Data Driven Contact Strategy

Deloitte for LUISS

Fabiana Caccavale, Marco Amadori, Lorenzo Meloncelli

MSc. Data Science and Management - A.Y. 2021/2022

Contents

1 Introduction

In the realm of Digital Marketing, personalizing sales and contacts basing on individual customers with Data Driven strategies is becoming increasingly important. It is essential to identify the target to address a specific product/offer through the most suitable channel. Advanced Analytics algorithms can support rules-based tools in making the personalization of marketing/caring campaigns more accurate.

The present work analyzes data of a company that deals with the sale of electricity, gas, and high energy efficiency solution (boilers, air conditioners, photovoltaics).

The goal is to develop, on an annual basis, a monthly contact strategy that maximizes the success of different marketing campaigns by avoiding contacting the customer excessively and distributing the contacts evenly over.

Specifically, there are two campaigns the company proposes: “cross-selling”, to offer a commodity to a customer that has a gas/power contract, and “solution”, to offer a highly energy efficient solution. These, in turn, are promoted through three communication channels: Direct Email Marketing (DEM), SMS and Teleselling (TLS).

Therefore, the monthly contact strategy will take the form of 6 csv files, each for every possible combination of marketing campaign and communication channel. In particular, the files must have one row per customer, and contain the following columns:

- ID: The customer’s ID
- Month_1, Month_2,...Month_12: they will take the value 0 if the customer will not be contacted for the campaign with that channel and 1 otherwise.

To fully achieve the desired result, there were two models to be implemented: a **propensity model**, to estimate the likelihood of customers’ positive responses to a specific campaign, and an **eligibility model**, to analyze the customers’ contactability.

Together these two models would allow to assign to each customer a campaign and a contact channel through which the company can send successful marketing communications.

2 Code

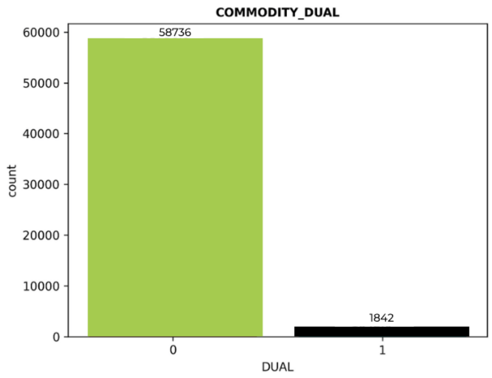
Click here in order to look at the Python code used to conduct the steps explained in the sections below.

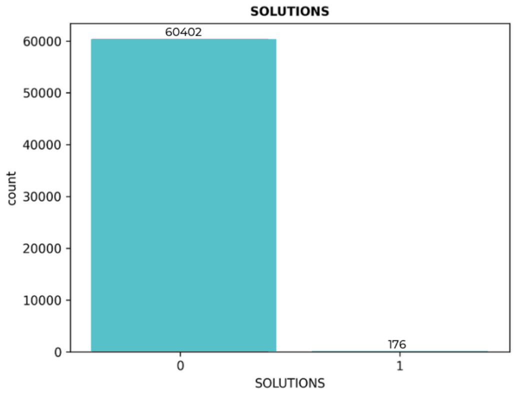
3 Exploratory data analysis

In order to discover patterns, to spot anomalies, to test hypothesis and to check assumptions, it is important to perform an exploratory data analysis with the help of summary statistics and graphical representations.

The dataset consisted of 64 features (columns) divisible into 6 main categories (customer, customer care interactions, behavior, products, churn and campaigns) which together portray the figure of 60578 customers (rows) very well.

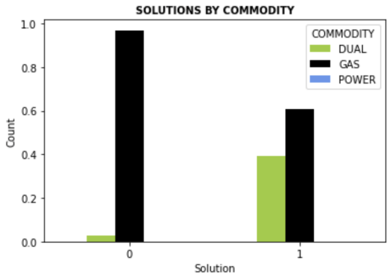
The exploratory data analysis was mainly related to the variables ”COMMODITY_DUAL” and ”SOLUTIONS”, since the assignment to the campaign type depends on them. The main discovery that was made concerns the dataset’s **imbalance**. As it is shown in the two countplots below, both in the case of cross-selling and solution, the majority of people had not yet purchased the contract. On the one hand this can be good, as it means more potential customers to contact. On the other hand, however, this bias can influence many machine learning algorithms, leading some to ignore the minority class entirely.

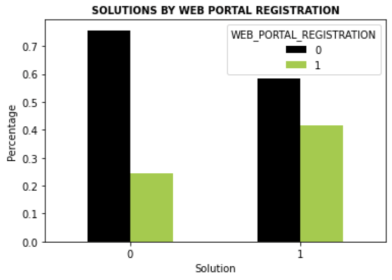


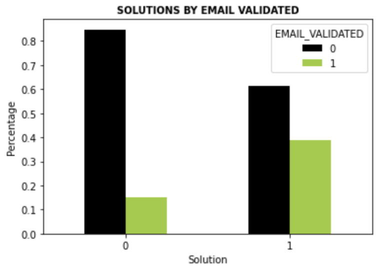


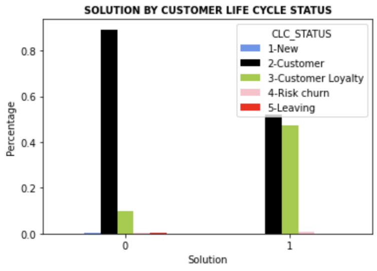
Then, the most descriptive features are plotted in relation to the variables ”COMMODITY_DUAL” and ”SOLUTIONS”. Below are displayed the most relevant plots.

Exploration related to ”SOLUTONS”:

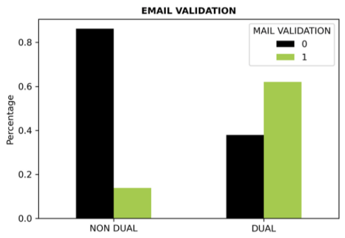


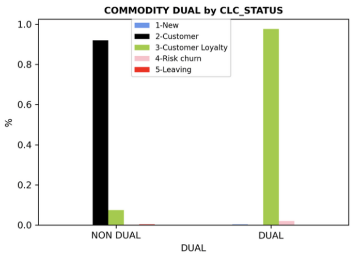


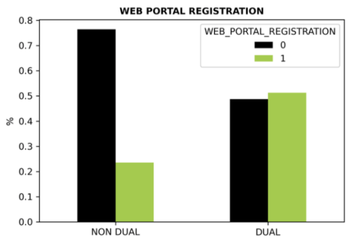




Exploration related to ”COMMODITY_DUAL”:







Afterwards, on the basis of this exploration, two ”persona” were built, the typical SOLUTION and DUAL customers, who exhibit these characteristics:

- **SOLUTION**: customers who have a solution contract are characterized by the fact of being loyal customers, of having validated their mail and registered to the web portal and of having a higher probability of having signed a dual contract.
- **DUAL**: these customers have characteristics similar to solution customers in terms of registration to the web portal, validation of the mail and customer life cycle.

4 Data preparation

Before implementing the models, data had to be manipulated into a form that can readily and accurately be analysed. The first things that had to be addressed were **missing values** and **outliers**. In particular, the rows containing the former were removed since they constituted a very small portion of the dataset and the latter were replaced with the respective median.

Then, to overcome the problem of the **unbalanced dataset** that was previously discussed, random samples were selected from the majority class and deleted from the training dataset (random undersampling) so that the resulting one contains the same number of observations for each class. This strategy may be better suited to datasets where there is a class imbalance, but there are enough examples in the minority class to fit a good model. As a matter of fact, as it will be shown later, the performances of the algorithms when applied in the case of cross-selling are far superior than in the case of solution.

Additionally, **the training dataset was made with all those customers who were not eligible to be contacted** (i.e. that we would never have reached, even if they were inclined, like those at risk of being churned), so as to maximize the number of those potentially contactable. That is because, if data from eligible consumers were used to train the model, they couldn't be used to predict their propensity (since you cannot predict on training data!). As a result, we would have missed the opportunity to contact eventual customers willing to sign the contract.

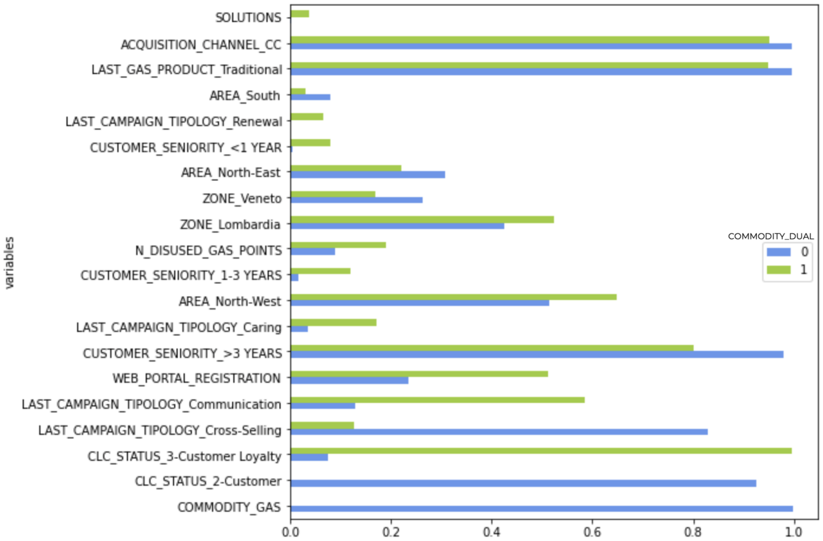
5 Propensity Model

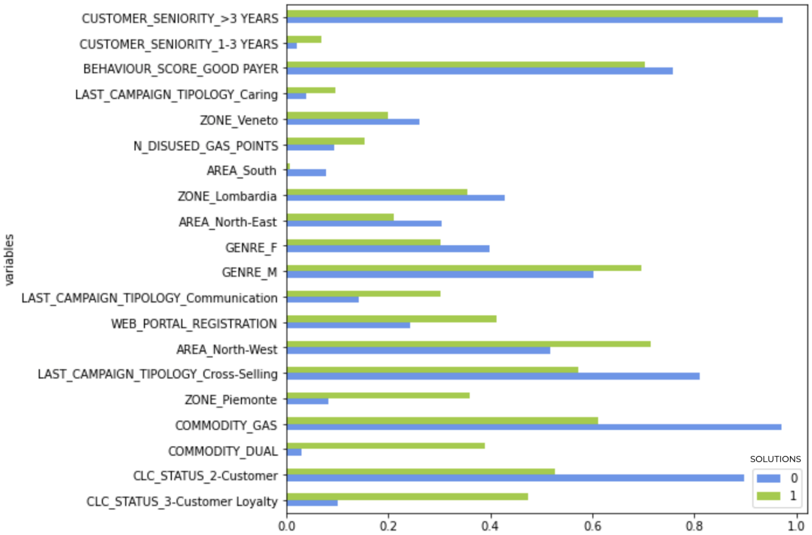
5.1 Feature Selection

In order to develop a model that would have helped to select which customers are most likely to purchase a particular contract, the features that could have most influenced consumers' choice were firstly identified. To do this, a **”mean analysis”** was firstly developed. Both for solution and for cross-selling, the dataset was divided into those who have already purchased the contract and those who have not. Then, for both groups, the average value of each of the features present in the dataset was computed (with the exception of features referring to the customer care interactions category, which were hypothesized not be relevant in this case).

The results were finally sorted in descending order on the basis of the difference between the means of the two groups.

In this way, a ranking of the variables whose values differed most between those who had already purchased the contract and those who have not was obtained. These are displayed below.

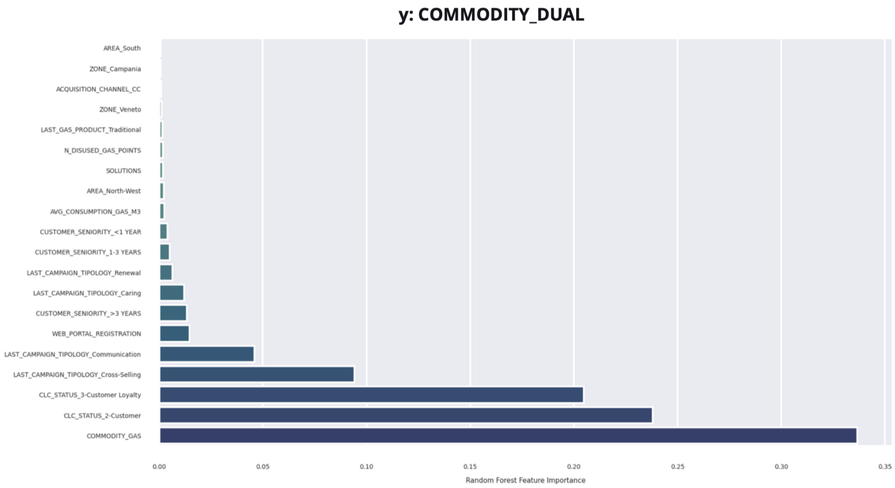


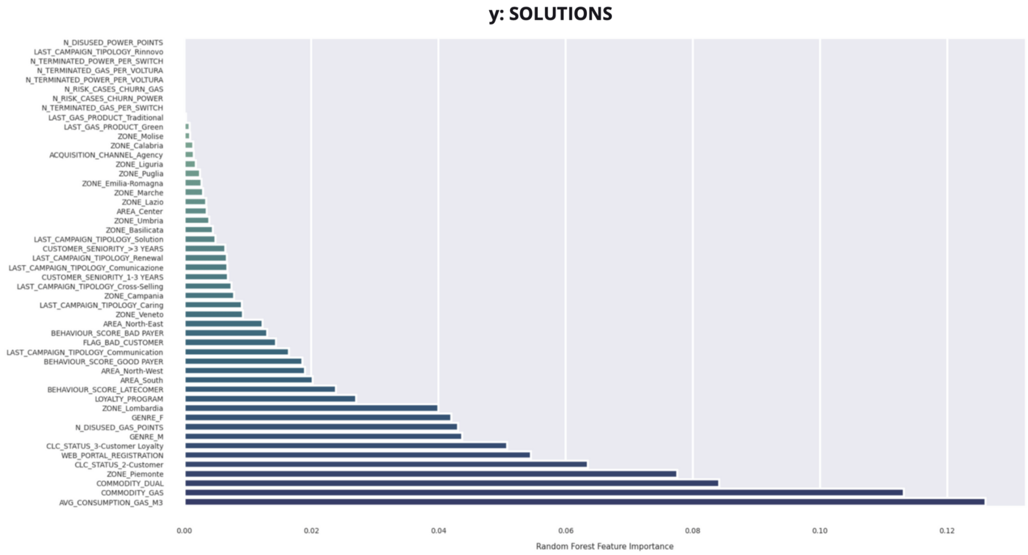


These variables are considered to be the ones that mostly influence customers' propensity.

To get a "second opinion", a **Random Forest** was also implemented by building two models in which the dependent variables are, respectively, "SOLUTIONS" and "COMMODITY_DUAL". Like its name implies, Random Forest consists of a large number of individual decision trees that operate as an ensemble. In general random forests provide a good predictive performance, low overfitting, and easy interpretability. This interpretability is given by the fact that it is straightforward to derive the **importance of each variable** on the tree decision. In other words, it is easy to compute how much each variable is contributing to the decision and, thus, which are the most relevant features that allow to predict whether a customer is a "dual" or "solution" customer. The following barcharts represent the results of the Random Forest feature importance, both for "SOLUTIONS" and for "COMMODITY_DUAL".

Interestingly, as it is shown in the two graphs below representing the features in an ascending order of importance, the results deriving from the "mean analysis" have been almost completely confirmed.





At this point, after removing the variables that lacked an economic meaning together with the strongly correlated ones (e.g. the dummy variables deriving from the same original feature), it was time to actually develop the model.

5.2 Building the model

Once the important features are identified, they are used as predictors in the propensity model. But, **a propensity measure with 0/1 values, depending on whether the customer is inclined to sign a certain type of contract or not, had to be created**.

In order to capture customers' propensity, different models having "SOLUTIONS" and "COMMODITY_DUAL" as dependent variables were trained and tested. Then, the binary outputs of the classification algorithms were transformed into **calibrated probabilities**.

Calibrated probabilities means that the probability reflects the likelihood of true events. In other words, for each customer, their probability of being 1 (the probability of having signed the contract) was computed. By doing this, it was possible to decide the **threshold** within which an observation must belong to a particular class. If a customer exceeds a certain probability threshold of being SOLUTIONS=1 or COMMODITY_DUAL=1, it means that the propensity will be equal to 1. Otherwise, it is 0. All this was done through **Platt Scaling**.

Of course, **the higher the threshold, the lower the number of people that result likely to give a positive response to the campaign**.

The models that were used are the following:

- K-Nearest Neighbors
- Random Forest
- Support-Vector Machine
- Logistic Regression
- Ensemble method

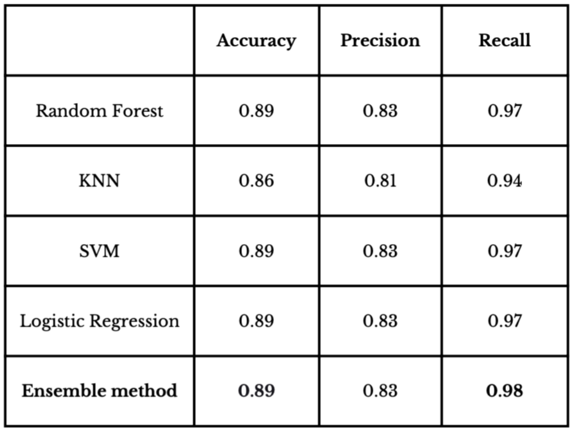
Before actually being trained, the algorithms went through a tuning phase in order to maximize their performance without overfitting or creating too high of a variance. To do this, **Grid Search** was used. It is a method that consists of manually defining a subset of the hyperparametric space and exhausting all combinations of the specified hyperparameter subsets. Each combination's performance is then evaluated using cross-validation and the best performing hyperparametric combination is chosen.

5.3 Results

Each of the models listed above were evaluated with three performance metrics: accuracy, precision and recall.

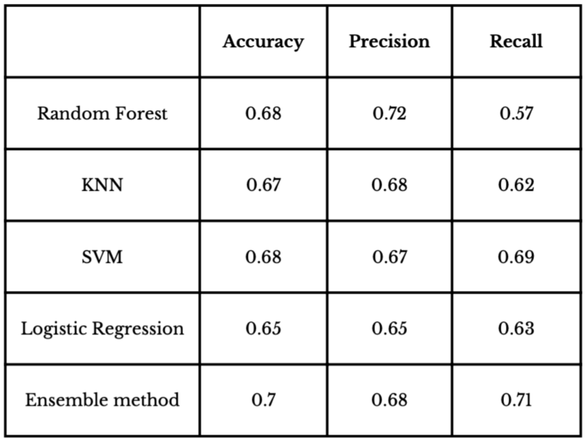
Among these metrics, **recall** is the one that was mainly considered because it is the measure that has the most importance when the cost of a false negative is high. This is the case, because the cost of NOT contacting a customer who is likely to give a positive response to the campaign is higher than the cost of contacting someone who is not willing to sign the contract.

As it is shown in the table below, the models' performance metrics are quite high when predicting the propensity for cross selling campaigns. The best model is the ensemble, with recall = 98%, accuracy = 89% and precision = 83%.



	Accuracy	**Precision**	**Recall**
Random Forest	0.89	0.83	0.97
KNN	0.86	0.81	0.94
SVM	0.89	0.83	0.97
Logistic Regression	0.89	0.83	0.97
Ensemble method	**0.89**	0.83	**0.98**

For solutions, instead, the models' performance metrics are a little lower, but this is mainly due to the fact the amount of data available was very low since the training dataset was balanced and, in the case of solutions, customers who had signed the contract were only 176. Here, again, the best model is the ensemble with recall = 71%, accuracy = 70% and precision = 68%.



	Accuracy	Precision	Recall
Random Forest	0.68	0.72	0.57
KNN	0.67	0.68	0.62
SVM	0.68	0.67	0.69
Logistic Regression	0.65	0.65	0.63
Ensemble method	0.7	0.68	0.71

6 Eligibility Model

Identifying the customers most likely to buy a contract is not enough. To design an effective contact strategy, only eligible customers must be contacted.

After having defined who are the customers that are willing to give a positive response to the campaign thanks to the best performing propensity model, **only the ones who have a propensity value which exceeds the threshold are feed into the eligibility model**. That is because contacting customers has a cost. Therefore, if the company contacts those who do not exceed a certain propensity value, it would probably waste money since these customers will probably reject the proposal.

Identifying eligible customers was possible by putting in code form a whole series of rules and indications provided by the company. The first step was to filter the consumers chosen by the propensity model on the basis of the so-called **"general rule"**: those customers whose last contact dates back to at least N months ago are eligible, where N = # months / # contacts per year. It should be emphasized that n can vary according to both the marketing campaign and the communication channel. This differentiation allowed to create the six csv files mentioned above.

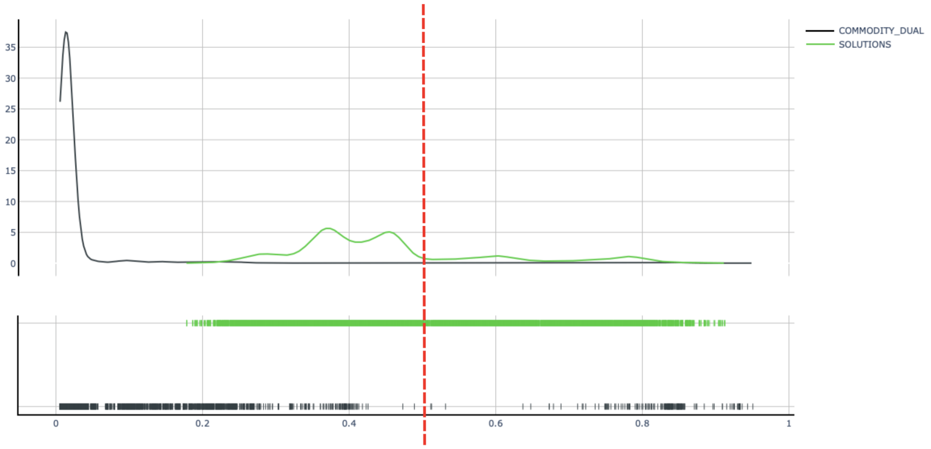
It was finally possible to plan the monthly contact strategy. The criterion used to do this was to **maximize the number of contacts, while remaining within the limits imposed by both the campaign rules and the cross-campaign rules**.

Afterwards, whether the consumer was contactable for both campaigns (solution and cross-selling), had validated his/her e-mail and his/her phone number had to be taken into consideration. Not doing this would have drastically reduced the number of possible interactions. For instance, one of the campaign rule was: "No overlaps between DEM and TLS in the same month for the same type of campaign". If the possibility that the consumer might not be "phone validated", the number of times the consumer could be contacted by e-mail or SMS would have been significantly reduced in vain. The same if the consumer had been contactable for only one of the two marketing campaigns. In this case, in fact, the limits deriving from the cross- campaign rules could be exceeded.

7 Conclusion

As seen through the implementation of the project, the model that was built and the consequent strategy allow to be quite **flexible in the number of people to contact** by moving the threshold after which a customer is predicted to give a positive response to the campaign.

Below, the probability distribution of the two models is represented. The model for the cross-selling campaign, having a high recall and accuracy, is very much certain, and the probabilities are mainly distributed at the edges. Instead, the probability distribution of the model for the solution campaign is more concentrated in the middle.



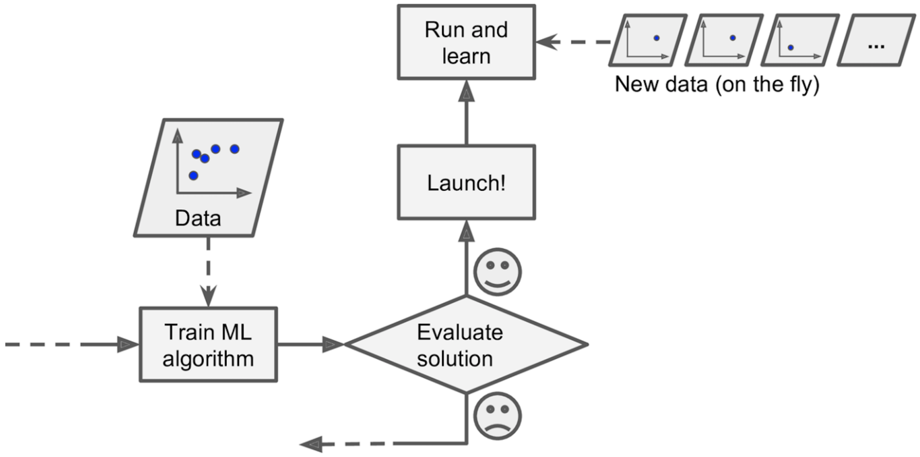
By setting the threshold at 0.5, it is possible to have a balanced strategy. If the threshold is moved to 0.6, the number of people to be contacted decreases and only customers with a higher propensity are considered.

The strategy followed throughout the project is **efficiency-oriented** because only customers with a high propensity are contacted. This strategy would not be linked to a predefined budget or number of people to be contacted, but it would depend on the number of people with a probability higher than the threshold.

By taking into consideration the unitary cost of of the type of contact channel, the number of people that would be contacted is approximately 5000€ (threshold 0.5).

On the other side, if the company had to not exceed a fixed budget or a number of people to be contacted, a good strategy could be to **sort the probabilities** related to the eligible customers so that only inclined customers are contacted until the budget is exhausted or the number of people to be contacted is reached.

Concerning future developments, the model after a certain period of time has to be retrained. This process can be automated by adopting an online way of learning, by continuously adjourning the model on the fly when new information enters the loop, feeding the model.



8 References

”Using Platt Scaling and Isotonic Regression to Minimize LogLoss Error in R”, Analytics Vidhya

”How to Calibrate Probabilities for Imbalanced Classification”, Machine Learning Mastery

“Undersampling and oversampling: An old and a new approach”, Analytics Vidhya

”sklearn.ensemble.VotingClassifier Documentation”, scikit-learn