

Response Model: A Real World Binary Classification Problem

Course Project of Machine Learning (NOVA IMS)

Cordially sponsored by prof. A. P. Melo (amelo@novaims.unl.pt)

Illya Bakurov

email ibakurov@novaims.unl.pt

Leonardo Vanneschi

email lvanneschi@novaims.unl.pt

28/03/2019

Abstract

“A response model can provide a significant boost to the efficiency of a marketing campaign by increasing responses or reducing expenses. The objective is to predict who will respond to an offer for a product or service” [1].

Introduction

The Company

Consider a well-established company operating in the retail food sector. Presently they have around several hundred thousands of registered customers and serve almost one million consumers a year. They sell products from 5 major categories: wines, rare meat products, exotic fruits, specially prepared fish and sweet products. These can further be divided into *gold* and *regular* products. The customers can order and acquire products through 3 sales channels: physical stores, catalogs and company’s website. Globally, the company had solid revenues and a healthy bottom line in the past 3 years, but the profit growth perspectives for the next 3 years are not promising... For this reason, several strategic initiatives are being considered to invert this situation. One is to improve the performance of marketing activities, with a special focus on marketing campaigns.

The Marketing Department

The marketing department was pressured to spend its annual budget more wisely. The CMO perceives the importance of having a more quantitative approach when taking decisions, reason why a small team of data scientists was hired with a clear objective in mind: to build a predictive model which will support direct marketing initiatives. Desirably, the success of these activities will prove the value of the approach and convince the more skeptical within the company.

The Objective

The objective of the team is to build a predictive model that will produce the highest profit for the next direct marketing campaign, scheduled for the next month. The new campaign, sixth, aims at selling a new gadget to the Customer Database. To build the model, a pilot campaign involving 2.240 customers was carried out. The customers were selected at random and contacted by phone regarding the acquisition of the gadget. During the following months, customers who bought the offer were properly labeled. The total cost of the sample campaign was 6.720MU and the revenue generated by the customers who accepted the offer was 3.674MU. Globally the campaign had a profit of -3.046MU. The success rate of the campaign was 15%. The objective of the team is to develop a model that predicts customer behavior and to apply it to the rest of the customer base. Hopefully the model will allow the company to cherry pick the customers that are most likely to purchase the offer while leaving out the non-respondents, making the next campaign highly profitable.

Moreover, other than maximizing the profit of the campaign, the CMO is interested in understanding to study the characteristic features of those customers who are willing to buy the gadget.

The Data

The data set contains socio-demographic and firmographic features about 2.240 customers who were contacted. Additionally, it contains a flag for those customers who responded the campaign (1), by buying the product. Table 1 describes the data set.

Objectives and Evaluation Rules

The main objective is to train a predictive model which allows the company to maximize the profit of the next marketing campaign and understand the characteristic features of respondents.

Objectives

Taking into account the Data Mining (DM) process and Machine Learning (ML) algorithms you have studied throughout this course, consider the following list of main steps when solving the project:

1. explore the data you are provided - be creative and pay attention to the details;
2. apply and compare different data preprocessing techniques;
3. apply and compare different feature engineering techniques;
4. apply and compare different ML techniques;
5. summarize the results of the benchmarks across different seeds (at least 5).

Feature	Description
AcceptedCmp1	1 if costumer accepted the offer in the 1 st campaign, 0 otherwise
AcceptedCmp2	1 if costumer accepted the offer in the 2 nd campaign, 0 otherwise
AcceptedCmp3	1 if costumer accepted the offer in the 3 rd campaign, 0 otherwise
AcceptedCmp4	1 if costumer accepted the offer in the 4 th campaign, 0 otherwise
AcceptedCmp5	1 if costumer accepted the offer in the 5 th campaign, 0 otherwise
Response (target)	1 if costumer accepted the offer in the last campaign, 0 otherwise
Complain	1 if costumer complained in the last 2 years
DtCustomer	date of customer's enrollment with the company
Education	customer's level of education
Marital	customer's marital status
Kidhome	number of small children in customer's household
Teenhome	number of teenagers in customer's household
Income	customer's yearly household income
MntFishProducts	amount spent on fish products in the last 2 years
MntMeatProducts	amount spent on meat products in the last 2 years
MntFruits	amount spent on fruits in the last 2 years
MntSweetProducts	amount spent on sweet products in the last 2 years
MntWines	amount spent on wines in the last 2 years
MntGoldProds	amount spent on <i>gold</i> products in the last 2 years
NumDealsPurchases	number of purchases made with discount
NumCatalogPurchases	number of purchases made using catalogue
NumStorePurchases	number of purchases made directly in stores
NumWebPurchases	number of purchases made through company's web site
NumWebVisitsMonth	number of visits to company's web site in the last month
Recency	number of days since the last purchase

Table 1: Meta-data table

More than blindly following the main steps of a DM process and applying the techniques you were taught, you should always justify why did you consider important to include a given technique into your DM process (a-priori considerations) and comment the results you have obtained, taking in mind your expectations (a-posteriori considerations). Besides considering taught techniques, you should prioritize exploration and usage of other, previously unseen, technique(s). In such a case, it is mandatory to include a clear description of the technique(s) with a proper bibliographic support.

Evaluation Rules

The evaluation of the student i of group g will be based on the following formula:

$$grade_i = .1 * report_g + .1 * presentation_i + .35 * profit_g + .45 * techniques_g,$$

where *techniques* stands for the diversity, complexity and originality of applied techniques and *profit* stands for the similarity, in terms of campaign's real profit (calculated on unseen data), towards the best group in the course. It is worth to notice that the quality of the source code will not be evaluated.

Submissions and deadlines

1. submit until 23:59 of 28/04/2018:

- (a) the source code (*.zip*) containing the *jupyter* notebooks (*.ipynb*) with the exploratory analysis and the DM pipeline (*the project*);
- (b) the predicted classes for Customer Database (*.xlsx*), which will be provided soon;
- (c) the report (*.pdf*);

2. submit until 23:59 of 02/05/2018: the presentation document (*.pdf/.pptx/etc.*), elaborated according to a predefined structure (to be provided);

3. throughout 03/05/2018: oral presentations (to be scheduled).

Follow [this](#) link to register your group (sheet *groups*). The schedule of oral presentations will be provided in the same document (sheet *presentationschedule*).

References

- [1] O. Parr-Rud. *Business Analytics Using SAS Enterprise Guide and SAS Enterprise Miner*. SAS Institute, 2014.