A Probabilistic Programming Approach to Predict The Success of Bank Telemarketing

Fatma Oztel (fatma.oztel@metu.edu.tr)
Graduate School of Informatics, Data Informatics

Abstract

This paper proposes a probabilistic programming approach to predict the success rate of telemarketing calls for long-term bank deposits. Telemarketing is one of the most common interactive techniques of direct marketing, widely used by financial institutions such as banks to sell long-term deposits. In this study, the Generalized Linear Models and Multilevel Models approaches were used to create a scientific model and to analyze potential client who subscribes to a long-term deposit, which is considered an output variable. The causal relationship between marketing attributes and outcomes was created and interpreted according to the model results.

Keywords: Bayesian; GLM; Multilevel Model; Telemarketing; DAG; Causality

1. Introduction

Marketing is the process of getting people interested in your company's product or service. This happens through market research, analysis, and understanding your ideal customer's interests. There are a lot of sectors and industries using marketing strategies to promote their products and services. The data-driven approaches support the marketing strategies by using meaningful information about the companies and customers. In this project, the success of bank telemarketing will be predicted by using proper statistical approaches. Telemarketing is the most preferable and very common interactive marketing way in the financial area. In this paper, the purpose is the analyzing effect the telemarketing actions and attributes on the customer subscription of the long-term deposit, which is considered an output variable.

The 'bank-additional' dataset of the "Bank Marketing Dataset" was used to analyze the effect of different variables in the marketing area. The bank dataset includes 10 numeric attributes and 11 categorical attributes. One of the categorical attributes is the target.

In previous studies, the data-driven models have been explored for modeling bank telemarketing success. The important attributes of the dataset are interpreted according to the model result and analysis. In this study, the probabilistic programming models are used by using the Python Pymc3 library to analyze the factors which have an effect on a potential client's subscription to a long-term deposit. The models help to discriminate which factors are important to predict the acceptance of the offer which is considered an output variable. The model results were interpreted and important features were detected.

In this project, the main area is interpreting the effect of the telemarketing actions on the target. The dataset has also the other types of attributes which represents the client info(age, jobs, etc.) and Social and economic(consumer price index, number of employees, etc.) values. They were not considered in this project.

2. Materials and Methods

2.1. Bank Telemarketing Data

This project focus on the telemarketing actions to sell long-term deposits. This dataset is publicly available for research. The details of the dataset are described in this paper (Moro et. al., 2014).

The 'bank-additional' dataset of "Bank Marketing Dataset" was used to analyze the effect of different variables in the selling area. The bank dataset includes 10 numeric attributes and 11 categorical attributes. One of the categorical attributes is the target. The dataset includes 4119 non-null observations. Each record includes the output target, the contact outcome ({"failure", "success"}), candidate input features, and several features about the campaign and social-economic status. There are several missing values in some categorical attributes, all coded with the "unknown" label. Input variables are;

Bank client data:

1. age (AG) (numeric)
2. job (JB) : type of job (categorical: "admin.", "blue-collar", "entrepreneur", "housemaid", "management", "retired", "self-employed", "services", "student", "technician", "unemployed", "unknown")
3. marital (MS) : marital status (categorical: "divorced", "married", "single", "unknown")
4. education (ED) : (categorical: "basic.4y", "basic.6y", "basic.9y", "high.school", "illiterate", "professional.course", "university.degree", "unknown")
5. default (DF) : has credit in default? (categorical: "no", "yes", "unknown")
6. housing (HL) : has housing loan? (categorical: "no", "yes", "unknown")
7. loan (LO) : has personal loan? (categorical: "no", "yes", "unknown")

Related to the last contact of the current campaign:

8. contact (CN) : contact communication type (categorical: "cellular", "telephone")
9. month (MO): last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
10. day_of_week (DW) : last contact day of the week (categorical: "mon", "tue", "wed", "thu", "fri")

11. duration (DU): last contact duration, in seconds (numeric).

Other attributes:

12. campaign (CA): number of contacts performed during this campaign and for this client (numeric, includes last contact)
13. pdays (PD): number of days that passed by after the client was last contacted from a previous campaign (numeric; 0 means client was not previously contacted)
14. previous (PR): number of contacts performed before this campaign and for this client (numeric)
15. poutcome (PO): outcome of the previous marketing campaign (categorical: "failure", "nonexistent", "success")

Social and economic context attributes:

16. emp.var.rate (EV) : employment variation rate - quarterly indicator (numeric)
17. cons.price.idx (CP) : consumer price index - monthly indicator (numeric)
18. cons.conf.idx (CC) : consumer confidence index - monthly indicator (numeric)
19. euribor3m (E3) : euribor 3 month rate - daily indicator (numeric)
20. nr.employed (EM): number of employees - quarterly indicator (numeric)

Output variable (desired target):

21. y: has the client subscribed a term deposit? (binary: "yes", "no")

2.2. Causality

In this work, to build Bayesian scientific models for the attributes relationships, Generalized Linear Models and Multilevel Models were used.

Before creating a scientific model, the causal relationship among variables can be expressed in terms of DAG representation. The directed acyclic graph (DAG) with a set of nodes (variables) and a set of arcs represents the causal relationship between connected nodes. There are 20 variables and a target variable, and the DAG is given in Figure 1 which displays the network to explain the causality.

By creating the DAG, some assumptions were made to explain the reason for the cause-effect relation between the features given in the dataset. These assumptions;

For the Bank client data:

• Age affects the education level, education depends on the age.

• Age affects the marital status, marital status depends on the age.

• Education level affects having a job and also the type of the job.

• Job and marital status have an effect on having a housing loan. For example, if the client is just married, can want to have a house. And the job status will affect the getting a housing loan.

• The relationship between having a housing loan, loan, and credits(default) is thought of like this; if the customers don't have a housing loan, can tend to have a loan, and if the customers don't have a loan, can tend to be using credits.

• The job status also affects the having loan and credits separately.

Data related to the last contact of the current and previous campaign:

• The number of previous contact(PR) affects the number of days that passed after the client was last contacted from a previous campaign. If there is so much contact between the client and agent, the time interval between the last contact and current contact will be affected by this.

• The number of previous contacts (PR) can be a sign of the customer's interest in the bank and offers from the bank. If the bank has a strong connection with the client, the probability of getting 'yes' response to the offer will increase. So the number of the previous contacts directly affects the previous marketing outcome.

• The number of days that passed by after the client was last contacted from a previous campaign can change the customer preferences for bank choices or deposit needs. During the elapsed time, customers can change decisions about the bank, can get different good offers from other banks, etc. So, it will affect the customer's current attention. The number of contacts performed during this campaign and for this client is an indicator of this.

• The number of contacts performed during this campaign and for this client will affect the last contact duration. How much clients allow to communicate with them will affect the communication duration.

• The duration of the last contact is an important indicator of customer interest. How much the agent can keep the customer held in connection, can increase the possibility of the customer acceptance.

• According to the client's workload, the client can be busier on the specific days of the week and can't give enough time to interest the offer. But some other days the situation can be different. It's just an assumption and should be verified. So the day of the week can affect the target.

• The default(has credit in default) value can change the interest and willingness of the customer. Because of this, the default value can affect the number of previous contacts.

Social and economic context attributes:

• Euribor 3m directly affects both consumer price index and target.

• If interest rates fall, inflation rises. Inflation affects the consumer price index.

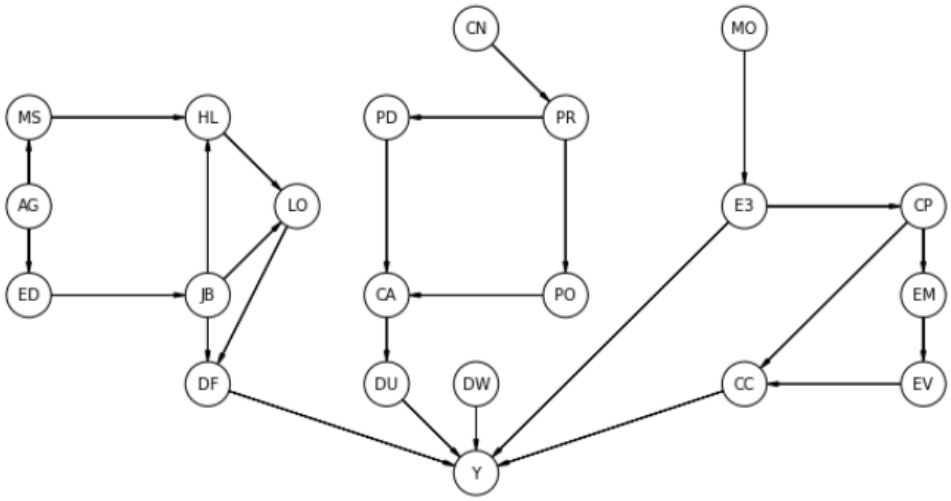

Figure 1: DAG

• If inflation increases, the number of employees decreases, which affects and increases the employment variation rate, which directly affects the consumer confidence index. The consumer price index is the indicator of whether the customers feel safe spending their money or need to save money. It directly affects the target. And also the consumer price index affects the customer's feeling of confidence.
• The Euro Interbank Offered Rate is a daily reference rate, published by the European Money Markets Institute, based on the averaged interest rates at which Eurozone banks offer to lend unsecured funds to other banks in the euro wholesale money market. Thus it will really affect the interest rate which is offered by the bank. It means that the interest rate is very important to decide whether the customer will invest their money in this bank. So euribor3m has a direct effect on the target variable. And it's affected by the month of the year. When considering the Euribor rate, one might think that a lower Euribor would result in a decline in savings rate since most European banks align their deposit interest rate offers with ECB indexes, particularly with the three month Euribor (Reilly, 2005).

According to the DAG and the aim of the project, these questions were analyzed and answered:
• What is the effect of the last call duration on the target?
• What is the effect of the previous campaign result on the current campaign result? Is there any relation between them?
• What is the effect of a previous number of calls on the current campaign result? Do more connections cause positive results for the current situation?
• Does the day of the week on contacted affect the customer results?

Firstly, before starting to create a scientific model, the dataset's descriptive analysis has been done to understand the attributes of the dataset deeply.

2.3. Scientific Models

In this work, to build a Bayesian model for attribute relationships, Generalized Linear Models and Multilevel Models approaches were used to create a scientific model. A generalized linear model (GLM) is much like linear regressions. It is a model that replaces a parameter of a likelihood function with a linear model. In this project, our target is the binary categorical variable. Because of this, creating a GLM model is more appropriate. For the parameters, normal and exponential distributions were used to explain the distribution of the features and their parameters. The priors were detected according to the logit link function which is the inverse of the logit function is the logistic.

And also, the Multilevel Model was used to analyze any clusters effect on the target. These models remember the features of each cluster in the data as they learn about all of the clusters. Depending upon the variation among clusters, which is learned from the data as well, the model pools information across clusters. This pooling tends to improve estimates about each cluster.

By taking a reference from other studies, firstly the causal effect of the duration on the target is a good starting point. If the DAG is analyzed, 'duration' has a direct relation to the target. There are no mid attributes between them. There are other attributes that have an effect on the duration. They will be analyzed in the following sections. The duration can be the main factor which can be the sign of the customer's willingness. As an assumption, if the customer and agent have long phone calls, the percentage of convincing the customer can increase.

In this dataset, each row represents a different customer. The Bernoulli distribution was used as a target distribution. Because the target only has 2 results; yes or no. And there is no more than one trial for the recent campaign. Thus, the Generalized Linear Models were used to interpret the effect. The GLMs use a link function to turn the bounded parameter into an unbounded space so that it can be

estimated by a linear function. The inverse of the logit link function was used to estimate the probabilities of each outcome of the Bernoulli distribution.

$$G_i \sim \text{Bernoulli}(p_i)$$
$$logit(p_i) = \alpha_{[i]} + \beta_{[i]} * Duration$$
$$\alpha_j \sim Normal(0, 1.5)$$
$$\beta_j \sim Normal(0, 1)$$

Figure 2: GLM Model

The model created with only the duration is given above. The other models were created by extending the GLM and are in the Jupyter Notebook.

The issue about the target that is important to research is the effect of the outcome of the previous marketing campaign on the acceptance of the current campaign. According to DAG given in the previous section, to find the direct effect of the previous outcome, the model was created which includes the previous number of contact and duration attributes in the logit link function. Especially, the previous number of contact was included to model to stratify the previous effect on the number of contact of the current campaign and the last call duration.

Another question of interest is the effect of the day of the week on the target. In the descriptive analyzing part in the code notebook, the count plot of the day of the week according to the target value stood out. This plot shows us that there are almost no differences between the day of the week. This assumption was analyzed by the modeling. The possible differences between the day of the week are counted as a cluster. The multi-level modeling was created. There are 5 different days in this dataset. Every day has its own distribution and variance to show how the days are different. Because of the centered version of the Multilevel Model has a lot of divergences, bad R-hat results, and a very low number of effective samples. The non-centered version was also created.

And also the last attribute of the marketing information, the 'campaign' was used to model. The 'campaign' is the number of contacts performed during this campaign and for this client. It is an indicator of the current campaign activities. The increase in the number of contacts can affect the target.

Before model creation, proper priors were estimated for the models, after model creation, the chains of the models were controlled according to priors. The posterior predictive distributions were controlled to see the mean and variance of the intercepts and features. Noneffective features were selected. The model comparison tables place in the following section were created to compare the performances.

3. Experiments and results

The models were created to understand the causal effect of the attributes on the target. The model results were evaluated and compared to each other. Table 1 includes the performance of the 4 different models which are included different features.

The 'waic' column contains the WAIC values as an indicator of the model performance. Smaller WAIC values indicate better models, and the models are ordered by WAIC, from best to worst.

According to the model comparison, the model included only duration (Dur), and the model included duration with the campaign (Dur-ca) have the lowest performance to explain the target. The model with the last call duration, the previous campaign outcome, and the number of contacts performed before this campaign is the highest performed model. It means that the additional attributes/features (poutcome and previous) gave more information and contribution to detecting and explaining the target. The model that takes into account the day of the week also has high performance because of the other features(poutcome, previous, and duration) contribution. Thus, the day of the week does not have any individual effect on the target.
Inferences based on model results:

• Except for the multilevel centered model, the other models have good R hat and effective sample sizes. And also Non-centered version of the model gives proper results. And the trace plots of the Markov chain of the models were clean, healthy Markov chains, both stationary and well mixing except for the centered model.
• The last call duration has a positive effect on the acceptance of the client. The elapsed time helps to convince the customer by providing more time to the agent to explain the details of the campaign. But the optimal call duration should be detected.
• According to the model results, if the customer rejected the previous campaign or does not have any contact for the previous campaign, the effect of these kinds of customers on the target, is negative. It means that if the customer rejects the previous campaign probably will reject the next one, or if the customer does not have any contact and collaboration with the bank, it will be hard to convince the customer for the next campaign. The interesting part is that the customer's acceptance of the previous campaign offer almost does not affect the next campaign result. It means that there are some other factors that have the more strong effect on the acceptance of the offer. One of them is the duration of the last call which is used in the models.
• The Multilevel Model showed that there is no differential effect between the day of the week on the target. Treating the day of the week as a cluster and adding them to the Multilevel Model is not an appropriate approach.
• The model results show that the number of contacts performed during this campaign for the client is not important for the customer's decision. For example, the customer can accept the offer in the first call or the second call, there should be another important factor from the number of calls.

Table 1: WAIC Scores of Models

	rank	waic	p_waic	d_waic	weight	se	dse	warning	waic_scale
Dur-pout-pr	0	436.378287	4.696171	0.000000	0.578289	34.784903	0.000000	False	deviance
Dur-pout-pr-dw(NC)	1	436.714460	6.392790	0.336173	0.421711	35.674755	2.879423	False	deviance
Dur	2	525.040829	2.304470	88.662543	0.000000	38.438428	22.774960	False	deviance
Dur-ca	3	1294.079029	2.840873	857.700743	0.000000	23.131636	37.502089	False	deviance

4. Conclusions

Marketing is an important factor for companies because it helps them sell their products or services. In banking, It is very important to develop good relationships with valued customers accompanied by innovative ideas which can be used as measures to meet their requirements. Telemarketing has become increasingly important in the banking industry, as the pressure is growing to make more profits and reduce operational costs. Capital is the product of the bank industry and companies should increase their capital. At this point, prediction models and data-driven decision support systems can help to improve the prediction and analyze the customer return, and to make the more proper decisions.

In this project, the models created with the Bayesian approach helps to choose more interested customer and reach the aim by reducing the operational cost. The goal is to model the success of subscribing to a long-term deposit concerning a set of marketing attributes. The client attributes were not used in the modeling because of the several studies done about the relations between age, jobs, marital status, etc. The focus of the project is on analyzing and predicting the effect of features of marketing actions. The number of calls during the previous and current campaigns, the outcome of the previous marketing campaign, and the duration of the last call are the main features.

According to created model's performance, the models were compared and the most influential attributes were detected. The most important factors which affect the client subscription decision are the duration of the last call and the result of the previous marketing campaign. The duration of the last call indicates the customer's interest and shows the success of the agent in keeping the customer on the call. If the last call duration is at the optimal level, it will positively affect the customer outcome.

The other important indicator is the outcome of the previous campaign. The model results show that if the client rejects the previous offer it has a high probability to reject the current offer, and if there is no previous connection between the bank and the client for the previous campaign, the current campaign results will be probably negative. And according to the outcomes, the number of calls of the previous campaign or current campaign has a slight effect on the target. Overall, these results show the important thing in telemarketing is not making a number of calls, is the making effective calls.

Additional studies were carried out besides the project. In these studies, the effect of the social and economic context attributes on the target variable is being analyzed. The significant effect of the 'The Euro Interbank Offered Rate' on the target is found.

References

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22-31.

Reilly, G. O. (2005). Information in financial market indicators: an overview. Quarterly Bulletin Articles, 4, 133-141.

Mehrotra, A., & Agarwal, R. (2009). Classifying customers on the basis of their attitudes towards telemarketing. Journal of Targeting, Measurement and analysis for Marketing, 17(3), 171-193.

Hosseini, S. (2021). A decision support system based on machined learned Bayesian network for predicting successful direct sales marketing. Journal of Management Analytics, 8(2), 295-315.