DATA DRIVEN CONTACT STRATEGY

# Michele Cibelli, Giacomo Flores, Khan Gony Jan

# Introduction

The scope of the project is to develop, on an annual basis, a monthly contact strategy for a company that deals with the sale of electricity, gas and high energy efficiency solution.

The goal of this strategy is to maximize the success of the campaigns by avoiding contacting the customer excessively and distributing the contacts evenly over both from the point of view of time, and from the point of view of the communication channels used (DEM, SMS, TLS).

The types of campaign carried out by the company are two: Cross-selling and Solution; the first one aim at offering to already existing clients a commodity in addition to the one they already have (gas/power) while the second one has the object of proposing to the client a new highly efficient energy solution.

# Methods and Experimental Design

The first task of our project was the data understanding, so we did an exploratory data analysis in order to understand all the different fields of the dataset and above all the relationship between them, to help us during this task we performed some data visualization to examine the correlation between the columns and the frequency of the most important observations.

Immagine che contiene testo, orologio

Descrizione generata automaticamente

At this point we had to clean the data, firstly dropping all the NaN cells and then conducting the contractability analysis, so removing from the dataset all the clients without the privacy consent, leaving or at risk of leaving.

We then needed to identify which communication channel was preferred by each customer, but, as our opinion was that none of the columns provided were suitable for this analysis, we decided to analyze the data and label manually a good percentage of the observations.

To do so we decided to consider specific fields in order to speed up the process and make it more precise and consistent, those that according to our opinion and to the analysis made in precedence were the most related to the propensity to one or another communication channel were: the costumer behavior, the costumer life cycle, the number and the type of campaigns received, the acquisition channel initially used to activate the service and number of customer care interactions that the costumer had in the last month, three months and year.

We also decided to create some new columns to investigate the relationship between the number of campaigns received and a multitude of other fields such as: the communications opened, the solution bought, the customer seniority, and the commodity already owned.

At this point we were ready to start out propensity prediction, firstly performing feature scaling to normalize the range of the features and then feeding the data to various classification algorithms such as Support Vector Machine, Random Forest, Logistic Regression and knearest neighbors Classifier to predict the propensity of the customers to the different channels. To avoid bias we performed the prediction with different percentage ratios of labeled and unlabeled data, and we noticed that after passing the 25% of labeled observations, the results of the various algorithms started to be very similar, and at some points identical, with also a very low variation between the results found with the different labelled data percentages ratios.

Immagine che contiene testo

Descrizione generata automaticamente

Being the SVM the model that, at least theoretically, should provide the most accurate prediction we decided to use its result to assign to every costumer it preferred communication channel.

The fields used for this analysis are multiple and in particular:

* LOYALTY\_PROGRAM: if the customer has joined the loyalty program
* LAST\_MONTH\_DESK\_VISITS: number of visits at help desk in the last month
* LAST\_3MONTHS\_DESK\_VISITS: number of visits at help desk in the last 3 months
* LAST\_YEAR\_DESK\_VISITS: number of visits at help desk in the last year
* LAST\_MONTH\_CC\_REQUESTS: number of requests at contact center in the last month
* LAST\_3MONTHS\_CC\_REQUESTS: number of requests at contact center in the last 3 month
* LAST\_YEAR\_CC\_REQUESTS: number of requests at contact center in the last year
* SOLUTIONS: if the customer bought a solution
* INBOUND\_CONTACTS\_LAST\_MONTH: Number inbound customer contact in the last month
* INBOUND\_CONTACTS\_LAST\_2MONTHS: Number inbound customer contact in the last 2 months
* INBOUND\_CONTACTS\_LAST\_YEAR: Number inbound customer contact in the last year
* WEB\_PORTAL\_REGISTRATION: if the customer is registered on the web portal
* N\_CAMPAIGN\_SENT: Number of campaigns sent
* N\_CAMPAIGN\_CLICKED: Number of campaigns clicked
* N\_CAMPAIGN\_OPENED: Number of campaigns opened

Lastly, we had to implement our contact strategy based on the propensity model and on the rules of the game:

Starting for the costumers with propensity to messages they had to receive the minimum number of phone contact, 1 for the solution campaign on the 6th month and , if possible, 1 for the cross selling campaign on the 12th month, filling all the other months with messages and email balancing one month for solution and the following months with cross selling

On the other hand the costumers with phone propensity had to receive the maximum number of phone contact, 2 for the solution campaign on the 1st and 7th month and , if possible, 2 for the cross selling campaign on the 2nd and 8th month, filling, as for the previous one, all the other months with messages and email balancing one month for solution and the following months with cross selling

Lastly we had to create a strategy for the less than hundred customer without the phone number validated and, being the email our only possibility, we decided to send one campaign per month balancing, as before, between solution and cross selling in order to send them an acceptable number of communications.

# Code Description

We started our code by importing all the libraries we needed for the task:

* Pandas: Data analysis and manipulation tool
* Matplotlib.pyplot : Collection of functions that make matplotlib work like MATLAB
* Seaborn: Data visualization library based on matplotlib
* Numpy: Wide variety of mathematical operations on arrays
* Sklearn: Tools for predictive data analysis

We then uploaded our dataset and dropped all the client without the privacy consent, leaving or at risk of leaving.

At this point we created all the dataset we needed for our modelling and analysis, firstly individuating all the customers with the phone number validated, then the ones with the email validated e lastly the customers that purchased only one of the two services (power and gas) and so suitable for the cross-selling campaign.

We then selected the columns that we needed to run the algorithms (as said in the previous section) and added the labeled data to the train dataset dropping also al the NaN cells.

Next, we performed the future scaling, trained the four models previously picked and printed the number of customers that each model has identified as more inclined to TLS. To do so firstly we imported the models, then we created the classifiers associating to the SVM a linear kernel, to the KNN 2 as the number of neighbors and to the Random Forest 100 as the number of estimators.

After getting our prediction we merged them with our dataset to assign to every costumer the right propension.

We then created our dummy datasets for the contact strategy with the IDs of the customers as index and a column for every month, we then merged again the predictions column (and for the cross-selling also the ‘commodity’ column) with this df and filled all the empty cells with zeros. We did the same also for the dataset of the customers with only the email validated with the only change being the column ‘commodity’ instead of the predictions.

At this point we started to compile the six datasets using for and while loops and a multitude of if sentences following the “rules of the game”.

The two SMS codes are both formed by 4 while loops with the only differences of the columns on which they make modifications and the presence of the column commodity in the cross-selling database.

Immagine che contiene tavolo

Descrizione generata automaticamente

Also the DEM codes are very similar being both their outputs mostly based on the SMS databases with the addition of the customers without the phone number validated

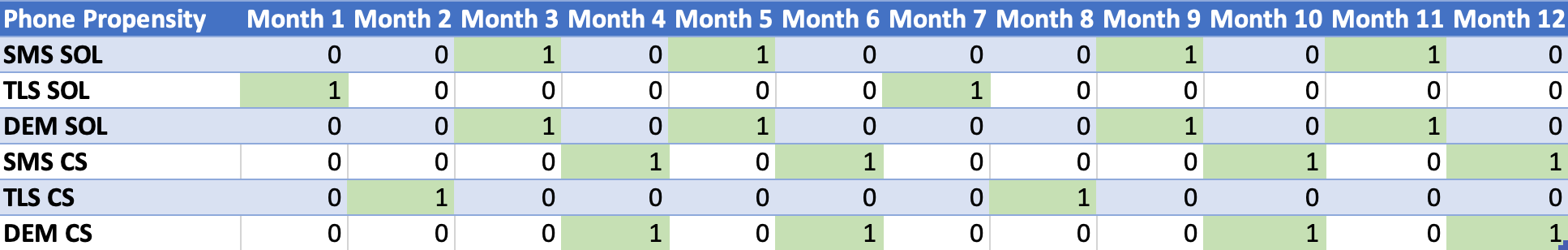
Immagine che contiene tavolo

Descrizione generata automaticamente

Immagine che contiene tavolo

Descrizione generata automaticamente

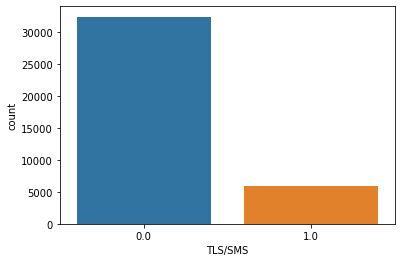
Lastly we have the TLS codes which just append the values in the right cells using a simple for loop



Finally the six databases are named in the proper way, cleaned of all the non-useful columns and downloaded as .csv files.

Results and conclusion

From our propensity prediction we reached the conclusion that just under 14% of the customers in our database prefer marketing campaigns via telephone, this is certainly not a very high percentage, but we think it may represent reality, since, in our experience, people tend to feel less disturbed by a written message that they can read at any time they want, than by a call that might disturb them at an inappropriate time of day, making them less likely to accept the proposed offer.



The implementation of our model in the existing process should be relatively easy and with new data flowing into the database new prediction could be made, computing again the propensity, checking its accuracy looking at the results of the campaigns and adapting the strategies to match better the behavior of every customer.

Finally in our opinion the model could be improved through Implementation of a feedback system to have more precise data and to avoid computing manually the propensity of the training dataset being sure of avoiding biased analysis.