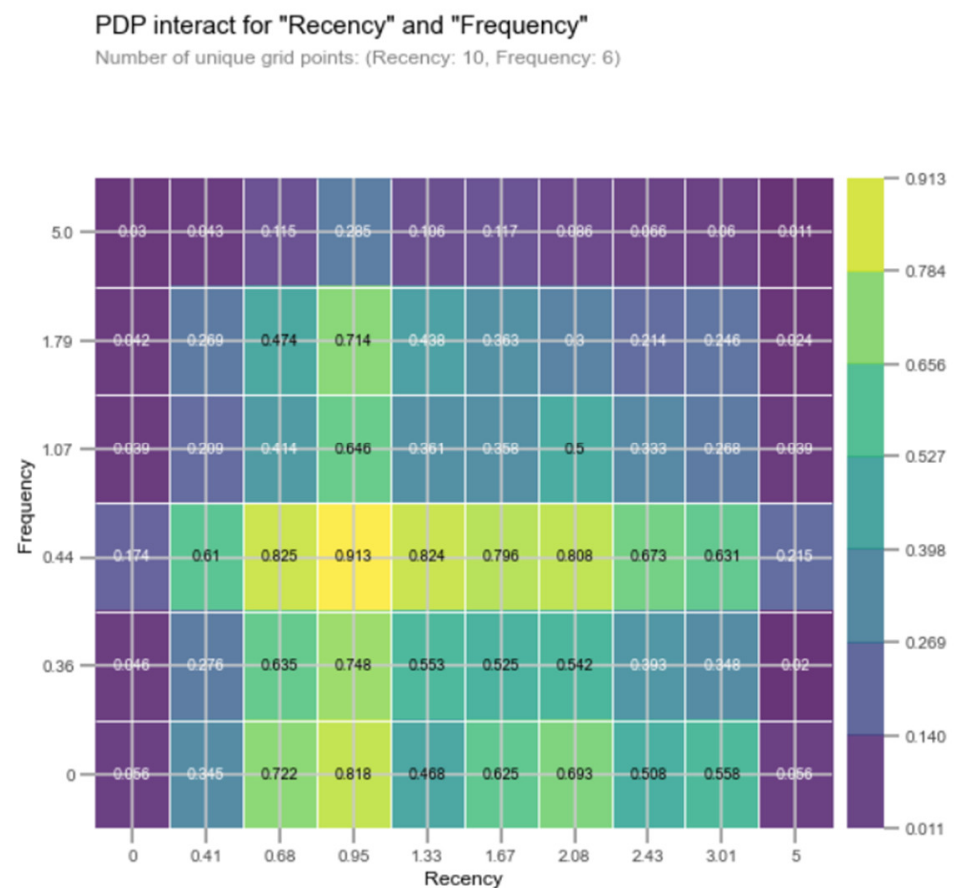Number of unique grid points: (Recency: 10, Frequency: 6)

Figure 22. Bivariate PD partial interaction diagram (frequency, recency).

Finally, we include in this section the ICE plots, which are similar to the PD plots but offer a more detailed view on the behavior of nearly similar clusters around the mean curve of the PD plot. The ICE algorithm provides insight into the various variants of conditional relationships estimated by the black box (Figure 23).
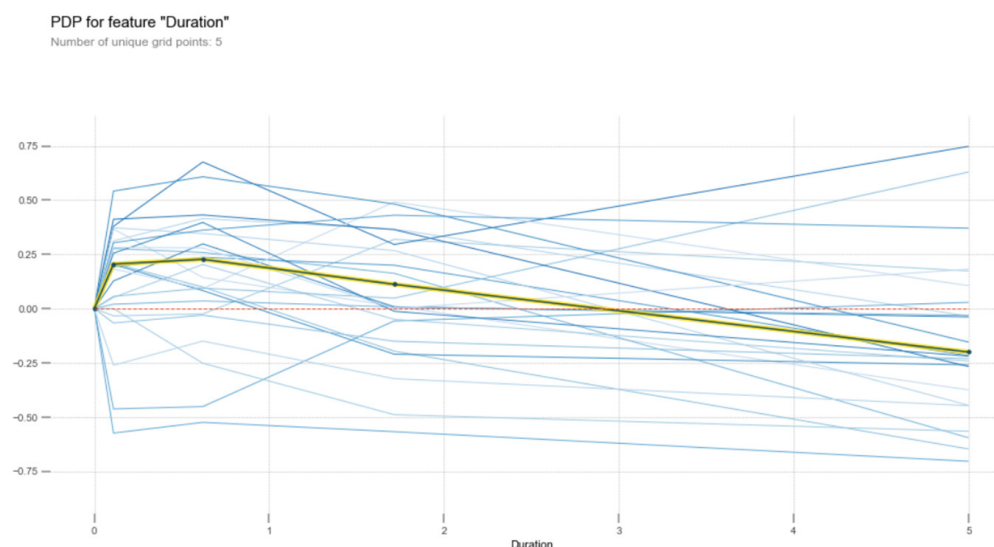
Figure 23. ICE graph (duration).

5.4.2. Feature Importance (ELI5)

The concept of the importance of characteristics is simple: it is a matter of assessing the importance of a given characteristic by calculating the increase in prediction error after making a permutation of this characteristic [82].

To make random tree predictions more interpretable, each model prediction can be presented as a sum of feature contributions (plus bias), showing how the features lead to a particular prediction. ELI5 does this by showing the weights of each feature, indicating their influence on the final prediction decision across all trees. This is a good step in the direction of agnostic interpretation of the model, but not fully agnostic, as we will see later, using LIME. The results obtained are shown below in Table 12.

Table 12. Feature importance (ELI5).

Weight	Feature
0.3610	Importance
0.2542	Frequency
0.1998	Duration
0.1849	Recency

According to the results obtained, the criterion importance has the greatest weight in the evaluation of the characteristics, followed by frequency, duration, and recency. However, as we have seen, the importance is biased towards the mean values within the whole sample.

5.4.3. Local Substitute (LIME)

LIME is a local model and works by checking what happens to the predictions when variations in the input data are introduced [52]. For this purpose, LIME generates new datasets with these variations, thus obtaining sets of predictions. The results applied to the model under study can be seen below.

In the first case, Figure 24, a record has been chosen in which there is a 99.64% success rate in the prediction of non-abandonment. In Figure 25, the prediction of abandonment is 75.49%, which corresponds to a partner who has left the partner channel.
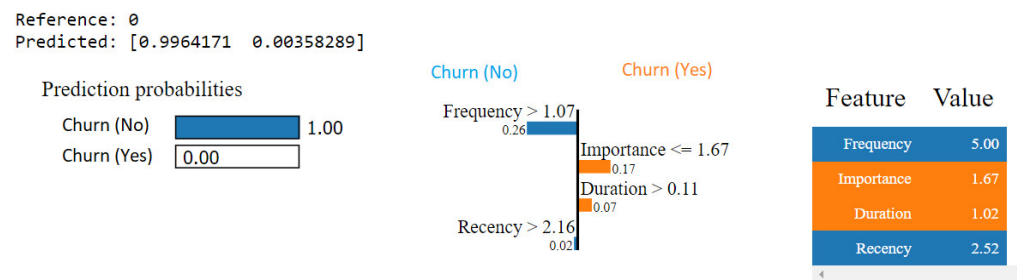
```
Reference: 0
Predicted: [0.9964171  0.00358289]
```


Figure 24. Attrition prediction based on characteristics (churn = no).

```
Reference: 1
Predicted: [0.24512208 0.7548779 ]
```
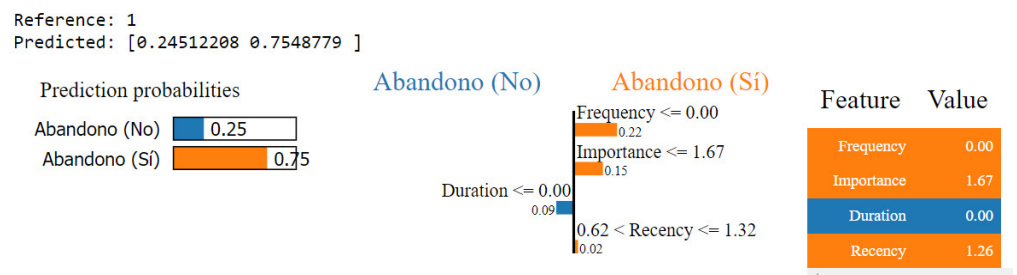

Figure 25. Attrition prediction based on characteristics (churn = yes).

5.4.4. SHAP Values

The objective of the SHAP interpretability model is to be able to provide an explanation for an instance x based on the contribution of each of the characteristics to the prediction (Figure 26) [54].
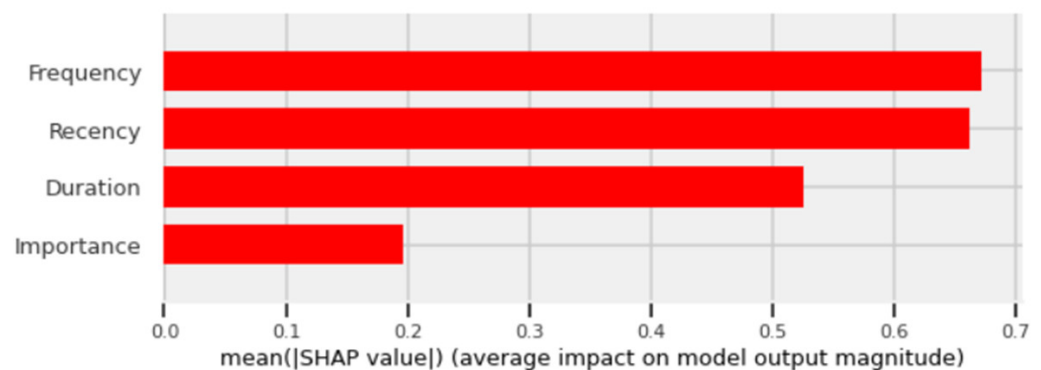

Figure 26. Features importance, SHAP.

As for the importance of the characteristics, the model is a priori more reliable than ELI5, since the importance measure is not a significant parameter for the model we are studying, remembering that the default value is 0.5 in most of the samples.

Figures 27 and 28 show the prediction of SHAP values for the partner represented, respectively, in Figures 24 and 25 by the LIME mode.

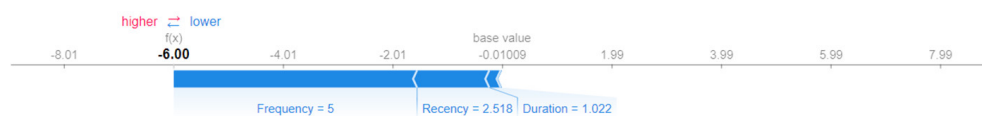**Figure 27.** SHAP value prediction (churn = yes).

Figure 28. SHAP value prediction (churn = no).

It is significant to note that, in the segmentation study, for the partner represented in Figures 25 and 27, it would be identified with cluster #1 and, therefore, it is a recently incorporated partner that needs to take its first steps to start and, with a high risk of abandonment, and in fact, it abandoned. On the contrary, the partner represented in Figures 24 and 28, would be identified with cluster #4. The profile corresponds to a partner with a large installed base of partners that uses the Contact Center to solve specific problems, therefore, with a low risk of abandonment.

5.4.5. Skater

Because of its interest in the use case under study, we have introduced Skater, since it allows both global and local interpretation; for global explanations, it is based on the use of PDP, and for local explanations it is based on LIME. It corresponds to a unified framework recently introduced and under development [83].

The results obtained are shown below.

In the graph in Figure 29, the forecasts obtained with Skater fit with those of the SHAP model in Figure 26. Analyzing in more detail, we obtain the dependence graphs, Figure 30, where the relationship between the classification variable with respect to each of the characteristics can be seen.
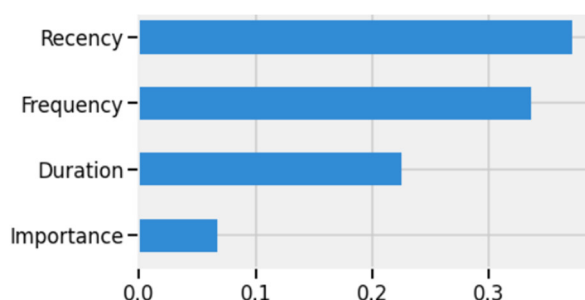

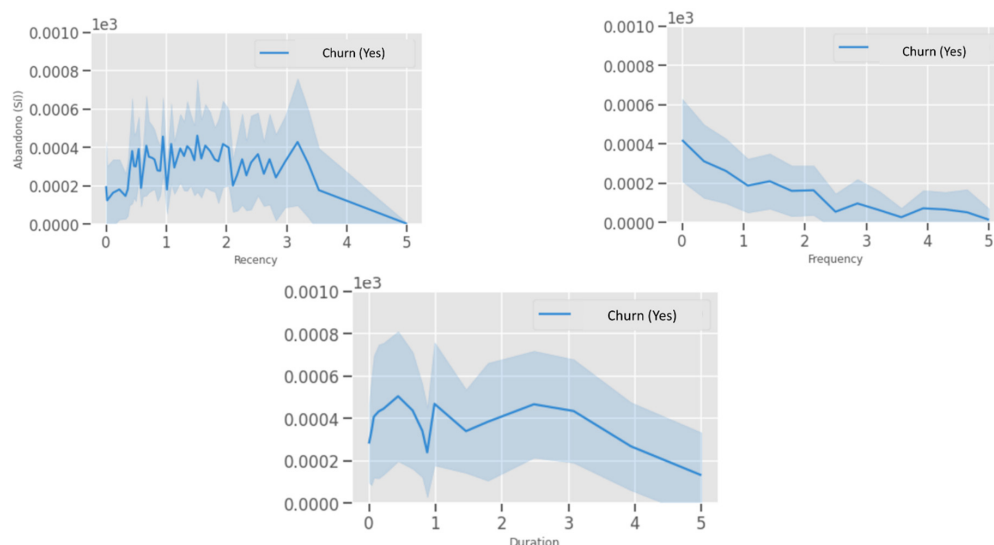**Figure 29.** Features importance, Skater.

Figure 30. Skater dependency graphs.

The graphs in Figure 30 confirm the trend seen in the PDP model. The tendency to abandon is centered on those partners with medium and low recency values (L, M), low frequency values (L), and with low, medium, and high duration of incidents (L, M, H). As mentioned above, they correspond to partners who have recently entered the channel and have not managed to mature sufficiently to be able to market and implement the software's solutions.

6. Discussion

The objective of this study is to complete a working methodology for the analysis of the interpretability in ML models, using agnostic models (global and local) that have been used for the analysis of the explainability of partner churn in a B2B environment.

Applying the methodological procedure designed in the RFID model, in a typical B2C environment, it is possible to achieve high levels of interpretability in the customer churn rate, since there is a direct dependence between the frequency, recency, duration, and importance of incidents and the churn rate. In the B2B model that has been proposed, as mentioned above, the relationship between the partner (distributor = customer) and the software company is very close. That is to say, the partner has had to develop a whole line of business and investment in its relationship with the software company, which translates into:

- Adaptation to the business plans established by the software company: number of people trained, sales commitment, and annual turnover.
- Highly demanding training process for each partner's technical, commercial, and pre-sales personnel.
- As the company grows, it becomes necessary to hire specialized personnel, with the consequent related economic cost.

Therefore, as the partner grows in sales, the relationship with the software company is closer and, consequently, the abandonment rate of that partner is lower (see Figure 11, Table 8). On the other hand, we find recently incorporated partners that do not reach the maturation process described above tend to abandon before their investment in the business model proposed by the software company is greater. Because of the above, and for the business case in question, the more interactions that take place between the partner and Contact Center, the lower the probability of abandonment (high frequency and recency). Logically, the partner who is more established in the channel has a greater number of customers to serve and, therefore, the greater the number of interactions with the software company.

Regarding the working methodology applied to solve the problem of interpretability in ML models, an in-depth study of agnostic interpretability models has been carried out and applied to the context of the problem to be addressed. This methodology helps us to interpret the decisions made by black box algorithms, and uses agnostic interpretability models, not dependent on the ML model, and therefore its flexibility and applicability to any type of learning model is guaranteed.

The innovation of this paper is based on three differentiating factors:

- Use of a customer assessment based on the RFID model, and on the set of interactions between the customer and the brand through the Contact Center.
- Application of ML models oriented to customer churn rate prediction in B2B environments. According to the literature review, in Section 2, there is not much research production in B2B environments.
- Application of a working methodology that provides an agnostic interpretability procedure, extensible to any predictive model in B2B or B2C environments in which we must predict a variable (classification or regression) based on the rest of the characteristics.

7. Conclusions and Future Work

In conclusion, the present work develops a completely new line of evaluation of the prediction of customer abandonment, from the point of view of the customer's interactions

with the Contact Center. Based on the RFID model, it is possible to determine the brand's evaluation of the customer through these interactions and, consequently, to analyze each of the characteristics that make up the model and their weight in the final evaluation of the binary target variable abandonment. An additional novelty is the application of the model to a B2B environment, for which the literature is scarce, in models that determine the prediction of abandonment, and consequently in the application of interpretability (XAI).

As an example, we have applied the implemented model to a dataset from a software manufacturing company with a large network of partners (customers) distributed around the world. A set of predictive models has been applied on unbalanced data; the churn rate represents 1.06% of the total sample. Therefore, data balancing techniques had to be applied, adjusting the behavior of the algorithm in unbalanced classification problems, in addition to simultaneously applying subsampling and oversampling algorithms to the dataset.

Then, the described working procedure has been applied, consisting of the application of successive interpretability techniques. The results obtained through the implemented interpretability methodology reveal that the conclusions are aligned with the clustering implemented with the RFID model. The more the partners interact with the Contact Center, the less propensity they have to abandon the relationship with the software company. The clustering (k-means) developed through the RFID model classifies partners into five groupings, and the abandonment prediction fits perfectly with the clusters in which the partner has a lower rating on the recency and frequency variables.

As future work, we propose the following:

- One metric of concern for Contact Centers is the employee attrition rate; turnover in the Contact Center is very high, mainly due to work and emotional demands [84]. Using the procedure designed in this paper to analyze, predict, and interpret Contact Center staff attrition rates would be a major challenge.
- Extend the model to any industry and any B2B and B2C environment, with a focus on retail, insurance, banking, and service delivery.
- Finally, consolidate the model with what customers think, i.e., contrast the model with the customer satisfaction score (NPS), the customer's assessment of the brand or of each of the customer satisfaction score (CSAT) interactions, or the customer effort score (CES) [13]. In addition to adding other factors such as the metrics introduced in the customer engagement value model, the following metrics can also be added (CEV) [85].
- In certain fields, such as decision making in image detection processes, it will be necessary to adapt the interpretability model described by incorporating the use of artificial neural networks [76,86,87]; in future works, an extension of the XAI model will be proposed by adapting these improvements.

Author Contributions: Conceptualization, G.M.D.; Data curation, G.M.D.; Formal analysis, G.M.D. and R.A.C.; Investigation, G.M.D.; Methodology, G.M.D. and R.A.C.; Project administration, G.M.D.; Resources, G.M.D., J.J.G. and R.A.C.; Software, G.M.D.; Supervision, R.A.C.; Validation, G.M.D., J.J.G. and R.A.C.; Visualization, G.M.D.; Writing—original draft, G.M.D.; Writing—review and editing, G.M.D., J.J.G. and R.A.C. All authors will be informed about each step of manuscript processing including submission, revision, revision reminder, etc., via emails from our system or the assigned Assistant Editor. All authors have read and agreed to the published version of the manuscript.

References

1. Jain, D.; Singh, S.S. Customer lifetime value research in Marketing: A review and future directions. *J. Interact. Mark.* **2002**, *16*, 34–46. [CrossRef]
2. Mulhern, F. Understanding and Using Customer Loyalty and Customer Value. *J. Relationsh. Market.* **2007**, *6*, 59–86. [CrossRef]
3. Marín Díaz, G.; Carrasco, R.A.; Gómez, D. RFID: A Fuzzy Linguistic Model to Manage Customers from the Perspective of Their Interactions with the Contact Center. *Mathematics* **2021**, *9*, 2362. [CrossRef]
4. Bridges, E.; Goldsmith, R.E.; Hofacker, C.F. Attracting and retaining online buyers: Comparing B2B and B2C customers. In *Advances in Electronic Marketing*; IGI Global: Hershey, PA, USA, 2005; pp. 1–27. [CrossRef]
5. Gordini, N.; Veglio, V. Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry. *Ind. Mark. Manag.* **2017**, *62*, 100–107. [CrossRef]
6. Kalwani, M.U.; Narayandas, N. Long-Term Manufacturer-Supplier Relationships: Do They Pay off for Supplier Firms? *J. Mark.* **1995**, *59*, 1–16. [CrossRef]
7. Eriksson, K.; Vaghult, A.L. Customer retention, purchasing behavior and relationship substance in professional services. *Ind. Mark. Manag.* **2000**, *29*, 363–372. [CrossRef]
8. Hughes, A.M. *Strategic Database Marketing*; Probus Publishing Company: Chicago, IL, USA, 1994; ISBN 1557385513.
9. Mirkovic, M.; Lolic, T.; Stefanovic, D.; Anderla, A.; Gracanin, D. Customer Churn Prediction in B2B Non-Contractual Business Settings Using Invoice Data. *Appl. Sci.* **2022**, *12*, 5001. [CrossRef]
10. Chang, H.H.; Tsay, S.F. Integrating of SOM and K-Mean in Data Mining Clustering: An Empirical Study of CRM and Profitability Evaluation. *J. Inf. Manag.* **2004**, *11*, 161–203.
11. Peker, S.; Kocyigit, A.; Eren, P.E. LRFMP model for customer segmentation in the grocery retail industry: A case study. *Mark. Intell. Plan.* **2017**, *35*, 544–559. [CrossRef]
12. Payne, A.; Frow, P. The role of multichannel integration in customer relationship management. *Ind. Mark. Manag.* **2004**, *33*, 527–538. [CrossRef]
13. Saberi, M.; Khadeer Hussain, O.; Chang, E. Past, present and future of contact centers: A literature review. *Bus. Process Manag. J.* **2017**, *23*, 574–597. [CrossRef]
14. Molnar, C. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*; Lulu.com: Morrisville, NC, USA, 2020.
15. Gunning, D.; Stefik, M.; Choi, J.; Miller, T.; Stumpf, S.; Yang, G.Z. XAI-Explainable artificial intelligence. *Sci. Robot.* **2019**, *4*, eaay7120. [CrossRef] [PubMed]
16. Reinartz, W.J.; Kumar, V. The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration. *J. Mark.* **2003**, *67*, 77–99. [CrossRef]
17. Garimella, B.; Prasad, G.V.S.N.R.V.; Prasad, M.H.M.K. Churn prediction using optimized deep learning classifier on huge telecom data. *J. Ambient Intell. Humaniz. Comput.* **2021**, 1–22. [CrossRef]
18. Chayjan, M.R.; Bagheri, T.; Kianian, A.; Someh, N.G. Using data mining for prediction of retail banking customer's churn behaviour. *Int. J. Electron. Bank.* **2020**, *2*, 303–320. [CrossRef]
19. Jamjoom, A.A. The use of knowledge extraction in predicting customer churn in B2B. *J. Big Data* **2021**, *8*, 1–14. [CrossRef]
20. Lismont, J.; Ram, S.; Vanthienen, J.; Lemahieu, W.; Baesens, B. Predicting interpurchase time in a retail environment using customer-product networks: An empirical study and evaluation. *Expert Syst. Appl.* **2018**, *104*, 22–32. [CrossRef]
21. Gopal, P.; MohdNawi, N. Bin A Survey on Customer Churn Prediction using Machine Learning and data mining Techniques in E-commerce. In Proceedings of the 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Brisbane, Australia, 8–10 December 2021; pp. 1–8.
22. Stevens, R.P. B-to-B Customer Retention: Seven Strategies for keeping your customers. *B2B Mark. Trends* **2005**, 1–13. Available online: http://www.ruthstevens.com/ (accessed on 18 September 2022).
23. Mora Cortez, R.; Johnston, W.J. The future of B2B marketing theory: A historical and prospective analysis. *Ind. Mark. Manag.* **2017**, *66*, 90–102. [CrossRef]
24. Venkatesan, R.; Kumar, V. A Customer Lifetime Value Framework for Customer Selection and Resource Allocation Strategy. *J. Mark.* **2004**, *68*, 106–125. [CrossRef]
25. Figalist, I.; Elsner, C.; Bosch, J.; Olsson, H.H. Customer churn prediction in B2B contexts. *Lect. Notes Bus. Inf. Process.* **2019**, *370*, 378–386. [CrossRef]
26. Janssens, B.; Bogaert, M.; Bagué, A.; Van den Poel, D. B2Boost: Instance-dependent profit-driven modelling of B2B churn. *Ann. Oper. Res.* **2022**, 1–27. [CrossRef]
27. Jahromi, A.T.; Stakhovych, S.; Ewing, M. Managing B2B customer churn, retention and profitability. *Ind. Mark. Manag.* **2014**, *43*, 1258–1268. [CrossRef]
28. Chen, K.; Hu, Y.H.; Hsieh, Y.C. Predicting customer churn from valuable B2B customers in the logistics industry: A case study. *Inf. Syst. E-Bus. Manag.* **2015**, *13*, 475–494. [CrossRef]
29. Zhang, Z.; Ravivanpong, P.; Beigl, M. Predicting B2B Customer Churn for Software Maintenance Contracts. In Proceedings of the 34th IBIMA Conference, Madrid, Spain, 13–14 November 2019; pp. 6593–6603, ISBN 978-0-9998551-3-3.
30. Hopmann, J.; Thede, A. Applicability of customer churn forecasts in a non-contractual setting. In *Innovations in Classification, Data Science, and Information Systems*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 330–337. [CrossRef]

31. Sheikh, A.; Ghanbarpour, T.; Gholamiangonabadi, D. A Case Study of Fintech Industry: A Two-Stage Clustering Analysis for Customer Segmentation in the B2B Setting. *J. Bus.-to-Bus. Mark.* **2019**, *26*, 197–207. [CrossRef]
32. De Caigny, A.; Coussement, K.; Verbeke, W.; Idbenjra, K.; Phan, M. Uplift modeling and its implications for B2B customer churn prediction: A segmentation-based modeling approach. *Ind. Mark. Manag.* **2021**, *99*, 28–39. [CrossRef]
33. Barfar, A.; Padmanabhan, B.; Hevner, A. Applying behavioral economics in predictive analytics for B2B churn: Findings from service quality data. *Decis. Support Syst.* **2017**, *101*, 115–127. [CrossRef]
34. Lee, H.; Choi, H.; Koo, Y. Lowering customer's switching cost using B2B services for telecommunication companies. *Telemat. Informatics* **2018**, *35*, 2054–2066. [CrossRef]
35. Liu, A.H.; Chugh, R.; Noel Gould, A. Working smart to win back lost customers the role of coping choices and justice mechanisms. *Eur. J. Mark.* **2016**, *50*, 397–420. [CrossRef]
36. D'Haen, J.; Van Den Poel, D.; Thorleuchter, D. Predicting customer profitability during acquisition: Finding the optimal combination of data source and data mining technique. *Expert Syst. Appl.* **2013**, *40*, 2007–2012. [CrossRef]
37. Schaeffer, S.E.; Rodriguez Sanchez, S.V. Forecasting client retention—A machine-learning approach. *J. Retail. Consum. Serv.* **2020**, *52*, 101918. [CrossRef]
38. Gattermann-Itschert, T.; Thonemann, U.W. How training on multiple time slices improves performance in churn prediction. *Eur. J. Oper. Res.* **2021**, *295*, 664–674. [CrossRef]
39. Marín Díaz, G.; Carrasco, R.A.; Gómez, D. Interpretability Challenges in Machine Learning Models. In *Moving Technology Ethics at the Forefront of Society, Organisations and Governments*; Universidad de La Rioja: La Rioja, Spain, 2021; pp. 205–217.
40. Britannica, E. MYCIN. 2018. Available online: https://www.britannica.com/technology/MYCIN (accessed on 18 September 2022).
41. Clancey, W.J. *The GUIDON Program*; MIT Press: Cambridge, MA, USA, 1987; ISBN 978-0471815242.
42. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef] [PubMed]
43. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **2019**, *267*, 1–38. [CrossRef]
44. Lipton, P. Contrastive Explanation. In *Royal Institute of Philosophy Supplement*; Cambridge University Press: Cambridge, UK, 2010; Volume 27, pp. 247–266. [CrossRef]
45. Kahneman, D.; Tversky, A. *The Simulation Heuristic*; The University of British Columbia: Vancouver, BC, Canada, 1981.
46. Nickerson, R.S. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Z. Neurol.* **1998**, *199*, 145–150. [CrossRef]
47. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic attribution for deep networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 3319–3328.
48. Bastani, O.; Kim, C.; Bastani, H. Interpreting blackbox models via model extraction. *arXiv* **2017**, arXiv:1705.08504.
49. Tan, S.; Caruana, R.; Hooker, G.; Lou, Y. Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, Online, 27 December 2018; pp. 303–310. [CrossRef]
50. Carvalho, D.V.; Pereira, E.M.; Cardoso, J.S. Machine learning interpretability: A survey on methods and metrics. *Electronics* **2019**, *8*, 832. [CrossRef]
51. Ribeiro, M.T.; Singh, S.; Guestrin, C. Model-Agnostic Interpretability of Machine Learning. *arXiv* **2016**, arXiv:1606.05386.
52. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Online, 13 August 2016; pp. 97–101. [CrossRef]
53. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
54. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *2017*, 4766–4775.
55. Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation importance: A corrected feature importance measure. *Bioinformatics* **2010**, *26*, 1340–1347. [CrossRef] [PubMed]
56. Goldstein, A.; Kapelner, A.; Bleich, J.; Pitkin, E. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *J. Comput. Graph. Stat.* **2015**, *24*, 44–65. [CrossRef]
57. Zafar, M.R.; Khan, N.M. DLIME: A Deterministic Local Interpretable Model-Agnostic Explanations Approach for Computer-Aided Diagnosis Systems. *arXiv* **2019**, arXiv:1906.10263.
58. Gurumoorthy, K.S.; Dhurandhar, A.; Cecchi, G.; Aggarwal, C. Efficient Data Representation by Selecting Prototypes with Importance Weights. In Proceedings of the 2019 IEEE International Conference on Data Mining (ICDM), Beijing, China, 8–11 November 2019; pp. 260–269. [CrossRef]
59. Leung, C.K.; Pazdor, A.G.M.; Souza, J. Explainable Artificial Intelligence for Data Science on Customer Churn. In Proceedings of the 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA), Porto, Portugal, 6–9 October 2021; pp. 1–10. [CrossRef]
60. Na, K.; Lee, J.; Kim, E.; Lee, H. A Securities Company's Customer Churn Prediction Model and Causal Inference with SHAP Value. *J. Bigdata* **2020**, *5*, 215–229.
61. Ullah, I.; Rios, A.; Gala, V.; McKeever, S. Explaining deep learning models for tabular data using layer-wise relevance propagation. *Appl. Sci.* **2022**, *12*, 136. [CrossRef]
62. Shafique, U.; Qaiser, H. A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). *Int. J. Innov. Sci. Res.* **2014**, *12*, 217–222.

63. Herrera, F.; Martínez, L. A 2-tuple fuzzy linguistic representation model for computing with words. *IEEE Trans. Fuzzy Syst.* **2000**, *8*, 746–752. [CrossRef]
64. Herrera, F.; Martínez, L.; Sánchez, P.J. Managing non-homogeneous information in group decision making. *Eur. J. Oper. Res.* **2005**, *166*, 115–132. [CrossRef]
65. Saaty, T.L. The Analytic Hierarchy. In *Process: Planning, Priority Setting, Resource Allocation McGraw-Hill*; International Book Co.: New York, NY, USA, 1980; 287p, ISBN 0070543712 9780070543713.
66. Khaira, A.; Dwivedi, R.K. A State of the Art Review of Analytical Hierarchy Process. *Mater. Today Proc.* **2018**, *5*, 4029–4035. [CrossRef]
67. Ishizaka, A.; Labib, A. Expert Systems with Applications Review of the main developments in the analytic hierarchy process. *Rev. Main Dev. Anal. Hierarchy Process* **2011**, *38*, 1–24.
68. Martínez, R.G.; Carrasco, R.A.; Sanchez-Figueroa, C.; Gavilan, D. An rfm model customizable to product catalogues and marketing criteria using fuzzy linguistic models: Case study of a retail business. *Mathematics* **2021**, *9*, 1836. [CrossRef]
69. Liu, H.; Cocea, M.; Gegov, A. Interpretability of computational models for sentiment analysis. *Stud. Comput. Intell.* **2016**, *639*, 199–220. [CrossRef]
70. Lipton, Z.C. The mythos of model interpretability. *Commun. ACM* **2018**, *61*, 35–43. [CrossRef]
71. Duval, A. *Explainable Artificial Intelligence (XAI) Explainable Artificial MA4K9 Scholary Report*; Mathematics Institute, The University of Warwick: Coventry, UK, 2019; pp. 1–53. [CrossRef]
72. Altman, D.; Yom-Tov, G.B.; Olivares, M.; Ashtar, S.; Rafaeli, A. Do customer emotions affect agent speed? An empirical study of emotional load in online customer contact centers. *Manuf. Serv. Oper. Manag.* **2021**, *23*, 854–875. [CrossRef]
73. Casella, G.; Fienberg, S.; Olkin, I. *An Introduction to Statistical Learning*; Springer: New York, NY, USA, 2013; ISBN 9780387781884.
74. Bradley, A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **1997**, *30*, 1145–1159. [CrossRef]
75. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36. [CrossRef]
76. Montavon, G.; Samek, W.; Müller, K.R. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process. A Rev. J.* **2018**, *73*, 1–15. [CrossRef]
77. Dorogush, A.V.; Ershov, V.; Gulin, A. CatBoost: Gradient boosting with categorical features support. *arXiv* **2018**, arXiv:1810.11363.
78. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3147–3155.
79. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *J. Assoc. Physicians India* **2016**, 785–794. [CrossRef]
80. Wang, C.; Deng, C.; Wang, S. Imbalance-XGBoost: Leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. *Pattern Recognit. Lett.* **2020**, *136*, 190–197. [CrossRef]
81. Batista, G.E.; Bazzan, A.L.; Monard, M.C. Balancing Training Data for Automated Annotation of Keywords. In Proceedings of the II Brazilian Workshop on Bioinformatics, Macaé, Brazil, 3–5 December 2003.
82. Fisher, A.; Rudin, C.; Dominici, F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.* **2019**, *20*, 1–81.
83. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable ai: A review of machine learning interpretability methods. *Entropy* **2021**, *23*, 18. [CrossRef] [PubMed]
84. Rameshbabu, A.; Reddy, D.M.; Fleming, R. Correlates of negative physical health in call center shift workers. *Appl. Ergon.* **2013**, *44*, 350–354. [CrossRef] [PubMed]
85. Kumar, V.; Aksoy, L.; Donkers, B.; Venkatesan, R.; Wiesel, T.; Tillmanns, S. Undervalued or overvalued customers: Capturing total customer engagement value. *J. Serv. Res.* **2010**, *13*, 297–310. [CrossRef]
86. Guo, H.; Zhuang, X.; Chen, P.; Alajlan, N.; Rabczuk, T. Stochastic deep collocation method based on neural architecture search and transfer learning for heterogeneous porous media. *Eng. Comput.* **2022**, 1–26. [CrossRef]
87. Guo, H.; Zhuang, X.; Chen, P.; Alajlan, N.; Rabczuk, T. Analysis of three-dimensional potential problems in non-homogeneous media with physics-informed deep collocation method using material transfer learning and sensitivity analysis. *Eng. Comput.* **2022**, 1–22. [CrossRef]