ICDSST 2020 on Cognitive Decision Support Systems & Technologies

Direct Email Marketing optimisation with a Random Forest based approach

Giulia De Poli, Maria Angélica Lobo Paulino, Stefania Tola, Manuela Bazzarelli, Leone De Marco, Matteo Bregonzio
3rdPlace SRL
Foro Buonaparte 71, 20121 Milan, Italy
giulia.depoli@3rdplace.com, maria.lobopaulino@3rdplace.com, stefania.tola@3rdplace.com, manuela.bazzarelli@3rdplace.com, leone.demarco@3rdplace.com, matteo.bregonzio@3rdplace.com
web-page: www.3rdplace.com

ABSTRACT

Optimisation in digital advertising is a complex task deemed to increase customers engagement and satisfaction. In more detail, optimisation involves not only identifying the right images, template and timing to engage a given customer, but also understanding the context and judge whether the message is actually relevant to the recipient. To be successful advertising optimisation requires taking into account lots of data coming from multiple digital sources such as Customer Relationship Management (CRM), web analytics, and advertising interactions. Although this process could be performed manually by marketing specialists, more recently data-driven methodologies have shown promising results. In this direction, our study proposes an automated system addressing advertising optimisation via a supervised learning approach where decision-making is performed accounting for the latest customers interactions in a near-real-time fashion. Specifically, this work presents a solution for direct email marketing (DEM) composed of three modules: monitoring, decision-making and automation. Monitoring is provided through a web dashboard showing historical performance of relevant Key Performance Indicators (KPI). The decision-making module computes a relevance score predicting how a given email message or sequence of messages are suitable for a specific customer or cluster of customers. Subsequently, this score is used to support the decision process within the automation module in order to deliver fully personalised messages. Experimental results confirm that the proposed DEM management system promotes customer satisfaction minimising perceived spamming. Moreover, DEM activities contribute to boost the revenue without sacrificing the customer's experience.

Keywords: Email marketing, Machine learning, Prediction, Email personalisation, Direct email marketing automation.

INTRODUCTION

Nowadays, digital advertising strategies aim to engage customers over multiple touchpoints where highly personalised content and messages are delivered. Within this context, direct email marketing (DEM) represents a crucial moment along the customer journey where a sequence of optimised messages could influence customer decisions [1]. Not only this optimisation process may lead to revenue growth but it also promotes a fruitful customer engagement and satisfaction over a long time period. To this end, one of the main challenges in DEM optimisation involves understanding customer needs and preferences in a timely fashion by observing his interactions with all digital sources. DEM optimization takes into account a variety of features including message copy, interactivity, promotions & rewards, illustration and timing. Different permutations of those features may be appealing to different customers: for instance, students tend to be more sensitive to price while business managers could be more interested in comfort [2]. However, when business managers book for their family they may be sensitive to price as well. Consequently, understanding the context also plays a significant role. Historically, this analysis has been performed manually by marketing specialists, while more recently, data-driven methodologies have been extensively tested and have shown promising results [3].

In this paper, we propose a machine learning based approach for DEM optimisation where a variety of data, generated by multiple digital sources such as Customer Relationship Management (CRM), web analytics and advertising tracking are analysed in near-real-time to account for behavioural changes in customer's preferences. This objective is achieved with a classification model that provides a relevance score for a given email message. This classification model is the core of the DEM managing system used in this experiment. The proposed DEM managing system is composed of three modules: monitoring, decision making and automation. Monitoring is provided through a web dashboard showing historical performance of relevant Key Performance Indicators (KPI) such as click-through rate (CTR), conversion rate (CR), and customer engagement. The decision-making module computes a relevance score predicting how a given email message or sequence of messages are suitable for a specific customer or cluster of customers. This score is then used to support the decisions process within the automation module in order to deliver fully personalised messages.

The paper is organised as follows. The next section describes the related work in the field of marketing automation. The third section presents the proposed DEM managing system with a focus on the classification model. Finally, the fourth section shows experimental results and the final section summarises the conclusions and achievements.

RELATED WORK

In the field of machine learning several papers tackle direct email marketing optimisation [3,4,5]. These studies show the possibility of identifying patterns in human behaviours and use them to predict customers interactions or optimise email marketing campaigns. Interesting results are reported by Flici et al. [6] when predicting customers interactions using a machine learning framework. Cui et al. [7] propose a comparison among different classifiers by testing a U.S. based catalogue of direct marketing emails; from this study emerges that a Bayesian Networks (BN) learned by evolutionary programming models reports the best performances. Ma et al. [8] presents a nonhomogeneous hidden Markov model of dynamics response and mailing optimization in the context of direct marketing. In [9] Ładyżyński et al. apply classification trees (CART), random forests (RF) and deep belief networks (DBN) to predict if a specific campaign and contact time will be effective over a customer. Zhu et al. [10] focus more on e-commerce applications where the challenge of purchase prediction is addressed applying a Logistic Regression model in a semi-supervised learning and multi-view learning. Besides the different goal, this paper contains interesting insights on data modelling procedures for the travel industry. A similar challenge is studied by Lang et al. [11] where a Recurrent Neural Network (RNN) was applied over a set of website sessions recording aiming to predict a purchase event. Despite positive results they remark how poor sessions tracking can compromise the quality of the model.

All the above-mentioned works show strong potential in the application of machine learning models for direct marketing optimisation.

OPTIMISED DEM MANAGING SYSTEM WITH A MACHINE LEARNING APPROACH

The proposed DEM managing system, composed of monitoring, decision making and automation modules, have been applied to two real world scenarios in the travelling sector which provided comparable datasets for processing.

Data

The used dataset [13] integrates multiple sources including: CRM, web tracking, advertising engagement and email marketing interactions. This information is then aggregated creating a rich customer profile describing the propension toward certain products, web navigation habits, and the level of advertising engagement. For instance, from the CRM are extracted features such as number of trips bought, average number of passengers per trip, percentage of trips during working days/holidays, favourite environment (first or second class) or usage of discounts. Complementary features come from web analytics including: website search behaviour, preferred web navigation time (morning, afternoon, evening, night), customer interests toward a product measured by time spent on a certain webpage. Regarding customer interaction with previous email marketing campaigns, features such as the number of emails received, number of emails opened, number of clicks and conversions are computed.

Monitoring dashboard

A web dashboard has been designed for continuous monitoring purposes and historical KPIs evaluation. Given a selected time period, the dashboard shows a variety of aggregated information including emails sent, A\B testing results, along with other KPIs (open rate, CTR, CR). Furthermore, it is possible to perform deep-dives for each single campaign. This dashboard is extensively used by analysts during planning activities and business intelligence explorations; Figure 1 proposes one of the available charts.

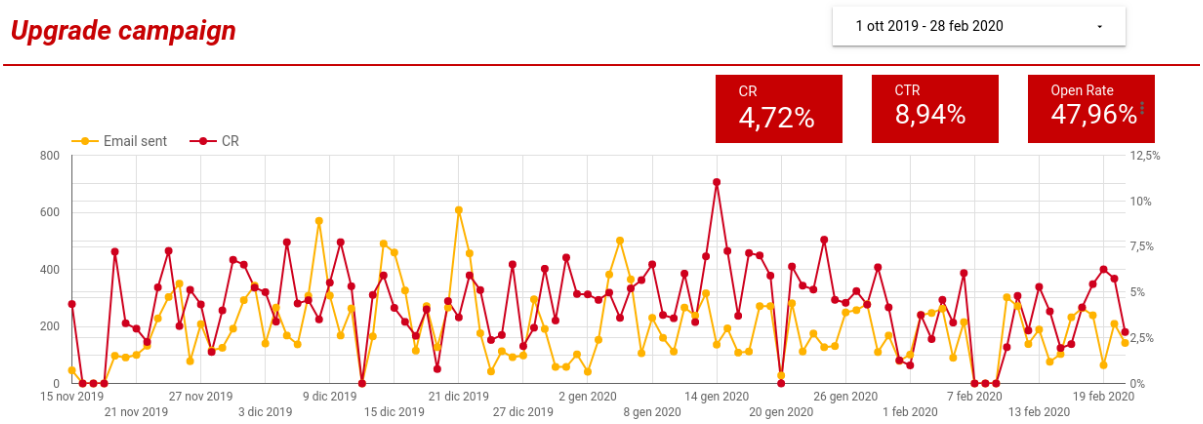

Figure 1: Example of one of the charts available in the monitoring dashboard

Classification model for decision-making

An ensemble learning classification model has been developed to support decision-making in the proposed DEM managing system. The model, a Random Forest classifier, provides a relevance score describing how a given email message could be of interest for the given customer (or cluster of customers). The main motivation behind Random Forest is to tackle an unbalanced and sparse dataset while avoiding overfitting. As validation step, Random Forest is compared against logistic regression, chosen as a baseline for its simplicity and robustness.

Aiming to improve the model accuracy without overfitting, a recursive backward feature elimination step is employed. Given the unbalanced dataset, different oversampling and undersampling techniques were applied, leading to the choice of SMOTE + Tomek Links since delivering better performances against the testing set. SMOTE is an oversampling method that synthesizes new plausible examples in the majority class; Tomek Links performs undersampling making the decision boundary less ambiguous. Throughout a k-fold cross-validation process the classifier hyperparameters, number of estimators and max depth, are selected and used to train the model. Similarly, to identify the optimal relevance score threshold used within the automation module, a selection process is performed. Importantly, the above-mentioned training routines are repeated weekly, in order to maintain the model up to date with respect to new user behaviours.

Within the Automation module the computed relevance score and the optimal threshold, in combination with the customer profile, are then iteratively consumed to: a) optimise the email parameters such as images, template, and sending time. b) Decide whether to fire or not the optimised email to a given customer.

Automation module

As part of a continuous integration pipeline, the automation module is deployed on a Cloud environment. As the first task, this module attempts to optimise an email message based on the customer profile (or cluster of customers) and verify whether the computed relevance score is above the acceptance threshold. If this check is positive the module schedules the email sending at an optimal time. Moreover, the automation module implements an anti-spam filter which controls the maximum number of marketing emails a user can receive during a time period and the time window between two emails in order not to overload user's mailbox. Finally, it offers an A/B testing solution to evaluate performances against a control group. As it can be seen, this automation module delivers a fast and dynamic decision-making approach towards DEM optimisation.

EXPERIMENTS

The proposed DEM management system has been validated and deployed on two real-world travel companies. The reported performances are evaluated over a 6 months period and importantly, in accordance with European GDPR [12] law, email personalisation is offered exclusively to customers having explicit marketing and profiling consent. Table 1 provides a summary of the experiment size.

Table 1: Summary of the monthly average number of users processed

	Monthly website users	**Users with marketing & profiling consent**
Company 1 Dataset	2,970,00	830,000
Company 2 Dataset	430,000	172,000

This paper evaluates two distinct email marketing campaigns applied in both travel companies: the first campaign consists of a reminder for an abandoned cart, the second is a discount for class upgrade. In detail, the abandoned cart reminder is sent to all customers who attempt to buy a ticket on the website (i.e. the ticket was in the cart) but didn't complete the purchase. The email is fired after around 15 minutes from the abandon and doesn't offer any discount. Differently, the class upgrade campaign offers a discount that expires within 3 days; is available only to a business cluster and the emails are fired at an optimal time based on customer habits. In both campaigns the goal consists of minimising the number of emails sent without sacrificing the revenue. In other words, the challenge involves finding the right balance between customer satisfaction (avoiding spamming) and revenue.

As presented in the previous section, the collected dataset appears unbalanced hence we use f1 score as metric for comparing different approaches. Table 2 presents the performances observed using a logistic regression classifier and a random forest.

Table 2. Comparison of classification performances on historical data valuated in terms of Accuracy (Acc), Precision (P), Recall (R), and F1 score (F1)

		Company 1				**Company 2**			
		Acc	**P**	**R**	**F1**	**Acc**	**P**	**R**	**F1**
Abandoned cart reminder	Logistic Regression	0.93	0.05	0.03	0.04	0.89	0.08	0.15	0.11
	Random Forest	0.94	0.17	0.09	0.12	0.91	0.18	0.24	0.2
Class upgrade	Logistic Regression	0.94	0.28	0.09	0.13	0.94	0.29	0.07	0.11
	Random Forest	0.94	0.29	0.09	0.14	0.95	0.31	0.08	0.12

Experimental results on historical data confirm that Random Forest based classification delivers a better performance compared to the benchmark and this solution is therefore adopted in the decision-making module.

For clarity, Figure 2 reports an example of feature importance for Company 1 dataset used in class upgrade. Interestingly, the three features with highest contribution are: number of trips, travelled in first or second class and number of opened emails with a discount offer. Those are coming from both CRM and email marketing interactions.

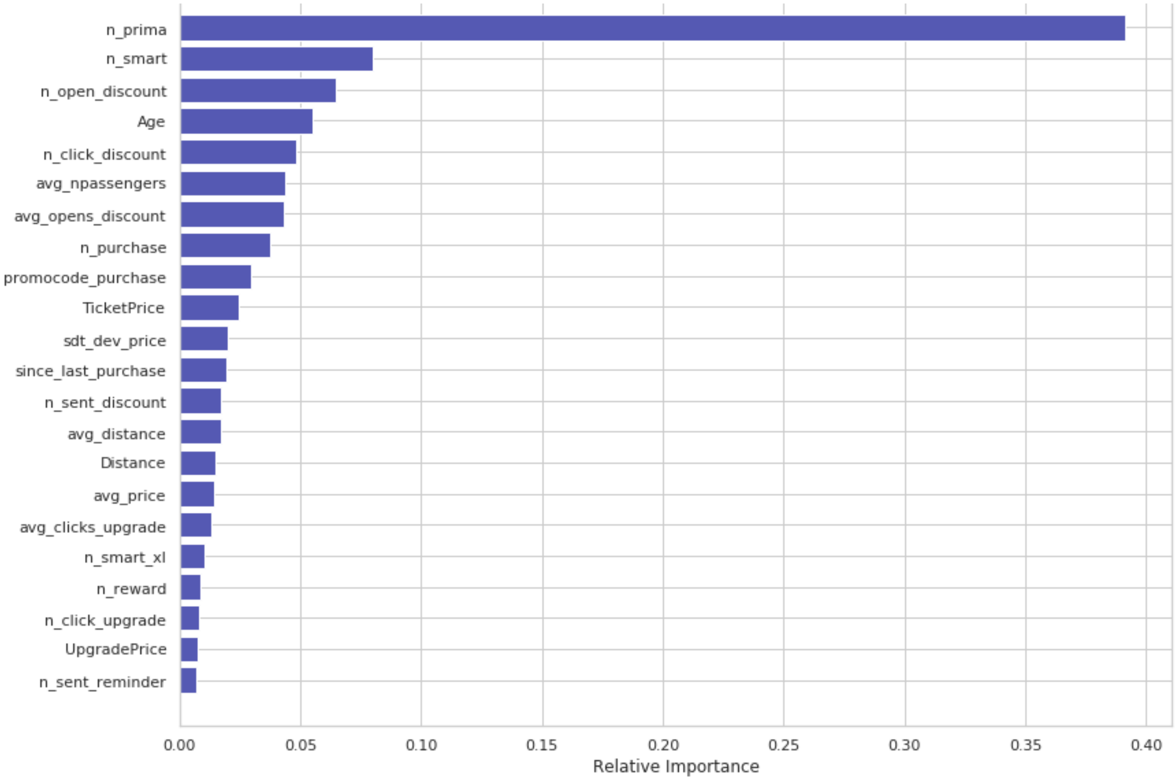

Figure 2: List of a subset of Company 1 dataset's features ordered by importance. Those features are used for class upgrade.

In order to verify the effectiveness of the proposed DEM managing system on live data, an A\B testing validation is performed by the automation module, setting as conversion goal the ticket purchase Specifically, marketing emails are also sent to a randomly selected control group representing 5% of the experiment population. Interestingly, as depicted in Table 3, the observed open and conversion rates appear notably higher when emails are delivered to customers selected by the proposed solution, which consists of 10% of the entire population for the reminder email and 12% for the class upgrade. This result confirms the effectiveness of the methodology highlighting that: 1) in both campaigns emails are delivered to actually interested customers. 2) A high open and conversion rates are associated with customer satisfaction and minimization of perceived spamming. 3) Revenue boost from DEM is achieved without sacrificing customer's experience.

Table 3. A\B Testing results on live data. Open Rate refers to opened email, Conversion Rate to purchased tickets

	Abandoned cart reminder		**Class upgrade**	
	Open Rate	**Conversion Rate**	**Open Rate**	**Conversion Rate**
(A) Selected customers	78.8%	12.4%	83.4%	18.7%
(B) Control group	44.2%	3.4%	48.3%	4.7%

CONCLUSIONS

This paper proposes an innovative DEM management system capable of personalising messages by selecting the most appropriate images, template, and sending time for a given customer. Additionally, a decision-making approach based on machine learning is introduced in order to judge whether an email message is actually relevant for a given customer within a specific context. This innovative DEM managing system is composed of three modules named monitoring, decision-making and automation. Experimental results on two real-world travel companies confirm the effectiveness of the system underlining that emails are delivered to customers actually interested in the advertisement. The observed open and conversion rates are expressions of customer satisfaction and minimization of perceived spamming. Finally, both direct marketing campaigns deliver an increase in revenue without sacrificing customers experience. Although the obtained results are promising, future improvements could be achieved by exploring different areas such as email wording optimisation, accounting for social network interactions and by testing more sophisticated classification algorithms as deep learning.

REFERENCES

1. Singh, G., Singh, H. & Shriwastav, S. Improving Email Marketing Campaign Success Rate Using Personalization. Advances in Analytics and Applications 77–83 (2019)

2. Yang, K., Min, J. H. & Garza-Baker, K. Post-stay email marketing implications for the hotel industry: Role of email features, attitude, revisit intention and leisure involvement level. Journal of Vacation Marketing vol. 25 405–417 (2019)

3. Abakouy, R., En-Naimi, E. M., El Haddadi, A. & Elaachak, L. Machine Learning as an Efficient Tool to Support Marketing Decision-Making. Innovations in Smart Cities Applications Edition 3 244–258 (2020)

4. Zhang, X. (alan), Kumar, V. & Cosguner, K. Dynamically Managing a Profitable Email Marketing Program. Journal of Marketing Research vol. 54 851–866 (2017)

5. Lawi, A., Velayaty, A. A. & Zainuddin, Z. On identifying potential direct marketing consumers using adaptive boosted support vector machine. 2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT) (2017)

6. Flici, A. A Conceptual Framework for the Direct Marketing Process Using Business Intelligence. (2011)

7. Cui, G., Wong, M. L. & Lui, H.-K. Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming. Management Science vol. 52 597–612 (2006)

8. Ma, S., Hou, L., Yao, W. & Lee, B. A nonhomogeneous hidden Markov model of response dynamics and mailing optimization in direct marketing. European Journal of Operational Research vol. 253 514–523 (2016)

9. Ładyżyński, P., Żbikowski, K. & Gawrysiak, P. Direct marketing campaigns in retail banking with the use of deep learning and random forests. Expert Systems with Applications vol. 134 28–35 (2019)

10. Zhu, G., Wu, Z., Wang, Y., Cao, S. & Cao, J. Online purchase decisions for tourism e-commerce. Electronic Commerce Research and Applications vol. 38 100887 (2019)

11. Lang, T. and Rettenmeier, M. "Understanding Consumer Behavior with Recurrent Neural Networks." (2017)

12. 2018 reform of EU data protection rules. European Commission. May 25, 2018. URL: https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf

13. Dataset sample, https://github.com/brego81/Direct-Email-Marketing-Dataset