loxz digital

Font Optimization Model

By Buwani Manuweera, Data Scientist

Table of Contents

ABSTRACT

This report describes the Font Optimization model developed by Loxz Digital Group. This model provides predictive analytics on optimizing the font styles in an email campaign **prior** to deploying it, in order to improve the engagement rates. The Font Optimization model provides recommendations for the size and the style of the font that will give the highest engagement rates based on the selected model parameters or inputs.

The current dataset contains 888 samples of data with features such as different age groups and vision impairment rates. The machine learning algorithm used in this model is Random Forest Regression which is an ensemble of decision trees. The Font Optimization model is able to provide the highest accuracy of 84.98% with our current dataset. We believe our model can achieve a higher accuracy given additional datasets from Email Service Providers and or large senders.

I . INTRODUCTION

We believe, that choosing a specific font for a targeted email is an important task in any email campaign. Based on the target group, you'll need to consider the font style, font size, and many other properties so the readers of your email feel comfortable and inclusive while reading it. In order to achieve a better user experience and increased engagement in an email campaign the campaign engineer needs to determine the optimal font properties. The Loxz Font Optimization Model helps campaign engineers determine the optimal font properties for their emails in a particular industry and in a particular type of campaign, albeit it an abandoned cart campaign or an engagement campaign.

For example, let's consider seniors who are viewing your emails. They may need a special type of font and a larger font size depending on their eyesight. The current model takes these factors into consideration and provides recommendations accordingly.

.

Currently, there is a higher number of emails for promotional campaigns since more promotional emails are being sent everyday in each industry. This distribution is shown as a pie chart in Figure 2. We plan to introduce additional datasets of emails for other campaigns as well to increase the list of campaign types in the future.

II. DATA SET

The model uses 888 emails across eight different industries and six different campaign types. Table 1 shows the number of emails across the industries and Figure 1 shows the percentages of emails as a pie chart. This set of emails was selected from a collection of proprietary emails containing the email font information. For the email data samples, a collection of carefully curated emails belonging to different campaigns and industries is used in the '.eml' format.

The types of campaigns are as follows.

- Abandoned Cart
- Engagement
- Newsletter
- Product Announcement
- Promotional
- Transactional

Currently, there is a higher number of emails for promotional campaigns since more promotional emails are being sent everyday in each industry. This distribution is shown as a pie chart in Figure 2. We plan to introduce additional datasets of emails for other campaigns as well to increase the list of campaign types in the future.

The model uses the body of the email to extract most of the features. The model features include the *font style, font size, gender of the target group, vision impairment rates according to US states, age group, industry type*, and *campaign type*. The target variables considered are *Click-To-Open Rate and Conversion Rate*. These target variables can be customized and should be considered when leveraging the model. The current target variable library plans to be expanded to include revenue per email, unsubscribe rates and average order volume, as needed.

For the Font model, *Open Rate* was disregarded as the email body appears after opening an email and won't affect the *Open Rate* by changing the font of the email body. We are also considering implementing *Revenue per email* in the next version of our model.

Due to the lack of publicly available data for the target variables, we refer to multiple resources of email campaign benchmarks to generate a set of data within predefined normalized distribution ranges [1–3, 6, 10]. This approach allows us to create our model for the potential use in email campaigns with non-synthetic data. Currently we use resources from Campaign Monitor, Ruler Analytics, ContentGrip, Listrak's Cross-Channel Marketing Automation Platform and CM Commerce. Benchmark Data is important for Loxz to establish a baseline. We then consider options to enhance the variance between each recommendation.

Industry	No. of Emails
Academic and Education	14
Energy	36
Entertainment	160
Financial	25
Healthcare	14
Hospitality	166
Real Estate	10
Retail	259
Software and Technology	188
Transportation	16
Total	**888**

Table 1. Email distribution in the dataset across the industries

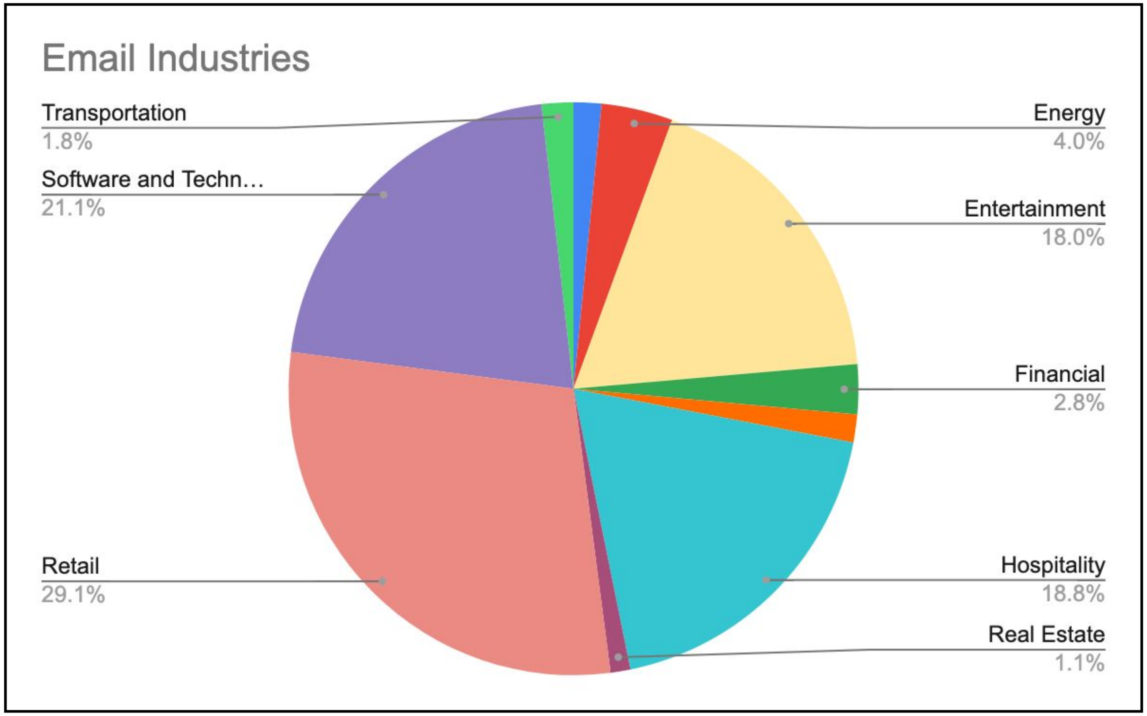


Figure 1. Percentages of emails distributed across industries in the dataset

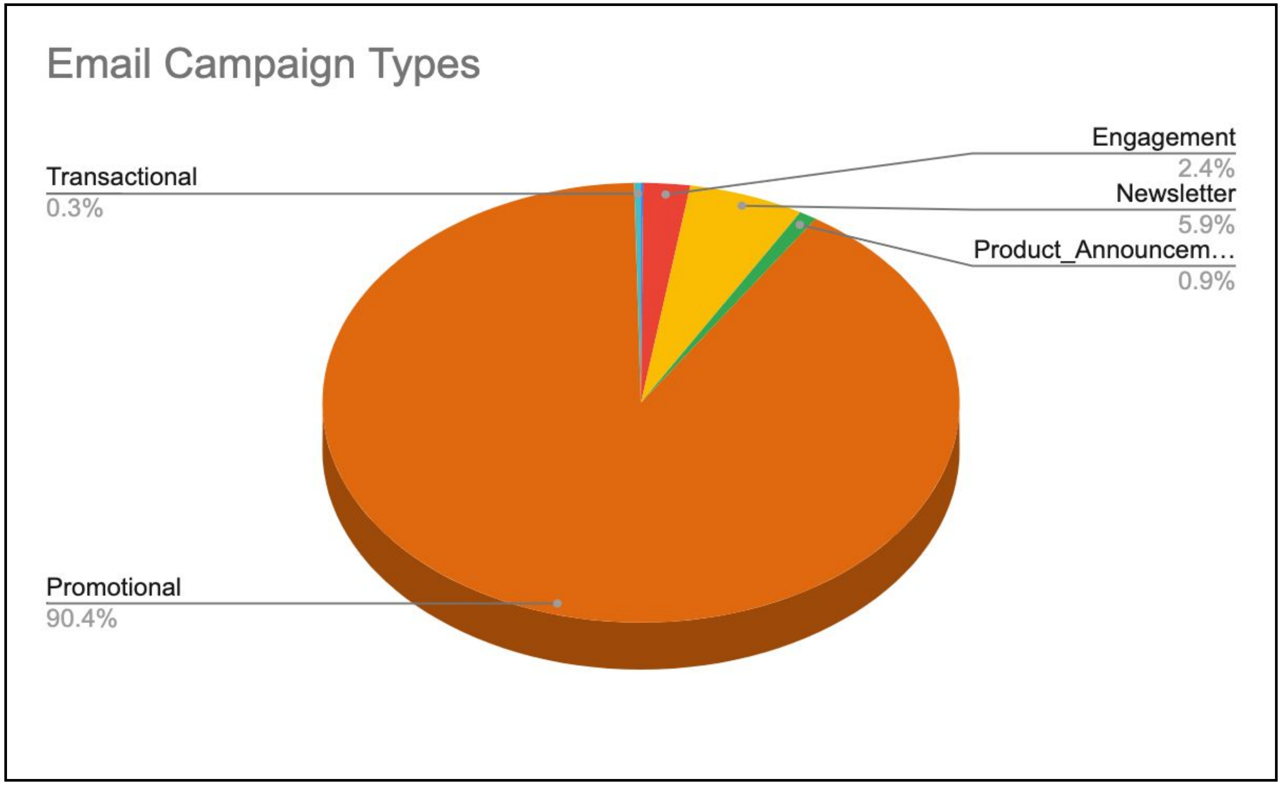


Figure 2. Percentages of emails distributed across different campaign types in the dataset

Ⅲ. FEATURE ENGINEERING

As mentioned in the previous section, the feature set consists of *industry type, campaign type, gender, state, age, font style* and *font size*. The *industry type* and *campaign type* information are directly available from the curated email collection. The emails are already categorized or labeled based on the industry and campaign type in the collection. Each feature is explained below with more information.

- industry type: assigned during email categorization
- campaign type: assigned during email categorization
- gender: 'male' or 'female' based on the targeted group
- state: assigned for each email that will reflect the vision impairment rate in that state
- age: a value between 15 to 90 is assigned
- font style: extracted from each email
- font size: extracted from each email

The age and gender information are added as synthetic data for the dataset. The age is added so as certain age groups prefer certain font styles or need certain font sizes to read. For gender, currently we consider statistics related to vision issues of each gender. For the state, the user can select a certain state they prefer to target. Based on the state they select, the model considers the corresponding average vision impairment rate for predictions.

For the *font style* and *font size*, the email content had to be parsed through carefully to identify the information. This is part of the data preparation phase. The filter first goes through the email which is in HTML format and identifies the sections with font styling. If those sections have the font style and size information used, they are collected for the features. Based on the list of font styles and their corresponding sizes, we select the font that is the most throughout the email and select it for font style and its size as features for the model. Following are the keywords we considered when filtering font styles and sizes.

- font-family: Specified for font style
- font-size: The size of the font (in px)
  (if the value is not is px, it is converted to px using a standard guide)

After filtering the sections corresponding to font information, the style and size values are extracted. The font information embedded within the section is extracted using *BeautifulSoup* web-scraping package [12].

IV. TOOLS REQUIRED

The implementation of the model was done in *AWS SageMaker* using Python programming [8]. For feature extraction, the *BeautifulSoup* web scraping package functionalities were used [12]. Machine learning tasks were implemented using the Scikit-learn package for Python [11].

V. MODEL DEVELOPMENT

After generating the features for the dataset, the next step is the model development. The model predicts *Click-to-open rate* and *Conversion rate* which are both continuous values requiring a regression model for predictions. The regression model takes the set of features and the target variables as inputs and trains the model for predictions. The model takes a total of seven features and two target variables. But the model is trained on one selected target variable at a time based on the user's inputs. For the machine learning model, a tree-based algorithm was considered for their simplicity. Both Random Forest and XGBoost algorithms were considered and based on their performance, Random Forest regression is used for model development [5, 9]. The algorithm is used with its default hyperparameter values except for random_state to maintain consistent results each time. The dataset is normalized with *L2-norm* before feeding into the regression model.

Apart from the predicted accuracy and current engagement rate, the model is further developed to give three recommendations for the user. Currently, the recommendations are selected from historical data and output upon run to the campaign engineer for the best font style and size combination for that particular email campaign.

Figure 3 outlines the model development process used for the font optimization model.

The user interface is developed with the option for the user to upload the HTML email for the campaign and select the parameters (industry, campaign, age, state and gender). The age, state and gender inputs are optional and the user can opt them out from predictions. All these parameters can vary in scope and number.

Then the user gets to select which engagement rate to be predicted by the model. The features are extracted from the uploaded email with the same process described in section 3.

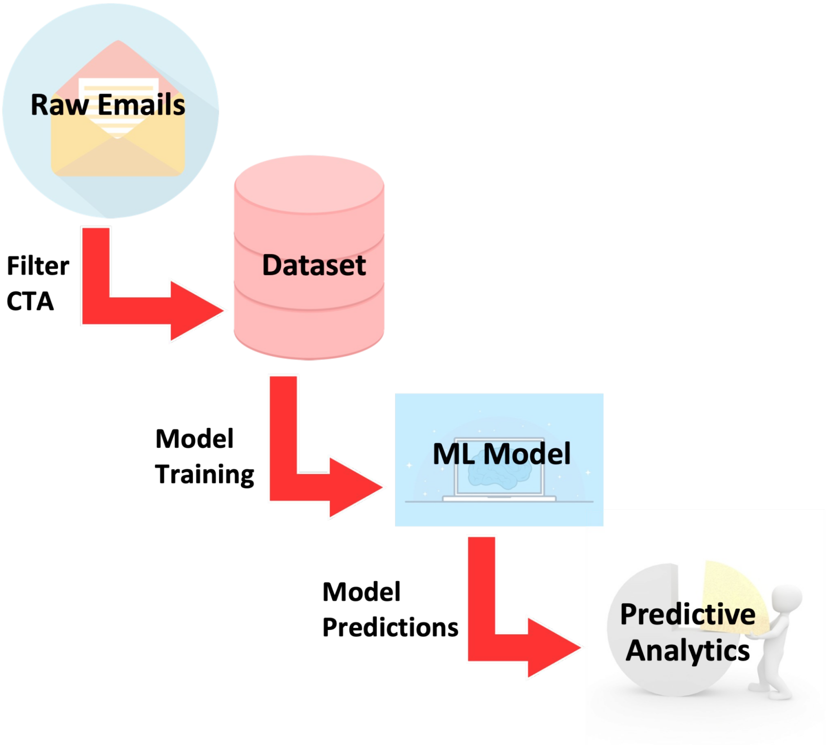


Figure 3. Outline of the model development process

VI. ALGORITHMS USED

For machine learning, the Font model uses Random forest regression algorithm [9]. The tree-based algorithms are easier to interpret than other algorithms. Random forest is a tree-based ensemble method that uses a bagging method where the model output is based on the majority prediction of the trees. The random forest regression model implemented in the Scikit-learn package is used directly for our model development.

VII. MODEL VALIDATION

To evaluate the performance of the prediction model, the regression model is trained with a subset of the dataset while it is validated with the remaining. This kind of partitioning ensures more samples for training while giving a sufficient number of samples for validation. Increased number of training samples can potentially reduce any over-fitting in the model.

To select a regression model with better performance both random forest and XGBoost algorithms were considered. Based on their accuracy values (explained in section 9), Random Forest is selected for the final model.

VIII. ASSUMPTIONS

There are several assumptions that had to be made during the entire process. When generating the dataset, we made assumptions on assigning the gender, age and state information to the email data. These assumptions were based on certain statistical information we researched and gathered.

Due to the lack of availability of data for the model, it refers to the email benchmark data, and assumptions are made to decide on the range of the distribution of target variables based on the average benchmark values.

IX. RESULTS

The accuracy of the model is measured by using the R^2 score. It represents how well the model fits the given data. The higher the value, the better the prediction is. So, using this metric the accuracy of the model is calculated.

First, the performances of both random forest and XGBoost algorithms were compared in order to select the best regression algorithm for the model. Figure 4 shows the performance comparison between the two models. Since Random Forest achieves the highest accuracy, it is selected for the model implementation.

The current model provides an accuracy score of 83.05% for click-to-open rate and 84.98% for conversion rate.

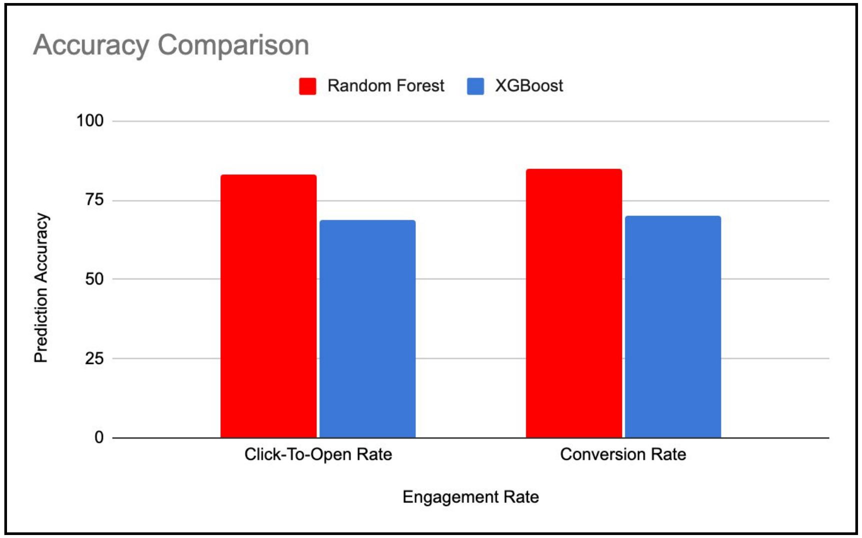
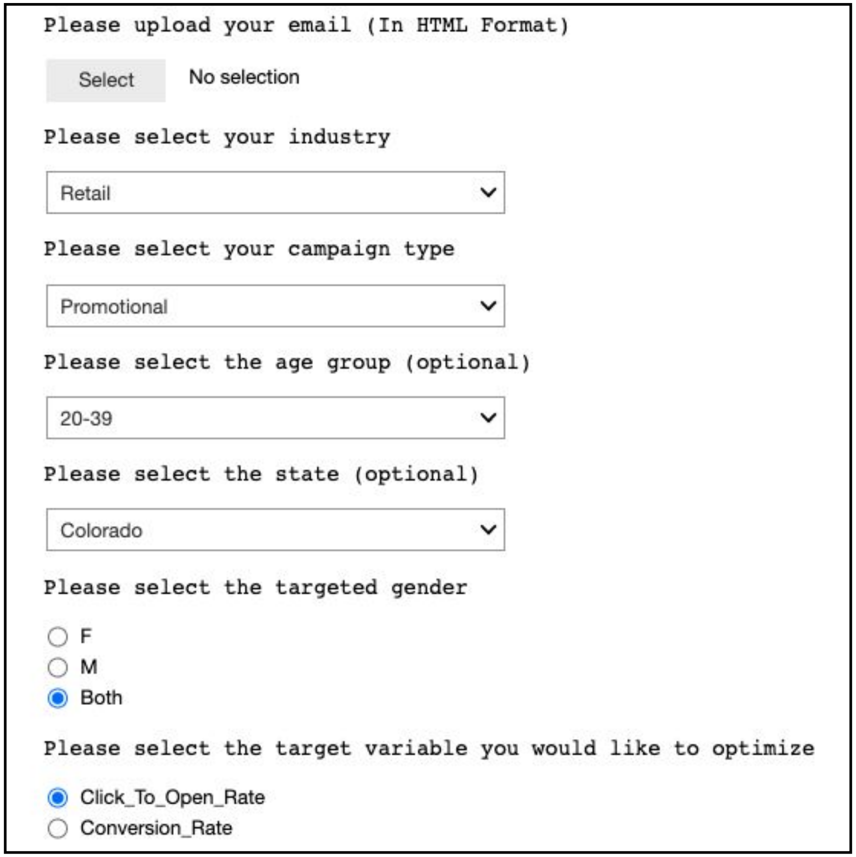


Figure 4. Performance comparison between Random Forest and XGBoost models

X. USE CASES

The Font Optimization model is developed for email marketing campaigns and is to be used by the campaign engineers within the workflow of the campaign and to identify ways to increase user engagement prior to deployment based on given parameters or inputs. Running two or more of our models at the same time can result in higher engagement rates. This would be considered in a multi-model campaign. The current model provides predictive analytics for click-to-open rate and conversion rate. The campaign engineer is able to upload the email for the campaign and select the industry and campaign type. Once again, these input elements can be optimized or increased. Then they can select the preferred engagement rate and the inputs for age, gender and state if preferred. This part of the model interface is shown in figure 5.



Please upload your email (In HTML Format)

No selection

Please select your industry

Retail

Please select your campaign type

Promotional

Please select the age group (optional)

20-39

Please select the state (optional)

Colorado

Please select the targeted gender

○ F
○ M
● Both

Please select the target variable you would like to optimize

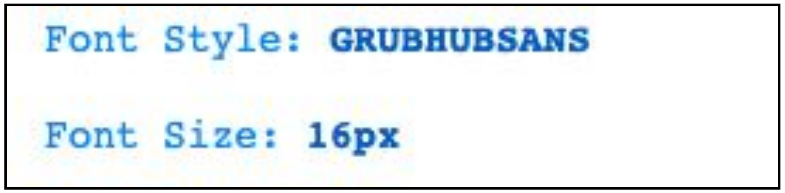
● Click_To_Open_Rate
○ Conversion_Rate

Figure 5. The set of parameters the user has to select for the model predictions

After selecting the parameters, the Font model will first scan the uploaded campaign email for the font information and display them to the user. Figure 6 shows a sample output. We firmly believe that the combination of font style and size as an output and base those outputs on recommendations will enhance campaign engagement rates.

Based on the parameters selected by the user, the model will run for the given campaign email to provide “three” best recommendations. We use three in our model examples, and provide the predicted rate with three recommendations to increase the intended rate by optimizing the font style and size. The model will also display the predicted engagement rate for the current font and the overall accuracy of the model. An example of recommendations and model accuracy is given in figure 7.

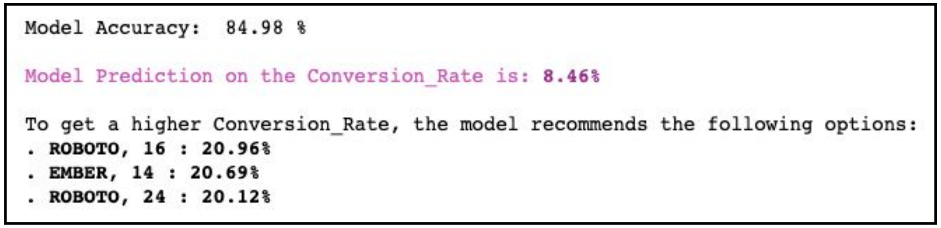
Based on these model outputs, the campaign engineer can decide the most suitable font properties to conform to the highest engagement rate. This is all done prior to deployment and ensures a potentially successful campaign. Using the font model, the campaign engineer can foresee the possible outcomes of the campaign and take any action to increase its outcome using various workflow tools in their UI.



```
Font Style: GRUBHUBSANS

Font Size: 16px
```

Figure 6. Sample output showing the current font style and size used in the email



```
Model Accuracy:  84.98 %

Model Prediction on the Conversion_Rate is: 8.46%

To get a higher Conversion_Rate, the model recommends the following options:
. ROBOTO, 16 : 20.96%
. EMBER, 14 : 20.69%
. ROBOTO, 24 : 20.12%
```

Figure 7. Model prediction accuracy, predicted engagement rate and the three recommendations

X I . DATA SCIENTIST OWN CONCLUSIONS

The Font optimization model is developed to provide predictive analytics for the email campaigns by giving recommendations to improve outcomes within a few seconds. According to the results from the model, it is able to provide higher accuracies for the target variables using the current dataset.

For the click-to-open rate, the model provides 83.05% of accuracy, and for the conversion rate 84.98% of accuracy which validates the performance of the Font model. As these results are based on our current dataset, this suggests our model has more room for improvement in the future with further tweaks and more data. These results show the potential of the model to be used in real-world scenarios. As these results are based on benchmark data, the model will gain further information and gain more performance when used in real-time.

X II . FUTURE CONSIDERATIONS

For immediate future work, the model will be extended to predict Revenue-per-email values. Additional email data will be introduced and will be added to the dataset to improve recommendations made by the model. The current machine learning model uses default hyperparameters and therefore, a hyperparameter tuning process will be done to further improve the performance. Furthermore, the options given for the user input will be improved to consider more genders and valuable eyesight data.

REFERENCES

[1] 2016. Ecommerce Email Marketing Benchmarks. https://cm- commerce.com/academy/email-marketing-benchmarks/

[2] 2018. Email Conversion Rate Benchmarks. https://www.listrak.com/ white-papers/2018-email-benchmarks

[3] 2021. Ultimate Email Marketing Benchmarks for 2021: By Industry and Day. https://www.campaignmonitor.com/resources/guides/email-marketing-benchmarks/

[4] James Bennett. 2022. Module contents — webcolors 1.12 documentation. https://webcolors.readthedocs.io/en/1.12/contents.html

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

[6] Katie Holmes. 2021. Average Conversion Rate by Industry and Marketing Source. https://www.ruleranalytics.com/blog/insight/conversion-rate-by-industry/

[7] Ameet V Joshi. 2020. Amazon's Machine Learning Toolkit: Sagemaker. In *Machine Learning and Artificial Intelligence*. Springer Nature, Chapter 24, 233–243. https://doi.org/10.1007/978-3-030-26622-6_24

[8] Yanli Liu, Yourong Wang, and Jian Zhang. 2012. New machine learning algorithm: Random forest. In *International Conference on Information Computing and Applications*. Springer, 246–252.

[9] Enricko Lukman. 2021. What is a good conversion rate for your business? https://www.contentgrip.com/conversion-rate-business-benchmark/

[10] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830. http://scikit-learn.sourceforge.net.

[11] Leonard Richardson. 2007. Beautiful soup documentation. *Dosegljivo: https://www. crummy. com/software/BeautifulSoup/bs4/doc/.[Dostopano: 7. 7. 2018] (2007).*