Building a Cloud-Based Regression Model to Predict Click-Through Rate in Business Messaging Campaigns

Alexandros Deligiannis, Charalampos Argyriou, and Dimitrios Kourtesis

Abstract—The goal of the research presented here is to describe an innovative approach to predicting the impact of a business messaging campaign, by estimating the percentage of message recipients who will engage with a message. The motivation is to facilitate business marketers to address the problem of estimating the return on investment coming from a potential messaging campaign. The presented solution relies on the processing of large scale business data, taking into account state-of-the-art predictive algorithms, GDPR compliance requirements, and the challenge of increased data security and availability. In this paper we discuss the design of the core functional components of a system that could make this possible, which encompasses predictive analytics, data mining and machine learning technologies in a cloud computing environment.

Index Terms—Marketing automation, predictive analytics, cloud computing, business messaging, data privacy, GDPR, machine learning, XGBoost, regression, IBM watson studio, SPSS modeler flow.

I. Introduction

Global marketing trends suggest that businesses will increasingly invest in technology that enables them to optimize the impact of their marketing campaigns, and allows them to deliver personalized services, interactions, and content adapted to their customers' unique preferences. This kind of perspective implies a series of benefits for them, such as more valuable time invested from marketers to each campaign, more efficient resource allocation, increased campaign effectiveness and, eventually, enhanced customer experience [1].

The research presented in this paper is part of a wider research project focused on developing new services for personalization of business-to-consumer mobile communications utilizing data from past purchase transactions and past exchanges of messages between the business and the customer. In the scope of this paper we focus on the second category of the aforementioned types of data. We present the design of a predictive model that uses past message exchange data to estimate the Click-Through Rate (CTR) of a message. Collecting, consolidating, and processing big data of this form poses challenges and risks for both the consumer and the company. These can be summarized into three main problems: (1) data are difficult to assemble, consolidate and process because of their large scale, (2) moving data between several system boundaries poses significant security risks, and (3) complying with data protection legislation can be particularly complex. The European Union's new legal framework for data protection (GDPR) [2] sets high standards for companies that want to enable advanced data processing and customization of communications with their customers.

Apifon is a telecommunication services provider that has set itself the goal of developing an innovative software platform, called PRIME [3]. This platform consists of a suite of services, which rely on predictive analysis, data mining, and machine learning technologies to support the personalization challenges outlined above, and enables businesses to drastically improve their direct communications with customers over text messaging channels, such as SMS and its successor RCS (Rich Communication Services), or OTT (Over the Top) IP-based messaging apps [4]. The motivation of the platform is to make communication more (1) personalized and relevant to each customer's preferences; (2) direct, interactive and content-rich; and (3) safe for both sides, by protecting consumers' personal data and ensuring GDPR compliance for the company.

The current study focuses primarily on defining the algorithmic approach of the CTR estimation as a regression problem, building on the type of messaging exchange data that would be typically available to any provider of business messaging services. This paper also presents a solution architecture to make the resulting predictive model available, as a service, on top of a cloud computing environment with advanced data processing and data modeling capabilities [5].

II. Business Case

A. Marketing Perspective

One of the most important challenges in everyday marketing practice is the lack of foresight on the effectiveness of a messaging campaign during planning. Marketers often have no way of knowing how effective a campaign will be before executing it. The predictive service introduced in this paper aims to produce a reliable estimation of a messaging campaign's CTR before it is actually sent. This is tied to predicting how many of the recipients will successfully receive the message and will subsequently engage with the content in an interactive manner, like opening the link inside the message body.

B. Research Objective

The challenge described above can be viewed as a problem of predicting the percentage of campaign recipients who will

The authors are with the Research & Development Department, Apifon, Greece (e-mail: a.deligiannis@apifon.com, c.argyriou@apifon.com, dk@apifon.com).

read the message content delivered to their mobile phone and follow the suggested call to action. This is a challenge of predicting a continuous value. Therefore, the predictive model can be built on top of a multivariate regression algorithm, with the goal being to minimize the MAE (Mean Absolute Error) between the predicted and the actual campaign CTR.

C. Related Personalization Tools

As stated in the introduction, personalization is the key to improve customer experience. This is also the aim of the CTR estimation solution presented in this paper, which is part of a bundle of services included in the PRIME platform [3]. These services were co-developed with companies that offer consumer products and services, through a process of interviews and product opportunity assessment.

The capabilities offered by the PRIME platform alongside CTR prediction, are the following:

- **Date & Time Optimization:** The goal of any business that promotes its products through direct marketing campaigns is to find the best time to send out the campaign content, so as to achieve the highest possible conversion rate. A corresponding use case that the PRIME platform supports, is to automatically determine the right day and time to send out a specific direct marketing campaign message to each customer individually, based on his/her unique profile.
- **Segment Recommendation:** A great challenge faced by many marketing professionals in their daily work has to do with how to choose the recipient list for a particular campaign message. Marketers often lack the tools to be able to specify a highly relevant target audience for a campaign. Messages are often sent out to large lists of non-relevant customers, resulting in low engagement rates and poor customer experience. Through automated segmentation, the PRIME platform enables marketers to reach those subscribers who are most likely to find the content of a specific campaign relevant and compelling enough to engage.
- **Keyword Suggestion:** One of the most important factors that define a campaign's success is its text content. Personalization means using the most appropriate vocabulary, in terms of keyword efficiency, to speak to each individual customer. PRIME offers personalized content enrichment for the message of an upcoming campaign by automatically suggesting effective words to be added to the campaign text.
- **Risk Factor Estimation:** Another challenge for marketers is knowing when a customer is highly likely to cease purchasing. Estimating the risk of losing a customer is about computing a reliable indicator of the client's risk of leaving at any given time. The PRIME platform can dynamically organize customers into groups, based on their risk factor. This enables marketers to take preventive measures, like sending special offers.

III. DATA COLLECTION AND PROCESSING

A. Messaging Analytics Data

Messaging service providers, like Apifon, are intermediaries in the communication between different companies and their customers. As such, they are in position to analyze the data generated during this messaging exchange (Fig. 1). This includes events like having the message delivered to the device, having the message opened by a user (in the case of OTT messaging apps like Viber), having a hyperlink inside the message clicked. In addition, it includes content and metadata like the content of a user's response, the type and sentiment of the response (i.e., positive or negative), the time elapsed between receiving a message and taking one of the actions outlined above, as well as user behavior (e.g., product purchase, e-shop visit).
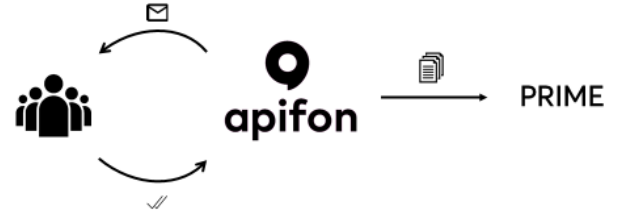

Fig. 1. Architecture of PRIME data flow.

B. Data Integration

The data to be processed consist of raw objects, including incoming messages from users, as well as events describing their status (sent, delivered, seen, opened etc). Messages can be sent through alternative messaging channels, such as SMS or Viber and the events reported by each messaging channel differ in terms of content. Table I presents the different kinds of data that are relevant to messaging analytics.

TABLE I: MESSAGING-RELATED DATA BEING PROCESSED

Message content/metadata	Messaging events
Message Id	Message Id
Account Id	Account Id
Campaign Id	Destination Id
Message text	Channel
Destination Id	Event type
Channel	Event timestamp
Message timestamp	

- **Message Id:** Message mapping key between messages and events.
- **Account Id:** Identifier referring to the company which produced the message.
- **(Optional) Campaign Id:** Campaign which the messages belong to.
- **Message text:** Content of the message sent.
- **Destination Id:** Hashed representation of the mobile number of the message recipient.
- **Channel:** Communication channel through which the message sent.
- **Event type:** Current status of the message.
- **SMS event values:** Sent, delivered, expired
- **Viber event values:** Sent, delivered, seen, opened, expired
- **Message & Event timestamp:** Exact date and time when the message/event object generated.

By the time each object of the above-mentioned entities

appears, a mining service, which is responsible for acquiring and transforming the data, stores them in a dedicated database, in a raw format. Additionally, the same service aggregates this information, and updates the database. Thus, the incoming objects are translated into a suitable representation in order to train the predictive model.

C. Data Processing and Feature Engineering

One of the features which are crucial for training the predictive model has to do with the message content itself. The text of every message sent out to users is processed in order to be assigned to a conceptual cluster. This procedure is based on the use of two supplementary models which have been trained with texts from past messages. These two models, referred to as Natural Language Processing (NLP), and Clustering model, respectively, form a pipeline, which transforms every incoming message text into a set of weights and numbers, providing the capability of dynamically assigning a new message to an existing cluster. (Fig. 2).
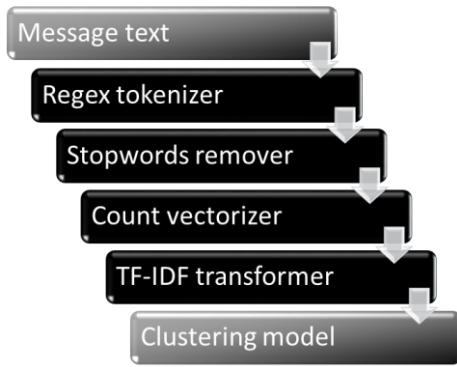

Fig. 2. Processing flow of the conceptual clustering of a new message.

In particular, the mentioned NLP model consists of four successive steps, starting with splitting the text into single words based on punctuation characters and spaces. This model function is called "Regex tokenizer". Following, the "Stopwords remover" function eliminates words in the text which do not have any conceptual significance. All these words are predefined by the model design.

The final two processes, referred to as "Count Vectorizer", and "TF-IDF transformer", are responsible for the conversion of words into numbers (i.e., bag-of-words representation), so as to be served as input to the clustering model. The resulting attributes describe the frequency of occurrence of each word in the whole set of words in the dataset, as well as their associated weights.

In our research, the clustering model was created utilizing the IBM SPSS Modeler Flow platform, which is described in the next section. It is worth mentioning that the clustering process is performed separately for each client company of the messaging services provider, and, consequently, each cluster is being segmented based on an unsupervised clustering model which is trained for each client company individually.

The clustering algorithm adopted in this research is the so-called "Auto Cluster" of IBM SPSS Modeler Flow, which actually applies a combination of several clustering algorithms, such as "TwoStep", "K-Means", and "Kohonen". Auto Cluster will automatically determine the best

combination of algorithms for each client company's dataset. Once this has been done, the system will calculate the optimal number of discrete conceptual categories, and assigns each campaign text to the closest cluster category. In this way, the trained model is able to cluster new message texts.

The rest of the features used for the training of the predictive model are created and modified on-the-fly, right after each new messaging event occurs. The complete list of features which are fed into the predictive algorithm are described in Table II.

TABLE II: DETAILED FEATURES TO FEED THE PREDICTIVE ALGORITHM

Feature	Explanation	Type
Average destinations delivered ratio	Mean of the average number of delivered to sent messages ratio per campaign recipient	Numerical
Average destinations seen ratio	Mean of the average number of seen to delivered messages ratio per campaign recipient	Numerical
Average destinations opened ratio	Mean of the average number of opened to seen messages ratio per campaign recipient	Numerical
Average destinations delivered opened ratio	Mean of the average number of opened to delivered messages ratio per campaign recipient	Numerical
Campaign delivered opened ratio	Average number of opened to delivered messages ratio of the campaign	Numerical
Cluster	Conceptual category assigned to the campaign text	Categorical
Season	Season within the year when the campaign propagated	Categorical

D. Data Privacy

One of the fundamental standards of the architecture of the CTR prediction service, is the compliance with the rules of the GDPR regulation. Having obtained the data subject's consent to processing is the precondition for any type of data processing happening on the platform. The way that this consent is obtained depends on the data privacy and marketing communication policies of the consumer product business that maintains the direct relationship with the customer (i.e., the data owner). Consent is therefore managed outside the whole PRIME platform.

Inside the platform, there are a number of data management services, which were built to support specific GDPR provisions, such as facilitating the data subjects' right of access, and the right to erasure. These terms describe the right of any customer (i.e., the data subject) to receive a copy of all the personalized information held in the platform, as well as to request that this information is permanently removed.

Additional measures to ensure GDPR compliance in the functionality of the CTR prediction algorithm include personal data encryption (e.g., recipient phone numbers), and data segregation. The whole platform architecture is designed such that no single processing node can ever have read access to the full set of data on an individual. Data where a person is present (i.e., identity information), and what their associated

profile is (e.g., predicted next actions) are kept physically separate.

IV. Predictive Model

A. Regression Algorithm Selection

As already mentioned, the CTR prediction service presented in this paper aims to estimate the percentage of the recipients of a message that will read the content, and follow the call to action embedded in the message (e.g., click on a link included in the message body). Hence, the problem can be mathematically formulated as a regression problem of trying to predict a continuous value, which is actually the above-mentioned percentage. This is considered as a reliable indicator that can be used in combination with the segment recommendation service outlined above, or as a stand-alone service to dynamically quantify how good the match is between a specific customer segment, and a particular message.

After testing a variety of machine learning algorithms, such as Logistic Regression [6], and Recurrent Neural Networks [7], we ended up to the XGBoost regression algorithm [8] because of its characteristics that render it ideal in many similar research scenarios [9].

In general, XGBoost belongs to the ensemble learning methods, which rely upon the results of more than one machine learning models. Ensemble learning offers a systematic solution to combine the predictive power of multiple learners. The resultant is a single model, which gives the aggregated output from several models. The models that form the ensemble could be either from the same learning algorithm, or from different learning algorithms. Bagging and boosting are two widely used ensemble learners. Though these two techniques can be used with several statistical models, their widest usage has been with decision trees [10].

XGBoost is a recently proposed algorithm, which is, by design, a scalable machine learning method based on the boosting approach. It incorporates several useful features, such as parallel processing, high flexibility, automated handling of incomplete values, inherent tree pruning and built-in cross-validation [11]. To explain how XGBoost regression works, it is necessary to decompose it into its subprocesses. As opposed to bagging, where trees are built in parallel, in boosting, the trees are built sequentially, such that each subsequent tree aims to reduce the errors of the previous one. Each tree learns from its predecessors, and updates the residual errors. Hence, the tree that grows next in the sequence will learn from an updated version of the residuals, and so on [12].

B. Deployment to the Cloud

The predictive service was implemented on top of the IBM Cloud infrastructure. Editing existing data to calculate the required features to feed the predictive algorithm, requires the processing of large amounts of data. To tackle this, we used Apache Spark for parallel data processing of costly operations, like the calculation of complex aggregations [13]. We also applied NLP techniques [14] for conceptual text parsing.

More specifically, the predictive model was built with the IBM SPSS Modeler Flow tool, provided by the IBM Watson Studio, utilizing the XGBoost linear regression algorithm, which is part of the available models section. It is worth mentioning that the XGBoost model in the IBM Watson Studio carries out the automatic encoding of the categorical variables [15]. After the completion of the feature engineering processes described in the previous section, we used the extracted features (Table II) to feed the XGBoost regression algorithm (Fig. 3).
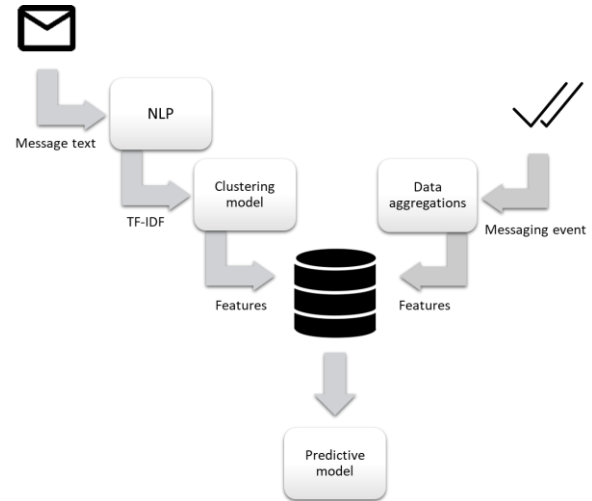

Fig. 3. Overview of the feature engineering procedures to train the model.

In particular, the independent/target variable is the 'delivered opened ratio' (Table II). The rest of the variables are the dependent ones. In the implementation presented by this study, 70% of the samples were used for training, 15% for validation, and the remaining 15% for testing, and the data were shuffled during the splitting process.

In addition, we utilized the Hyper-Parameter Optimization option, which enables parameter optimization based on Rbfopt, which automatically discovers the optimal combination of parameters, so that the model will achieve the expected, or lower, error rate on the sample data (Fig. 4).
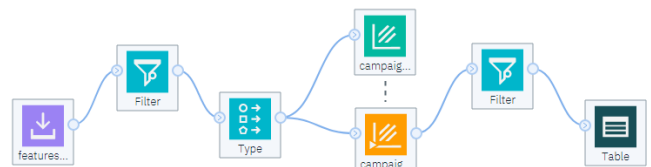

Fig. 4. Process representation in the SPSS modeler flow suite.

Eventually, the process comes up with the creation of a continuously available model, as a service, acting like a secure API (Application Programming Interface) endpoint, under the HTTPS (Hypertext Transfer Protocol Secure) protocol [15]. This API accepts requests in JSON (JavaScript Object Notation) format that must contain the exact features that used to train the model, and returns a float value that represents the predicted CTR in decimal format.

Regarding the model retraining strategy followed by the current study, the predictive model is being retrained at regular intervals, based on the number and rate of occurrence of events pertaining to each of the client companies using the introduced CTR predictor on the PRIME platform. Technically, this ad-hoc approach is implemented leveraging

the capability of IBM SPSS Modeler Flow suite to connect to multiple types of external databases, such as MySQL, and MongoDB, which are constantly being updated with new messaging events that finally are leveraged to create the regularly retrained model [16].

V. MODEL EVALUATION AND RESULTS

A. Scope of Pilot Testing

The data for testing the CTR prediction service were sourced from the messaging campaign history of one of Apifon's client companies. The selected client is a packaged consumer goods brand in baby care market. Apifon's platform has been serving as intermediary for the communication between the client and its customers for a sufficiently long time, such that all necessary messaging exchange data could be obtained.

B. Segmentation of Campaigns

A statistical analysis took place on all available past campaigns sent by Apifon's client. A descriptive statistical analysis of the particular client's profile, depicting some core information on its operational activity, in the context of utilization of the Apifon's platform, can be organized into the following Table III.

TABLE III: DESCRIPTIVE STATISTICS OF A CLIENT'S CAMPAIGNS

Statistical measure	Value
Number of campaigns	165
Average recipients per campaign	3220
Standard deviation of recipients	5817
Minimum number of recipients	20
25% of recipients distribution	129
50% of recipients distribution	471
75% of recipients distribution	2994
Maximum number of recipients	30823

The table shows that approximately ¼ of the campaigns appeared to have less than 130 recipients, and another ¼ of them exceeded 470 receivers, while almost half of the total campaigns had less than 500 recipients.

Taking into account both of the above investigation outcomes, as well as the high cost of larger campaigns, the evaluation process focuses on the campaigns having more than 470 recipients (i.e., 39 campaigns), discarding the rest.

C. Evaluation Approach

The evaluation of the CTR prediction service was performed based on backward testing. This kind of testing was applied to the selected campaigns based on the criteria described in the previous section. Moreover, in order to avoid biased results and overfitting, the whole evaluation process run in a form of simulation in terms of time. The predictive model was continuously re-trained with new campaigns. On this basis, the evaluation metric that appropriately quantifies the difference between the predicted, and the actual CTR, is the Mean Absolute Error (MAE) [17].

In more detail, the MAE measures the average size of errors in a set of forecasts, without considering their direction, in the sense of the origin of the error. In other words, it is the mean of the test sample of the absolute differences between the prediction, and the actual observation, where all the individual differences are of equal weight, and the

corresponding mathematical formula is given by the following equation:

$$MAE = \frac{1}{n} \sum_{j=1}^{n} \left| y_j - \hat{y}_j \right|$$

where n is the sum of the observations, y is the actual value and $\hat{y}$ is the predicted value. The MAE value can range from 0 to ∞, irrespective of the direction of the errors, as noted earlier. Practically, these are negative-oriented results (i.e., the lower prices are the better ones).

D. Results

Based on the above descriptive analysis, the average MAE metric was calculated for several campaign segments, and specifically for campaigns with number of recipients greater than 100, 200, 400, 470, 1000, and 2000, and is illustrated in the following diagram (Fig. 5).
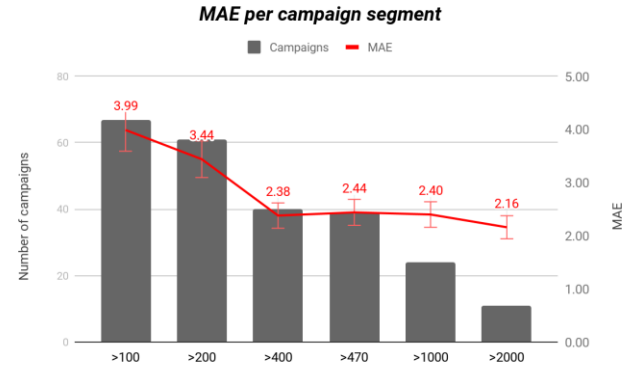

Fig. 5. CTR prediction deviation for multiple campaign segments.

The above diagram reveals that the minimum error value is approximately 2.4 percent points, and it is achieved for campaigns with more than 400 recipients, a condition that is met for 40 historical campaigns.

VI. CONCLUSIONS

The services which are part of the PRIME platform allow businesses which offer consumer products and services to improve the way they communicate with their customers throughout the lifecycle of the customer relationship. Marketers can have a trustworthy prediction of the expected CTR of their campaign by simply providing the message text and the desired customer segment. The intended benefit is improved customer experience and marketing effectiveness while promoting responsible innovation.

This paper showcases a CTR prediction approach of a potential business messaging campaign, and highlights it as a critical indicator of marketing success. Simultaneously, this research analyzes the associated algorithmic perspective describing how it is actually implemented utilizing the XGBoost regression algorithm, along with its deployment as a service on top of a cloud infrastructure. In addition, it also addresses the inherent real-world challenges (e.g., large scale data computation, data privacy).

As shown through pilot testing with backward measurement, our solution was able to predict the CTR of

messaging campaigns having more than 400 recipients with an average deviation of 2.4 percentage points from the actual CTR values.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

A.D. and C.A. analyzed the data, designed the model along with the associated computational framework, and carried out the implementation. A.D. wrote the manuscript with input from all authors. D.K. was in charge of the overall direction and planning; all authors have approved the final version.

REFERENCES

[1] A. Manrai, "New emerging business models, frameworks, and trends in global marketing," *Journal of Global Marketing*, vol. 29, no. 4, pp.171-173, 2016.

[2] D. George, K. Reutimann, and A. Tamò-Larrieux, "GDPR bypass by design? transient processing of data under the GDPR," *SSRN Electronic Journal*, pp. 6-17, 2018.

[3] A. Deligiannis, C. Argyriou, and D. Kourtesis, "Predictive personalization of conversational customer communications with data protection by design," in *Proc. IEEE/WIC/ACM International Conference on Web Intelligence on - WI '19 Companion*, 2019.

[4] J. Braun, "Going over the top: Online television distribution as sociotechnical system," *Communication, Culture and Critique*, vol. 6, no. 3, pp.432-458, 2013.

[5] X. Zhu, H. Li, and F. Li, "Privacy-preserving logistic regression outsourcing in cloud computing," *International Journal of Grid and Utility Computing*, p. 144, 2013.

[6] K. Yilmaz and S. Belbag, "Prediction of consumer behavior regarding purchasing remanufactured products: A logistics regression model," *International Journal of Business and Social Research*, vol. 6, no. 2, 2016.

[7] T. Lang and M. Rettenmeier, "Understanding consumer behavior with recurrent neural networks," *Workshop on Machine Learning Methods for Recommender Systems*, 2017.

[8] T. Çakmak, A. Tekin, C. Şenel, T. Çoban, Z. Uran, and C. Sakar, "Accurate prediction of advertisement clicks based on impression and click-through rate using extreme gradient boosting," in *Proc. the 8th International Conference on Pattern Recognition Applications and Methods*, 2019.

[9] T. Neller, "AI education matters," *AI Matters*, vol. 4, no. 2, pp. 5-7, 2018.

[10] B. Sumana and T. Santhanam, "Optimizing the prediction of bagging and boosting," *Indian Journal of Science and Technology*, vol. 8, no. 35, 2015.

[11] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in Proc. *the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, ACM, pp.785-794, 2016.

[12] A. D. Myttenaere, B. Golden, B. L. Grand, and F. Rossi, "Mean absolute percentage error for regression models," *Neurocomputing*, vol. 192, pp. 38-48, 2016.

[13] S. Ko and J. Won, "Processing large-scale data with apache spark," *Korean Journal of Applied Statistics*, vol. 29, no. 6, pp.1077-1094, 2016.

[14] A. Svyatkovskiy, K. Imai, M. Kroeger, and Y. Shiraito, "Large-scale text processing pipeline with Apache Spark," in *Proc. 2016 IEEE International Conference on Big Data*.

[15] P. Dhoolia, P. Chugh, P. Costa, N. Gantayat, M. Gupta, N. Kambhatla, R. Kumar, S. Mani, P. Mitra, C. Rogerson, and M. Saxena, "A cognitive system for business and technical support: A case study," *IBM Journal of Research and Development*, vol. 61, no. 1, pp. 74-85, 2017.

[16] T. Morelli, C. Shearer, and A. Buecker, "IBM SPSS predictive analytics: Optimizing decisions at the point of impact," 2010.

[17] A. Botchkarev, "A new typology design of performance metrics to measure errors in machine learning regression algorithms," *Interdisciplinary Journal of Information, Knowledge, and Management*, vol. 14, pp. 045-076, 2019.

Alexandros Deligiannis obtained his MSc. in big Data science from the Queen Mary University of London, UK and his BSc. in mathematics from the Aristotle University of Thessaloniki, Greece. Currently, he is working as data engineer in the Research and Development department of Apifon, Greece in the field of personalized communication in order to increase sales and improve customer experience. His main research interests include model driven systems engineering in several application areas and ad-hoc process optimization algorithms creation.

Charalampos Argyriou obtained his BSc. in applied informatics from the University of Macedonia, Thessaloniki, Greece. Currently, he is working as a research and development engineer in Apifon, Greece. His main research interests include, amongst others, machine learning, data mining, natural language comprehension and information retrieval, especially in the field of customer experience management.

Dimitrios Kourtesis holds a PhD in computer science and an MSc in software engineering & telecommunications from the University of Sheffield. He has been working at the intersection of technology research and product innovation since 2006, taking over a variety of roles as R&D software engineer, PhD researcher, technology and product development consultant, founder of technology businesses and advisor to fast-growing companies. He has worked on a range of software engineering research and product development projects involving cloud service architectures, data intelligence and IoT.