Received: 11 April 2021 | Revised: 20 August 2021 | Accepted: 23 August 2021

DOI: 10.1002/nvsm.1724

RESEARCH ARTICLE

WILEY

What drives a donor? A machine learning-based approach for predicting responses of nonprofit direct marketing campaigns

Davide Cacciarelli[1] | Marco Boresta[2]

[1]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Lyngby, Denmark

[2]Department of Computer, Control and Management Engineering, Sapienza University of Rome, Rome, Italy

Correspondence
Davide Cacciarelli, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Lyngby, Denmark.
Email: dcac@dtu.dk

Abstract

Direct marketing campaigns are one of the main fundraising sources for nonprofit organizations and their effectiveness is crucial for the sustainability of the organizations. The response rate of these campaigns is the result of the complex interaction between several factors, such as the theme of the campaign, the month in which the campaign is launched, the history of past donations from the potential donor, as well as several other variables. This work, applied on relevant data gathered from the World Wide Fund for Nature Italian marketing department, undertakes different data mining approaches in order to predict future donors and non-donors, thus allowing for optimization in the target selection for future campaigns, reducing its overall costs. The main challenge of this research is the presence of thoroughly imbalanced classes, given the low percentage of responses per total items sent. Different techniques that tackle this problem have been applied. Their effectiveness in avoiding a biased classification, which is normally tilted in favor of the most populated class, will be highlighted. Finally, this work shows and compares the classification results obtained with the combination of sampling techniques and Decision Trees, ensemble methods, and Artificial Neural Networks. The testing approach follows a walk-forward validation procedure, which simulates a production environment and reveals the ability to accurately classify each future campaign.

KEYWORDS
direct marketing, machine learning, target selection

1 | INTRODUCTION

World Wide Fund for Nature (WWF) is one of the largest nonprofit organizations in the world, supporting the safeguard of the environment and the conservation of species. It was founded in 1961 and it is currently operating in more than 100 countries, including Italy. Like any other NGO in the world, fundraising and the support of volunteers are crucial for the maintenance of its activities. Among all the ongoing fundraising activities performed by WWF, direct marketing is easily one of the most important. Direct marketing campaigns are activities where the organization spontaneously reaches out to a target of receivers, expressly asking for a donation. At WWF Italia this usually means targeting about 65,000 recipients by sending a letter and communicating the risky situation within a specific habitat or species, and soliciting a contribution to the ongoing conservation activities undertaken by WWF.

Each year, WWF Italia organizes approximately six large mailing campaigns, targeting people who are currently supporting the organization or those who used to support it within a time. These mailings may differ or follow a "seasonality" in terms of the relevant theme

(e.g., polar bear for the Christmas campaign), or may simply adhere to a topic that is being currently supported by WWF on a global scale (e.g., worldwide campaign to save tigers in Asia). Those who receive a letter may respond in several ways. The most common is to fill and send back a pre-compiled postal form stating how much they wish to donate (they can stick to one of the standard/suggested amounts or write any other quantity). On average, about 3% of the total letters sent come back as a donation. This means that out of the typical 65,000 items sent, roughly 2000 of them will produce a return and the remaining 63,000 will not. If sending a letter might not sound like a huge cost to control, for an organization that lives upon donations and sends nearly 390,000 letters a year (of which 378,000 do not produce a return), finding ways to send fewer letters while keeping the same amount of income represents a terrific opportunity for reducing operational costs.

In the past, the organization has made ad hoc analyses of a specific campaign target, in an attempt to remove those who were not likely to respond based on previous campaign statistics or guided by their experience. That is how the urgency for a more standardized and scientific way to approach the matter came into place. Consequently, after observing how Machine Learning has come in useful in classifying patterns in almost every field, it was decided that building a classification model to predict the outcome of any campaign and reduce its overall cost would be beneficial. An obvious obstacle to it that will be properly stressed in the upcoming sections, is the critical ratio of responses on the total letters sent, resulting in a heavy abundance of instances labeled as "0." The aforementioned problem has been referred to in the literature as "class imbalance" or "imbalanced data." This problem is commonly encountered in various businesses, such as engineering and medical applications in which experts deal with the prediction of rare events. Some notorious examples are fraud detection and default prediction in finance (Bolton & Hand, 2002), spam and intrusion detection in computer science (Cieslak et al., 2006), anomaly detection in engineering process control (Isermann, 1997), and medical diagnosis studies (Khalilia et al., 2011). With regards to the marketing context, several scenarios like churn prediction in recurring donations or the response rate when recruiting participants to live events and the conversion rate postevent share the same issues. Root causes of the class imbalance issue are generally domain-specific. Nonetheless, there may be situations in which the imbalance is introduced by errors made during the data collection procedure. Imbalanced classes represent a problem for machine learning algorithms since most of them are accuracy-based, meaning that they aim to minimize the error rate as the percentage on incorrect prediction, ignoring the diverse types of misclassification errors. This is one of the reasons that explain why most of the standard machine learning models tends to be biased toward the majority class when applied to imbalanced data (Ganganwar, 2012).

Thus, the main objective of this work will be to show the performance of the suggested methods in building a valuable classification model, optimizing the target selection for future direct marketing campaigns.

2 | RELATED WORKS

Despite the technological advances, direct marketing campaigns remain one of the main fundraising sources for nonprofit organizations and their effectiveness is crucial for the sustainability of the organizations. Many studies have been focusing on the prediction of marketing campaigns' outcomes. In particular, the importance of understanding the determinants of donation amounts in the nonprofit sector has been investigated by several scholars (Breeze & Jollymore, 2017; Ki & Oh, 2018; Pentecost & Andrews, 2009; Rupp et al., 2014).

As in the case of the private sector, where accurate prediction of consumer responses has become a priority and challenge for marketing managers (Bodenberg & Roberts, 1990; Gönül & Shi, 1998), optimizing the target selection for direct marketing campaigns and identifying in advance which customers or donors are more likely to respond is a topic of extreme interest. Luckily for them, the amount of data that both profit and nonprofit organizations have at their disposal is increasing year after year, thus allowing researchers to develop direct marketing response models using consumer data. The type of models used for predicting customer responses has changed over the years. A traditional one is the recency, frequency, monetary (RFM) model (Berger & Magliozzi, 1992), where the likelihood of consumers responding to a direct marketing promotion is predicted based on the recency of their last purchase, their frequency of purchases over the past years, and the monetary value of a customer's purchase history. Statistical models like discriminant analysis and logistic regression have been widely used in the past (Berger & Magliozzi, 1992), with their limitations being discussed (Bhattacharyya, 1999). Other proposed models include tree models like classification and regression trees (CART) and chi-square automatic interaction detection (CHAID; Haughton & Oulabi, 1997), the beta-logistic model (Rao & Steckel, 1995), and the hierarchical Bayes random-effects model (Allenby et al., 1999). Recent years have seen the proliferation of the application of machine learning models in fields like handwriting recognition (Such et al., 2018), stock-exchange prediction (Singh et al., 2017), and anomaly detection (Ahmed et al., 2007) just to name a few. Some attempts of using neural networks for predicting consumer responses have been tried in the past years (Baesens et al., 2002; Cui et al., 2006), not always with better results than the ones obtained with simpler models (Zahavi & Levin, 1999).

Tree-based methods like Random Forests have also been tested with the same aim. Asare-Frempong and Jayabalan (2017) tested the performances of tree-based methods and Artificial Neural Networks in predicting customer subscriptions of bank term deposits, obtaining good results, especially with Random Forests. Likewise, Ayetiran and Adeyemo (2012) and Ladyzynski et al. (2019) tried in predicting the output of financial direct marketing campaigns. Ladyzynski et al. (2019) in particular focused their work on deep learning and Random Forests. Apampa (2016) also worked on classification algorithms for banking direct marketing campaigns, dealing with the class imbalance problem and obtaining satisfactory results with the use of Decision Trees.

Direct marketing data often contain only a small proportion of donors due to the low response rate of the campaigns (Cui & Wong, 2004). This imbalance problem represents one of the main difficulties in the application of machine learning models for the prediction of consumer response and has been widely treated in literature.

Imbalanced classes represent a problem for machine learning algorithms since most of them are accuracy-based, meaning that they aim to minimize the error rate as the percentage on incorrect prediction, ignoring the diverse types of misclassification errors. This is one of the reasons explaining why most of the standard machine learning models tend to be biased toward the majority class when applied to imbalanced data (Ganganwar, 2012). In particular, when dealing with Decision Trees, as affirmed by Yanmin et al. (2011), "the split action may be terminated before the branches for predicting small classes got detected" or again "branches used for minority class may be pruned as being susceptible to overfitting," so there is a significant probability that the edges useful to classify instances from the less populated class finally got replaced by a leaf labeled with the target from the majority class. This is what urged intervention in order to adapt data and model for facing this particular problem. Among the various techniques that can help in trying to tackle the imbalance, sampling methods are one of the first approaches suggested (Chawla et al., 2004; Ganganwar, 2012; Yanmin et al., 2011). A further improvement can be pursued with the help of ensemble methods. The use of these methods to face the imbalance problem has been widely proposed in the literature: Yanmin et al. (2011), suggested, between the algorithm-level approaches, the use of boosting algorithms (AdaBoost); Chen and Breiman (2004) focused on the use of adjusted Random Forests (Balanced Random Forest and Weighted Random Forest); Galar et al. (2012) exposed the potential of combining ensemble algorithms to data preprocessing techniques (sampling methods). In this framework, the suggested approaches will be tested in this work with two objectives. Firstly, to see whether these methods can help in improving the classification performance on a heavily imbalanced dataset and secondly, to observe if the approaches that are accurate in the banking industry can be useful also in a nonprofit context. Eventually, the chosen classification model should be able to identify the crucial parameters in determining the output of the campaigns.

3 | DATA AND METHODS

The analyzed dataset contains information and output of 16 direct marketing campaigns, undertaken by WWF Italy in 2016, 2017, and 2018. The dataset counts nearly 1 million observations, with both numeric and categorical variables. Table 1 reports the variables used to build and test predictive models. The observed parameters, that have been fed as input to the examined models, can be grouped in different explanatory families, according to the type of information they deliver.

- Variables from 1 to 4 are details about the campaign, expressing the main topic and background of the campaign. The theme expresses the main subject of the campaign, this can be a specific species that is repeatedly advocated by WWF or a more general topic (e.g., pollution or plastic). The scope of the campaign is a simplification of the previously expressed theme, grouping the themes into two broad categories: environment and animal species. The nationality states whether the campaign refers to a national campaign or it is part of an international task force. Finally, the month indicates the launch period of the campaign, intending to unveil possible seasonality in the responses.
- Variables from 5 to 7 concern the personal information of the receivers.
- Variables from 8 to 20 reveal the marketing history of the receivers, indicating all past actions undertaken by each of them, with different levels of detail. The seniority says when the potential donor has supported the organization for the first time or when he/she became a prospect if he has never donated. The segment corresponds to the internal classification of donors. The flag variable customer indicates if the past transactions of the potential donor are only related to the purchasing of products and never to donations.
- Variables 21 and 22 are concerning recent actions, indicating whether the recipient has responded to one of the two last direct marketing campaigns. This information is different from the previous block of variables (from 15 to 20) since the latter can include donations received by WWF through any possible channel (not only direct marketing).

TABLE 1 Dataset summary

#	Name	Type
1	Campaign theme	Categorical
2	Campaign scope	Categorical (binary)
3	Campaign nationality	Categorical (binary)
4	Campaign month	Categorical
5	Age	Numeric
6	Gender	Categorical
7	Region	Categorical
8	Seniority	Numeric
9	Segment	Categorical
10	Flag customer	Categorical (binary)
11	Number of donations	Numeric
12	Number of purchases	Numeric
13	Average donation amount	Numeric
14	Average purchase amount	Numeric
15	Donation in last 12 months	Categorical (binary)
16	Donation in last 24 months	Categorical (binary)
17	Donation in last 36 months	Categorical (binary)
18	Purchase in last 12 months	Categorical (binary)
19	Purchase in last 24 months	Categorical (binary)
20	Purchase in last 36 months	Categorical (binary)
21	Output of 1st previous campaign	Categorical (binary)
22	Output of 2nd previous campaign	Categorical (binary)
23	Output	Categorical (binary)
24	Amount donated	Numeric

TABLE 2 Models and parameters

Classification methods and hyperparameters space
Decision Tree
A single CART tree with the following combinations of parameters explored
• Maximum depth of the tree ranging from 1 to 50, it can also be described as the length of the longest path from the tree root to a leaf.
• Maximum number of features to consider when looking for the best split ranging from 5 to all the features available.
• Minimum number of samples required to split an internal node ranging from 2 to 1000, if it increases the tree must consider more samples at each node.
• Minimum number of samples required to be at a leaf node ranging from 1 to 1000.
RUSBoost
A boosting (AdaBoost) ensemble of trees, sequentially trained on randomly under-sampled bootstrap replica of the original training set. Conversely to the Random Forest, the "sequentiality" of the boosting scheme is expressed by means of a learning rate, stating the contribution of each classifier in influencing the weights of each instance (that control the probability of each instance to be in the training set fed to the next tree of the ensemble).
• Number of trees constituting the ensemble, ranging from 10 to 300.
• Learning rate ranging from 10−5 to 1.
Balanced Random Forest
An ensemble of trees trained on randomly under-sampled bootstrap replica of the original training set. Additional parameters to the ones indicated for the Decision Tree needed to be tuned.
• Number of trees constituting the forest, ranging from 10 to 300.
• Different sampling techniques have been tested, affecting majority class only or both and including replacement or not.
Artificial Neural Network
A feed-forward network structure with multiple hidden layers and one output layer.
• Number of processing units (neurons) in the first layer ranging from 10 to 50.
• Number of internal layers ranging from 1 to 5, with a number of neurons from 5 to 50.
• Three alternative activation functions: hyperbolic tangent, rectified linear unit and sigmoid.
• Two alternative optimizers: adam and adagrad.
• Balanced batch (randomly under-sampled) are fed into the network structure, with a size ranging from 100 to 1000.
• A number of epochs spacing from 5 to 50.

- Variable 23 represents the output of the current campaign, thus the variable that the tested models will try to predict. Variable 24 is the amount corresponding to the donation.

The data gathering process has followed a comprehensive approach, namely, almost all the available information concerning donations and campaign have been collected. Some of the variables considered in this study represent the typical parameters of a RFM segmentation model. As affirmed by Olson (2012) and according to many other studies (Ayetiran & Adeyemo, 2012; Baesens et al., 2002; Fader et al., 2005), this model is extremely important and useful in predicting future responses of customers solicited by a direct

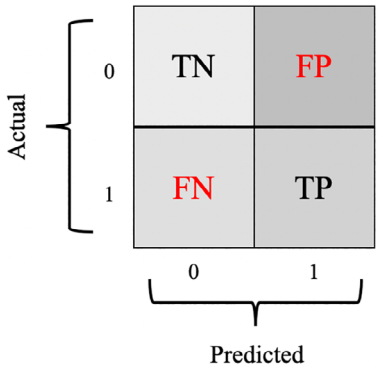

FIGURE 1 Confusion matrix

marketing campaign. This importance will be tested in the nonprofit context when the feature importance of the predictive models will be shown. In general, the RFM model aims at describing the customer (in this case donor) behavior with three main parameters:

Recency: is the time passed since the last purchase/donation, this parameter can be found in the dataset in variables ranging from 15 to 22;

- Frequency: is the total number of purchases/donations over a certain period or since the first one, in this dataset represented by variables 11 and 12;
- Monetary: is the economic value of the purchases/donations, in this dataset represented by variables 13 and 14.

Besides the RFM variables, other metrics like consumer lifetime and transaction variables have been used in the modeling stage as suggested in Bhattacharyya (1999), and Venkatesan and Kumar (2004).

The methods and techniques that have been applied to the available data are mainly the ones that have been suggested in the two streams of analyzed literature. First, the methods used in a comparable context, like Decision Trees (Asare-Frempong & Jayabalan, 2017; Olson, 2012), ensemble methods (Apampa, 2016; Haupt et al., 2018), and Artificial Neural Networks (Haupt et al., 2018; Zakaryazad & Duman, 2015). Second, those techniques suggested by researchers strictly focusing on the imbalanced data issue. Among these techniques, sampling methods unveiled a significant potential in improving the generalization ability, in particular when combined with ensemble algorithms. One of the most successful attempts has been represented by the use of a Balanced Random Forest (Chen & Breiman, 2004), an evolution of a Random Forest (Breiman, 2001) that has been shown to improve the prediction accuracy of the minority class. Another approach that has led to fairly good results is the RUSBoost (Seiffert et al., 2010). Similarly to the Balanced Random Forest example, RUSBoost represents an attempt to improve the classification performance of the AdaBoost algorithm (Yoav & Robert, 1996) when dealing with skewed data. Chen and Breiman (2004) exhaustively explained the importance of combining sampling techniques with tree-based ensembles. Indeed,

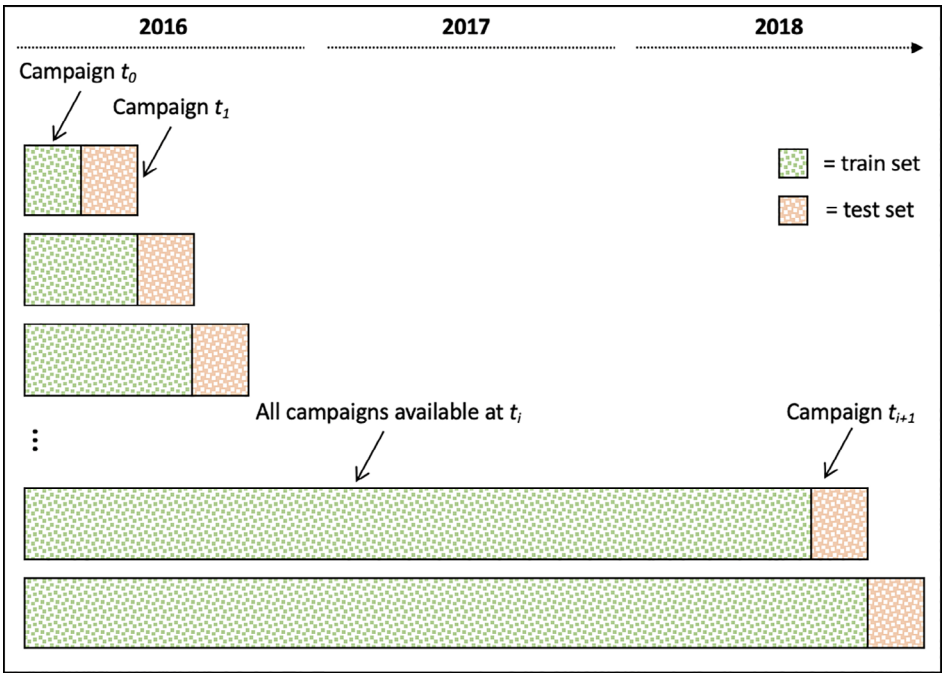

FIGURE 2 Walk-forward validation procedure

dealing with a heavily imbalanced dataset, there is a significant probability that the trees build on bootstrap replica of the original training set, may eventually contain few or no instances of the minority class, resulting in poor performance of the classifier. This concept may be also extended to the training of Artificial Neural Networks. In particular, the balancing matter has been taken into consideration when determining the proper structure of the batches fed into the network while training. This method has been proposed in particular for image data classification (Shimizu et al., 2018) but can be also implemented in other classification problems. It aims to balance the class ratio of training samples that are passed to the Artificial Neural Network while training, before updating the internal model parameters. Hence, the model has been constructed giving the same importance to both the minority and majority classes. In the upcoming section, the aforementioned methods will be tested, showing their ability in predicting donor responses. Most of the techniques that include sampling methods have been implemented with the Python library Imbalanced-learn (Lemaitre et al., 2017), Artificial Neural Networks with the support of TensorFlow and Keras (Chollet, 2015) while other machine learning models (like Decision Trees) with the use of Scikit-learn (Pedregosa et al., 2011). To obtain the best configuration for each model, a robust hyperparameters optimization procedure has been followed. A wide space of possible configurations has been explored using a sequential model-based optimization with a tree parzen estimator, a method based on Bayesian Optimization (Bergstra et al., 2012). This technique has been implemented through the use of the Hyperopt library (Bergstra et al., 2015). Similarly to the work of Haupt et al. (2018), Table 2 presents the space of hyperparameters that has been explored when looking for the best configuration of our models.

In order to properly take into account, the class imbalance when assessing the performances of the different models, two different measures have been used. The first measure is the ROC curve (Receiving Operating Characteristics) and the area under it. This metric, widely used in literature when dealing with imbalanced classes (Apampa, 2016; Asare-Frempong & Jayabalan, 2017; Chawla et al., 2004; Haupt et al., 2018), provides a summary of the performance of a classifier for no fixed threshold, showing the trade-off between true positive rate and false positive rate. The area under the ROC curve is called AUC (namely area under the curve) and represents a measure that may be useful when comparing several models. However, how properly explained by Powers (2012), AUC should not be relied on as a single definitive measure and, when possible, information concerning the relative cost of positive and negative classes in the studied application should be used. This is what we investigated through the use of Net Incremental Income. Before describing the procedure that led us to this score, a summary of Confusion Matrix and its derived measures is proposed.

The matrix in Figure 1 shows how a binary classification model performs by comparing the predicted labels and the actual ones. Possible cases are true negatives (TN), false positives (FP), false negatives (FN), and true positives (TP). The main metrics that can be drawn from this table are:

- $\text{Sensitivity} = \dfrac{\text{TP}}{\text{TP}+\text{FN}}$
- $\text{Specificity} = \dfrac{\text{TN}}{\text{TN}+\text{FP}}$
- $\text{Precision} = \dfrac{\text{TP}}{\text{TP}+\text{FP}}$

The ROC curve is obtained by plotting the true positive rate against the false positive rate, the first measure corresponds to the sensitivity and the latter is obtained as 1 – specificity.

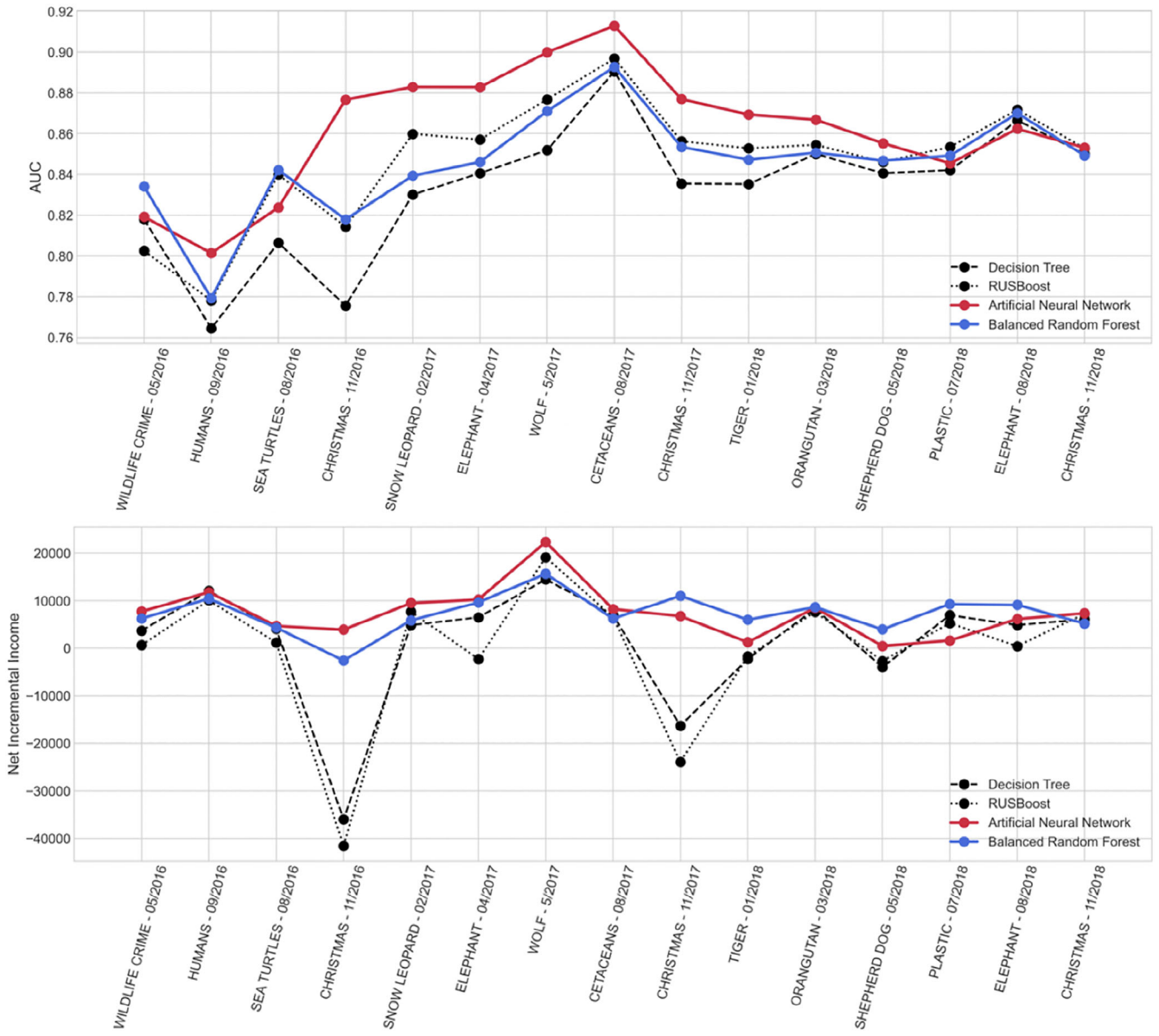

FIGURE 3 AUC and net incremental income results on different test sets

TABLE 3 Summary of results

Model	Average AUC	Standard deviation	Average net incremental income	Standard deviation	Total net incremental income
Artificial neural network	0.862	0.029	$7364.86	$5144.97	$110,472.89
Balanced random forest	0.847	0.024	$7265.31	$3971.21	$108,979.67
RUSBoost	0.846	0.029	−$425.46	$14,184.06	−$6381.85
Decision tree	0.833	0.031	$1345.83	$12,193.87	$20,187.47

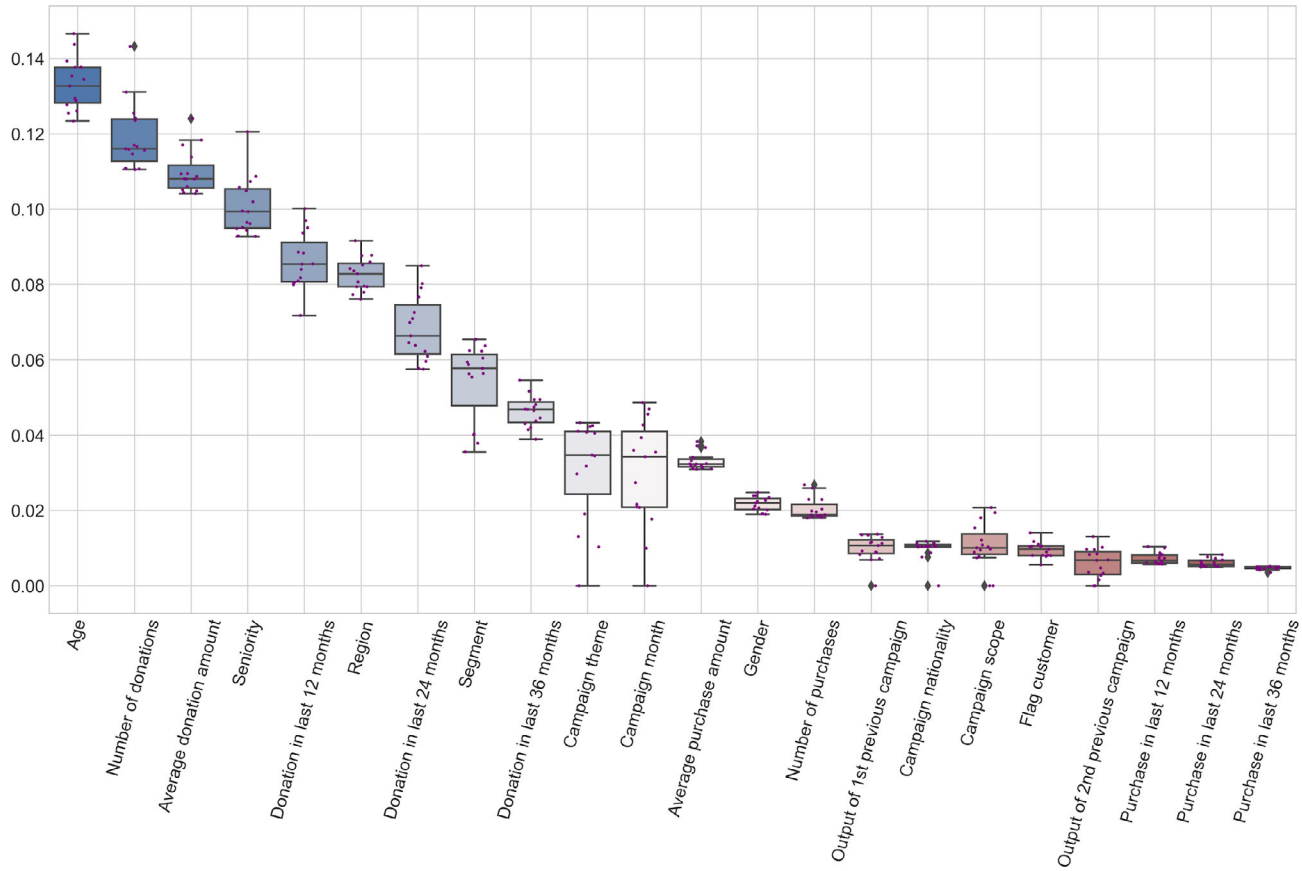

FIGURE 4 Feature importance in balanced Random Forest classifier

The ad hoc measure represented by Net Incremental Income is obtained by firstly computing the income of each campaign as the sum of the amount donated by each donor and secondly subtracting the total costs, represented by the costs for sending the donation request to the whole target and the thank you letter to those that eventually responded. This will give us the as-is net income, before the use of our proposed Machine Learning approach. The potential net income (to-be) after the implementation of predictive models is obtained by subtracting the cost of sending the request to all those predicted as 1 (TP + FP) and then the cost of the thank you letter only to those who donated (TP) to the amount actually donated by the correctly predicted donors (TP). Hence, it is important to highlight the fact that false positive and false negative instances have quite different costs in this application. A high number of false negatives means wasting money by sending letters to non-donors but, on average, the cost of sending one extra letter is 60 times lower than the value corresponding to the missed opportunity of classifying a donor as a non-donor. This problem has been approached in the hyper-parameters tuning phase. Indeed, when looking for the best configuration of each model, the maximized objective function was represented by the difference between the net income as-is and the net income to-be.

4 | RESULTS AND DISCUSSION

Given the high amount of data available and the temporal distribution of the campaigns, models performance has been rigorously assessed with the use of out-of-sample and out-of-time test sets (Haupt et al., 2018). The main idea behind this approach is to simulate a production environment, where the model is trained on currently available data and tested on future campaigns (that the model should be able to predict in real-life conditions). Hence, the approach is similar to a walk-forward validation procedure (Stein, 2007), where the time frame and sample size dedicated to testing are fixed and dictated by the upcoming campaign. This scheme is summarized in Figure 2.

Following this procedure, it will be possible to highlight the average performance of different models but keeping track of their ability to predict each campaign, singularly. The performance evaluation makes use of the two measures presented in Section 3, namely AUC and Net Incremental Income. In such a way, the plots in Figure 3 provide a comprehensive overview of how the different models perform both with Machine Learning and marketing-oriented KPIs. While AUC still maintains a significant role in the overall assessment, the second metric becomes much more interesting from a marketing decision-making perspective.

It is possible to observe how from an AUC point of view all the models share very similar behavior. De facto, in terms of predictability they often overlap if we consider the average performance and its standard deviation (see table 3). That is why observing the economic performance becomes even more important in discerning the best models. From the second plot, reporting the Net Incremental Income, it is immediately possible to discard two models, namely the Decision Tree and the RUSBoost ensemble, from the alternatives. Indeed, these two models dramatically fail in two campaigns (Christmas 2016 and Christmas 2017), leading to a situation where the savings associated with the reduced number of letters sent are overwhelmed by the missed donation opportunities. On the other hand, the Artificial Neural Network model and the Balanced Random Forest achieve to deliver compelling results in almost every campaign.

The necessity highlighted by Powers (2012) of considering also an application-based KPI while evaluating classification models becomes much more evident in Table 3. The two ensemble methods Balanced Random Forest and RUSBoost are extremely close in terms of AUC performance. Nonetheless, their ability to correctly classify the most valuable donors is dramatically different. That said, all the models overlap in terms of one standard deviation when it comes to AUC. Therefore, it will be much beneficial to drive the model selection procedure by the Net Incremental Income results. Indeed, it is possible to observe how the Artificial Neural Network and the Balanced Random Forest represent the two best alternatives, allowing an average improvement of $7K on each campaign, with an overall saving of $110K circa. The network-based model has the slight advantage of never leading to negative results, namely a worsening from the as-is situation (this only happens once with the forest algorithm – Christmas 2016). Conversely, this worsening happens more often with the two other tree-based methods rising, in two cases, serious concerns. Indeed, these two models missed about $40K and $20K in 2016 and 2017 Christmas campaigns, respectively. The forest and the network offer performances more robust to large variations in the response rate of the campaigns (as it happens during those two Christmas campaigns where the response rate rises up to 5.3% from a grand mean equal to 3.3%).

Hence, despite being different in their topology structure, the average performance of the Balanced Random Forest and the Artificial Neural Network is quite similar. Following the benchmark of the models in predicting the responses of the test campaigns, a feature importance study is hereby proposed. This analysis has the scope of unveiling the most influencing parameters in classifying donors and nondonors. Hence, these parameters might eventually play a crucial role in the marketing strategy formulation for future campaigns. In order to have a robust estimate of this importance, according to the walk-forward approach that has been followed, feature importance has been computed for each of the 15 trained models.

The feature importance analysis is performed on the balanced random forest, by all means a more interpretable model than the artificial neural network. The summary of the obtained results is exposed in Figure 4. Since this classifier belongs to the tree-based family, feature importance will be computed using Mean Decrease Impurity (MDI; Breiman, 2001; Louppe et al., 2013). Indeed, dealing with a (Balanced) Random Forest, the importance of each variable is calculated as the weighted decrease of a given impurity measure, for all the nodes of a tree and is finally averaged over all the trees constituting the forest. When the Gini index (1) is used as the impurity measure, the MDI is called Gini importance.

- $G = 1 - \sum_{i=1}^{n} p^2(c_i)$ (1)

where $p(ci)$ is the probability/percentage of ith class in a node. The impurity decrease is meant as the observed decrease of the importance of a certain node, after its split into child nodes. The importance of a node j is expressed as ni_j.

Assuming two child nodes only, it could be explained as:

- $ni_j = w_j C_j - w_{\text{left}(j)} C_{\text{left}(j)} - w_{\text{right}(j)} C_{\text{right}(j)}$

Where:

- $ni_j =$ importance of node j
- $w_j =$ number of samples reaching node j divided by the total number of samples
- $C_j =$ impurity (Gini) of node j

Subsequently, the importance of the ith feature is calculated as:

- $fi_i = \frac{\sum_{\text{node } j \text{ splits on feature } i} ni_j}{\sum_{k \in \text{all nodes}} ni_k}$

Finally, this measure is normalized to 1 dividing it by the sum of the importance of all the features used by the model.

5 | CONCLUSIONS

The analysis shows how the most compelling results are obtained with the Balanced Random Forest and the Artificial Neural Networks with balanced batches. This confirms how the classification performance of two widely used algorithms, in the case of imbalanced data, can be significantly improved when they are combined with sampling methods. On a less technical note, the case study reports an interesting application of machine learning in a nonprofit context. It surely does not represent the first example but we hope that the quite promising economical results may be encouraging to foster the adoption of such tools even in small-scale organizations. Indeed, the results suggest how the cost benefits may be particularly helpful especially when operating on a tight budget.

Another important insight coming out from this analysis is the confirmation of the importance of RFM variables. Indeed, feature importance analysis shows how these variables are crucial in determining customer behavior, even in a nonprofit context. Among the first parameters, a great influence is delivered by the columns containing data that express the attitude of the potential donor toward the organization. Strictly speaking, knowing if the recipient has

contributed to previous campaigns over the last 3 years and perceiving the number of donations and their average amount seems to be much more discriminant than the theme of the campaign itself. Nonetheless, a crucial role in understanding their fidelity for the themes sponsored by the organization is played by the age of the donors and the years passed since the first time they supported WWF (seniority). This can be particularly useful in supporting the choice of appropriate techniques to implement a future segmentation of the donor base.

ACKNOWLEDGMENT

None.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from WWF Italia. Restrictions apply to the availability of these data, which were used under license for this study. Data are available from the author(s) with the permission of WWF Italia.

ORCID

Davide Cacciarelli https://orcid.org/0000-0001-6664-9038

Marco Boresta https://orcid.org/0000-0002-9356-204X

REFERENCES

Ahmed, T., Oreshkin, B., & Coates, M. (2007). Machine learning approaches to network anomaly detection. *Proceedings of SysML.*, *7*, 1–6.

Allenby, G., Leone, R., & Jen, L. (1999). A dynamic model of purchase timing with application to direct marketing. *Journal of The American Statistical Association.*, *94*, 365–374.

Apampa, O. (2016). Evaluation of classification and ensemble algorithms for bank customer marketing response prediction. *Journal of International Technology and Information Management*, *25*(4), 85.

Asare-Frempong J. M., & Jayabalan, M. (2017). Predicting customer response to bank direct telemarketing campaign. *2017 IEEE The International Conference on Engineering Technologies and Technopreneurship (ICE2T 2017)*.

Ayetiran, E., & Adeyemo, A. (2012). A data mining-based response model for target selection in direct marketing. *International Journal of Information Technology and Computer Science*, *1*, 9–18.

Baesens, B., Viaene, S., Van den Poel, D., Vanthienen, J., & Dedene, G. (2002). Bayesian neural network for repeat purchase modelling in direct marketing. *European Journal of Operational Research*, *138*, 191–211.

Berger, P., & Magliozzi, T. (1992). The effect of sample size and proportion of buyers in the sample on the performance of list segmentation equations generated by regression analysis. *Journal of Direct Marketing*, *6*(1), 13–22.

Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., & Cox, D. (2015). Hyperopt: A python library for model selection and hyperparameter optimization. *Computational Science & Discovery*, *8*, 014008. https://doi.org/10.1088/1749-4699/8/1/014008

Bergstra, J., Yamins, D., & Cox, D. (2012). Making a science of model search. *JMLR Workshop and Conference Proceedings*, *28*, 115–123.

Bhattacharyya, S. (1999). Direct marketing performance modeling using genetic algorithms. *INFORMS Journal on Computing*, *11*, 248–257.

Bodenberg, T. M., & Roberts, M. L. (1990). Integrating marketing research into the direct-marketing testing process: The market research test. *Journal of Advertising Research*, *30*(5), 50–60.

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science.*, *17*(3), 235–255.

Breeze, B., & Jollymore, G. (2017). Understanding solicitation: Beyond the binary variable of being asked or not being asked. *International Journal of Nonprofit and Voluntary Sector Marketing*, *22*(4), e1607.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Chawla, N., Japkowicz, N., & Kolcz, A. (2004). Editorial: Special issue on learning from imbalanced data sets. *SIGKDD Explorations*, *6*(1), 1–6.

Chen, C., & Breiman, L. (2004). *Using random forest to learn imbalanced data*. University of California.

Chollet, F. (2015). Keras. Retrieved from https://keras.io

Cieslak, D., Chawla, N., & Striegel, A. (2006). Combating imbalance in network intrusion datasets. *2006 IEEE International Conference on Granular Computing* (pp. 732–737).

Cui, G., & Wong, M. L. (2004). Implementing neural networks for decision support in direct marketing. *International Journal of Market Research.*, *46*(2), 235–254.

Cui, G., Wong, M. L., & Lui, H. K. (2006). Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. *Management Science*, *52*, 597–612.

Fader, P., Hardie, B., & Lee, K. (2005). Rfm and clv: Using iso-value curves for customer base analysis. *Journal of Marketing Research American Marketing Association*, *4*, 415–430.

Galar, M., Fernandez, A., Barrenechea, E., Sola, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, *42*(4), 463–484.

Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*, *2*, 42–47.

Gönül, F., & Shi, M. (1998). Optimal mailing of catalogs: A new methodology using estimable structural dynamic programming models. *Management Science*, *44*, 1249–1262.

Haughton, D., & Oulabi, S. (1997). Direct marketing modeling with CART and CHAID. *Journal of Direct Marketing*, *11*(4), 42–52.

Haupt, J., Bender, B., Fabian, B., & Lessmann, S. (2018). Robust identification of email tracking: A machine learning approach. *European Journal of Operational Research*, *271*, 341–356.

Isermann, R. (1997). Supervision, fault-detection and fault-diagnosis methods—An introduction. *Control engineering practice.*, *5*(5), 639–652.

Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risk from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making*, *11*(1), 51.

Ki, E. J., & Oh, J. (2018). Determinants of donation amount in nonprofit membership associations. *International Journal of Nonprofit and Voluntary Sector Marketing*, *23*(7), e1609.

Ladyzynski, P., Bikowski, K. Z., & Gawrysiak, P. (2019). Direct marketing campaigns in retail banking with the use of deep learning and random forests. *Expert Systems with Applications*, *134*, 28–35.

Lemaitre, G., Nogueira, F., & Aridas, C. (2017). Imbalanced-learn: A python tool-box to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, *18*, 1–5.

Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013). Understanding variable importances in forests of randomized trees. *Advances in Neural Information Processing Systems*, *26*, 431–439.

Olson, D. (2012). Direct marketing decision support through predictive customer response modeling. *Decision Support Systems*, *54*(1), 443–451.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J. A., Passos, D. C., Brucher, M., Perrot, M., Duchesnay, E. & Louppe, G. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Pentecost, R. D., & Andrews, L. (2009). Differences between students and non-Students' willingness to donate to charitable organizations. *International Journal of Nonprofit and Voluntary Sector Marketing*. *15*(2), 122–136.

Powers, D. M. W. (2012). The problem of area under the curve. *IEEE International Conference on Information Science and Technology*.

Rao, V., & Steckel, J. (1995). Selecting, evaluating, and updating prospects in direct mail marketing. *Journal of Direct Marketing*, *9*(2), 20–31.

Rupp, C., Kern, S., & Helmig, B. (2014). Segmenting nonprofit stakeholders to enable successful relationship marketing: A review. *International Journal of Nonprofit and Voluntary Sector Marketing*, *19*(2), 76–91.

Seiffert, C., Khoshgoftaar, T., Van Hulse, J., & Napolitano, A. (2010). Rusboost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, *40*(1), 185–197.

Shimizu, R., Asako, K., Hiro, O., Morinaga, S., Hamada, M., & Kuroda, T. (2018). Balanced mini-batch training for imbalanced image data classification with neural network. *Proceedings 1st IEEE International Conference on Artificial Intelligence for Industries*, *1*, 27–30.

Singh, N., Khalfay, N., Soni, V., & Vora, D. (2017). Stock prediction using machine learning a review paper. *International Journal of Computer Applications*, *163*, 36–43.

Stein, R. (2007). Benchmarking default prediction models: Pitfalls and remedies in model validation. *Journal of Risk Model Validation*, *1*, 77–113.

Such, F., Pillai, S., Brockler, F., Singh, V., Hutkowski, P., & Ptucha, R. (2018). Intelligent character recognition using fully convolutional neural networks. *Pattern Recognition*, *88*, 604–613.

Venkatesan, R., & Kumar, V. (2004). A customer lifetime value framework for customer selection and resource allocation strategy. *Journal of Marketing*, *68*(4), 106–125. https://doi.org/10.1509/jmkg.68.4.106.42728

Yanmin, S. A., Wong, A., & Kamel, M. S. (2011). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, *23*(4), 687–719.

Yoav, F., & Robert, S. (1996). Experiments with a new boosting algorithm. Machine Learning*: Proceedings of the Thirteenth International Conference* (pp. 148–156).

Zahavi, J., & Levin, N. (1999). Applying neural computing to target marketing. *Journal of Direct Marketing*, *11*(4), 76–93.

Zakaryazad, A., & Duman, E. (2015). A profit-driven artificial neural network (ann) with applications to fraud detection and direct marketing. *Neurocomputing*, *175*, 121–131.

How to cite this article: Cacciarelli, D., & Boresta, M. (2021). What drives a donor? A machine learning-based approach for predicting responses of nonprofit direct marketing campaigns. *International Journal of Nonprofit and Voluntary Sector Marketing*, e1724. https://doi.org/10.1002/nvsm.1724