Applying Learning Machine to predict results of a banking marketing campaign and find trends and patterns for future campaigns

José Manuel Castaño

Abstract

We analyze the data related with a direct marketing campaign of a Portuguese bank, based on phone calls, from May 2008 to November 2010. The campaign objective was that the client subscribes a term deposit (deposit that a bank or a financial institution offers with a fixed rate, in which your money will be returned back at a specific maturity time). We initially performed an exploratory analysis of the data; we evaluate the incidence of the different variables in the target success and we analyze trends and behavior patterns of customers that will allow us to segment customers and provide recommendations to improve the results of future marketing campaigns. Finally, we create a classification model based on the Gradient Boosting Classifier to predict the result of the campaign, which achieves an accuracy of 0.9064.

Keywords: bank marketing, supervised learning, Gradient Boosting Classifier,

I.- INTRODUCTION

Many companies carry out telephone marketing campaigns among their clients to get new subscriptions, extend them or launch new products.

Prior to the start of the campaign, both to get the largest number of subscriptions, and to optimize the resources used in the campaign[1], it would be very convenient:

- Know the trends and patterns of customer behavior
- Segment customers based on probability of success
- Determine the type of campaign, frequency, intensity, when and how it is carried out, etc, to achieve better results.

Likewise, for the bank it's very important to predict a right number of subscriptions[2], since such prediction would allow:

- To know the total amount they will obtain and they can use for investments
- Assess whether the cost-benefit ratio compensates
- Decide on the resources and investments appropriate to the campaign.

II.- STATE OF ART

There are 5 authors who have carried out works with the dataset object of this study. As a summary, we will highlight that they have scaled using MinMaxScaler or StandardScaler, they have coded the categorical variables using numbers (ordinal), they have balanced the dataset using SMOTE and the models that have obtained the best results are RandomForest and DecisionTree.

The best of the predictions presents an accuracy of 0.89 with RandomForest and the number of False Positives is in all cases greater than 1.6%.

III.- DATASET

The dataset comes from a direct marketing campaign of a Portuguese bank, based on phone calls, from May 2008 to November 2010. It is provided by kaggel: https://www.kaggle.com/datasets/hariharanpavan/bank-marketing-dataset-analysis-classification?resource=download.

The dataset has 45,211 instances, 16 attributes and an output target (yes/no). There are 7 features numerical and 9 features categorical (3 of which are binary). Qualitatively, we would group the variables according to:

- Client-specific, such as age, marital status, education and job
- Economic, as balance, personal loan, housing loan and credit in default
- Campaign-specific, as communication type, last contact month, last contact day, last contact duration, total contacts, days from last campaign, contacts previous and outcome of the previous campaign.

It's a highly imbalanced dataset, target has 11.7% of positives and 88.3% of negatives.

IV.- METHODS

A. Data preparation

We initially perform an exploratory analysis of the data. We observe that there are no null values or duplicate values. We study the distributions of the variables: we analyze the symmetries, the tails and the outliers in terms of quantity and distance from the mode. We also study the correlation between the variables and with the target.

We transform binary variables into (0/1) to improve processing. We scale the numeric variables using RobustScaler because features have many outliers and far from the mode. Finally, we convert the categorical variables to dummy variables.

B. Building a prediction model

We divide the dataset into a train set (70%) and a test set (30%).

Since the target is highly unbalanced (0.12/0.88), from the train set we create another more balanced dataset using the undersampling technique (0.30/0.70), in order to subsequently select the dataset that obtains the best results[3].

We are interested in predicting the greatest number of successes and with the least number of errors in the positive predictions, because it generates many problems to predict income that will not be produced later, and therefore, investments cannot be made.
So, the metrics used to evaluate the model are:

- Accuracy: percentage of correct predictions over the total sample
- Few false positive: few positive predictions corresponding to negative cases

We tested 7 supervised classification learning models with the standard parameters. To evaluate and select a model, we use the accuracy metric together with the number of false positives (we intend low values). From the analysis, we conclude:

- Imbalanced dataset has better results in all metrics comparing with balanced dataset
- The best models are **Random Forest**[4] and **GradientBoostingClassifier**[5]

We apply Cross Validation[6] on the models to ensure that overfitting is not occurring and finally we adjust the hyperparameters[7] to maximize accuracy.

C. Trends and patterns

To find out the trends and behavior patterns, we analyze the distribution of the variables and then calculate the percentage of target for each of the components of the variables. In the case of numerical variables, we previously create bins and calculate the target percentage for each bin.

V. RESULTS

A. Trends and patterns

We can segment customers into 3 groups:

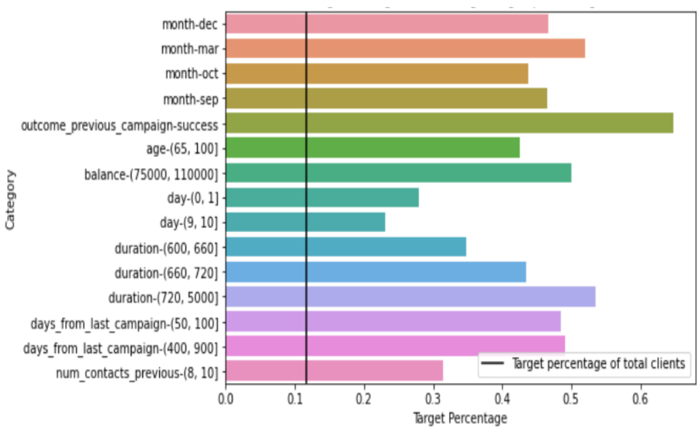
- **Clients most sensitive to the marketing campaign**, with the highest subscription ratios. They are characterized by having subscribed a product in the previous campaign, being over 65 years old, having a balance over 75,000 Eu or being students.
- **Clients less sensitive to the marketing campaign**, with the lowest subscription ratios. They are clients with a credit in default, clients with negative balance, clients with a personal loan or with a blue-collar job.
- **Rest of clients**. They have a moderate sensitivity to marketing campaigns, with results around the average. It would depend on the type of campaign, intensity, when and how it is carried out, etc. to achieve better results.

The following recommendations for the design and execution of the campaign can significantly increase the subscription ratio:

- Campaign frequency. There are 2 optimal windows to carry out campaigns: between 50 and 100 days or between 400 and 900 days
- The best months to carry out campaigns are March, September, October and December

- The best days of the month are the 1st and 10th
- Maintain contact with the client between the different campaigns. The optimal number of contacts is between 8 and 10
- Try to make the duration of the call as long as possible

Fig.1. Categories with high target percentage



B. Prediction model

Initially we tested 7 models of classification supervised learning with the standard parameters and we obtain the following metrics and confusion matrix.

Table I. Testing models metrics

	Accuracy	Precision	Recall	F1	Roc_Auc
Logistic Regression	0.8991	0.3498	0.6277	0.4493	0.7729
Decision Tree	0.8746	0.5022	0.4690	0.4850	0.7010
Random Forest	0.9028	0.3893	0.6429	0.4850	0.7828
AdaBoostClassifier	0.8991	0.3912	0.6106	0.4769	0.7666
GradientBoostingClassifier	0.9025	0.4201	0.6279	0.5034	0.7770
K-Nearest Neighbor	0.8919	0.3442	0.5666	0.4282	0.7418
Balanced Bagging Classifier	0.8548	0.8639	0.4401	0.5832	0.7097

Tabla II. Testing models confusión matrix

	TP	TN	FP	FN
Logistic Regression	558	11638	331	1037
Decision Tree	801	11062	907	794
Random Forest	621	11624	345	974
AdaBoostClassifier	624	11571	398	971
GradientBoostingClassifier	670	11572	397	925
K-Nearest Neighbor	549	11549	420	1046
Balanced Bagging Classifier	1378	10216	1753	217

TP*: True positive, predict positive result as positive*
TN*: True negative, predict negative result as negative*
FP*: False positive, predict negative result as positive*
FN*: False negative, predict positive result as negative*

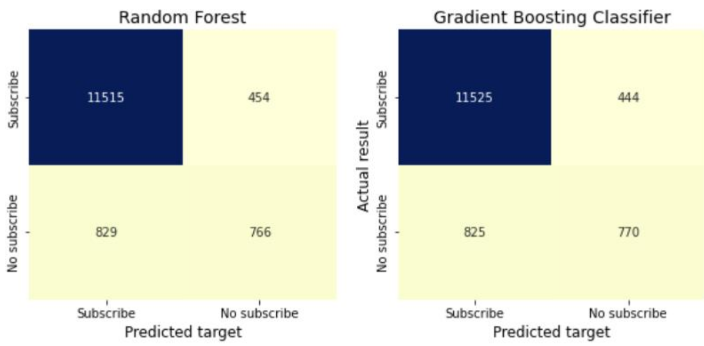
The 2 models that obtain the best metrics are Random Forest and Gradient Boosting Classifier. We adjust the hyperparameters to optimize the accuracy and we obtain the following results:

Table III. Metrics after hyperparameter tuning

	Accuracy	Precision	Recall	F1	Roc_Auc
Random Forest	0.9028	0.3893	0.6429	0.4850	0.7828
GradientBoostingClassifier	0.9025	0.4201	0.6279	0.5034	0.7770
Random Forest Optimized	0.9045	0.4740	0.6232	0.5385	0.7777
GradientBoostingClassifier Optimized	0.9064	0.4828	0.6343	0.5482	0.7837

Table IV. Confusion matrix after hyperparameter tuning

	TP	TN	FP	FN
Random Forest	621	11624	345	974
GradientBoostingClassifier	670	11572	397	925
Random Forest Optimized	756	11512	457	839
GradientBoostingClassifier Optimized	770	11525	444	825



Gradient Bosting Classifier Optimized obtain the best results in all metrics and also get more True Positive and less False Positive than Random Forest Optimized. Therefore, the model finally chosen is **Gradient Boosting Classifier with an accuracy of 0.9064 and 1.4% of false positives**.

VI. CONCLUSIONS

Conducting telephone marketing campaigns among a company's clients is an excellent way to increase sales or launch new products. But they have the disadvantage of being expensive, they can often be annoying for customers and we also have no predictions of their outcome.

In this project we have managed to segment customers based on the probability of success in the campaign, while at the same, time we have established some recommendations that would improve the subscription ratio.

In addition, we have achieved a prediction of the campaign result with an accuracy greater than 0.906 and with a moderate level of false positives (1.4%).

This will allow, among other improvements, to optimize resources, select the target clients for the campaign, carry out more personalized campaigns and finally, predict the capital of the deposits that they will obtain with the campaign, that would help design and select future investments.

REFERENCES

[1] Hentzen, J.K., Hoffmann, A., Dolan, R. and Pala, E. (2021), "Artificial intelligence in customer-facing financial services: a systematic literature review and agenda for future research", International Journal of Bank Marketing, Vol. 40 No. 6, pp. 1299-1336, doi: 10.1108/IJBM-09-2021-0417.

[2] Mogaji, E. and Nguyen, N.P. (2021), "Managers' understanding of artificial intelligence in relation to marketing financial services: insights from a cross-country study", International Journal of Bank Marketing, Vol. 40 No. 6, pp. 1272-1298, doi: 10.1108/IJBM-09-2021-0440

[3] Jake Vanderplas. "Python Data Science Handbook: Essential Tools for Working with Data". O'Reilly Media, 2016.

[4] Joaquín Amat Rodrigo. "Random Forest con Python". Octubre 2020. Available online: https://www.cienciadedatos.net/documentos/py08_random_forest_python.html.

[5] Joaquín Amat Rodrigo. "Gradient Boosting con Python". Octubre 2020. Available online: https://www.cienciadedatos.net/documentos/py09_gradient_boosting_python.html

[6] Mohammed Alhamid. "What is Cross Validation?". December 24.2020. Available online: https://towardsdatascience.com/what-is-cross-validation-60c01f9d9e75

[7] Jeremy Jordan. "Hyperparameter tuning for machine learning models". November 2017. Available online: https://www.jeremyjordan.me/hyperparameter-tuning/