University of Warsaw
Faculty of Economic Sciences

Working Papers
No. 15/2020 (321)

COMPARISON OF TREE-BASED MODELS PERFORMANCE IN PREDICTION OF MARKETING CAMPAIGN RESULTS USING EXPLAINABLE ARTIFICIAL INTELLIGENCE TOOLS

Marcin Chlebus
Zuzanna Osika

Warsaw 2020

WORKING PAPERS 15/2020 (321)

Comparison of tree-based models performance in prediction of marketing campaign results using Explainable Artificial Intelligence tools

Marcin Chlebus*, Zuzanna Osika

Faculty of Economic Sciences, University of Warsaw
** Corresponding author: mchlebus@wne.uw.edu.pl*

Abstract: The research uses tree-based models to predict the success of telemarketing campaign of Portuguese bank. The Portuguese bank dataset was used in the past in different researches with different models to predict the success of campaign. We propose to use boosting algorithms, which have not been used before to predict the response for the campaign and to use Explainable AI (XAI) methods to evaluate model's performance. The paper tries to examine whether 1) complex boosting algorithms perform better and 2) XAI tools are better indicators of models' performance than commonly used discriminatory power's measures like AUC. Portuguese bank telemarketing dataset was used with five machine learning algorithms, namely Random Forest (RF), AdaBoost, GBM, XGBoost and CatBoost, which were then later compared based on their AUC and XAI tools analysis – Permutated Variable Importance and Partial Dependency Profile. Two best performing models based on their AUC were XGBoost and CatBoost, with XGBoost having slightly higher AUC. Then, these models were examined using PDP and VI, which resulted in discovery of XGBoost potenitial overfitting and choosing CatBoost over XGBoost. The results show that new boosting models perform better than older models and that XAI tools could be helpful with models' comparisons.

Keywords: direct marketing, telemarketing, relationship marketing, data mining, machine learning, random forest, adaboost, gbm, catboost, xgboost, bank marketing, XAI, variable importance, partial dependency profile

JEL codes: C25, C44, M31

1. Introduction

As information - thanks to the Internet - is becoming more accessible to everyone, customers are becoming more sophisticated, better informed and more demanding. It causes companies to put more effort into maintaining relationship with their customers in order to keep their business going. Direct marketing has become an important way of communication with customers as it involves direct contact with customer to ensure he or she is happy with offered services or goods and is open to new offers. It targets existing customers and it was improved by advancements in technologies like data mining, data warehousing and campaign management software (Rygielski, et al., 2002). With increasing amount of data every day, marketing campaigns can perform better, be more efficient and offer customers individual approach to their needs by extracting insights from available data. As the amount of data is too large to be explored by a human, data mining techniques are used in marketing departments and Customer Relationship Management (CRM) departments to identify customers, who should be targeted for a certain campaign and to predict customers' behaviour. Data mining can improve performance of direct marketing campaign, increase its success rate and optimise resources used to run a campaign. With data mining, organizations can identify specific patterns in customers behaviour and offer deals for customers, who are most likely to buy them. In data mining, there is an extensive use of statistical analysis, mathematical modelling, artificial intelligence and machine learning algorithms (Apampa, 2016). Machine learning uses past data to make accurate predictions of future events or samples (Bishop, 2007) and offers flexible models for datasets of different structure and information. With insights gathered from machine learning algorithms, direct marketing campaigns can be more successful and offer the best solutions to the customers.

This work uses telemarketing campaign data from a Portuguese bank from May 2008 to November 2010 (Moro, et al., 2014) to predict customer's response with five different tree-based models, namely Random Forest (RF), AdaBoost, GBM, XGBoost and CatBoost. Database used offers features describing customers' and campaigns' characteristics, as well as macroeconomic indicators. Telemarketing campaign was targeted at existing bank's customers to cross-sell long-term deposits. Machine learning models were used to predict outcome of the call – if the product was bought by the customer or not. Predicting whether the customer would buy a product, if contacted, can be beneficial for the business, especially at financial institutions, which have access to large amount of data about their customers and the market. In the analysis, the newest boosting algorithms were used alongside older tree-based and boosting models like

Random Forest and AdaBoost. In this paper, we try to prove that the newest boosting algorithms perform better and should be used for heterogenous data problems. Performance of the models was later compared to select the best one to predict customers' decisions. To ensure reliability of the model and its quality of prediction, tools of explainable AI (XAI) were used. Variable Importance (VI) and Partial Dependency Profile (PDP) analysis were applied to understand how *black-box* models made their predictions. Aim of the research is to choose the best model, not only basing on its performance (AUC measure), but also on its stability, interpretability and overfit. With XAI tools, models, which performed the best can be compared and analysed in terms of aforementioned properties. It could be the case that model, which had the best discriminatory power is too overfitted and should not be used in business environment. Therefore, the research examines the hypothesis that XAI tools may be additional crucial indicators, to assess which model is better.

This paper is organised as follows: the first section presents overview of marketing techniques used by companies and overview of literature, which used Portuguese bank's database to research marketing strategies; the second section describes models used in this work; the third describes data used for the analysis; the fourth and fifth shows results from the models and explanatory analysis. The work ends with a summary of the research.

2. Literature overview

2.1. Mass campaigns and directed campaigns

In order to sell, banks or insurance companies, must advertise and promote services and products. It can be done using two different perspectives: mass marketing and direct marketing (Ling & Chenghui, 1998). Nowadays marketing strategies change fast as companies seek ways to increase performance of campaigns using their limited resources efficiently. More and more companies try to use marketing to be closer to their customers, who are becoming more demanding. They tend to create a relationship with customers to win their loyalty rather than operating mass campaigns. This is the reason why direct marketing plays an important role these days. It can be done by combining traditional media (print, mail, telephone) and on-line services to sell products and services by offering individualized offers to existing and potential customers. This offers a measurable campaigns' results and a long-term relationship with the customer (Kotler, 2002, p. 6). In order to run a successful marketing campaign, a thorough marketing research must be done. It involves defining the problem and research objective,

developing research plan, collecting information, analysing the information and presenting the findings. Within developing a research plan, data must be gathered. The amount of data produced every day is growing constantly so it is crucial for the companies to be able to use it efficiently. Especially within direct marketing, marketers should use the power of the database marketing. With a customer database, company can achieve better precision in targeting campaign's efficiency. One of the most powerful ways to use data in direct marketing is data mining. Data mining is the analysis of large, observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner (Hand, et al., 2001). Pattern of customer's behaviour can be explored with integration of machine learning, statistics and visualisation to select target group of customers for a certain campaign. The most important part of data mining is machine learning, which studies automatic techniques for learning to make accurate predictions based on past observations (Koumetio, et al., 2019). The main reason of using data mining in marketing is to increase return on investment (ROI) or net profits. Advanced data analysis allows campaigns to be targeted at certain group of customers, who are most likely to buy the product. It significantly lowers cost of promotions, which is beneficial for ROI. However, results obtained with data mining still need to be verified by business. Data mining helps business analysts to generate hypotheses, but it does not validate the hypotheses (Rygielski, et al., 2002). What is more, machine learning models, which are the most popular data mining applications, are very complex and incomprehensible for human. For the machine learning models to be applied in business, additional tools must be used to enable interpretation of models' results. By using Explainable Artificial Intelligence (XAI) tools with machine learning models in organisations, findings from complex machine learning algorithms can be used in straightforward way in organisations.

In recent years, a new way of direct marketing became popular – relationship marketing. One of the main reasons behind its popularity is the cost of acquiring new customers - it is easier and more cost efficient to keep relationship with existing customers and offer them products, especially when data about their behaviour is gathered. Relationship marketing is a mean of direct marketing. Techniques of data mining used to maintain a better relationship with a customer are usually used within Customer Relationship Management (CRM) department in companies. They focus on gathering knowledge and understanding of company's customers and the market, on which it is operating. CRM's main objective is to identify customers, who are profitable – worth targeting – and those who are not. It is also about selecting which product

to sell to which customers and which channel to sell it through (Rygielski, et al., 2002). Customer Relationship Management links data mining with campaign management in order to achieve competitive advantage.

2.2. Data Mining on Marketing Campaigns

Over the past decade, publicly shared dataset from Portuguese bank (Moro, et al., 2011) has been widely used to research customers' behavioural patterns through different data mining technique. The dataset contains information about clients included in bank's telemarketing campaign and campaign data itself.

S. Moro et al. (2014) use Portuguese bank data to show data mining algorithms can help in making business decisions. The analysis was based on a telemarketing campaigns selling long-term deposit to the clients of a Portuguese bank and aimed to build a model to predict a probability of buying the advertised product. The dataset used consists of calls' and clients' characteristics as well as social and economic indicators for each observation. In order to choose relevant characteristics to be used in the models, feature engineering process was applied. Out of 150 features in the original database, 22 were chosen to be included in the models. With chosen attributes, 4 models were evaluated: logistic regression, decision tree, neural network and support vector machine. Neural network model has outperformed logistic regression, decision tree and SVM models with AUC of 0.8 in the rolling window evaluation phase. Using analysis of feature importance, research has shown that *Euribor 3-month rate* was the most relevant attribute (importance around 17%). Variables with high importance were also direction of the call and agent's experience. Client's attributes turned out to be less relevant. Prior the research from 2014, S. Moro et al. (2011) used the same Portuguese bank data from different time span and less attributes (database didn't include macroeconomic data) to showcase implementation of Cross Industry Standard Process for Data Mining (CRISP-DM) methodology. In the analysis, 3 models were evaluated - Naïve Bayes, Decision Tree and Support Vector Machine. The highest AUC and Lift values were obtained with SVM (AUC higher than 0.9 and Lift higher than 0.85). The most relevant features in SVM were called *duration* (importance around 20%) and *month of contact* (importance around 10%).

To show how data mining can help a bank to improve outcomes of direct marketing campaign, Prusty (2013) used Naïve Bayes and Decision Tree algorithms to classify customers (using Portuguese bank dataset with less attributes). He compared models with balanced and unbalanced data. Based on overall Accuracy rate and lift values for the models, Decision Tree

with balanced data performed the best. Using balanced dataset, *k means* clustering showed that the customers, who subscribed to the product were customers employed in management, who are married and obtained tertiary education.

Using the same data, Elsalamony H.A. (2014) evaluated and compared four machine learning algorithms, which predicted whether bank's customer would subscribe to a long-term deposit. Models, which were applied were: multilayer perception neural network (MPLNN), Naïve Bayes (TAN), logistic regression (LR), and C5.0 decision tree classification model. Models were evaluated based on accuracy, sensitivity and specificity measures. C5.0 model had the highest values of all three measures for training sample, and the highest specificity for testing sample. MPLNN and logistic regression performed slightly worse. Additionally, for all the models, importance analysis was applied. For C5.0, MPLNN and LR variable *Duration* (duration of the call) had the highest importance.

Reminiscing work has been done to evaluate performance of four different models, using telemarketing data (Jayabalan & Asare-Frempong, 2017). Machine learning algorithms applied consisted of: Multilayer Perceptron Neural Network (MLPNN), Decision Tree (C4.5), Logistic Regression and Random Forest. The models were evaluated based on accuracy and AUC values. Random Forest classifier had the best results, with AUC of 0.92 and accuracy of 0.86. It was followed by decision tree (AUC of 0.87 and accuracy of 0.84) and logistic regression (AUC of 0.90 and accuracy of 0.83) models. Based on cluster analysis of customers, characteristics of customers, who are most likely to subscribe to a product were identified. People on managerial positions, who are married are more likely to buy long-term deposit from a telemarketer.

O. Apampa (2016) applied Cross Industry Standard for Data Mining (CRISP-DM) methodology to the same dataset to study whether random forest algorithm improved performance of decision tree (CART) algorithm. Performance of logistic regression and Naïve Bayes models are also compared. Two experiments were conducted: with unbalanced and balanced data. For unbalanced data, random forest had lower AUC than decision tree (AUC for Random Forest was 0.57, whereas Decision Tree model obtained AUC of 0.678). Random Forest had lower AUC than logistic regression (AUC of 0.657) and Naïve Bayes (AUC of 0.627). For balanced data (4 763 instances of "yes" and "no" responds each), Random Forest algorithm improved its performance, compared with the results obtained with unbalanced data (AUC of 0.742). However, algorithm still performed worse than the Decision Tree (AUC 0.766), logistic regression (AUC 0.757) and Naïve Bayes (AUC 0.756) models. When

analysing contribution of data features, variable *duration* (of the call) turned out to be the most important contributor to success of the campaign. Other important features were *poutcome, month* and *contact*.

A. Nachev (2015) focused on three stages of CRISP-DM (data preparation, modelling and evaluation) to address gaps in previous studies (problems of under- and over-fitting, data saturation, variable selection). He used double testing procedure, which uses cross-validation and multiple runs over selection of hyperparameters and partitions. Testing at different levels of data saturation, four models were compared. Models used for comparative analysis include: Neural Network, Logistic Regression, Naïve Bayes and Linear and Quadratic Discriminant Analysis. Analysing different levels of saturation, Neural Network algorithm outperformed other models (performance measured in AUC value). Only for lowest levels of data saturation, on the level of 10% and 20%, Linear and Quadratic Discriminant Analysis showed better results. Additionally, Neural Network design's effect on the model performance was explored to find the optimal size of hidden layers.

Kim et al. (2015) have explored Portuguese bank's database to classify customers with deep learning algorithm – Deep Convolutional Neural Network (DCNN). The model is mainly used in image and audio recognition and usually outperforms other techniques. The dataset used consisted of 16 attributes on customer's characteristics and campaign's characteristics. Authors highlighted the fact, that correlation coefficients obtained from correlation analysis of the data are rather low (between -0.05 and 0.05). The correlation analysis indicated that the relationship between financial-related attributes is not enough to use for recommendations and that the relationships are local. Deep Convolutional Neural Networks exploits convolution to raise the features from near nodes. When compared with different classifiers (Decision Tree, K Nearest Neighbours, Naive Bayes, Logistic Regression, Multilayer Perceptron, Support Vector Machine), DCNN obtained highest accuracy (0.76).

Data Mining as a tool for direct marketing has its limitations. Problems like imbalanced class distribution of dependent variable, predictive accuracy not suitable for evaluating learning methods and too large number of examples (Ling & Chenghui, 1998) are common when applying data mining into direct marketing solutions. Ling X. and Li C. address these problems using data from Canadian bank on loan product promotion, life insurance company for Registered Retirement Savings Plan campaign and company, which deals with "bonus program". They presented ways to solve these problems, which include using learning algorithms, which classify with a confidence level, for example produce probability to rank the

testing examples; using the lift and ROC measures instead of the predictive accuracy as the evaluation criterion; reducing negative examples in training dataset by oversampling with replacement the positive examples or reducing training set. In order to reduce variation error of classifiers they also applied AdaBoost to build models (Naïve Bayes algorithm and C4.5 - decision tree learning algorithm).

Table 1. presents summary table of papers, which used Portuguese bank dataset, with information on machine learning models used and which algorithms were concluded by the authors to be the best. Since the last analysis conducted, new machine learning algorithms have been introduced. Recently, boosting algorithms have been used on large scale as they perform well with different data. Gradient Boosting Machine, XGBoost and CatBoost are the most popular boosting algorithms, which have been recognised for their performance. For this reason, it is worth to examine how these algorithms will perform with Portuguese bank data. Also, most papers used accuracy measure to compare models. It is not a good approach as accuracy is based on cut-off point, which determines success and is not optimised for discrimination quality in most of the algorithms. With this in mind, we propose to compare models using AUC value, permutation-based Variable Importance (VI) and Partial Dependency Profile (PDP) analysis. Permutation-based VI not only presents relative importance of variables, but also enables comparison of variable's importance between different models. When using VI with loss function *one_minus_auc*, it is possible to estimate importance against measure, which is being optimised, in comprehensible units. PDP enables better understanding of dependencies between features and class variable, which can verify correctness of the results and possibly identify overfitting, when profile is too rough. Additionally, most of the authors do not consider features' characteristics, which are used in the models. For the model to be used by the business, it needs to use data available before the call is made, so the campaign can be planned up front. Most of the analysis use *duration* of the call as an independent feature, which is not available *apriori* and cannot be used in prognostic model.

Table 1. Results from different papers obtained using Portuguese bank dataset

Authors	Models used	Performance measure	Best model
Moro, Cortez & Rita (2014)	Logistic Regression, Decision Tree, Neural Network, Support Vector Machine	AUC, ALIFT	Neural Network
Moro, Cortez & Laureano (2011)	Naïve Bayes, Decision Tree, Support Vector Machine	AUC, lift	Support Vector Machine
Prusty (2013)	Naïve Bayes, Decision Tree	accuracy, lift	Decision Tree
Elsamony (2014)	Multilayer Perception Neural Network, Naïve Bayes, Logistic Regression, Decision Tree (C5.0)	accuracy, sensitivity, specificity	C5.0 decision tree
Jayabalan & Asare-Frempong (2017)	Multilayer Perceptron Neural Network, Decision Tree (C4.5), Logistic Regression, Random Forest	accuracy, AUC	Random Forest
Apampa (2016)	Decision Tree, Random Forest, Logistic Regression, Naïve Bayes	AUC	Decision Tree
Nachev (2015)	Neural Network, Logistic Regression, Naïve Bayes, Linear and Quadratic Discriminant Analysis	AUC	Neural Network
Kim, Lee, Jo & Cho (2015)	Decision Tree, K Nearest Neighbours, Naive Bayes, Logistic Regression, Multilayer Perceptron, Support Vector Machine, Deep Convolutional Neural Network	accuracy	Deep Convolutional Neural Network

Source: own preparation.

3. Methods & Materials

3.1. Random Forest for classification

Random forest algorithm is one of the most powerful and popular supervised machine learning algorithms. It can be used both for classification and regression, but this article focuses on Random Forest for classification. Within the model, number of decision trees are built on bootstrapped training sample. The more trees are built in the forest, the more robust the prediction is and the higher the accuracy. Each decision tree in the forest classifies new object based on the provided attributes. As a result, every tree in the forest predicts a certain class for the object. Random forest algorithm predicts the class, which has the most "votes" from the trees.

The biggest advantage of random forest algorithm is that it deals well with missing values in data and maintains accuracy in such case. The model also does not overfit the data. It can handle large amounts of data and is less sensitive to outliers (Kuhn & Johnson, 2013).

Although the model consists of decision trees, which are very transparent, RF is a *black-box* model. It does not provide straightforward explanation of how the prediction is made.

3.2. AdaBoost

AdaBoost is the first practical boosting algorithm presented in machine learning. Boosting is an approach to machine learning based on the idea of creating a highly accurate prediction rule by combining many relatively weak and inaccurate rules (Schapire, 2013). AdaBoost was introduced by Freund And Schapirev (1997) as an algorithm, which adaptively adjusts to the errors of the weak hypothesis returned by learning algorithm (*WeakLearn*). The accuracy of the final hypothesis depends on accuracy of all hypothesis returned by *WeakLearn*. The algorithm gives a clear method for handling real-valued hypothesis (Freund & Schapire, 1997). Weak hypothesises are combined by summing their probabilistic predictions to calculate accuracy of final hypothesis. The most common way to use AdaBoost is with decision trees as with low number of nodes as *WeakLearn*s.

The biggest advantage of AdaBoost is that it uses weak classifiers to predict data with high precision. Additionally, it considers weight of each classifier.

However, it does not work well with unbalanced data and outliers as it mainly focuses on correcting errors.

3.3. Gradient Boosting Machine (GBM)

Gradient Boosting Machine (GBM) was created using the idea of AdaBoost algorithm, which was connected to statistical concept of loss function, additive modelling and logistic regression (Kuhn & Johnson, 2013). Gradient boosting is based on sequential ensemble formation. GBM involves building successive models (base-learner models) in multiple iterations, where each model learns and improves regarding the error of the ensemble. Although it is possible to use a few different models in GBM, it is most common to use decision trees as base-learners. As base learners must be weak, trees used in GBM are not deep. GBM continuously improves models to minimize the loss function. Friedman (2002) proposed improvement in accuracy and execution speed to gradient boosting by including randomization in the process (stochastic gradient boosting). He suggested, that for each iteration, a subsample should be drawn at random without replacement from the training set. The subsample is used, instead of training sample, to fit the base learner and update loss function in the specific iteration. With such approach, robustness against overcapacity of the base learner is increased (Friedman, 2002).

Gradient boosting does not require pre-processing of the data and performs well with categorical as well as numerical values. Additionally, it handles missing data well.

Improving model in each iteration can lead to overfitting. What is more, it is computationally expensive as it requires high number of trees, which can be memory consuming and extend computational time.

3.4. XGBoost

XGBoost (eXtreme Gradient Boosting) is a scalable and effective tree bosting system. It is an implementation of gradient boosting. In terms of computational time and performance, it is better than GBM (Chen & He, 2020). It is due to the fact, that XGBoost algorithm is optimized with innovations. XGBoost is based on GBM, but has two main additions, which improve GBM. These additions are weighted quantile sketch for efficient proposal calculations and novel tree learning algorithm to deal with sparse data (Chen & Guestrin, 2016). The algorithm is widely used by data scientists and provides state-of the-art results for many problems (Chen & Guestrin, 2016). Authors of XGBoost proposed a regularization parameter in the loss function to smooth the final learnt weights to avoid overfitting.

The biggest advantage of XGBoost is its scalability. It runs 10 times faster than previous boosting solutions thanks to parallel and distributed computing.

3.5. CatBoost

Catboost, short for Categorical Boosting, is an implementation of gradient boosting created by Yandex researchers. It was introduced as a solution to prediction shift, caused by a special kind *of target leakage*, which appeared in prior boosting algorithms (Prokhorenkova, et al., 2018). Catboost uses boosting methodology but implements advances in algorithm such as permutation-driven ordered boosting and categorical feature support. It uses full binary trees, which are always symmetrical as base predictors, in contrast to XGBoost, which builds trees layer by layer and then prunes them. Within the trees, algorithm introduces innovative way to process categorical features to best split the data. It calculates statistics based on category and category plus label value. It also allows feature combinations to split the node.

The algorithm works the best with heterogeneous data and is easy to use. It is stable to parameter changes and does not need advanced parameter tuning to improve its performance. It works well with small and noisy data with complex dependencies. For larger datasets, it is four times faster than XGBoost, for smaller dataset it takes the same amount of time.

3.6. Performance assessment

3.6.1. AUC

To evaluate performance of chosen tree-based models, AUC measure was used. It was shown in empirical research (Huang & Ling, 2005) that AUC is a better measure of performance than accuracy as it is statistically consistent and more discriminating.

AUC is the area under the Receiver Operating Characteristics (ROC) curve. ROC curve plots true positive rate (number of positives correctly classified divided by total number of positives) on the Y axis and false positive rate (number of negatives incorrectly classified divided by the number of total negatives) on the X axis. Area under the ROC curve measures degree of a discrimination between successes and failures.

AUC is an equivalent to probability that a randomly chosen negative example will have a smaller estimated probability of belonging to the positive class than a randomly chosen positive example (Huang & Ling, 2005). For binary classification it can be calculated as (Hand, et al., 2001):

$$\hat{A} = \frac{S_o - \frac{n_o(n_o+1)}{2}}{n_o n_1}, \tag{1}$$

where n_o is the number of negative examples, n_1 is the number of positive examples and S_o is the sum of ranks of positive examples in the ranked list. AUC of 1 denotes perfect performance of the model while AUC of 0.5 denotes random classification of the model.

3.6.2. Explainable Artificial Intelligence

Predictive models like neural networks or ensembles based on decision trees are *black-box* models. They predict certain value or a class, but provide very complex logic behind their decisions. Usually, we only know what the input values for the models are and what are the outputs produced by the models, but model's results do not explain interference process. Unlike econometric models (linear or logistic regressions), which show explicitly how each feature impacted the final predicted value, machine learning algorithms are complex and hard to understand. Complexity of these models, caused by number of parameters and hyper parameters used, makes it incomprehensible to most of human beings. On the one hand, large number of parameters makes predictive models more elastic, but on the other it makes them hard to interpret. Increasing amounts of data and growing number of areas of human activity, in which predictive modelling is used, contribute to the fact that predictive models are becoming more sophisticated and their predictions more accurate. There is a trade-off between complexity of the models and its predictive accuracy. The more complex and the less interpretable the model is, the better accurate its predictions may be (Biecek, 2018).

Even though accurate predictions are highly desired from business, the need of explainability is also crucial Every business decision made entails financial cost for the company, so it is important for stakeholders to have all the information before deciding to spend money on certain action. Interpretability cannot be sacrificed also because of legal requirements or possibility of unfair decisions. It also brings light on model's predictions – it can provide information on the reason behind poor predictions. When results are easy to explain, it increases trust in model's predictions not only within technical people, but also business partners. From statistical point of view, interpretability can help to detect bias, protect from overfitting and reduce hidden debt in machine learning models (Biecek, 2018).

Latest researches provide solutions to interpretability problem for machine learning algorithms. With solutions like Partial Dependence Plots, Accumulated Local Effects Plots, Merging Path Plots, Break Down Plots, Permutational Variable Importance Plots or Ceteris Paribus Plots, *black-box* models can be explored.

To understand how black-box models made their predictions, XAI tools were used to compare models – Variable Importance and Partial Dependency Profiles were built.

3.6.3. Variable Importance

Variable Importance (VI) methodology was first introduced for Random Forest by Breiman (2001) and was later extended to other models (Fisher, et al., 2019). VI tools describe how much a prediction model's accuracy depends on the information in each covariate (Fisher, et al., 2019).

Fisher et al. (2019) proposed improvement to previously introduces VI tools. They highlighted the fact that when different models with equally good prediction accuracy are being analysed for variable importance, different variables can be important for different models, which could lead to discrepancies in results between models. Thus, they introduced model class reliance (MCR) for variable importance tools across models within specified class. MCR captures the range of explanations, or mechanisms, associated with well-performing models (Fisher, et al., 2019). Variable Importance measures how much model's fit changes when given variable was to be removed from the model and it is permutation-based. After permutation of each variable, model's performance is computed. For important variable, model's performance decreases. Variable's importance level is measured with size of the decline in performance.

The algorithm to measure importance for X^j starts with computing loss function for the original dataset: $L = \mathcal{L}(\tilde{y}, y)$, Loss function calculates goodness of fit of the model $f()$ based on predictions for the modified data $\tilde{y}$ and observed values y. Variable X^j is permutated, such that vector of observed values x^j is replaced with vector of permutated values x^{*j} . Then, loss function for modified data is computed: $L^{*j} = \mathcal{L}(\tilde{y}^{*j}, y)$. It calculates goodness of fit of the model $f()$ based on predictions for the modified data $\tilde{y}^{*j}$ and observed values y. Variable importance is quantified by computing difference or ratio between loss functions: $VI_{Diff}(x^{*j}) = L^{*j} - L$ or $VI_{Ratio}(x^{*j}) = L^{*j}/L$ (Biecek & Burzykowski, 2020).

3.6.4. Partial Dependency Profile

Partial Dependency Profile (PDP) shows relationship between expected value of predicted variable and a selected independent variable. It is created by aggregating Ceteris Paribus Profiles (CPP) by averaging a set of individual CPPs. PDP can be constructed for all observations from the dataset or for groups of instances. A single Ceteris Paribus Profile shows

how the expected value of the prediction changes as a function of a selected variable on instance level. Partial Dependency Profile uses CP profiles and shows the relationship for a set of observations.

A Ceteris Paribus profile $h()$ for model $f()$ of j-th independent variable and x_* observation vector can be defined as:

$$h_{x_*}^{f,j}(z) \coloneqq f\left(x_*^{j|=z}\right), \tag{2}$$

where $x_*^{j|=z}$ is vector, where value of j-th element of x_* was changed to a scalar z.

Partial Dependency Profile is an expected value of the Ceteris Paribus Profile for explanatory j-th variable, X^j, over joint distribution of independent variables other than j-th variable, X^{-j}:

$$\hat{g}_{PD}^{f,j}(z) = E_{X^{-j}}\left[f\left(X^{j|=z}\right)\right] \tag{3}$$

Unknown distribution of X^{-j} can be estimated by empirical distribution of N (number of observations in dataset) resulting in Partial Dependency Profile as a function:

$$\hat{g}_{PD}^{f,j}(z) = \frac{1}{N} * \sum_{i=1}^{N} f(x_i^{j|=z}). \tag{4}$$

PDP can be useful for comparing different models. Discrepancies between variables' profiles across models can be a sign of over-fitting. PDPs are easy to understand and to explain (Biecek & Burzykowski, 2020).

3.7. Data

Data used for the research was provided by Moro et al. (2014) and is available for public use. Database was downloaded from UCI Machine Learning Repository (http://archive.ics.uci.edu/).

The dataset has information on direct marketing campaigns conducted by Portuguese banking institution from May 2008 to November 2010 among bank's clients. The marketing campaigns were based on phone calls to sell bank long-term deposit.

Database was first published by Moro et al. in 2011 with 16 input attributes describing clients' characteristics and current and previous campaigns. In 2014 database was updated with additional social and economic features (national wide indicators). The available dataset has 41 188 instances and 20 attributes (clients' characteristics, campaign's attributes and economic

features). For the case of this research, 19 attributes were used. Variable *duration* was not used in the analysis as it passes information on the duration of the phone call – information which is known after the contact is made. Aim of this study is to create a solution for marketing teams to know what attributes increase probability that a client will purchase a product prior contacting the client. Additionally, duration of the call can be affected by external factors, such as telemarketer experience.

Finally, the data used in the research is of 41 188 instances and 20 variables. Description of 19 input variables and 1 class label, *y* can be found in Table 1. Dependant variable, *y* is binary and denotes, whether customer has subscribed to the advertised product or hasn't. Data is unbalanced, only 4 640 (11.26%) of the phone calls ended with success (Figure 1.).

For modelling and evaluation purpose, data was split into training and testing set (70/30 proportion). Models were built using training dataset (28 832 instances) and their performance was evaluated on testing dataset (12 356 instances). Testing data contains of 1 352 successful phone calls (10.94%) and training data of 3 288 (11.40%).

All models were built in R using caret package (Kuhn, 2019), which enables to streamline the model training process and compare the models. Every model described in the paper was trained with caret and specific package for the model. Table 3. presents models trained with applied packages to caret. Comparison between models was possible as the same configuration for parameter tuning algorithm was applied in *caret* pipeline. Models were trained with adaptive cross-validation with 10 folds and three repeats. Adaptive resampling (Kuhn, 2014) does not run full set of resamples for each model. It uses futility analysis to asses, which models have low probability of being optimal ones and reduces runtime. For adaptive resampling, additional parameters must be set in *trainControl* pipeline. These are: minimum number of resamples used before models are removed, confidence level of the one-sided intervals used to measure futility, method to measure futility (either generalized least squares or Bradley-Terry model) and whether when a single parameter value is found before the end of resampling should the full set of resamples be computed for that parameter (TRUE for case of this research). For this research, minimum number of resamples was set to two, with confidence level of 0.05 and method *gls*.

To search tuning parameters, random search procedure was applied. *TuneLength* parameter, which denotes the amount of granularity in the tuning parameter grid was set to default as large values of this parameter did not improve performance of the models. Default

value of this parameter is the maximum number of tuning parameter combinations that will be generated by the random search (Kuhn, 2019).

Table 2. Description of variables used

Name	Type	Description	Values
age	numerical	Customer's age	[17, 98]
job	categorical	Type of job	'admin.','blue-collar', 'entrepreneur', 'housemaid', 'management','retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown'
marital	categorical	Marital status	'divorced', 'married', 'single', 'unknown'
education	categorical	Education level	'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown'
default	categorical	Has credit in default?	'no', 'yes', 'unknown'
housing	categorical	Has housing loan?	'no', 'yes', 'unknown'
loan	categorical	Has personal loan?	'no', 'yes', 'unknown'
contact	categorical	Communication type of contact	'cellular', 'telephone'
month	categorical	Month of contact	'jan', 'feb', 'mar', ..., 'nov', 'dec'
day_of_week	categorical	Day of the week of contact	'mon', 'tue', 'wed', 'thu', 'fri'
pdays	numerical	Number of days since last contact with the client	999 if client wasn't previously contacted
Campaign	numerical	Number of contacts with the client performed for this campaign	[1, 56]
previous	numerical	Number of contacts with the client before	[0, 7]

		this campaign	
poutcome	categorical	Outcome from the previous campaign	'failure','nonexistent','success'
emp.var.rate	numerical	Employment variation rate *quarterly indicator*	[-3.4, 1.4]
cons.price.idx	numerical	Consumer price index *monthly indicator*	[92.2, 94.77]
cons.conf.idx	numerical	Consumer confidence index *monthly indicator*	[-50.8, -26.9]
euribor3m	numerical	Euribor 3-month rate *daily indicator*	[0.634, 5.045]
nr.employed	numerical	Number of employees *quarterly indicator, in thousands*	[4964, 5228]
y *dependant*	**binary**	**Has the client subscribed a term deposit?**	**'yes', 'no'**

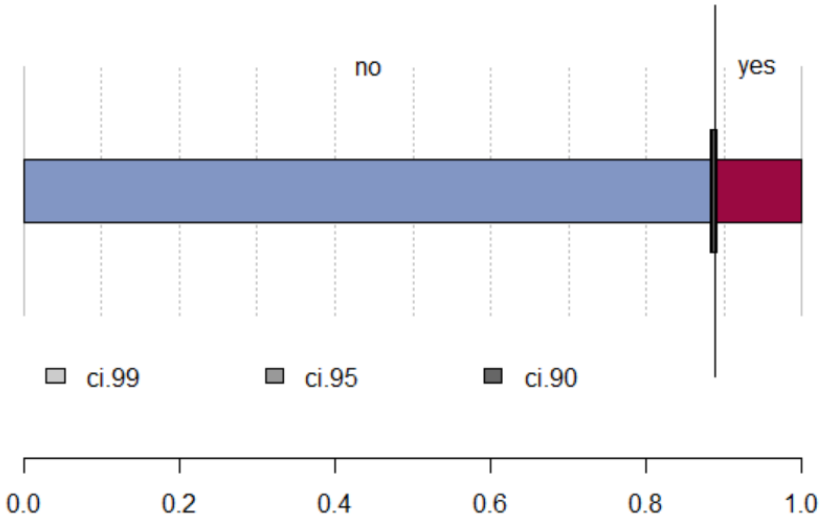
Source: own preparation.

Table 3. Packages used with caret for different models

Models	**R Package used within caret**
Random Forest	ranger
AdaBoost	adaboost
GBM	gbm
XGBoost	xgbTree
CatBoost	catboost

Source: own preparation.

Figure 1. Distribution of campaign's success (class variable y)



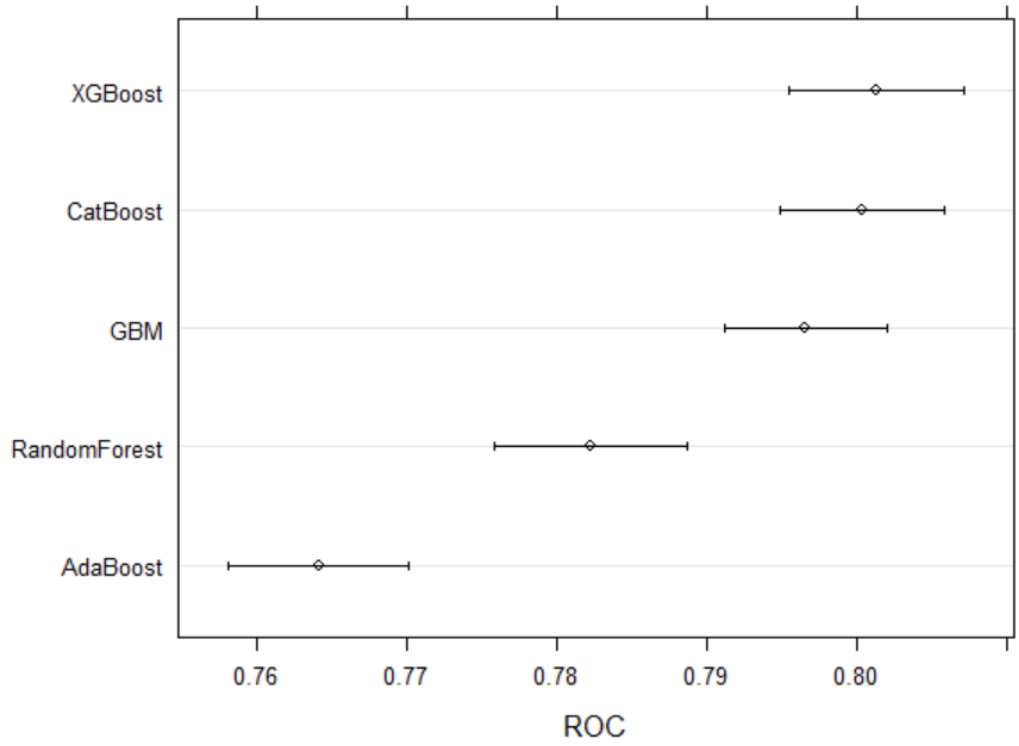
Source: own preparation using R.

4. Results

4.1. Performance results for all models

Caret package provides comparison of the models, which were built with the same tuning parameter search configuration. All models share common set of resampled data set as the same resampling specifications were used for models and random seed was set before modelling phase. Caret's *resamples* function enables comparison between models based on their cross-validation statistics. Figure 2. Shows that XGBoost and Catboost perform the best out of five models built in this research. These models have highest AUC values, which were calculated with resampling data set from training data. Mean AUC value for XGBoost is 0.8012, while mean AUC for CatBoost is 0.8003. GBM has slightly worse mean AUC, of 0.7965. Confidence levels for GBM, XGboost and CatBoost are similar, with XGBoost's and CatBoost's confidence levels almost identical. Random Forest and AdaBoost models performed the worst as mean AUC for Random Forest is 0.7822 and for AdaBoost 0.7648.

Figure 2. Models' performance comparison with caret on training data. AUC was used as a measure of performance.



Source: own preparation using R.

In order to evaluate possible performance differences between the algorithms, pair-wise inferences have been tested (Hothorn, et al., 2005). Statistical test has been conducted to assess, whether differences between AUC values for each two models are equal to zero (null hypothesis). Based on the results (table 4.), models which do not show difference in performance are XGBoost, GBM and CatBoost (*p-values* equal to 1). Null hypothesis is rejected for Random Forest and AdaBoost. The results show that advanced boosting algorithms perform better.

Additionally, to compare performance of the models, testing dataset was used to compute AUC measures for each model (table 5.). XGBoost algorithm was the most successful in discrimination between classes for testing set, with AUC of 0.8025. GBM and CatBoost performed slightly worse, AUC for GBM was 0.7955 and for CatBoost 0.7952. As previously, Random Forest and AdaBoost performed the worst. However, Random Forest's performance is far better than AdaBoost performance.

Table 4. Statistical tests' results (p-values), where H_0: Difference between AUC values for each two models equals 0

	Random Forest	GBM	XGBoost	AdaBoost	CatBoost
Random Forest					
GBM	0.0009				
XGBoost	0.0001	1.0000			
AdaBoost	0.0019	<0.0001	<0.0001		
CatBoost	0.0008	1.0000	1.0000	1.055e-08	

Source: own preparation.

Table 5. Comparison of performance of the models based on AUC for the test sample.

Model	AUC
XGBoost	0.8025
GBM	0.7955
CatBoost	0.7952
Random Forest	0.7907
AdaBoost	0.7667

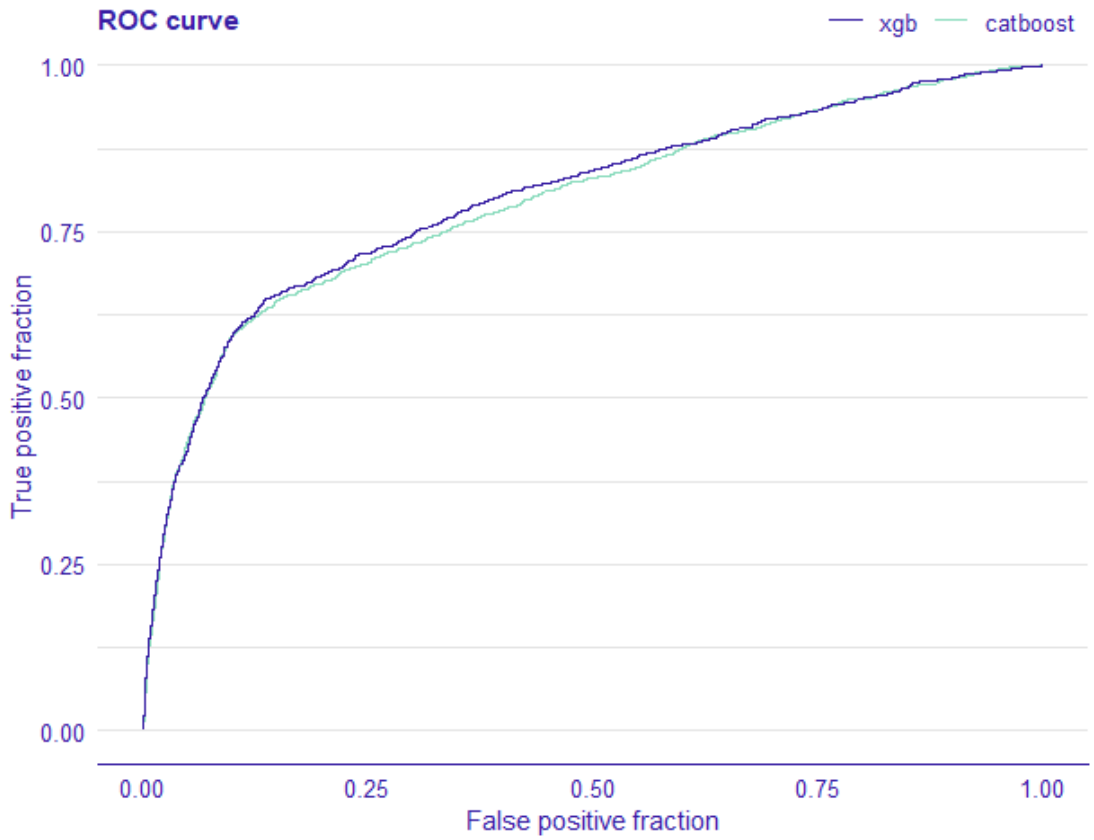
Source: own preparation.

Based on the performance results, two models, which performed the best are the most advanced and complex boosting algorithms with decision trees as weak learners: XGBoost and CatBoost. When performance of these two final models is compared on ROC curve (figure 3.), XGBoost ROC curve is slightly closer to (0,1) point of the graph than CatBoost as XGBoost has better AUC.

Results above show that two best models, XGBoost and CatBoost, have almost identical discriminatory power, although XGBoost has slightly better average results on cross-validation and test samples. Important aspect to consider is whether better performance of XGBoost may be connected with overfitting or simply by better way of describing non-linearities, this can be analysed with XAI tools.

For further analysis, two best performing models (XGBoost and CatBoost) will be used to explain how the decisions were made by them.

Figure 3. ROC curves for XGBoost and CatBoost



Source: own preparation in R.

4.2. Models explanation

To learn which features influence decision made by Portuguese bank customers and how CatBoost and XGBoost made the predictions, Variable Importance and Partail Dependency Profile analysis was conducted. Models' explainers have been computed with *Dalex* (Biecek, et al., 2020) package and PDP and feature importance plots were built with *Ingredients* (Biecek, et al., 2020) package in R.

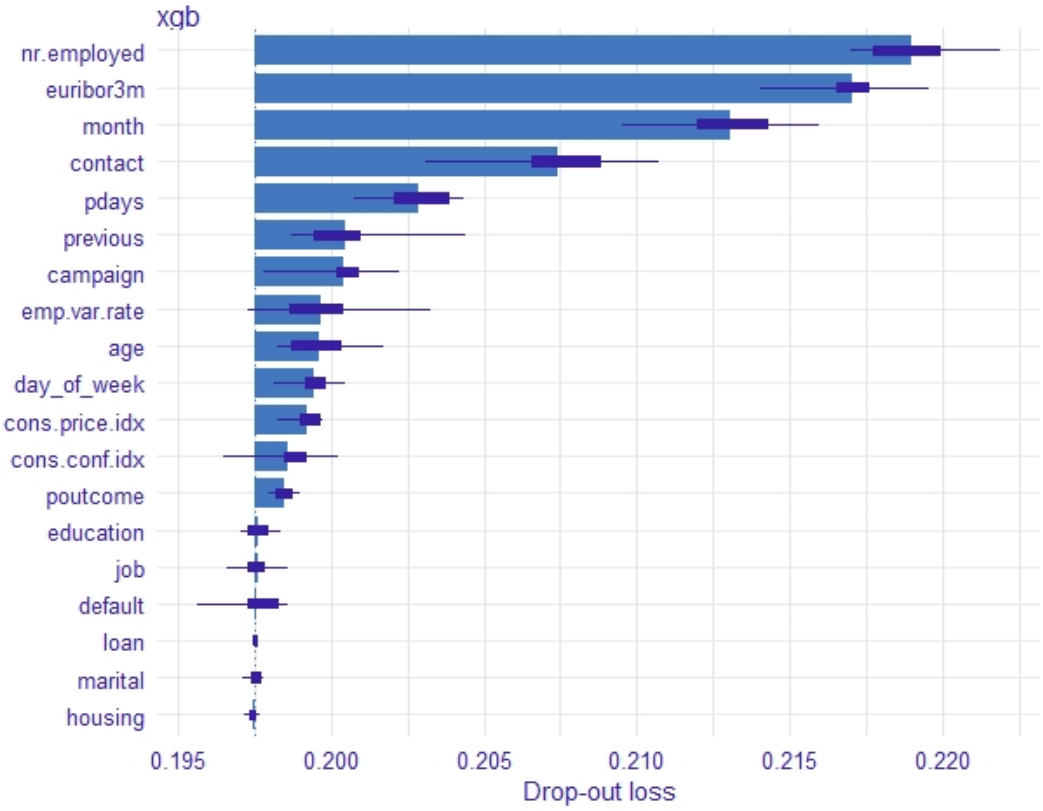
4.2.1. Variable Importance

Permutation-based variable importance values can be compared between different model structures. Figures 4. and 5. present VI plots for XGBoost and CatBoost models. In terms of loss function value, *L,* better results were obtained with XGBoost model. For XGBoost, the most important feature was number of employed people indicator (for CatBoost it was the fourth most important variable) and for CatBoost it was Euribor 3-month rate (for XGBoost it was the second most important variable). The second most important variable for CatBoost, consumer price index, was not the case for XGBoost, as this feature did not affect model's

performance much. For both models, month of the call was important aspect to influence the decision. Type of contact, number of days since the last contact with the client and number of contacts with the client performed for this campaign were equally important for both models (medium important). For XGBoost the least important features turned out to be client's credit situation (whether the client has loan or mortgage), and client's characteristics like job, education, marital status. For CatBoost it was similar case, although it also indicated day of the week of the call to not be important, while for XGBoost it was of medium importance.
Generally, variable importance was similar for both models. Only some macroeconomic indicators (consumer price index and number of employed people indicator were differently important between models).

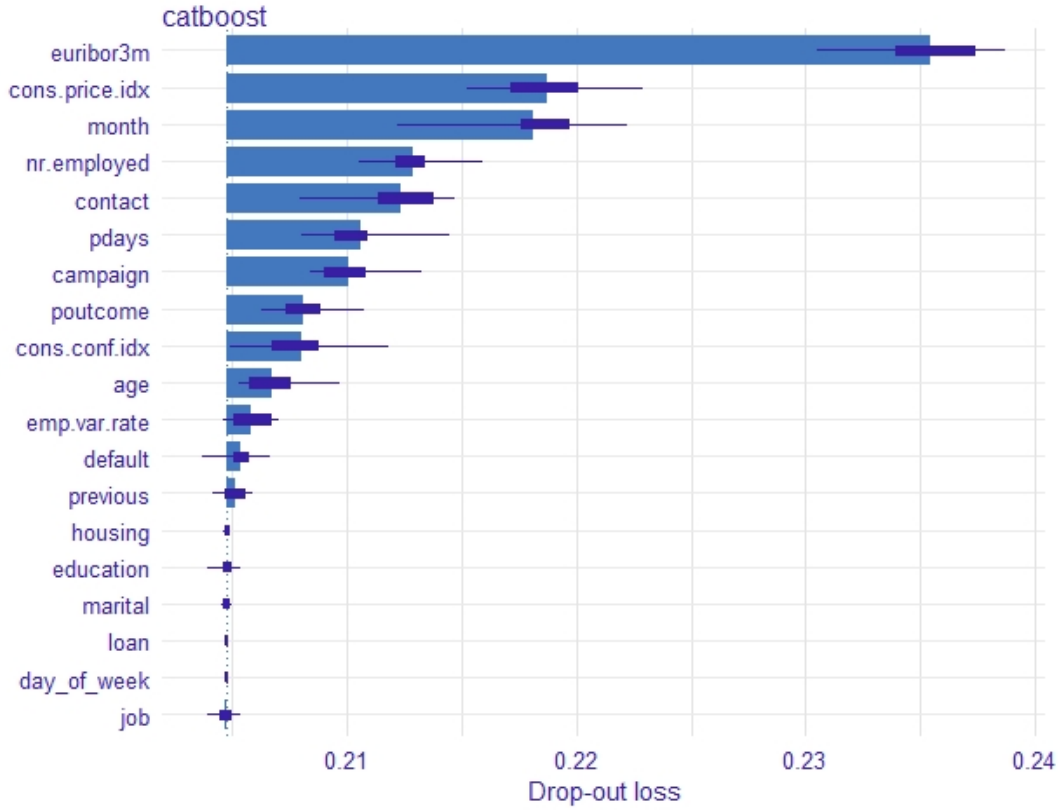
Considering results from both models, it seems that macroeconomic variables influence the decision whether to buy the deposit advertised or not the most. Surprisingly, client's characteristics do not affect the decision to a large extent. This result was also obtained by Moro et al. (2014). This could lead to a conclusion that external effects and not client's situation should be taken into consideration when planning a marketing campaign.

Figure 4. Variable Importance Plot for XGBoost.
Note: Bars indicated average loss in model's performance and the navy-blue boxes are confidence levels.



Source: own preparation in R.

Figure 5. Variable Importance Plot for CatBoost.
Note: Bars indicated average loss in model's performance and the navy-blue boxes are confidence levels.



Source: own preparation in R.

4.2.2. 8 Partial Dependency Profile

Partial Dependency Profiles were calculated for CatBoost and XGBoost using DALEX and ingredients packages in R. Plots aggregated Ceteris Paribus Profiles for all observations from testing set. For each model, PDPs were constructed for continuous variables (figure 6.) and categorical variables (figure 7.).

Overall, it seems that for most of the continuous variables, profiles for both models have the same general direction of relation between predicted probability and independent variables' values. However, XGBoost provides flatter profiles for almost all continuous variables except of number of employed people and Euribor 3-month rate. For most of the variables, the biggest difference between models' profiles are at the edges of independent variables' scales. It seems that XGBoost "shrinks" predicted probability for extreme values of independent variables towards the average prediction. This is the case for age of the customer contacted; number of contacts with the customer for the campaign, number of days since last contact, number of

contacts with the customer for previous campaign, consumer price index and Euribor 3-month rate.

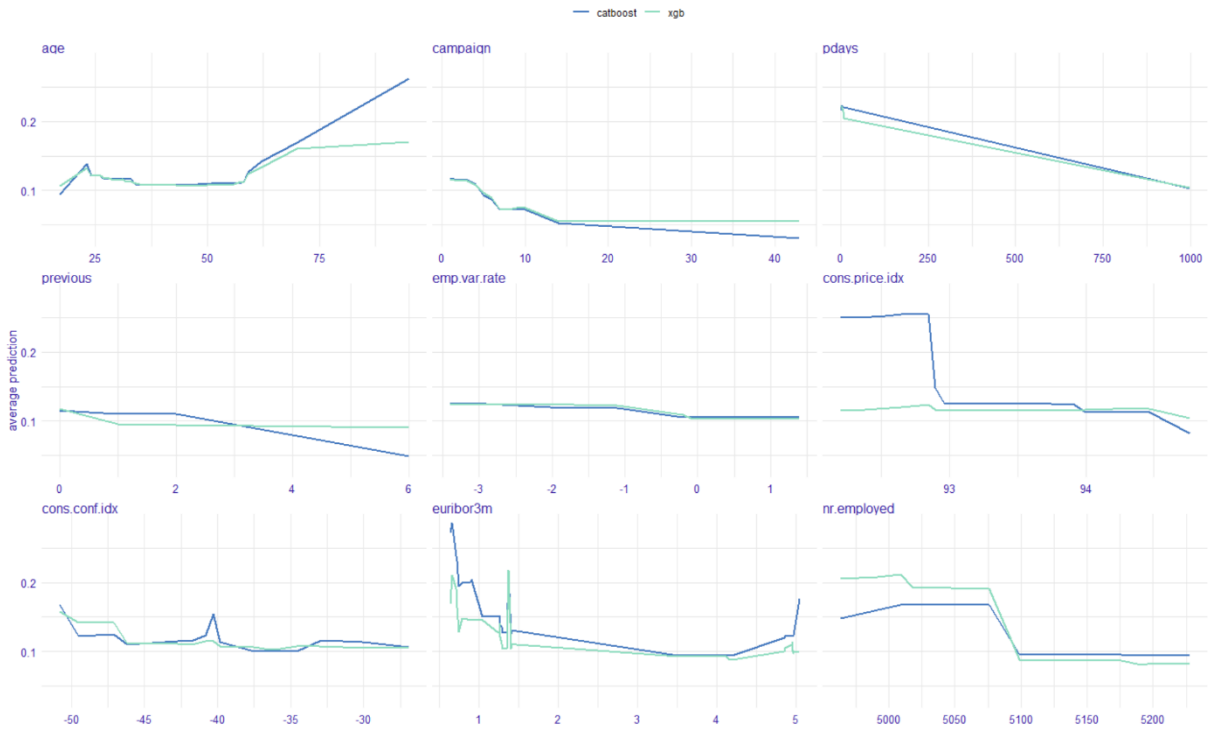
Variables, which turned out to be the most important during VI analysis for XGBoost were number of employed people indicator and for CatBoost it was Euribor 3-month rate. It is interesting as for only these variable PDPs are not smoother for XGBoost. Especially for number of people employed, CatBoost's profile is much flatter. It is worth noticing that profile for Euribor 3-month rate, which was the second most important variable for Xgboost is anomalous. The fact that XGBoost tends to overfit for the most important variables may be the reason why its performance is slightly better than CatBoost. This finding is crucial as it can be the reason for choosing CatBoost over XGBoost, even though CatBoost was second best model. It is also worth noting that profile for the second most important variable for CatBoost, consumer price index, is much rougher than XGBoost's profile.

Relationship between the most important variables for both models and predicted probability of subscribing to the product, shows that when number of employed people in the country is greater than 510 000, predicted probability remains constant. There is also non-monotonic relationship between value of Euribor 3-month rate and predicted probability: for values of rate less than 1.5%, the probability of success lowers, the higher the Euribor is and for value of rate above 1.5%, probability seems to achieve constant level, only to rise when the rate is above 4%. For low values of rate (lower than 1.5%) it seems that relationship is very complex, as even though it has general declining trend, for certain values of rate, value of probability peaks. When exploring customer's characteristics, the most important variable was age of the customer. PDP shows nonlinear relationship between customer's age and predicted probability. For the youngest and the oldest customers of the bank, probability of subscribing the deposit increases with age, but for customers of age between 25 and 55 the probability is quite constant, about 10%.

For categorical variables, most of features have similar relationship with predicted probability for both models. It is different only for customers, who are illiterate and for December, March and September - months of the call. For CatBoost, for education variable, people who are illiterate had higher probability of subscribing to the product than for XGBoost model. However, this variable turned out not to be important for both models during VI analysis phase. On the other hand, month of the call was one of the most important variables for both models. For all three months, for which there where discrepancies between probabilities from the models, probabilities for CatBoost were decidedly higher than for XGBoost.

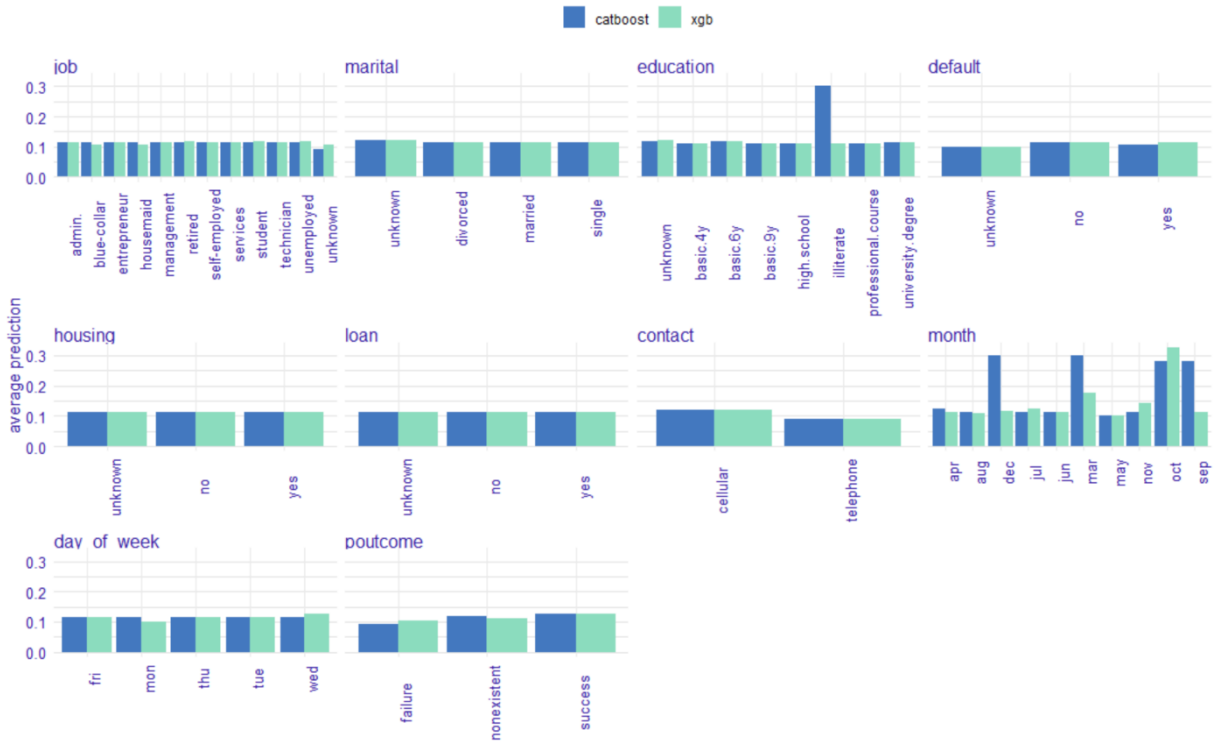
When it comes to categorical variables, the most important variables from previous section for both models were month of the call and type of the call. When taking into consideration results from both models, the highest probability that the customer will buy the advertised product is in March and October. For CatBoost, high probability is also obtained during the month of December and September. The lowest probability of subscribing the product for both models is in May. Also, contacts made via cellular phone are more likely to be successful. For other categorical values, there is no significant difference of probabilities for different classes of the variables.

Figure 6. Partial Dependency Plots for continuous variables for XGBoost and CatBoost models



Source: own preparation in R.

Figure 7. Partial Dependency Plots for categorical variables for XGBoost and CatBoost models



Source: own preparation in R.

Based on the results from XAI gathered in the analysis above it can be concluded that most of the features that impact outcome of the call cannot be controlled by the bank. It seems that macroeconomic situation mostly influences people's decisions. Euribor, a reference interest rate of deposits and credits on the European interbank market was proven to be one of the most important features for both models. It is a conclusion in line with previous research (Moro, et al., 2014). However, the relationship between Euribor and predicted value was found vague. Previously (Moro, et al., 2014) it was concluded that 2008 financial crisis changed how Euribor rate influenced people's decisions. Prior the financial crisis, lower value of Euribor rate would result in lower savings rate. Financial crisis changed people's approach to saving and reversed the relationship. For results presented in this paper this kind of relationship can be weakly observed. Number of employed people was also shown to be important factor to impact call's outcome (Moro, et al., 2014). This can lead to a conclusion that marketing and CRM managers should make their decisions and adjust their marketing strategies considering situation on the market and they should monitor the most important macroeconomic indicators.

However, some of the factors can be regulated by the bank. They include the month of the contact, type of telephone and number of days since previous contact. These aspects should be

considered by the decision makers when planning the campaign as they can be controlled internally.

Surprisingly, client's characteristics do not influence the decision to a large degree. Only age seems to be important factor for call's outcome and the older the customer, the higher the probability of subscribing to the product. Unfortunately, this result does not help stakeholders to target the customers based on their characteristics.

Most of the previous researches used Portuguese's bank dataset consisting of less (Kim, et al., 2015) features (without macroeconomic indicators), but with variable *duration* included in the models therefore it is hard to compare the results obtained with the results from previous papers. However, authors have chosen complex machine learning algorithms like Neural Network (Nachev, 2015; Moro, et al., 2014; Kim, et al., 2015) when used, which is compatible with our results as we also have chosen more complex algorithms as the best ones. What is more, most reasearches preferred tree based models like Decision Tree (Prusty, 2014; Elsamony, 2013; Apampa, 2016) and Random Forest (Jayabalan & Asare-Frempong, 2017) and CatBoost and XGBoost uses decision trees and base-learners.

5. Summary

With constant developments in technology and growing number of new, revolutionary products and services, customers are becoming more demanding. To attract a customer, company must use more sublime ways to communicate with clients. In today's world, company's marketing approach must focus on maintaining relationship with clients. For this reason, companies like financial institutions build special CRM departments to deal with relations with customers. To optimise plans regarding communication with the clients, advanced data mining techniques can be used. With growing amount of data gathered on almost every aspect of our lives, data can be useful and solid way to gather insights. Combining data with direct marketing techniques can maintain good relationship with customers and optimize campaign's resources. It is important to choose the best statistical technique, which could provide explanation on how the decisions are made by the customer.

In this study, direct marketing campaign data was explored with tree-based algorithms to select which customers are likely to buy advertised product and what are the factors affecting the decision. Analysis was conducted using Portuguese bank's telemarketing campaign data. The campaign sold long-term deposits to bank's customers from 2008 to 2013. Dataset consisted of 19 features describing client's characteristics, campaigns characteristic's and

macroeconomic situation and one class variable, which indicated if the product was purchased by customer. The aim of the research was to choose the best tree-based model, which predicted the outcome of the call, based not only on performance measure, but also by analysing overfitting aspects with XAI tools. Tree-based algorithms are good for heterogenous data, so they are good technique for CRM departments to use with their data. With the growing number of machine learning models, it is important to study their performance and prediction mechanism. All aspects must be examined by analysts, so that decision makers have data of the best quality. For these reasons, predictions were made using Random Forest and four tree-based boosting algorithmsAdaBoost, GBM, XGBoost and CatBoost. To measure their performance, AUC values was used. Comparisons were made using training and testing data. For both cases, XGBoost performed the best with average AUC of 0.8012 for cross-validation sets and AUC of 0.8025 for testing set. The second-best performing model was CatBoost with average AUC from cross-validation sets of 0.8001 and 0.7952 AUC for testing set. These results show that advanced boosting algorithms, thanks to gradual improvements in every iteration, modified loss function and new tree schema are better at discriminating than less complex models. With these models, predicted outcome of the campaign for each customer prior the contact is made can be used to optimise marketing campaigns.

For both boosting models, knowledge extraction techniques were used in order to better understand how models were impacted by specific features. As boosting algorithms are black-box models, which alone cannot be easily understood by a human being, Variable Importance (VI) and Partial Dependency Profile (PDP) were used to study which independent variables influenced each model the most and how expected probability prediction of class variable changes for different values of each independent variable. Variables, which were important differed between two models – the most important variable for CatBoost was Euribor 3-month rate, which was also the case in previous research (Moro, et al., 2014); while for XGBoost it was number of employed people indicator. PDPs for continuous variables were smoother for XGBoost, especially at extreme values for the variables, which could mean that XGBoost "shrinks" predicted values at the edges of variable's numerical scale to the average predictions. The most interesting PDP for both models was from 3-month Euribor rate (the most important variable for CatBoost and second most important variable for XGBoost). For low values of interest rate, predicted probability of subscribing to the product seems very unstable. PDP for XGBoost is much rougher than for CatBoost, which could lead to a conclusion that this is the reason of slightly better performance of XGBoost and may lead to a decision that Catboos

would be preffered by a business as profile for XGBoost is not very intuitive and the model itself may overfit the data. Thanks to the explanatory knowledge, models can be better understood by analysts and managers.

Future work on this matter should include analysis of data from banks operating in different countries around the globe to study whether the results would be robust. Also, datasets with more attributes should be used to study impact of other characteristics on the final decision of a customer. Using data mining techniques and constant research of customers' behaviour is important now more than ever. Next steps for research should also include quantifying how global pandemic changed customers' approach to products offered by banking institutions, especially in savings area.

References:

Ali, J., Khan, R., Ahmad, N. & Maqsood, I., 2012. Random Forests and Decision Trees. International Journal of Computer Science Issues(IJCSI)., 9(5), pp. 272-278.

Apampa, O., 2016. Evaluation of Classification and Ensemble Algorithms for Bank Customer Marketing Response Prediction. Journal of International Technology and Information Management, 25(4), pp. 85-100.

Barman, D., Shaw, K. K., Tudu, A. & Chowdhury, N., 2016. Classification of Bank Direct Marketing Data Using Subsets of Training Data. s.l., Springer India.

Biecek, P., 2018. DALEX: Explainers for Complex Predictive Models in R. Journal of Machine Learning Research, 19(1), pp. 1-5.

Biecek, P., Baniecki, H. & Izdebski, A., 2020. Package 'ingredients'. [Online] Available at: https://cran.r-project.org/web/packages/ingredients/ingredients.pdf [Accessed 01 05 2020].

Biecek, P. & Burzykowski, T., 2020. Explenatory Model Analysis. [Online] Available at: https://pbiecek.github.io/ema/preface.html [Accessed 23 4 2020].

Biecek, P., Maksymiuk, S. & Baniecki, H., 2020. Package 'Dalex". [Online] Available at: https://cran.r-project.org/web/packages/DALEX/DALEX.pdf[Accessed 01 05 2020].

Bishop, C. M., 2007. Pattern Recognition and Machine Learning (Information Science and Statistics). 1 ed. s.l.:Springer.

Breiman, L., 2001. Random Forest. Machine Learning, 45(1), pp. 5-32.

Chen, T. & Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. San Francisco, ACM.

Chen, T. & He, T., 2020. xgboost: eXtreme Gradient Boosting. [Online] Available at: http://cran.fhcrc.org/web/packages/xgboost/vignettes/xgboost.pdf [Accessed 22 04 2020].

Elsalamony, H. A., 2014. Bank Direct Marketing Analysis of Data Mining Techniques. International Journal of Computer Applications , 85(7), pp. 12-22.

Fisher, A., Rudin, C. & Dominici, F., 2019. All Models are Wrong, but Many are Useful: Learning a. Journal of Machine Learning Research, 20(1), pp. 1-81.

Freund, Y. & Schapire, R., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences , 55(1), pp. 119-139.

Friedman, J. H., 2002. Stochastic gradient boosting. Computational Statistics & Data Analysis, 38(1), p. 367–378.

Friedman, J., Hastie, T. & Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting. Annals of Statistics, 28(2), pp. 337-407.

Hand, D., Heikki, M. & Padhraic, S., 2001. Principles of Data Mining. Boston: The MIT Press.

Hand, D. H. & Till, R., 2001. A Simple Generalisation of the Area under the ROC Curve for Multiple Class Classification Problems. Machine Learning, 45(1), pp. 171-186.

Hothorn, T., Leisch, F., Zeileis, A. & Hornik, K., 2005. The Design and Analysis of Benchmark. Journal of Computational and Graphical Statistics , 14(3), pp. 675-699.

Huang, J. & Ling, C., 2005. Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on Knowledge and Data Engineering , 17(3), pp. 299-310.

James, G., Witten, D., Trevor, H. & Tibshirani, R., 2013. An Introduction to Statistical Learning with Applications in R. Nowy Jork: Springer Science+Business Media .

Jayabalan, M. & Asare-Frempong, H., 2017. Predicting Customer Response to Bank Direct Telemarketing Campaign. Kaula Lumpur, 2017 IEEE The International Conference on Engineering Technologies and Technopreneurship, pp. 330-341.

Kim, B., Lee, Y. & Choi, D.-H., 2009. Construction of the radial basis function based on a sequential sampling approach using cross-validation. Journal of Mechanical Science and Technology volume, 23(1), p. 3357–3365.

Kim, K.-H., Lee, C.-S., Jo, S.-M. & Cho, S.-B., 2015. Predicting the Success of Bank Telemarketing using Deep Convolutional Neural Network. Fukuoka, IEEE.

Kotler, P., 2002. Marketing Management, Millenium Edition. Custom Edition for University of Phoenix ed. Boston: Pearson Custom Publishing.

Koumetio, S. C., Cherif, W. & Silkan, H., 2019. A data modeling approach for classification problems: application to bank telemarketing prediction. Rabat, Association for Computing Machinery.

Kuhn, M., 2014. Futility Analysis in the Cross–Validation of Machine Learning Models. Groton, Pfizer Global R&D.

Kuhn, M., 2019. The caret Package. [Online] Available at: http://topepo.github.io/caret/index.html [Accessed 01 05 2020].

Kuhn, M. & Johnson, K., 2013. Applied Predictive Modeling. 5 ed. Ney York City: Springer.

Ling, C. X. & Chenghui, L., 1998. Data Mining for Direct Marketing: Problems and Solutions. KDD'98: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, pp. 73-79.

Martens, D. & Provost, F., 2011. Pseudo-Social Network Targeting from Consumer Transaction Data. NYU Working Paper No. CEDER-11-05.

Moro, S., Cortez, P. & Laureano, R., 2011. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. s.l., Proceedings of the European Simulation and Modelling Conference..

Moro, S., Cortez, P. & Rita, P., 2014. A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62(1), pp. 22-31.

Nachev, A., 2015. Application of Data Mining techniques for direct marketing. [Online] Available at https://pdfs.semanticscholar.org/df19/ecf7f2b27ac84f0f337d2ef2ca2a862a9cff.pdf [Accessed 22 04 2020].

Palaniappan, S., Mustapha, A., Foozy, C. F. M. & Atan, R., 2017. Customer Profiling using Classification Approach for Bank Telemarketing. International Journal of Infromatics Visualization, 1(4-2), pp. 214-217.

Probst, P., Wright, M. & Boulesteix, A.-L., 2019. Hyperparameters and Tuning Strategies for Random Forest. [Online] Available at: https://arxiv.org/pdf/1804.03515.pdf [Accessed 22 04 2020].

Prokhorenkova, L. et al., 2018. CatBoost: unbiased boosting with categorical features. Montreal, NeurIPS.

Prusty, S., 2013. Data mining applications to direct marketing: identifying hot prospects for banking product. s.l., Web Data Mining (ECT 58429) DePaul University,Chicago.

Rygielski, C., Wang, J.-C. & Yen, D. C., 2002. Data mining techniques for customer relationship management. Technology in Society, Issue 24, pp. 483-502.

Schapire, R. E., 2013. Explaining AdaBoost. [Online] Available at: http://rob.schapire.net/papers/explaining-adaboost.pdf [Accessed 20 04 2020].

Soleimani Neysiani, B., Soltani, N. & Ghazalash, S., 2015. A Framework for Improving Find Best Marketing Targets Using a Hybrid Genetic Algorithm and Neural Networks. 2nd International Conference on Knowledge-Based Engineering and Innovation , IEEE.

Vermont, J. et al., 1991. Strategies for graphical threshold determination. Computer Methods and Programs in Biomedicine, 35(2), pp. 141-150.

Zhao, Y. & Cen, Y., 2014. Data Mining Applications with R. Waltham: Elsevier Inc..

University of Warsaw
Faculty of Economic Sciences
44/50 Długa St.
00-241 Warsaw
www.wne.uw.edu.pl