information

Article

Understanding Online Purchases with Explainable Machine Learning

João A. Bastos * and Maria Inês Bernardes

Lisbon School of Economics and Management (ISEG) and REM, Universidade de Lisboa,
1649-004 Lisboa, Portugal; inesbernardes@aln.iseg.ulisboa.pt
* Correspondence: jbastos@iseg.ulisboa.pt

Abstract: Customer profiling in e-commerce is a powerful tool that enables organizations to create personalized offers through direct marketing. One crucial objective of customer profiling is to predict whether a website visitor will make a purchase, thereby generating revenue. Machine learning models are the most accurate means to achieve this objective. However, the opaque nature of these models may deter companies from adopting them. Instead, they may prefer simpler models that allow for a clear understanding of the customer attributes that contribute to a purchase. In this study, we show that companies need not compromise on prediction accuracy to understand their online customers. By leveraging website data from a multinational communications service provider, we establish that the most pertinent customer attributes can be readily extracted from a black box model. Specifically, we show that the features that measure customer activity within the e-commerce platform are the most reliable predictors of conversions. Moreover, we uncover significant nonlinear relationships between customer features and the likelihood of conversion.

Keywords: customer profiling; conversion; direct marketing; explainable artificial intelligence; SHAP value; accumulated local effects

check for updates

Citation: Bastos, J.A.; Bernardes, M.I. Understanding Online Purchases with Explainable Machine Learning. *Information* **2024**, *15*, 587. https://doi.org/10.3390/info15100587

Academic Editors: Elias Dritsas and Maria Trigka

Received: 7 August 2024
Revised: 20 September 2024
Accepted: 23 September 2024
Published: 26 September 2024

1. Introduction

Customer profiling in e-commerce is a valuable tool for understanding customer behavior during their interactions with a company's website. This knowledge empowers organizations to create personalized offers through direct marketing. A key objective of customer profiling is to understand whether a website visitor will purchase one or more products, thereby generating revenue. When a customer generates revenue, we say that the browsing session resulted in a "conversion" or, alternatively, that the customer has "converted". In order to predict the likelihood of conversion, we can analyze the navigational patterns of visitors and their engagement with the content provided by the online platform. Sociodemographic data and historical behavior may also define the customer profile for registered users. Furthermore, certain web design features may influence the probability of conversion [1].

One of the most direct approaches to creating a customer profile is by using clickstream data—the unique fingerprint of customers as they navigate a website. Previous studies have demonstrated that the manner in which visitors interact with a website plays a significant role in determining their likelihood of generating revenue. For example, Moe [2] conducted a study showing that visitors to an online store can be categorized based on their observed navigational patterns. Similarly, Moe and Faden [3] showed that a parametric model incorporating dynamic behavior derived from clickstream data can predict the probability of a purchase. Another study by Sismeiro and Bucklin [4] suggested that decomposing the purchase process into a sequence of tasks also aids in predicting the likelihood of a purchase.

Clickstream data provide information about traffic origin, the time visitors spent on a given page, which content they engaged with, and where they went next. Information about the browsing device and the operating system is usually also available. Once a customer profile is established, we can use it to predict the probability that they will generate revenue. This is a binary classification problem—the task of identifying which of two classes an individual belongs to. Here, we want to classify customers into the class of those who have converted and the class of those who have not. This decision is based on their profiles. Most classification algorithms provide a number bounded to the interval [0, 1] called the *score*. If the class of customers who have converted is encoded as 1, and the class of those who have not is encoded as 0, a well-trained classifier will provide scores close to 1 for customers with a high likelihood of conversion.

The most simple tool to model the probability of a conversion is binary choice models, such as logistic regression. This model specifies that probability is the logistic function of a linear combination of predictors. Because the maximum likelihood estimators of the parameters are asymptotically normal, this model has good inference properties, and customer conversions can be easily explained in terms of customer attributes. Goodness-of-fit measures, such as likelihood ratio, are also provided by the optimization procedure. Van den Poel and Buckinx [5] used logistic regression to predict whether or not a purchase is made during the next customer visit to a wine retailer's website. They use clickstream data, customer demographics, and past purchase behavior as predictors. The estimated Wald statistics for the coefficients and the chi-squared statistics score allow them to identify the most important predictors using variable selection procedures. Olbrich and Holsing [6] used a logistic regression model to understand which factors are most significant for predicting consumer purchasing behavior within social shopping communities. Lo et al. [7] also used logistic regression to predict the intention to purchase by tracking user activity on a content discovery platform.

Due to its fixed functional form, logistic regression does not have enough flexibility to fit complex data and is often less accurate than the more sophisticated machine learning models, such as ensemble methods and deep neural networks [8]. Furthermore, the low visit-to-purchase rate [9] means that small improvements in identifying those customers more likely to buy can lead to substantial increases in sales revenue. Therefore, several studies explored machine learning models to predict customer conversions in e-commerce platforms. For instance, Kim et al. [10] compared different strategies for combining the outputs of three neural networks trained to predict purchases. They found that a combination strategy based on a genetic algorithm provides the best accuracy. Mokryn et al. [11] trained different ensembles of decision trees to understand how the popularity of a product affects the likelihood of it being purchased at the end of a visit. Esmeli et al. [12] used ensembles of decision trees to predict customer buying intention from the first interactions with an online platform. Martínez et al. [13] trained a gradient boosting machine and a feedforward neural network to predict if a customer is going to purchase within a certain time frame in the near future. Chaudhuria et al. [14] considered online engagement and demographic attributes to train deep neural networks to predict a customer's decision to purchase. Esmeli et al. [15] used ensembles of decision trees, deep neural networks, and contextual customer characteristics—such as location, operating system, time of the visit, and previous browsing data—to predict the purchase intention before the user interacts with the e-commerce platform.

Due to the "black-box" nature of machine learning models, companies may be reluctant to use them for marketing purposes. If a model is opaque and the most relevant customer attributes are not understood, then it may fail to deliver actionable insights. Therefore, companies may prefer to sacrifice prediction accuracy and use simpler models where the customer attributes that led to a purchase are easily understood. In this study, we show that companies do not have to sacrifice prediction accuracy to understand their customers. By using website data from a multinational communications service provider, we show that the most relevant customer attributes can be easily extracted from a black-box model. Therefore, the marketing team is provided with the most relevant information to elaborate personalized offers without compromising targeting efficiency while maximizing the return on investment. This information also improves communication with the customer, promotes the company's efficient use of resources, and helps improve the e-commerce channel. To the best of our knowledge, this is the first study analyzing online purchasing decisions in the telecommunications sector. The navigation data were obtained from Google Analytics, a tool that measures what happens in a website, providing audience and acquisition reports. Audience reports provide data regarding device technology and visitor age, gender, and region. Acquisition reports provide information about how the users reached the website and the origin of their traffic. Our black-box model is a gradient boosting machine [16,17]. This is a powerful nonparametric model consisting of a "committee" of decision trees. It is one of the best off-the-shelf algorithms for a wide range of predictive problems and a winner of major data science competitions [16].

First, we show that the gradient boosting machine performs better than the parametric logistic regression in our dataset since it is better at discovering complex dependencies in the data. Then, we propose three model-agnostic techniques for explaining the decisions of the black box model. The first is based on random permutations of the predictor variables. In this approach, we randomly permute the values of a given variable, train the model, and evaluate how the prediction error changes. The random permutation breaks the relationship between this variable and the output. A regressor is "important" if, after shuffling its values, the model error increases substantially and "unimportant" if the model error does not change significantly. The second technique is called Shapley additive explanations [18,19], which is based on co-operative game theory. This approach evaluates the importance of a predictor by measuring its impact on the model predictions when it is present in or absent from all possible 'coalitions' of predictors. The third technique is the accumulated local effects plot [20]. It does not evaluate the overall importance of a regressor but provides a visualization of how the target variable changes with the input variables. For instance, it indicates whether this change is positive or negative, linear or nonlinear, convex or concave. By using these techniques, we show that the black box model identifies important relationships between customer features and the probability of conversion that the logistic regression model ignores. For instance, we learn that for the black box model, the attributes measuring customer activity in the e-commerce platform—such as the average page views per day and the number of distinct days in which there were visits—are the best predictors of conversions. Furthermore, the accumulated local effects plots uncover highly nonlinear relationships between predictors and the probability of conversion.

The remainder of this paper is structured as following. The next section describes the dataset used in this study. This is followed by an explanation of the variable selection procedure. In Section 4, we describe the models for predicting customer conversions and evaluate their performance. Section 5 compares the most important determinants of conversions according to the black box and logistic regression models. Section 6 provides some concluding remarks.

2. Data

The dataset used in this study was obtained from a multinational telecommunication services provider. The data are statistics collected by Google Analytics on the company's website. This service measures user traffic and engagement across websites and apps, such as session duration, pages per session, visited pages, and browsing devices [21]. A session is a unique visit to the website. The collected data are structured into several reports, each describing different information about the user's browsing behavior. Each report contains a unique customer identifier that allows information to be merged across different reports.

Each entry in our dataset describes the browsing behavior of a unique customer in the studied period. The first set of variables is related to the origin of the website traffic generated by a given customer. The origin of the traffic is characterized by a "source/medium" pair—for instance, "Google/cost-per-click paid search" or "Bing/organic search". For each customer, we recorded the following variables during the period under study:

- the total number of entrances from each source/medium pair;

- the total number of pageviews for sessions originating from each source/medium pair;
- the average time on pages for sessions originating from each source/medium pair;
- the total number of distinct source/medium pairs used by the customer.

Then, for specific pages on the company website, we recorded the following:

- the total number of pageviews in the analyzed period;
- the total number of times the customer started a browsing session on the page;
- the average time spent on the page.

In particular, we constructed these variables for special pages, such as those showing equipment content (e.g., smartphones, tablets, and accessories), special campaigns, and those related to customer loyalty programs.

We also recorded information regarding the device category and operating system used to access and navigate through the website. The device category variable is either desktop, mobile, or tablet. The operating system variable is either Windows, OSX, Linux, Android, or iOS. For each user, we created a dummy variable that indicates whether a given device/OS pair was used in the analyzed period and the total number of pageviews from each device/OS pair. In addition to the variables obtained from the company's digital channel, we also included variables describing sociodemographic characteristics, service engagement, and whether the customer was targeted by campaigns.

Finally, we obtained information about the total revenue generated by the user in the analyzed period. This revenue includes both purchases on the company website and in-app transactions. Since we are interested in understanding the likelihood of conversion, we created a dummy variable that indicates whether the user converted (i.e., generated revenue) or not. Customers who were converted were encoded as 1, and those who were not were encoded as 0. The final dataset contains almost 250 variables describing around 7000 customers who did convert. The conversion rate is about 4%. This value is higher than the 2.1% rate for online shoppers in the United States observed in the first semester of 2022 [9]. Because the two classes are rather unbalanced, we tested several undersampling and oversampling techniques, but these results are omitted for brevity as those techniques did not improve the accuracy of the models.

3. Selection of Customer Features

The first step to understand which predictors or customer "features" are the most important (in the sense of having the greater discrimination power of the target variable) is achieved by using univariate analysis. This step allows us to select the most relevant customer features to be included in the model, making it more robust and less prone to overfitting the training data. It also reduces the computational complexity of the model and increases its scalability for deployment. The "weight of evidence" and "information value" are two common tools for exploratory data analysis and variable screening when dealing with binary classifiers. These metrics let us identify which variables are ill-conditioned or do not contain useful information to predict the target variable.

Suppose that the observations belong to two classes: the positive and the negative. In binary classification, the positive class usually corresponds to the 'presence of something', which, in our analysis, is the conversion of a customer. The weight of evidence (WoE) tells us how confident we are that a variable, X, will help us discriminate between the two classes. It compares the empirical density function of X for the two classes for different bins:

$$\text{WoE}_k = \log \frac{f(X_k|+)}{f(X_k|-)}, \quad k = 1, \ldots, \#bins. \tag{1}$$

If X has low discrimination power, $f(X|+)$ and $f(X|-)$ will overlap significantly. Therefore, $f(X_k|+)/f(X_k|-) \approx 1$ and $\text{WoE}_k \approx 0$ across all bins, k. On the other hand, if X separates the two classes well, we will observe bins with different proportions of positive and negative observations and obtain values for WoE deviating from zero.

The information value (IV) aggregates the values of WoE at different bins into a single metric:

$$\mathrm{IV} = \sum_{k=1}^{\#bins} \mathrm{WoE}_k \times [f(X_k|+) - f(X_k|-)]. \tag{2}$$

Note that the term $f(X_k|+) - f(X_k|-)$ has the same sign of WoE_k, hence ensuring that the IV is always a positive number. The IV ranks the predictors in terms of the discrimination power of the two groups, indicating which should be included in the model. A heuristic rule states that predictors with 'medium' discrimination power have an IV of greater than 0.1, whereas those with strong discrimination power have an IV of greater than 0.3.

For the modeling phase, we selected the set of 17 variables with an IV of greater than 0.2. These are reported in Table 1. The variables that measure how active a customer was in the analyzed period are those that better discriminate the customers who have converted from those who have not. Indeed, they reflect the interest and engagement of the user with the website content. For instance, the variable that better discriminates the two groups is the average page views per day. Variables such as the total number of pageviews, the total number of pageviews related to equipment or brands, the average time a user spends visualizing pages, and the number of distinct days on which the user accessed the website are also important. The most surprising fact is that many customer features unrelated to browsing behavior do not discriminate well between customers who have converted from those who have not. The only features of this type that do so are the dummy variables indicating if the user searched the company's website for any of the three payment methods. If a user has decided to purchase, it is natural that they will search for the available payment options. Less surprising is the fact that neither the browsing device nor the operating system has a significant impact on conversions.

Table 1. Description and information value (IV) of predictor variables to be included in the models. These variables have the highest discrimination power for predicting customer conversions according to the IV criterion.

Variable Name	Variable Description	IV
Avgpageviews_day	Average number of viewed pages per day	1.56
Equip_pageviews_total	Total number of viewed pages related to equipment content	1.22
Pageviews_total	Total number of viewed pages	0.97
Average_time_page_total	Sum of average time spent on page	0.86
Purchase_method_1	1 if the user searched for the type 1 purchase method; otherwise, 0	0.73
Equip_type1_pageviews	Total number of viewed pages of related to equipment of type 1	0.62
Distinct_days	Number of distinct days on which the user accessed the website	0.59
Distinct_sources	Number of distinct sources/mediums from which the user accessed the website	0.54
Equip_type1_avgtime_sum	Sum of average time spent on equipment type 1-related pages	0.41
Source1_pageviews	Number of viewed pages through the source/medium of type 1	0.41
Purchase_method_2	1 if the user searched for the type 2 purchase method; otherwise, 0	0.35
Brand2_pageviews	Number of viewed pages related to brand 2	0.30
Equip_type2_pageviews	Number of viewed pages related to equipment type 2	0.27
Brand1_pageviews	Number of viewed pages related to brand 1	0.21
Equip_pageviews_other	Number of viewed pages not related to the equipment categories	0.21
Equip_type1_entrances	Total number of entrances in equipment type 1-related pages	0.21
Purchase_method_3	1 if the user searched for the type 3 purchase method; otherwise, 0	0.20

4. Models

4.1. Logistic Regression

Logistic regression is one of the most simple binary classification models. Let Y denote a Bernoulli random variable that equals 1 if the customer converted and 0 otherwise. Logistic regression explains the probability that a client will convert as a function of the logistic function of a linear combination of variables, $\boldsymbol{X} = \{X_1, X_2, \ldots, X_p\}$, describing the customer features,

$$\Pr(Y = 1|\boldsymbol{X}) = \frac{1}{1 + \exp\left[-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)\right]}, \tag{3}$$

where $\boldsymbol{\beta} = \{\beta_0, \beta_1, \ldots, \beta_p\}$ is a vector of coefficients. If the observations from the two classes are not linearly separable in the regressor space, it is possible to estimate $\boldsymbol{\beta}$ by maximizing the Bernoulli log-likelihood function. The estimators for $\boldsymbol{\beta}$ have good inference properties because they are asymptotically distributed according to a normal distribution. In order to convert a logistic regression model into a classifier, we impose a threshold value on $\Pr(Y = 1|\boldsymbol{X})$. If a given customer has $\Pr(Y = 1|\boldsymbol{X}) \geq 0.5$, we predict that they will convert.

Table 2 shows the estimated logistic regression model for our data. Because the logistic function is strictly monotonic, the sign of the coefficient provides the direction of the partial effect. For instance, the variable “Source1_pageviews”, measuring the number of viewed pages for visitors coming from source/medium 1, has a positive impact on the probability of conversion. Surprisingly, the variable “Pageviews_total”, which measures the total number of viewed pages in the analyzed period, and the variable “Equip_pageviews_other”, which measures the total number of viewed pages of equipment that are not within the main categories (e.g., smartphones, accessories, and tablets), have a negative impact on the probability of conversion when we control for other variables.

Table 2. Output of the logistic regression model.

	Coefficient	Std. Error	z-Stat	p-Value
Constant	−4.860	0.034	−143.002	0.000
Brand2_pageviews	−0.002	0.006	−0.376	0.707
Purchase_method_2	0.395	0.037	10.685	0.000
Equip_type1_avgtime_sum	−0.014	0.019	−0.751	0.453
Source1_pageviews	0.010	0.001	11.673	0.000
Distinct_sources	0.218	0.014	15.776	0.000
Distinct_days	0.084	0.007	12.958	0.000
Equip_type1_pageviews	−0.006	0.005	−1.178	0.239
Purchase_method_1	1.111	0.037	29.746	0.000
Average_time_page_total	0.128	0.008	16.067	0.000
Pageviews_total	−0.013	0.001	−11.671	0.000
Equip_pageviews_total	0.016	0.003	4.792	0.000
Avgpageviews_day	0.038	0.001	30.259	0.000
Purchase_method_3	0.534	0.055	9.727	0.000
Equip_type1_entrances	−0.040	0.012	−3.225	0.001
Brand1_pageviews	0.007	0.005	1.383	0.167
Equip_pageviews_other	−0.036	0.008	−4.483	0.000
Equip_type2_pageviews	0.073	0.011	6.946	0.000

The last column gives the p-value for the null hypothesis that regressor X_j has no effect on the probability of conversion: $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$. For the usual significance levels, there are four regressors that have no effect on conversions (when we control for other variables in the model). “Type 1” is a popular type of equipment. In the multivariate model, when we control for the number of entrances in pages showing content for this type of equipment (Equip_type1_entrances), the sum of the average time spent on these pages (Equip_type1_avgtime_sum) and the total number of viewed pages (Equip_type1_pageviews) are not statistically significant. The pages related to specific brands (Brand1_pageviews and Brand2_pageviews) also become nonsignificant when we control for other variables that measure user activity.

In order to measure the importance of variables on conversions, we should not only consider their statistical significance but also the magnitude of the corresponding coefficients. However, looking at these coefficients alone ignores the variation of the explanatory variables in the data. Table 3 shows the sample standard deviation of the statistically significant variables in the model. Variables such as the average page views per day or the total pageviews in the analyzed period have large variations in the data. On the other hand, dummy variables, such as those indicating if a payment method was searched for by the user have lower variation. An approximate measure of variable importance is the product of the absolute value of the coefficient estimates with the sample standard deviation of the variables in the data,

$$\text{Importance of variable } j = |\beta_j| \times \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_{ji} - \bar{X}_j)}, \tag{4}$$

thereby accounting for the difference in variance between the variables [22]. The last column in Table 3 shows the customer feature importance given by this metric. For the logistic regression model, the average number of viewed pages per day is the most important determinant of the probability of conversion. If a user has decided to purchase, they might search for available payment options. According to logistic regression, the "Purchase_method_1" is the purchase option that leads to more conversions. On the other hand, user activity on specific pages related to equipment is the least important feature of this model.

Table 3. Sample standard deviation of the statistically significant variables and customer feature importance are given by the logistic regression model. Feature importance is measured as the product of the absolute value of the estimated coefficients with the sample standard deviation of the corresponding predictors.

Customer Feature	**Std. Dev.**	**Importance**
Avgpageviews_day	9.014	0.343
Purchase_method_1	0.288	0.320
Average_time_page_total	2.034	0.260
Pageviews_total	17.434	0.227
Distinct_days	2.411	0.202
Distinct_sources	0.918	0.200
Purchase_method_2	0.326	0.129
Equip_pageviews_total	7.605	0.120
Source1_pageviews	10.102	0.103
Purchase_method_3	0.160	0.085
Equip_type2_pageviews	0.880	0.065
Equip_pageviews_other	1.233	0.045
Equip_type1_entrances	0.874	0.035

4.2. Extreme Gradient Boosting Machine

Our black box model is a "gradient boosting machine" [17]. Boosting machines combine several base models to produce a powerful "committee" of models. Typically, the base models are decision trees [23,24]—a set of if-then-else conditions on the features of the observations that lead to a decision. For instance, a hypothetical branch of a decision tree may be:

IF Avgpageviews_day $>$ 15 **AND** Distinct_days $>$ 5 **THEN** customer converts.

A typical decision tree may have hundreds of these branches, each with dozens of sequential tests on the features.

The prediction of a gradient boosting machine, $\hat{Y}$, is the sum of the predictions given by a set of K decision trees $\{f_k(\boldsymbol{X})\}_{k=1}^{K}$,

$$\hat{Y} = \sum_{k=1}^{K} f_k(\boldsymbol{X}). \tag{5}$$

The first tree, $f_1(\boldsymbol{X})$, is a normal decision tree trained on the original data. The following decision trees, $\{f_k(\boldsymbol{X})\}_{k=2}^{K}$, are added sequentially to the committee. However, each added tree is trained on the residuals generated by the trees that are already in the committee—it "corrects" the errors made by the current committee. A well-known implementation of gradient boosting is extreme gradient boosting, also known as XGBoost [16]. This is one of the best off-the-shelf algorithms for a wide range of predictive tasks. Indeed, about 60% of the winning solutions posted on Kaggle in 2015 and the best solutions in the KDD Cup in 2015 used XGBoost [16]. We optimized the 'hyper-parameters' (parameters that are not learned in the training process) of our XGBoost model using grid search.

4.3. Evaluating Model Performance

In order to evaluate the performance of the models, we consider the F1-score and the area under the ROC curve since these metrics are more appropriate in scenarios where the classes are unbalanced [25]. Let TP denote the "True Positives"—customers who have converted and were correctly classified as such—FP denote the "False Positives"—customers who have *not* converted but were incorrectly classified as having converted—TN denote the "True Negatives"—customers who have *not* converted and were correctly classified as such—and FN denote the "False Negatives"—customers who have converted but were incorrectly classified as not having converted. "Recall" (or true positive rate) measures the proportion of positive observations that were correctly identified:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{6}$$

"Precision" is the ratio between the number of correctly predicted positives and the total number of observations predicted as positive:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{7}$$

The F1-score combines the precision and recall metrics into a single metric. Because precision and recall are rates, the F1-score is defined as their harmonic mean:

$$\text{F1-score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}. \tag{8}$$

A receiver operating characteristic (ROC) curve plots the true positive rate (or recall) against the false positive rate

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \tag{9}$$

We vary the threshold value for $\Pr(Y = 1|\boldsymbol{X})$, which is used to classify an observation as positive. A classifier that cannot discriminate the two classes well will have a true positive rate similar to the false positive rate as we vary this threshold. The ROC "curve" will be a straight line with a slope of 1, and the area under this curve will be about 0.5. A classifier with good discrimination power will have a true positive rate greater than the false positive rate regardless of the threshold value, and the area under the ROC curve will be greater than 0.5. The higher the area under the ROC curve, the better the discrimination power of the classifier is. A perfect classifier would have an area under the ROC curve of 1.

We randomly split the data into a training set with 80% of the original data and a validation set with 20% of the original data, ensuring that the proportion of conversions was roughly equal in both samples. The validation data were used to estimate the out-of-sample accuracy. Table 4 shows the out-of-sample precision, recall, F1-score, and AUC in the validation data for the logistic regression and the gradient boosting machine. The precision and recall were calculated using a cutoff value for $\Pr(Y = 1|\boldsymbol{X})$ of 0.5. Of course, the optimal cutoff value is an empirical matter; it is defined by the amount of false positives and false negatives expected by the marketing team. When looking at the precision metric, we might suspect that logistic regression is the best model. However, high precision comes at the cost of poor recall; logistic regression generates many false negatives, failing to detect customers who have actually converted. When we combine these metrics into the F1-score, the gradient boosting machine is the clear winner. The AUC metric gives a precision metric that is independent of the cutoff value for $\Pr(Y = 1|\boldsymbol{X})$. According to this metric, the gradient boosting machine is the best model.

Table 4. Model performance metrics on the validation data.

Classifier	Precision	Recall	F1-Score	AUC
Logistic Regression	0.46	0.08	0.14	0.86
Gradient boosting machine	0.28	0.43	0.34	0.88

4.4. Gain and Lift Analysis

In addition to evaluating the performance metrics above, it is convenient to inspect how a predictive model benefits the business when compared to a situation in which the model is not used, meaning when the customers are randomly targeted. Gain and lift analysis is often used in marketing to evaluate campaign performance and help identify the best customers, thereby improving the prospect of future campaigns. In order to calculate the gain and lift, we calculate the scores given by the model for all customers in the test data. We recall that the score gives the probability that a customer converts. Then, we sort those scores in descending order. Afterward, we divide the data into deciles. Finally, we calculate the number of conversions in each decile and the cumulative number of conversions up to a decile. The gain is calculated as

$$\text{Gain} = \frac{\text{Cumulative number of conversions up to a decile using model}}{\text{Total number of conversions in the data}}, \tag{10}$$

whereas the lift is computed as

$$\text{Lift} = \frac{\text{Cumulative number of conversions up to a decile using model}}{\text{Cumulative number of conversions up to a decile using random guessing}}. \tag{11}$$

Table 5 shows the gain and lift values given by the gradient boosting machine for the test data. Suppose that we consider the 10% of observations with the highest scores. This sample contains about 59% of the customers who converted. Likewise, if we consider the 20% of observations with the highest scores, we cover about 78% of the customers who converted. In business terms, this means that if those 20% of customers are targeted, we expect that about 78% of them will convert. Regarding the lift, the value for the second decile is 3.91, which means that by covering 20% of the data using the gradient boosting machine, the probability of predicting customers who converted is 3.91 times higher than by randomly selecting 20% of the data, that is, without using the model.

Table 5. Gain and lift values given by the gradient boosting machine for the test data.

Decile	# Cases	# Responses	Cumulative	% Events	Gain	Lift
1	3696	859	859	59.16	59.16	5.92
2	3696	277	1136	19.08	78.24	3.91
3	3695	136	1272	9.37	87.61	2.92
4	3696	69	1341	4.75	92.36	2.31
5	3695	54	1395	3.72	96.08	1.92
6	3695	32	1427	2.2	98.28	1.64
7	3697	21	1448	1.45	99.73	1.42
8	3233	4	1452	0.28	100.01	1.25
9	4089	0	1452	0	100.01	1.11
10	3765	0	1452	0	100.01	1.00

5. The Determinants of Conversions

The previous section shows that the gradient boosting machine is more accurate than logistic regression. On the other hand, we also remarked that it is rather straightforward to understand which variables contribute the most to the predictions provided by logistic regression. If we wish to understand which variables contribute to the predictions of a black box model, we need to use techniques to explain machine learning models. These methods are also known as eXplainable artificial intelligence (XAI) techniques.

5.1. Permutation Feature Importance

Perhaps the most straightforward approach to measure the importance of a predictor in model output is to randomly permute its values and then observe any deterioration in model accuracy. The random permutation breaks the relationship between the inputs and the output. An input is "important" if, after shuffling its values, model accuracy decreases considerably and "unimportant" if the model accuracy does not change significantly. If the accuracy decreases, then the model relies on the variable in question to generate its predictions. On the other hand, if the accuracy does not change, the model "ignored" that variable.

Figure 1 shows customer features in decreasing order of importance when we apply this technique to our gradient boosting machine. The total number of viewed pages related to equipment (Equip_pageviews_total) is by far the most important customer attribute for determining conversions. In contrast, this attribute is only the eighth-most important variable for logistic regression (Table 3). "Purchase method 1" is the second-most important feature for both models. Average page views per day is the third-most important feature for the gradient boosting machine, whereas it is the most important for logistic regression. On the other hand, the variables that are not statistically significant at usual significance levels for logistic regression appear at the bottom of the permutation feature importance chart.

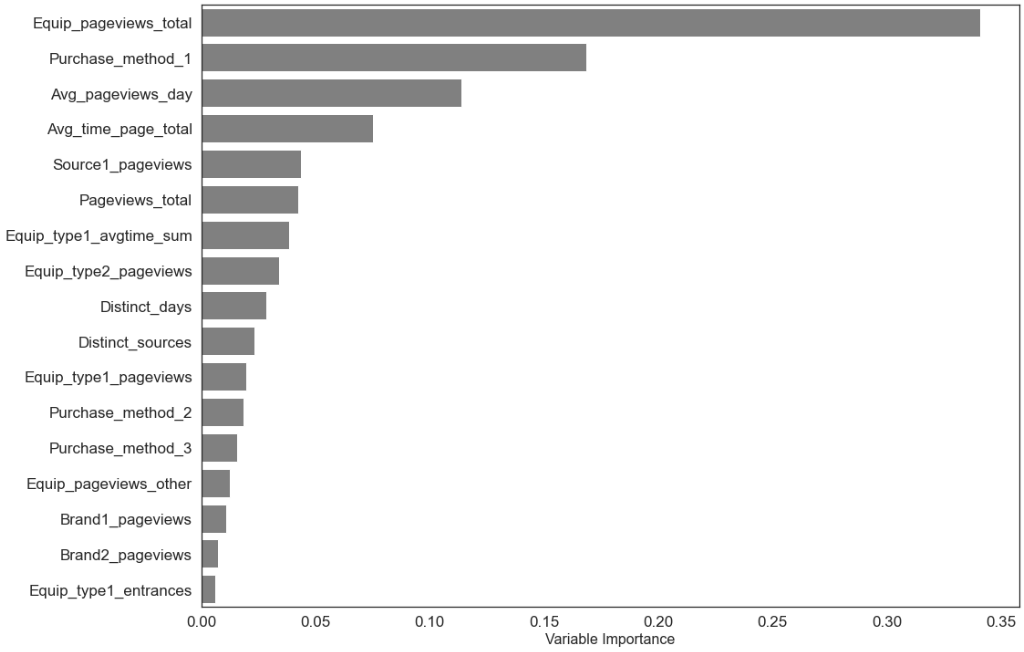

Figure 1. Permutation feature importance for the gradient boosting machine. The vertical axis shows the selected determinants of customer conversions, and the horizontal axis gives the variable importance.

5.2. Shapley Additive Explanations (SHAP)

Permuting the values of a predictor also destroys the relationship with other covariates. Therefore, importance measures based on permutation also take into account the effect of variable interactions. This may be a disadvantage of this method since the strength of the interaction between two variables contributes to the importance of both variables. Shapley additive explanations (SHAP) [18] is a technique for explaining black box models that do not assume any independence of the predictors. It is based on Shapley values—a concept for predicting which strategies are adopted by players in a co-operative game. Shapley values give a solution to the following problem: a coalition of players co-operates to obtain a certain gain from that co-operation; however, some players may contribute more to that gain than others; how is it possible to fairly distribute the gain among the players in any particular game?

Lundberg and Lee [18] adapted Shapley values to evaluate the importance of the input variables of a black box model. In their framework, the game predicts the model output for observation. The players are the input variables that collaborate to receive the final gain; the importance of an input variable is measured by how much it contributes to the model output, and the final gain is the prediction minus the average prediction for all observations. Feature importance is measured by the extent to which it affects model predictions when it is present or absent from all possible 'coalitions' of regressors. This approach is computationally expensive when the number of variables is high, and the algorithm must be modified to obtain approximate SHAP values more quickly. However, Lundberg et al. [19] proposed a variation of the original algorithm that computes exact SHAP values quickly when the machine learning model is based on decision trees, as is the case here.

The results of a SHAP analysis can be summarized in a "SHAP summary plot", where the vertical axis shows the model inputs, ordered by their importance, and the horizontal axis represents the SHAP values for each observation. The shade of the dots represents the feature value, allowing us to analyze the sign of the relationship between this feature and the model output. If the points extending towards the right are increasingly darker (lighter), the variable has a positive (negative) effect on conversions. Figure 2 shows the SHAP summary plot given by the gradient boosting machine for the customer attributes. Comparing the SHAP summary plot for the gradient boosting machine with the most important inputs for logistic regression, as shown in Table 3, we note that there is a reasonable agreement on which customer features have a greater impact on customer conversions. For both models, the most important variable is the average page views per day. The effect of this variable on the probability of conversion is positive, meaning that customers who are consistently more active on the company website are more likely to purchase. Furthermore, the variables that are not statistically significant at the usual significance levels for the logistic regression appear at the bottom of the SHAP summary plot. However, there are also some disagreements. For example, the total number of viewed pages related to equipment (Equip_pageviews_total) and the number of viewed pages by customers originating from the type 1 source/medium (Source1_pageviews) are the second and fourth most important variables for the gradient boosting machine, respectively; however, these appear in the bottom half of the most important variables for logistic regression. Perhaps more surprising are some of the differences between SHAP and the feature permutation technique (Figure 1). For instance, average page views per day represent only the third-most important feature according to the permutation technique. This indicates that some customer features interact to jointly determine the probability of conversion. When there are differences between the feature permutation technique and SHAP, we should consider the latter.

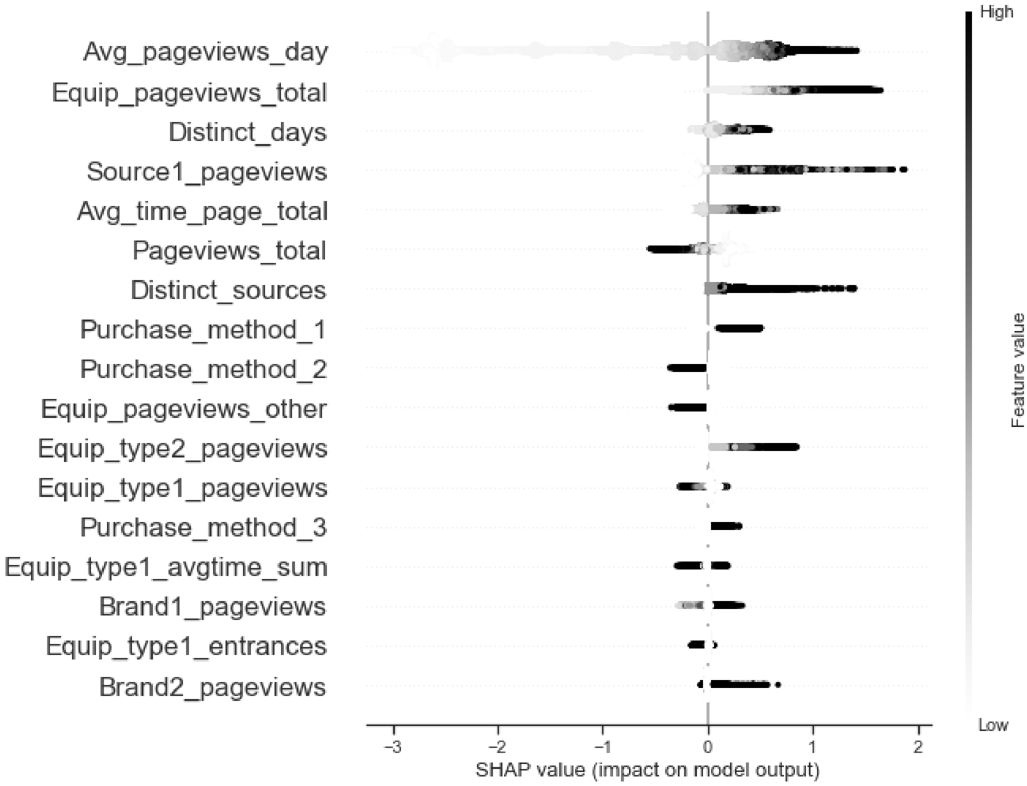

Figure 2. SHAP summary plot for the gradient boosting machine. The vertical axis shows the selected determinants of customer conversions, and the horizontal axis gives the Shapley values for each website visitor.

5.3. Accumulated Local Effect Plots

The previous techniques for explaining black box models just provide the overall importance of the customer's attributes. Accumulated local effects (ALEs) plots [20] provide a visualization of how conversions change with the model input variables. For instance, it indicates whether this change is positive or negative, linear or nonlinear, or convex or concave. The idea of ALE plots is simple yet powerful. For a given regressor, X_j, we first divide its range using a grid with a certain number of bins. Typically, the bin limits are chosen using the quantiles of the empirical distribution of X_j. Then, we compute how much the model output (i.e., the probability of conversion) changes, on average, in each of these bins. These changes give the local average effect of X_j on the model output. Afterward, these changes are summed ("accumulated") from the first bin up to a given value of X_j. The ALE plot is a representation of these sums as a function of X_j, providing a visualization of how the model output depends on X_j across its range. The average change is usually subtracted from the individual sums, and therefore, these are centered at zero.

Figure 3 shows the ALE plots for the four most important customer features according to the SHAP values. These plots show complex dependencies between these features and the probability of conversion. They show that the probability of conversion is stable when the average number of pageviews is lower than 5. For higher values of average pageviews, it increases linearly. The probability of conversion reaches a maximum when the feature "Equip_pageviews_total" is between 8 and 10 and reaches a plateau when the number of distinct days the customer browsed the website is greater than 7. They also show a highly nonlinear behavior of the probability of conversion as a function of the feature "Source1_pageviews".

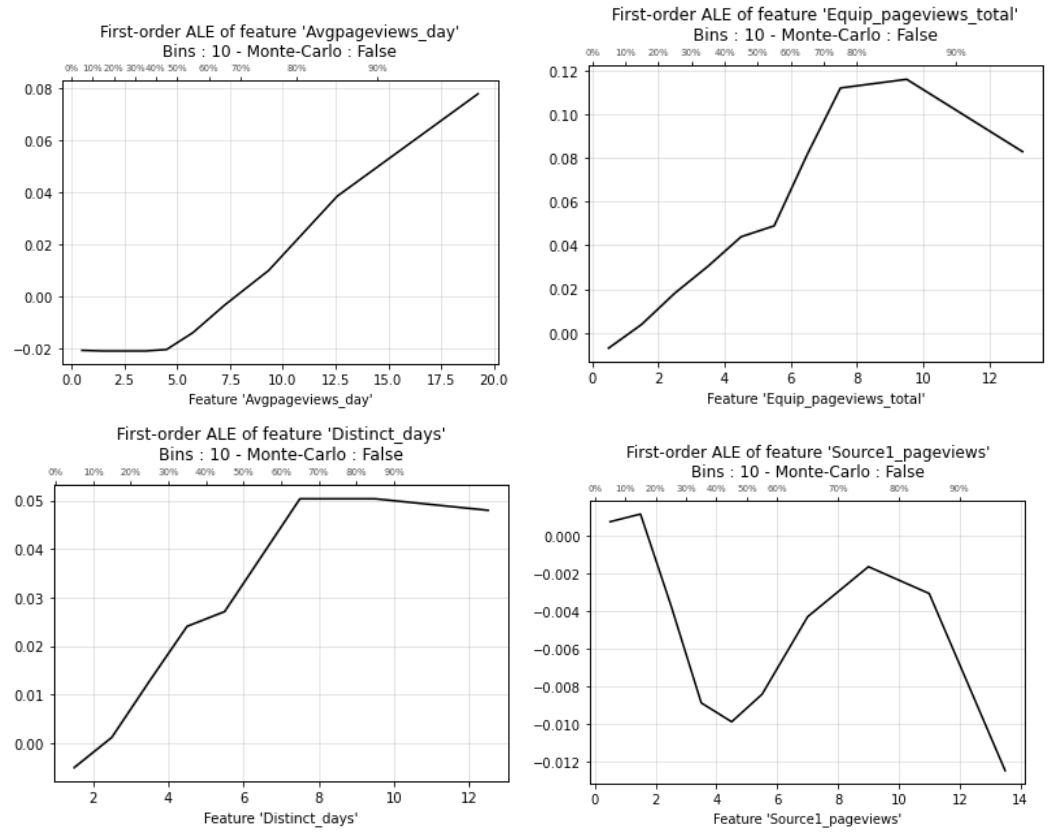

Figure 3. ALE plots for the most important customer features according to the SHAP values.

6. Conclusions

In this study, we showed that companies with e-commerce channels can rely on black box models for customer profiling despite their opaqueness. By using website data from a multinational communications service provider, we showed that the most relevant customer features can be easily extracted from a black box model, providing the marketing team with the most appropriate information about elaborate personalized offers without compromising targeting efficiency. Therefore, companies do not have to sacrifice prediction accuracy to understand their customers. In order to achieve this goal, we trained a glass-box model—logistic regression—and a black box model—a gradient boosting machine—to predict customer conversions. We showed that, as expected, the black box model had greater accuracy than the glass-box model. By using explainable machine learning techniques, we showed that the gradient boosting machine identified the most important customer features in the logistic regression model. Furthermore, the gradient boosting machine identified important relationships between customer features and the probability of conversion that logistic regression failed to capture. The black box model suggested that attributes measuring customer activity in the e-commerce platform of the communications service provider—such as the average page views per day and the number of distinct days in which there were visits—are the best predictors of purchases. Additionally, the accumulated local effects plots revealed complex relationships between predictors and the probability of conversion.

Author Contributions: Conceptualization, J.A.B. and M.I.B.; Methodology, J.A.B.; Software, M.I.B.; Formal analysis, M.I.B.; Data curation, M.I.B.; Writing—original draft, J.A.B. and M.I.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the FCT—Fundação para a Ciência e a Tecnologia [grant number UIDB/05069/2020].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. McDowell, W.C.; Wilson, R.C.; Kile, C.O., Jr. An examination of retail website design and conversion rate. *J. Bus. Res.* **2016**, *69*, 4837–4842. [CrossRef]
2. Moe, W.W. Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *J. Consum. Psychol.* **2003**, *13*, 29–39. [CrossRef]
3. Moe, W.W.; Fader, P.S. Dynamic conversion behavior at e-commerce sites. *Manag. Sci.* **2004**, *50*, 326–335. [CrossRef]
4. Sismeiro, C.; Bucklin, R.E. Modeling purchase behavior at an e-commerce web site: A task completion approach. *J. Mark.* **2004**, *41*, 306–323. [CrossRef]
5. Van den Poel, D.; Buckinx, W. Prediction online-purchasing behavior. *Eur. J. Oper. Res.* **2005**, *166*, 557–575. [CrossRef]
6. Olbrich, R.; Holsing, C. Modeling consumer purchasing behavior in social shopping communities with clickstream data. *Int. J. Electron. Commer.* **2011**, *16*, 15–40. [CrossRef]
7. Lo, C.; Frankowski, D.; Leskovec, J. Understanding behaviors that lead to purchasing: A case study of Pinterest. In Proceedings of the KDD ’16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 531–540.
8. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*, 2nd ed.; Springer: New York, NY, USA, 2021.
9. Statistica Inc. Conversion Rate of Online Shoppers in the United States from 2nd Quarter 2021 to 2nd Quarter 2022. 2022. Available online: https://www.statista.com/statistics/439558/us-online-shopper-conversion-rate (accessed on 22 September 2024).
10. Kim, E.; Kim, W.; Lee, Y. Combination of multiple classifiers for the customer’s purchase behavior prediction. *Decis. Support Syst.* **2003**, *34*, 167–175. [CrossRef]
11. Mokryn, O.; Bogina, V.; Kuflik, T. Will this session end with a purchase? Inferring current purchase intent of anonymous visitors. *Electron. Commer. Res. Appl.* **2019**, *34*, 100836. [CrossRef]
12. Esmeli, R.; Bader-El-Den, M.; Abdullahi, H. Towards early purchase intention prediction in online session based retailing systems. *Electron. Mark.* **2020**, *31*, 697–715. [CrossRef]
13. Martínez, A.; Schmuck, C.; Pereverzyev, S., Jr.; Pirker, C.; Haltmeier, M. A machine learning framework for customer purchase prediction in the non-contractual setting. *Eur. J. Oper. Res.* **2020**, *281*, 588–596. [CrossRef]
14. Chaudhuria, N.; Gupta, G.; Vamsi, V.; Bose, I. On the platform but will they buy? Predicting customers’ purchase behavior using deep learning. *Decis. Support Syst.* **2021**, *149*, 113622. [CrossRef]
15. Esmeli, R.; Bader-El-Den, M.; Abdullahi, H. An analysis of the effect of using contextual and loyalty features on early purchase prediction of shoppers in e-commerce domain. *J. Bus. Res.* **2022**, *147*, 420–434. [CrossRef]
16. Chen, T.; Guestrin, E. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
17. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
18. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In Proceedings of the NIPS’17: Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4768–4777.
19. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef] [PubMed]
20. Apley, D.W.; Zhu, J. Visualizing the effects of predictor variables in black box supervised learning models. *J. R. Stat. Soc. Ser. B* **2020**, *82*, 1059–1086. [CrossRef]
21. Cutroni, J. *Google Analytics: Understanding Visitor Behavior*; O’Reilly Media: Sebastopol, MA, USA, 2010.
22. Masís, S. *Interpretable Machine Learning with Python*; Packt Press: Birmingham, UK, 2021.
23. Breiman, L.; Friedman, J.; Stone, C.J.; Olshen, R.A. *Classification and Regression Trees*; Wadsworth: Belment, CA, USA, 1983.
24. Quinlan, J.R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
25. Jeni, L.A.; Cohn, J.F.; De La Torre, F. Facing imbalanced data recommendations for the use of performance metrics. In Proceedings of the Humaine Association Conference on Affective Computing and Intelligent Interaction, Geneva, Switzerland, 2–5 September 2013; pp. 245–251.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.