

# Boosting Email Marketing Campaigns with Machine Learning

## Abstract

This project uses Machine Learning to increase the possibility of engaging customers taking into consideration the two classic problems to solve in Email Marketing: content and delivery date and time. Here we will use Machine Learning to innovate and add value to the traditional marketing approach on that matter.

## Introduction to the Project

Traditionally, when conducting email marketing campaigns, basic techniques are used to increase engagement based on content changes: A/B Testing on the subject of the email to see which one generates the most open and Multivariable Testing on the content to identify the best version of the text of the body, its images, and distribution.

## The Perfect Moment

However, one of the main factors that determine the opening of the mail is the moment in which it is sent. Before Machine Learning, there was no exact way to learn the best delivery time for each group of recipients, but intuition and observation. This is a classification problem, and we will explore a few models to solve it.

Based on information provided by an Argentinian nonprofit entity, this project takes the challenge of classifying groups of recipients by its most essential characteristics and builds a delivery schedule distributed in as many sub-groups as desired.

## The Relevant Content

There's even more space to improve with Machine Learning. We can get insights to understand the balance between the things we are communicating, and the thing companies are interested in. That's a very valuable information useful to understand how to plan our content marketing strategy and it's easily implemented using Python and Natural Language Processing.

## Challenge 1: Getting insights from the Content

### Data Gathering, Feature Engineering

As we will discover which keywords are important to our audience, we need to gather that information from the existent campaigns. In this case, we analyze 185 campaigns and the engagement to understand which subjects could be essential to generate engaging content.


The feature `total_engagement` is used as a measure to weight the individual keywords. This weight will be visually compared to the weight represented by the number of times that we use that word in different campaigns.

### Using NLP to get insights

We used the NLP `tokenize` library to create the keywords and count the number of times the keyword is used in our campaigns with the library `FreqDist`.

Finally, we create the dataframe to contain both the number of times the word appears in the campaign and the engagement obtained.

To finish, we need to put the engagement score in each Keyword along with the number of times it appears in the campaigns.



	<b>keyword</b>	<b>occurrences</b>	<b>engagement</b>
<b>0</b>	productive	10	7072
<b>1</b>	simplification	11	9946
<b>2</b>	department	29	43192
<b>3</b>	newsletter	29	19560
<b>4</b>	industry	6	1605

## Visualizing the insights with Tableau

In the report generated in Tableau, we can see many opportunities to deliver engaging content. For example, we can create content more related to credits, expiration policy of certificates, registration to programs, among others. We can also note that we are talking too much about international trade and people is not engaged with the subject. This kind of ideas are valuable tools to plan the content marketing of the company and this is an easy way to know where to go in a glimpse.



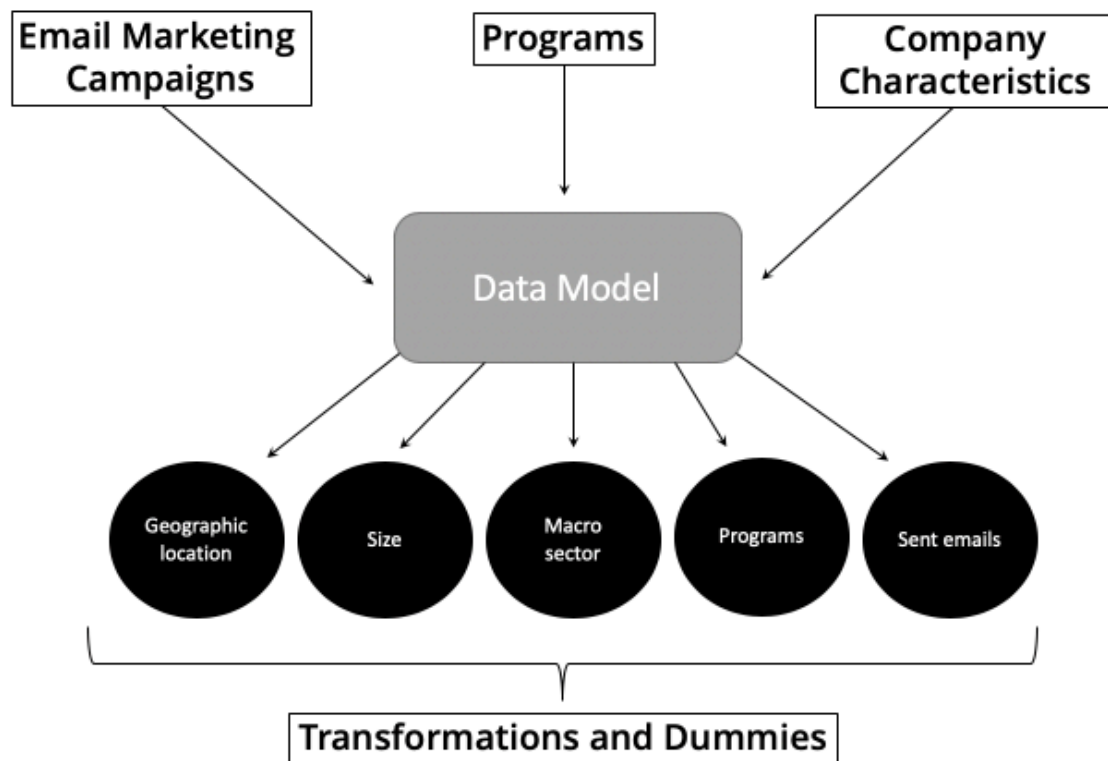
## Challenge 2: Choosing the best date and time to send the campaign

Reaching out the people at the right moment it's probably one of the most important factors involved in engaging the audience. Nowadays, there are too many emails waiting for the user on inbox and chances are we will increase the open rate if we send our email when the user is online. How is it possible to increase the probability of reach the audience at the best date and time of the week?

We need to train a classifier model of Machine Learning using all the relevant features available and decide the appropriate moment for each company.

## The Data

In this case, the nonprofit entity has many departments holding information about the companies enrolled in different programs and benefits. All the details about the companies are stored in the State database and, all the information on the email campaigns are stored in another place.



## Data Gathering and Cleaning

In this step, we obtained the relevant data from email campaigns and the particularities of each sending. Also, all the information related to the company that could be relevant such as:

Important feature descriptions for email campaigns:

- 'campaignId': is the ID of the campaign.
- 'subscriberEmail': is the subscriber email, in this case, is encrypted.
- 'cuit': is the company ID given by the government. In Canada is also known as Business Number.
- 'deliveryStatus': we have interested only the unbounced emails.
- 'lastOpenDate': when this field is not NAN means the mail was open at this date and time.

Note about data of email campaigns: The data was extracted from the email marketing system database. The original input came on CSV files. To protect privacy, the subscriberEmail was encrypted.

Important feature descriptions for companies:

- `cuit`: is the company ID given by the government. In Canada is also known as Business Number.
- `province`: is the province where the company is located
- `lat`: geographic latitude
- `longit`: geographic longitude
- `is_employer`: 0 or 1 describing if the company has employees.
- `size`: if the size of the company. Could be MICRO, MEDIUM or LARGE.
- `macro_sector`: is the macro sector of the economy in which the company fits.
- `fast_growth`: 0 or 1 describing if the company is considered as fast growth.
- `importing`: 0 or 1 describing if the company imports.
- `exporting`: 0 or 1 describing if the company exports.
- `is_client`: : 0 or 1 describing if the company is existent client.

Note about data of email campaigns: The data was extracted from the State database, pre-processed in SQL and extracted to a CSV. The SQL file with the query is `data_empresas_target.sql`.

### Modeling the solution and feature engineering

In order to be able to classify the best delivery date and time, it is necessary to analyze the deliveries with open emails as a result. We will select the messages sent to all the companies that have all the attributes that we will use in the classification and in turn, have had open the mail.

To classify the specific day and time of the week on which the campaign should be sent, it is necessary to conceptualize the week in intervals.

During the week, from Monday to Friday the morning and the afternoon are recognized, each being half a day. Saturday and Sunday have no division between morning and afternoon. That should be 12 intervals per week to classify.

Before trying this division, we tried to divide the day into six sections of 4 hours for greater accuracy, but the result was not right in terms of classification. In this case, we had 42 intervals per week.

### Dummies

It's necessary to create dummy variables for `macro_sector`, `province`, and every `DateTime`, interval.

## Choosing the right classifier to solve the problem

Prior to getting the right classifier, we tried the following classifiers in a loop with different hyperparameters and regularizations:

- Logistic Regression
- KNN
- Decision Trees
- Support Vector Machine Classifier

Finally, and after trying Support Vector Machine Classifier also with different hyperparameters, the best configuration for the whole model was the following.

To start, we should say that the probability of getting the right date and time is  $1/12 = 0,0833$

In the strange case that we wouldn't apply any filter of size, macro\_sector or province for example just with SVC we would have 0,1534.

This is 1.84 more chances for that mail to be sent in the right moment and increase the possibilities for opening

But this is not the most common scenario. The typical scenario is at least filter two variables as Province and Macro Sector. We can see the chances of increment six times the probability of sending the campaign at the right moment.