Vrije Universiteit Amsterdam

Host organization: Vandebron

Master Thesis

Leveraging XAI and IAI to comprehend email marketing

Author: Abel van Gennep (2621604)

Company supervisor:	Jiri Brummer	(Vandebron)
1st supervisor:	Dr. Karine Miras	
2nd reader:	Dr. Frank van Harmelen	

A thesis submitted in fulfillment of the requirements for the VU Master of Science degree in Business Analytics

September 14, 2023

Abstract

In the competitive landscape of green energy providers, customer retention is crucial. For Vandebron, a Dutch company, it's vitally important to comprehend and enhance the effectiveness of their retention period. This research explores the effectiveness of supervised machine learning models, specifically tree-based models, in predicting email conversion. The study is focused on comprehending the influence of variables affecting prediction outcomes, leveraging interpretable and explainable AI (IAI and XAI). Our findings indicate a trade-off between model comprehensibility and predictive power. A Decision Tree model emerged as highly interpretable yet less effective, whereas a Histogram Gradient Boosting Classifier showed superior performance, albeit with increased complexity and reduced interpretability.

remainder is confidential

Keywords: Retention, email conversion, XAI, IAI, SHAP, tree-based, Decision tree

Contents

List of Figures

List of Tables

1

Introduction

Vandebron, a Dutch energy provider, is mainly engaged in the buying and selling of sustainable generated energy. Therefor the customer base of Vandebron are energy consumers. This study centers on the retention of those customers who are currently already in the customer base of Vandebron. Among the diverse types of energy contracts, this research narrows its focus to subscription-based contracts for house holds, having a duration of either one or three years.

In the context of customer retention tasks, machine learning has become a field of interest demonstrating efficacy in discerning patterns and processing voluminous data sets. The high performance in terms of predictive power makes machine learning models attractive for different machine learning task. Often these models operate as 'black boxes', offering little insight into the reasoning behind predictions. This lack of transparency impedes the transformation of predictions into actionable strategies, an essential requirement to provide decision-makers with insights.

Customer renewal rates, managed by a dedicated retention team, serve as crucial performance metrics. The balance of higher retention rates and new customer acquisition directly impact the company's growth trajectory. Thus, enhancing retention is key to maintaining Vandebron's competitiveness. Customers can renew there contract within 150 days before there current contract ends. This time frame of 150 days is therefor called the retention period. Within this retention period contracts are often send to customers via email.

1.1 Problem statement

The retention team wants to investigate possibilities to optimize the retention strategy by using data-driven insights. The sending of contract offers via email marketing can potentially be adjusted for all customers, sub-groups or even individuals if deemed relevant. Before investigating the effects of adjusting the retention strategy, the marketing team is looking for data-driven insights into potentially interesting directions. To generate these insights the following two research questions are developed:

- How do machine learning models differ comparing predictive power and comprehensibility in the context of predicting email conversion?
- Does the combined use of IAI and XAI methods enhance our understanding of explanatory variables on the prediction outcome of email conversion?

1.2 Objectives

The aim of this study is to investigate, which types of mailings yield higher conversion rates and understand why for these emails conversion rates are higher. Therefor machine learning models are employed to investigate historical data, regarding customers and emails send to customers. Insights enable the retention team to set the course of action, therefor insights that can be translated into actionable strategies hold greater business value. Specifically those that can enhance email conversion rates.

1.3 Organization context

Vandebron is a Green tech energy provider and is part of the Essent group, which is one of the 3 biggest energy suppliers in the Netherlands. The main goal of Vandebron is generating smart solutions that could accelerate the energy transition. To finance the energy transition and stay financially stable they make margin on there products, like for example energy contracts.

This project is executed by the Commercial analytics team, which is one of the two data analytics teams. The data analytics team is responsible for maintaining a data stack, supplying different domains with insights and provide prove of concepts to support data driven decision making. This project is designed to support the marketing domain, within Vandebron with insights on historical data.

1.4 Related work

The potential of machine learning for marketing purposes is underlined, by several papers [1, 2, 3]. In literature this is accredited to the ability of machine learning methods to process large-scales unstructured data, having flexible model structures and yielding strong predictive performance [2]. Statistical testing is often preferred in the field of marketing, mainly because machine learning models rely on engineered features and flexible structures, which can results in models becoming "black-boxes" that excel in prediction but lack in delivering insights [2]. This threat is not only a threat for marketing purposes, but also for other high-stake scenarios were explaining 'black-box' models is not deemed sufficient [4].

More recently understanding how, when, and why predictions are made has become very popular. This interest led to a lot of research into the field, leading to a variety of techniques focused on interpreting and explaining machine learning models [5, 4]. The terms explainability and interpretability are often used interchangeably, but in this study we adhere to the following distinction. Interpretability refers to the ease with which a human can understand why a model makes certain decisions. It's about understanding the model's overall functioning and therefor it can only performed on so called 'white-box' models [6]. Paradigms underlying these type of problems fall within the field of interpretable AI (IAI). Explainability is about understanding the predictions made by complex models and making the decision process understandable for humans. The goal is to provide clear explanations that are comprehensible for the end users understanding why decisions were made by the model, without the need to understand the inner functioning of a model [7]. This allows for investigating 'black-box' models using different techniques. Within the field of explainable AI (XAI) a sub-group of methods are applicable over a wide-range of different (machine leaning) models, called model-agnostic methods [5].

So far, there is no scientific research on retention approaching the task as an email conversion problem using XAI or IAI. The task of retention has received some attention investigating customer renewal (or churn) in subscription-based companies using different methods [8, 9, 10]. Research in this area of practice is tailored towards explaining, black-box models. While marketing research is calling for researchers to extend the methods to improve interpretability, to make machine learning more suitable [2]. This stresses the need for more scientific research on applying the field of IAI and XAI for marketing purposes. An other factor causing resistance in the field of marketing towards employing machine learning is the difficulty of proving causality [2]. This problem is despite recent developments in the field of IAI [11], recognized in this paper and considered a limitation.

1.5 Report structure

In this report the background of different machine learning methods and techniques will be explained. After which the methodology provides information regarding the data, preprocessing, selection of features, applied models and the evaluation of the used methods. The result chapter evaluates the different models and focuses on comprehension of the tuned models applying XAI and IAI techniques. In the final chapter results are discussed, the work is concluded and the limitations are discussed among some suggestion for further research directions.

2

Background

In this chapter the background on techniques used for modeling are explained diving into tree based-models, data transformations and a selection of machine learning models. In the second section of this chapter the aim is to explain the background of methods used for understanding the decision process.

2.1 Tree-based models

Tree-based models systematically divide the space of variables into several subsets, leading to discretization of continuous explanatory variables. Their structure is characterized by a series of straightforward IF-THEN statements, which lend simplicity and interpretability to the resulting decision trees. Tree-based models excel in situations where the relationship between explanatory and predictive variables exhibits a moderate level of complexity, and where data trends are non-linear and non-monotonic. These characteristics render them highly suitable for a multitude of tabular machine learning tasks [5].

Interestingly, tree-based models often function effectively without necessity for heavy engineering for features, allowing the model to perform well without data transformations scaling and missing values. This is because the division of data within these models relies on inequality (i.e. whether a explanatory variable's value is greater or lesser than a certain threshold), rather than being determined by the precise value of the explanatory variable itself [5, 12].

Despite their fundamental simplicity, tree-based models often deliver impressive results. They have been observed to outperform neural networks for regression or classification on tabular tasks (data displayed in columns or tables), showing their efficacy and robustness [13].

2.2 Data transformations

The need for specific transformations is often model-dependent, tree-based models function effectively with minimal transformations. This is because the division of data within these models relies on inequality (i.e. whether a features value is greater or lesser than a certain threshold), rather than being determined by the precise value of the explanatory variable itself [5, 12].

Categorical encoding is a crucial transformation that converts categorical data into a format that can be ingested by machine learning algorithms to enhance predictive accuracy, particularly when there is no ordinal relationship among the variables. This process involves the numerical representation of categories containing string values.

The preparation of data for machine learning algorithms can involve various transformations such as imputing missing data, transforming categorical data, and scaling of measurements. Transforming data serves the goal of converting the data into a usable and interpretable format and often lead to faster convergence of machine learning models [12]. Different models perform well with the necessity for different transformations. Decision trees often perform well with a minimal amount of transformations, for example scaling is often not needed because of the discretization.

Categorical encoding is used to convert categorical data into a format that can be provided to machine learning algorithms to improve prediction, when their is no ordinal relation between the variables. Therefor data containing labelled data are transformed into a suitable format. The first method is suitable for low cardinality and called one-hot encoding, which generates a sparse matrix generating a explanatory variable for every label. The explanatory is assigned a positive or negative label dependent on the pressence of the label in the original data [12]. For explanatory variables containing a high cardinality this can create explanatory variables, with low variance and therefor this is often not desirable. A technique called frequency encoding is used to replace each category by the frequency or proportion of occurrences of that category in the explanatory variable column. Frequency encoding can help preserve the information about the frequency of categories [14].

Various strategies for addressing missing values in machine learning data have been developed, their applicability hinging significantly on the reason behind the absence of the value [15]. If the data is Missing Completely at Random (MCAR), meaning the missingness has no correlation to any other variables, elementary techniques (E.g. mean, mode, or median imputation) are often sufficient. On the other hand, when data is classified as Missing at Random (MAR), where missingness is tied to certain observed explanatory variables but not the unobserved ones, more sophisticated methods become necessary. A prime example of such a method is the indicator technique, which is typically used when the missingness can signal specific characteristics, like the non-completion of a survey [16]. In the case of Missing Not at Random (MNAR) data, which suggests missingness is associated with the target variable or the missing data itself, handling becomes more intricate. Such missingness can introduce biases into classification predictions, necessitating advanced imputation or modeling techniques. Simple imputation approaches can still be applied, albeit with caution due to their potential to distort the full distribution of the data [15]. These methodologies allow for more accurate handling and interpretation of missing data, contributing significantly to the robustness of machine learning predictions [17].

The scaling of explanatory variables is commonly applied to standardize datasets and enhance the efficiency of model training and lessen the impact of outliers. However, it could obscure model interpretability by introducing another layer of complexity. Therefore, the necessity for scaling frequently depends on the specific algorithm being employed. For instance, in the context of tree-based models, scaling is often optional because these models partition data based on value comparisons, but not the absolute values themselves [18]. Yet, it's worth noting that scaling can influence impurity measures, potentially impacting the determination of splits and the assignment of explanatory variable importance. Several standardization methods exist to cater different data characteristics and requirements. The *Minmaxscaler*, for example, re-scales explanatory variables to fit within a 0 to 1 range without making distribution assumptions, making it a good fit for general standardization needs. On the other hand, the *RobustScaler* employs statistics that are impervious to outliers for explanatory variable scaling. Specifically, it subtracts the median from each data point and divides by the Interquartile Range (IQR), confining most scaled values within the 0 to 1 range. This approach is especially beneficial for explanatory variables with extreme outliers, as it prevents these anomalies from overly distorting the model's understanding of the data [19].

2.3 Feature selection methods

Feature selection is the process of selecting a subset of engineered features out of all explanatory variables and is often implemented in machine learning, offering multiple benefits such as enhancing model performance, reducing the risk of overfitting, increasing prediction performance and reducing model complexity [20]. The cumulative impact of these benefits results in model interpretability. In contexts where model understanding is of special interest, such as this research, the improved interpretability of models holds particular importance and relevance.

Explanatory variable selection techniques are generally classified into three broad categories: filter methods, wrapper methods, and embedded methods. Filter methods, in the context of decision trees, involve a pre-processing step where each explanatory variable is individually evaluated for its relevance and predictive power (E.g. evaluation of correlations or variance within the explanatory variable). Wrapper methods conceptualize the selection of an optimal explanatory variable set as a search problem, examining and contrasting diverse explanatory variable combinations. This search process is informed by the performance of a specific machine learning model, hence the name 'wrapper'. One such strategy is *Forward selection*, an iterative method that begins with an empty explanatory variable set. During each iteration, it incorporates the explanatory variable that yields the maximum enhancement in model performance. This process is sustained until no further improvement is witnessed or a predetermined explanatory variable limit is reached [21]. *Recursive Feature Elimination* (RFE), also an iterative procedure, but uses explanatory variable importance to remove the least important explanatory variables based on their importance scores. This elimination process is reiterated until only a specified number of explanatory variables persist. Both forward selection and RFE provide a practical and systematic way to navigate the search space, optimizing the balance between model performance and complexity. Embedded explanatory variable selection is, when the selection is incorporate is the model training phase, this is often automatically the case with Desicion trees, since it can automatically decide not to select certain explanatory variables.

2.4 Classification models

This chapter delves into the foundations and working mechanisms of four tree-based models; Decision Trees, Random Forests, Adaptive Boosting (AdaBoost), and the Histogram-based Gradient Boosting Tree (HistGradientBoosting).

Classification and Regression Trees (CART) is a highly intuitive and versatile machine learning method that is applicable to both categorical (classification) and continuous (regression) dependent variables. This popular decision tree implementation models the data decisions based on specific feature values [22], the method has proven it's effectiveness over many years and applications in different fields [23, 24].

The primary principle behind CART is binary recursive partitioning, wherein the data is split into two child nodes at each decision node, this is done according to a certain measure (E.g. Gini or entpy importance). The splitting process starts from a root node that includes the entire dataset, and it proceeds by making the split that results in the largest possible reduction in heterogeneity (classification) or variation (regression) within the child nodes [22]. CART is prone to overfitting, especially with complex trees. Therefore, overfitting techniques are often used. A possible solution is control of the tree size and complexity, enhancing the model's generalization capabilities [22].

Random Forests, an extension of decision trees, addresses some of the limitations of single decision trees. By constructing a multitude of decision trees and averaging their predictions, Random Forests mitigate the risk of overfitting and offer a more stable and generalized model [25].

Random Forests are particularly efficient with large datasets with high dimensionality. They can manage a large number of features without selection, by using a technique called bootstrap aggregating or bagging to increase robustness and performance compared to Decision trees [25].

As an ensemble model, Random Forests often provide superior performance compared to individual models; however, this comes at the cost of interpretability as the decision-making process within a forest of trees is far more complex than that of a single tree [26].

Adaptive Boosting (AdaBoost), operates by iterative adjusting the weights of trees during training. It is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified, this puts special attention to the weaker classifiers. Subsequently, a new model is trained on these re-weighted instances. In doing so, it creates a sequence of weak models that gradually improve at the classification task [27]. On the downside, AdaBoost can be sensitive to noisy data and outliers as it strives to correct all misclassifications, including those that may stem from noise [28]. Moreover, AdaBoost demands careful tuning of various hyper-parameters, such as the number of iterations, and may be more computationally expensive than some other methods [29].

Histogram-based Gradient Boosting (HistGradientBoosting), signifies another progression in the development of boosting techniques. By binning continuous input explanatory variables into discrete bins and subsequently constructing the histogram of the data, HistGradientBoosting considerably accelerates the training process and allows the algorithm to utilize integer-based data structures [30]. Advantages are proficiently managing of large datasets and provides exceptional predictive performance, requiring less hyper-parameter tuning than traditional Gradient Boosting [30]. It is also designed to handle missing values without prior data imputation, making it more user-friendly [30].

2.4.1 Interpretable AI: Decision Tree

The easy-to-follow visual representation of decision trees significantly contributes to their interpretability. The tree root, located at the top, signifies the initial condition, while the leaves at the bottom denote the final results. The journey from the root to any leaf constitutes a chain of decisions leading to a particular outcome, thereby explaining the prediction process at it's full extent [22]. Also the decision tree visualization allows to follow the support in the train data, this allows to explain the correlation between the explanatory variables and prediction. Therefor this method is recommended for research looking for interesting and actionable explanatory variables [31].

2.4.2 Explainable AI: SHAP

Explainable AI techniques, on the other hand, are used when humans are made to enable a human to understand why a predictions was made by the AI. This allows to even explain models considered 'black box' models [5]. In this research we will be looking at a specific method, which belongs to the class of 'model agnosic' methods, meaning that the method can be applied on many different types of AI's.

Shapley Additive Explanations (SHAP) is a unifying measure that allocates the change in prediction expectation, when considering a particular feature, to that feature itself. The method was first introduced by Lundberg and Lee [32], based on the cooperative game theory concept the Shapley value.

To compute the SHAPley values, an initial average prediction for the entire dataset is calculated. Subsequently, for each instance, every potential feature combination is considered, leading to a power set of features ($2^{features}$). This extensive exploration is due to the underlying idea of Shapley values, which is based on the idea that the outcome of all feasible combinations of variables should be taken into account to determine the contribution of a single variable.

For each feature combination, the model prediction is compared when including and excluding the feature of interest. The difference between these two predictions reflects the marginal contribution of that feature. The Shapley value for each feature is then calculated by averaging these marginal contributions over all possible feature combinations.

Due to the exponential explosion, if the number of features grow, calculating exact Shapley values can be computationally heavy. As a result, popular approximation methods are designed. One of these methods is KernelSHAP, which is allowed to use for both trees and ensemble methods. This method also provides the possibility to account for interactions among features [33].

Shapley values are used to make individual prediction, but an interesting aspect that certain implementations of SHAP provide is to account for interaction values. Similar to the calculation of Shapley values, this process averages over all possible feature coalitions. The end result is a square matrix of SHAP interaction values for each instance. This gives a symmetric matrix, with the main effects on the diagonal and the interactions effects off diagonal.

3

Methodology

confidential

4

Results

confidential

5

Conclusion and Discussion

confidential

References

[1] Vinicius Andrade Brei et al. **Machine learning in marketing: Overview, learning strategies, applications, and future developments**. *Foundations and Trends® in Marketing*, **14**(3):173–236, 2020.

[2] Liye Ma and Baohong Sun. **Machine learning and AI in marketing–Connecting computing power to human insights**. *International Journal of Research in Marketing*, **37**(3):481–504, 2020.

[3] Joseph F Hair Jr and Marko Sarstedt. **Data, measurement, and causal inferences in machine learning: opportunities and challenges for marketing**. *Journal of Marketing Theory and Practice*, **29**(1):65–77, 2021.

[4] Cynthia Rudin. **Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead**. *Nature machine intelligence*, **1**(5):206–215, 2019.

[5] Uday Kamath and John Liu. *Explainable artificial intelligence: An introduction to interpretable machine learning*. Springer, 2021.

[6] Finale Doshi-Velez and Been Kim. **Towards a rigorous science of interpretable machine learning**. *arXiv preprint arXiv:1702.08608*, 2017.

[7] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. **Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI**. *Information fusion*, **58**:82–115, 2020.

[8] Gabriel Marín Díaz, José Javier Galán, and Ramón Alberto Carrasco. **XAI for Churn Prediction in B2B Models: A Use Case in an Enterprise Software Company**. *Mathematics*, **10**(20):3896, 2022.

[9] Carson K Leung, Adam GM Pazdor, and Joglas Souza. **Explainable artificial intelligence for data science on customer churn**. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE, 2021.

[10] Kamil Matuszelański and Katarzyna Kopczewska. **Customer Churn in Retail E-Commerce Business: Spatial and Machine Learning Approach**. *Journal of Theoretical and Applied Electronic Commerce Research*, **17**(1):165–198, 2022.

[11] Jonathan G Richens, Ciarán M Lee, and Saurabh Johri. **Improving the accuracy of medical diagnosis with causal machine learning**. *Nature communications*, **11**(1):3923, 2020.

[12] Shichao Zhang, Chengqi Zhang, and Qiang Yang. **Data preparation for data mining**. *Applied artificial intelligence*, **17**(5-6):375–381, 2003.

[13] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. **Why do tree-based models still outperform deep learning on tabular data?** *arXiv preprint arXiv:2207.08815*, 2022.

[14] Patricio Cerda and Gaël Varoquaux. **Encoding high-cardinality string categorical variables**. *IEEE Transactions on Knowledge and Data Engineering*, **34**(3):1164–1176, 2020.

[15] Maytal Saar-Tsechansky and Foster Provost. **Handling missing values when applying classification models**. 2007.

[16] A Rogier T Donders, Geert JMG Van Der Heijden, Theo Stijnen, and Karel GM Moons. **A gentle introduction to imputation of missing values**. *Journal of clinical epidemiology*, **59**(10):1087–1091, 2006.

[17] Donald B Rubin. **Inference and missing data**. *Biometrika*, **63**(3):581–592, 1976.

[18] Anthony J Myles, Robert N Feudale, Yang Liu, Nathaniel A Woody, and Steven D Brown. **An introduction to decision tree modeling**. *Journal of Chemometrics: A Journal of the Chemometrics Society*, **18**(6):275–285, 2004.

[19] Md Manjurul Ahsan, MA Parvez Mahmud, Pritom Kumar Saha, Kishor Datta Gupta, and Zahed Siddique. **Effect of data scaling methods on machine learning algorithms and model performance**. *Technologies*, **9**(3):52, 2021.

[20] Isabelle Guyon and André Elisseeff. **An introduction to variable and feature selection**. *Journal of machine learning research*, **3**(Mar):1157–1182, 2003.

[21] Ron Kohavi and George H John. **Wrappers for feature subset selection**. *Artificial intelligence*, **97**(1-2):273–324, 1997.

[22] Wei-Yin Loh. **Classification and regression trees**. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, **1**(1):14–23, 2011.

[23] Mohammad M Ghiasi, Sohrab Zendehboudi, and Ali Asghar Mohsenipour. **Decision tree-based diagnosis of coronary artery disease: CART model**. *Computer methods and programs in biomedicine*, **192**:105400, 2020.

[24] Abraham Itzhak Weinberg and Mark Last. **Selecting a representative decision tree from an ensemble of decision-tree models for fast big data classification**. *Journal of Big Data*, **6**(1):1–17, 2019.

[25] Leo Breiman. **Random forests**. *Machine learning*, **45**:5–32, 2001.

[26] Gérard Biau and Erwan Scornet. **A random forest guided tour**. *Test*, **25**:197–227, 2016.

[27] Yoav Freund and Robert E Schapire. **A decision-theoretic generalization of on-line learning and an application to boosting**. *Journal of computer and system sciences*, **55**(1):119–139, 1997.

[28] Philip M Long and Rocco A Servedio. **Random classification noise defeats all convex potential boosters**. In *Proceedings of the 25th international conference on Machine learning*, pages 608–615, 2008.

[29] Robert E Schapire. **The boosting approach to machine learning: An overview**. *Nonlinear estimation and classification*, pages 149–171, 2003.

[30] Tianqi Chen and Carlos Guestrin. **Xgboost: A scalable tree boosting system**. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[31] Matthew J Vowels. **Trying to outrun causality with machine learning: Limitations of model explainability techniques for identifying predictive variables**. *arXiv preprint arXiv:2202.09875*, 2022.

[32] Scott M Lundberg and Su-In Lee. **A unified approach to interpreting model predictions**. *Advances in neural information processing systems*, **30**, 2017.

[33] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. **From local explanations to global understanding with explainable AI for trees**. *Nature machine intelligence*, **2**(1):56–67, 2020.