SUBSCRIPTION FOR TERM DEPOSIT PREDICTION

Marcos Paulo Pereira Moretti

Escola Politécnica da USP
PCS5787 - Tópicos Especiais em Ciência dos Dados e Big Data (2022)
São Paulo - SP, Brazil

ABSTRACT

This project consists of building a predictive model to determine if a client will subscribe to a term deposit or not, given that he was contacted via telephone by a marketer during direct marketing campaigns of a Portuguese banking institution. We use data from UCI Machine Learning Repository, transform and model it in order to do the classification task. We use different modeling techniques, ranging from traditional models like Tree models (Random Forests, Boosting models), Nearest Neighbors, until more sophisticated models, like Neural Network models (Multilayer Perceptron) to do the classification task. The best result we achieved was 0.6249 of ROC-AUC for the model Random Forest. We discovered that trimestral Euribor index, consumer price index (CPI) and quantity of calls during the current campaing were the most important to determine if a client would subscribe or not. Moreover, we discovered that roughly 43 years-old clients, having a university degree and no credit default or loans, are more susceptible to subscribe to a term deposit according to the model.

***Index Terms*—** term deposit, classification, bank, modeling, ROC-AUC.

1. INTRODUCTION

This project aims to build a classifier that is able to classify if a client is going to subscribe a term deposit or not, given that he was contacted via telephone or cellular by a marketer during direct marketing campaigns of a Portuguese banking institution.

That's a very common problem in the industry. [1] made a bibliographic review which cites 18 different papers, each one applying a different technique to solve the problem of bank telemarketing prediction using machine learning.

It shows that this is an important topic, and a real-life problem which frequently envolves banking institutions that want to contact their clients in order to try to strengthen relationship and sell products.

1.1. Bibliographic review

We have studied three papers in order to understand what had already been developed in the topic of bank telemarketing.

[2] applies the techniques of decision tree (DT) and rough set theory (RST) to the same dataset. He detailed the obtained tree and the most important features in terms of information gain, which were *Duration* (roughly 10.8%), *Poutcome* (roughly 3.8%), and *pdays* (roughly 3.6%).

[3] applies the techniques Logistic Regression, Naïve Bayes, K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Decision trees and Random Forest to the same dataset, and got *Duration* as the most important feature, based on impurity-based importance. Random Forest won against other models in the metrics accuracy (94.02%), precision (90.41%), recall (98.49%) and F1-score (94.28%), using deliberatively 0.5 as threshold and showing no ROC-AUC metric. [1] proposes a framework to deal with Bank Telemarketing Prediction problems. It does a bibliographic review on the topic, and suggests a data flow for training and scoring future data, presenting accuracy and ROC-AUC as the most important metrics for model selection.

That said, and considering the recency of these publications, we see that it's a really important and actual topic, which deserves attention and suggestions on new approaches.

1.2. Sections of the paper

The section 2 presents the methodology that was followed in the development of this project. We present the results obtained in the section 3. Then, we interpret the results as a conclusion in the section 4.

2. METHODOLOGY

In this section, we are going to show the methodology used to develop the proposed solution.

2.1. Choice of the database

We used the free dataset available in the UCI Repository [4] named as "Bank Marketing" [5], more specifically the one

that contains all the 41,188 examples and 22 attributes, ordered by date from May 2008 to November 2010.

The dataset contains many variables, with different types and origins of data, compiled in a unique dataset. These variables may be divided into categories, according to their concept:

- **Bank client personal data**: age, job, marital status, education, preferred contact type (cellular or telephone);
- **Bank client credit data**: has credit in default, housing, loan;
- **Contact data**: month of the call, day of week of the call, duration of the call, number of contacts during this campaign;
- **Historic of contact data**: number of days that passed by after the client was last contacted, number of contacts performed before this campaign, outcome of the previous marketing campaign;
- **Social and economic context attributes**: employment variation rate, consumer price index, consumer confidence index, euribor 3 month rate, number of employees;
- **Response variable**: if the client has subscribed a term deposit (yes) or not (no).

These data range from categorical to numerical data, while the target variable is categorical.

2.2. Exploratory Data Analysis

First of all, we analysed the features of the dataset in order to understand the database we are using.

2.2.1. General characteristics

We analysed the proportion of the target variable in the dataset. We see that we have 11,3% of positive examples in the full dataset, which configures a certain degree of imbalancement. It's plotted in the figure 1.

Moreover, we analysed the balance of the target variable for each reference in time, and it's shown in the figure 2. We observe that there are some missing references, like September 2008, and also much imbalancement regarding the references. That will affect our train-test division.

2.2.2. Analysis against the target

We analysed each field against the target in order to understand the database, the data types, and if there seems to have any relationship between the independent variables and the response variable.

We divided the analysis into 5 groups of variables, which are used as well in the UCI Repository: client personal data, client credit data, contact data, campaigns data and external indices data.

- **Client personal data:** these data are about the personal qualities of the person:
  - **Age (in years):** we observed that this field brings not much separability of the target variable, once we've got mean ages of 39.9 years for the negative target and 40.9 years for the positive target. We can see the distribution in the image 3;
  - **Marital status:** we observed that married and single are the most numerous (1,352 and 638 respectively) in terms of subscription. We can see the distribution in the image 4;
  - **Education degree:** we observed that clients having university degree or high school have more subscriptions in absolute terms (1,670 and 1,031 respectively) and proportional terms (13.7% and 10.8% respectively). We can see the distribution in the image 5;
  - **Job:** we observed that clients working as administrators or blue-collars have more subscriptions in absolute terms (1,352 and 638 respectively). We can see the distribution in the image 6.
- **Client credit data:** these data are about the personal credit situation:
  - **Default (binary):** we observed that clients without defaults are the most numerous (4,197) in terms of subscription. We can see the distribution in the image 7;
  - **Has housing loan (binary):** we observed no separability of the target. We can see the distribution in the image 8;
  - **Has personal loan (binary):** we observed that clients that don't have personal loan more subscriptions in absolute terms (3,850). We can see the distribution in the image 9.
- **Contact data:** these data are about the phone contact made by the marketer to the client:
  - **Contact type:** we observed that clients called by cellular are the most numerous in subscriptions in absolute terms (3,853) and in proportional terms (14.7%). We can see the distribution in the image 10. Moreover, we see that there's an evolution in the usage of cellular compared to that of telephone, which is presented in the figure 11;

  - **Day of week:** we observed no separability of the target. We can see the distribution in the image 12;
  - **Duration of the call (in seconds):** we observed that there is high separability of the target in the figure 13, and also we have in average 220.8 seconds in a call that don't lead to a subscription in comparison to 553.2 seconds for that lead to subscriptions.

- **Campaign data:** these data are about the current and past campaigns the company had:
  - **Contacts made before the current campaign:** we observed that recalling the client in the current campaign lead to a proportion of at least 10%. We can see the distribution in the image 14;
  - **Days passed since the last call:** we observed no separability of the target. We got average of 6.2 days for non successful calls and 5.9 days for successful calls. We can see the distribution in the image 15;
  - **Contacts done before the current campaign:** we observe that after being called once, twice or three times, we have a high percentage (greater than 20%) of subscription, far better from no recall (zero times). We can see the distribution in the image 16;
  - **Outcome of the previous marketing campaign:** we observe that success in the last campaign leads to more subscriptions (65.1% of success in subscription compared to 14.2% of non success). We can see the distribution in the image 17.

- **External indices data:** these data are about external market indices that may influence the interest in investments:
  - **Employment variation index (quarterly):** we observed that when the variation is smaller, we tend to have more subscriptions (-1.233 in the positive case compared to 0.249 for ne negative case). We can see the distribution in the image 18;
  - **Confidence consumer indicator (monthly) (CCI):** we observed no separability of the target. We got average of -40.593 for non successful calls and -39.790 for successful calls. We can see the distribution in the image 19;
  - **Consumer price index (monthly) (CPI):** we observed good separability of the target. We got average of 93.604 for non successful calls and 93.354 for successful calls. We can see the distribution in the image 20;
  - **Euribor (quarterly) (Euro Interbank Offered Rate):** we observed good separability of the target. We got average of 46.503 for non successful calls and 271.248 for successful calls. We can see the distribution in the image 21.

2.2.3. Correlation analysis

We did also a correlation analysis with the target variable, in order to identify most related variables that should potentially be used in a model.

We did the encoding of categorical variable as to be described in section 2.3.3.

The figure 22 shows the Pearson correlation values. In green, we've highlighted the variables selected for the modeling pipeline.

We can draw some conclusions from that correlations:

- The variable *duration* is very correlated with subscription, however it cannot be used to predict the subscription, because we don't have it in the time of prediction, only after the call;
- Time variables like *anomes*, *year*, *month_*, *campaign* are difficult to analyse, because they are very imbalanced in terms of quantity of examples. As we saw, we have much more examples in the first months and year than we have in the last months;
- The index *euribor3m* is important because this rate provides the basis for the price or interest rate of all kinds of financial products, like interest rate swaps, interest rate futures, saving accounts and mortgages. We see high positive correlation, because high Euribor guarantees higher returns from the investment;
- The index *emp.var.rate* has negative correlation with subscription, because in general high values of variation of employment lead to more uncertainty in the market behaviour, and then less prone to subscribe to an investment;
- The index *cons.price.idx* is the CPI (inflation rate), and also with negative correlation with subscription. That's because term deposits do not keep up with inflation. When inflation is high, investors tend not to subscribe to term deposits;
- The index CCI (confidence rate) has negative correlation with subscription. The CCI measures the confidence of the consumer related to the market. As we can see, higher confidence in the market indicates more subscriptions;
- The variable measuring the quality of past campaigns *poutcome_success* is also important, meaning that

if we had success with this client in past marketing campaigns, we may have success in future campaign calls as well (note the negative correlation with *poutcome_nonexistent*, and no correlation with *poutcome_failure*);

- The contact type also seems to be relevant, as we get high correlation with *contact_cellular* and low correlation with *contact_telephone*. We may infer that the clients tend to subscribe more when contacted via cellular;
- Having no default is also important (*default_no*);
- Having no loans seems not to be important (*loan_no*).

2.3. Pipeline of modeling

We followed a sequence of steps of treatment of the dataset, so as to model it and analyse the results.

We present each step with further details in the subsections.

2.3.1. Pre-manipulation of the data file

We added the column *year* manually according to the sequence of the registers. We could do that because it's informed that the registers are ordered by date [5].

2.3.2. Load the file

We loaded the file as downloaded from the website of UCI [5]. That's a CSV (Comma-Separated Values) file containing 41,188 examples and 22 attributes.

2.3.3. Feature engineering

We encoded the variables *y*, *loan*, *housing*, *default* replacing "yes" per 1.0 and "no" per 0.0.

We encoded the *day of week* with a One Hot Encoder, and removed the variable 'day_of_week_mon' to avoid collinearity.

From now on, we have a fully numerical database, that can be learnt from a model.

After this step, we ended up with the following variables: *default*, *housing*, *loan contacted_by_cellular*, *campaign*, *pdays*, *previous*, *poutcome*, *emp.var.rate*, *cons.price.idx*, *cons.conf.idx*, *euribor3m*, *day_of_week_fri*, *day_of_week_mon*, *day_of_week_tue*, *day_of_week_wed* and *y*.

2.3.4. Feature selection

We selected the independent features analysing their correlations with each other (see the figure 23).

We removed the features *poutcome* and *emp.var.rate* because of their high correlation with *pdays* and *euribor3m* respectively. The choice of the remotion was done based on the correlation of each pair of variables related to the target variable.

2.3.5. Dataset division

We used an out-of-time division of the dataset in order to divide it into a training and a testing dataset.

As we have a very imbalanced dataset in terms of date references, we chose to select the period from May 2008 to December 2008 as our train dataset and January 2009 to December 2009 as our test dataset. We decided that so as to try to balance the most the target between the two dataset. Actually, we got 4,8% of positive targets in the train dataset, and 19,5% in the test dataset, ending up with datasets with 27,690 and 11,450 examples respectively.

2.3.6. Models' training

We trained 4 different models: Random Forest, XGBoost, KNN (K-Nearest Neighbors) and Multi-Layer Perceptron.

We used 15 predicting variables to model the subscription. For each model, a set of parameters was chosen and defined using Grid Search to find the setup with best ROC-AUC among all the possible setups. We used cross-validation with 5 folds to average the metric result for every setup.

We needed no imputations, since there were no missing values.

For distance models and neural networks, we applied normalization to make the convegence of the models possible (that was not needed for tree models).

We used the following parameters, and in **bold** is the chosen configuration:

- Random Forest:
  - *n_estimators*: [100, **250**, 500]
  - *max_depth*: [**5**, 10, 15, 20]
  - *min_samples_split*: [3, 5, **7**]
  - *min_samples_leaf*: [**3**, 5, 7]
- XGBoost:
  - *n_estimators*: [5, 10, **15**]
  - *max_depth*: [3, 4, **5**]
  - *min_child_weight*: [1, **5**, 10]
  - *colsample_bytree*: [0.6, 0.8, **1.0**]
- KNN:
  - *n_neighbors*: [5, 10, 15, 20, 25, 30, 50, 100, **150**]
- Multi-Layer Perceptron:
  - *alpha*: [**0.0001**, 0.001, 0.01, 0.1]
  - *beta_1*: [0.9, 0.99, **0.999**]

2.3.7. Models' evaluation

We used the Area Under the Curve of the ROC curve (ROC-AUC). That metric measures the separability of the model for any threshold we set to the probabilities given by the model to a test dataset.

In this case, we used the test dataset to evaluate the performance, and got the ROC curve for each model, as we can see in the figure 24.

We can see that the Random Forest got best results against the other models (ROC-AUC = 0.6249). That's the model we'll choose as our best model.

3. RESULTS

After we trained the models and tested the generalization power of the models with the metric ROC-AUC, we evaluated the importance of variables and deduced a subscriber's profile by means of the predictions. In this section we present these results.

3.1. Importance of variables

We plotted the best features and displayed them in figure 25.

We can see that the features *euribor3m*, *cons.price.index* and *campaign* dominated the importance, in accordance to what we exposed in section 2.2.3, where we discovered that they have high correlations with the target.

3.2. Subscriber's profile

We divided the prediction according to the score that was assigned by the model. We assigned every call to a group, numbered as the ceil of the score of the call.

Then, we averaged the numeric attributes and computed the mode of the categorical attributes (figure 26). So, we found for each group a predominant characteristic for the clients in that group.

For the scores 3.0, 4.0 and 8.0, we have a small public, so we need to be careful about the conclusions we derive from them. Let's ignore these groups.

For the score 7.0 (the highest), we get this public profile:

- about 42 years old in average;
- majority of administrators and married people, with university degree;
- majority having no defaults, no housings and no loans;
- majority was contacted via cellular, with more than 3 minutes of call;
- roughly 2 contacts performed during the campaign (campaign) (smaller than the other categories), roughly 6 days passed after the last campaign (*pdays*) (greater than the other categories), almost no previous contact performed before this campaign (*previous*). That shows a characteristic that the clients tend to be more receptive to subscribe after waiting for some time to be recalled, and tend to subscribe more when the calls are longer;
- lower *CPI* (inflation) mean rate, higher *CCI* (confidence rate) and higher *Euribor* rate.

4. CONCLUSION

We were able to get good results in terms of the metric ROC-AUC for the Random Forest model (0.6249). It's clear that these results are really inferior compared to the ones shown in the references in [1], but we see that most of the cited articles use the variable *duration* in the prediction, what is wrong, except if the interest is to discover the most important independent variables related to the target.

We could find a profile that is very prone to subscribe to a term deposit. That information could be used to implement a campaign to engage the public that doesn't subscribe frequently, or even deepen the engagement with the public that is prone to the subscription.

It's worth saying that this project was a good opportunity to put into practice the concepts dealt in the data lifecycle learnt in the discipline.

As an improvement, we should compute the metrics of accuracy, precision, recall and F1-score as helpers together with the ROC-AUC to choose the best model.

One can have access to the project development accessing the link `https://github.com/mpereiramoretti/pcs5787-projeto`

5. REFERENCES

[1] H. Toulni et. al. S. C. K. Tékouabou, S. C. Gherghina, "A machine learning framework towards bank telemarketing prediction," *Journal of risk and Financial Management*, , no. 15, 2022.

[2] S. Abbas, "Recognition of spoken languages from acoustic speech signals using fourier parameters," *International Journal of Computer Applications*, , no. 3, 2015.

[3] A. Alqaddoumi S. E. Saeed, M. Hammad, "Predicting customer's subscription response to bank telemarketing campaign based on machine learning algorithms," *International Conference on Decision Aid Sciences and Applications (DASA)*, , no. 3, 2022.

[4] G. Casey D. Dheeru, "UCI machine learning repository," 2017.

[5] P. Cortez S. Moro and P. Rita, “A data-driven approach to predict the success of bank telemarketing. decision support systems,” 2014.

Appendices

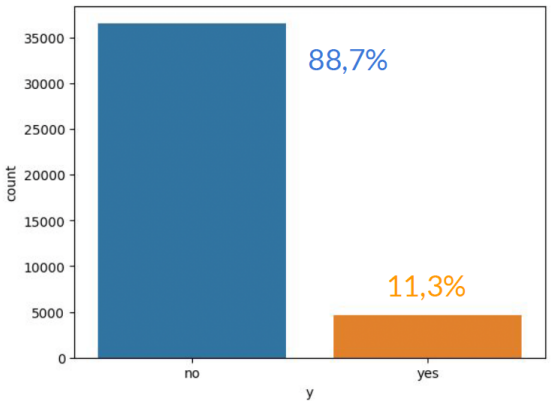

Fig. 1. Distribution of the target.

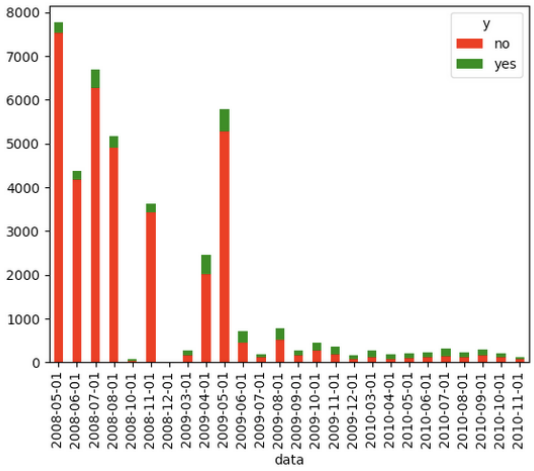

Fig. 2. Distribution of the target in time.

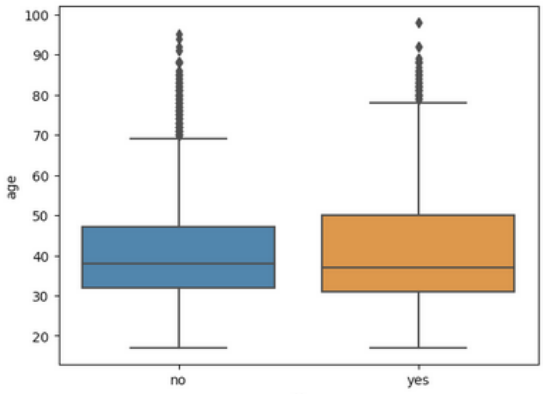

Fig. 3. Distribution of the age.

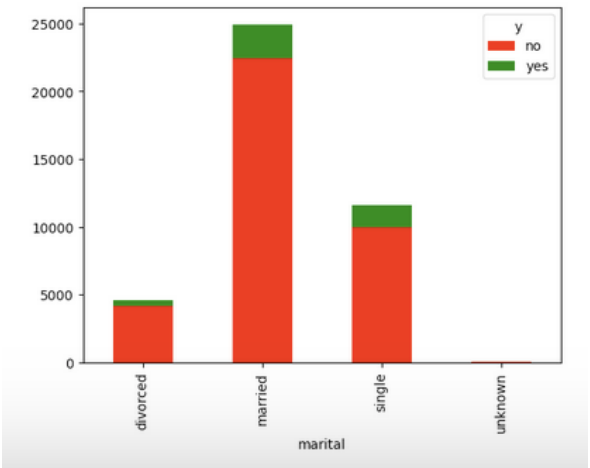

Fig. 4. Distribution of the marital status.

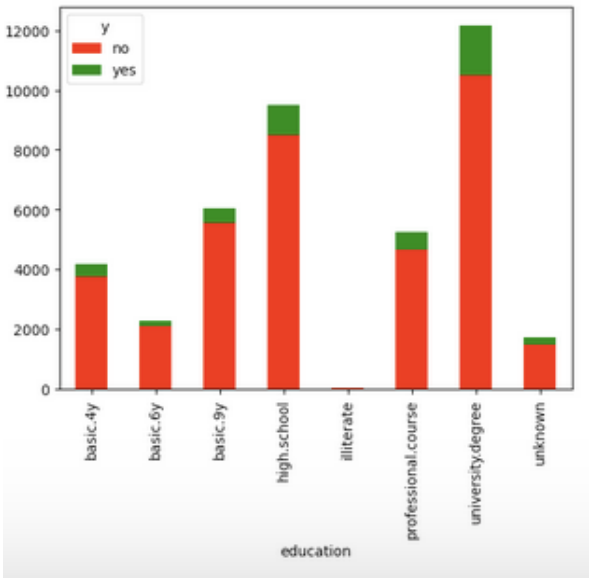

Fig. 5. Distribution of the education degree.

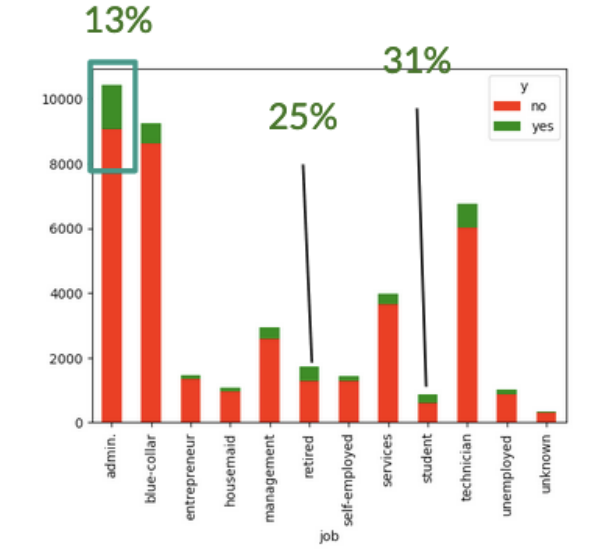

Fig. 6. Distribution of job.

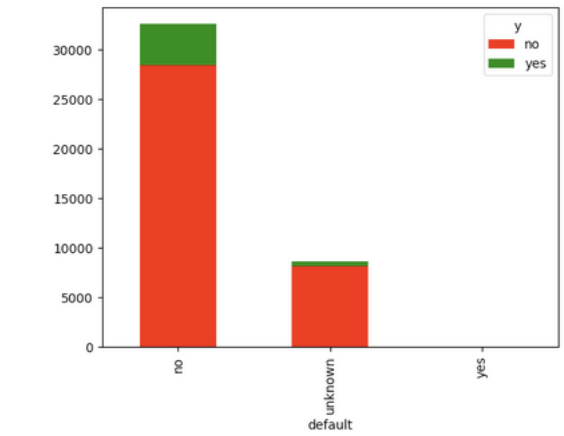

Fig. 7. Distribution of default clients.

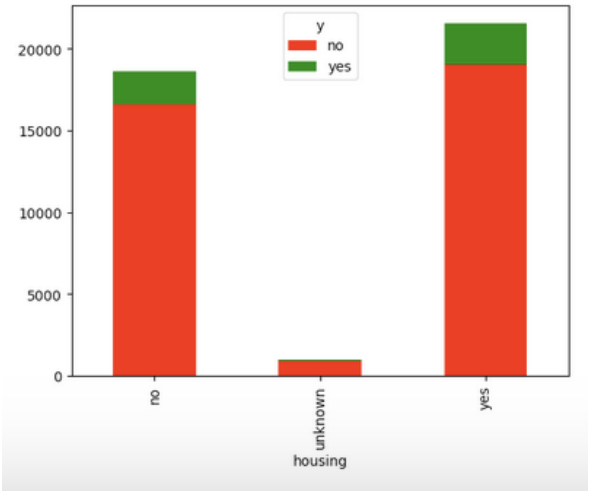

Fig. 8. Distribution of clients with housing loan.

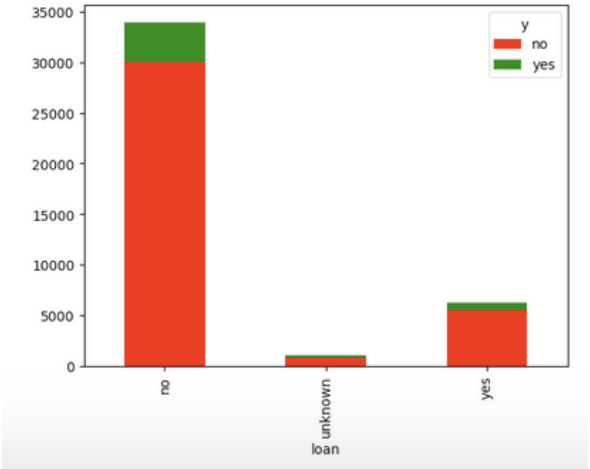

Fig. 9. Distribution of clients with personal loan.

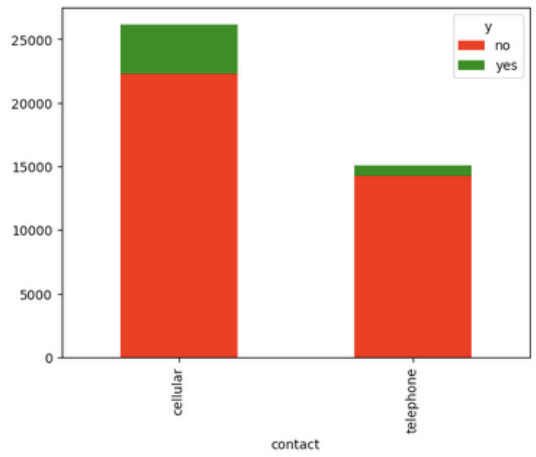

Fig. 10. Distribution of clients contacted by each contact type.

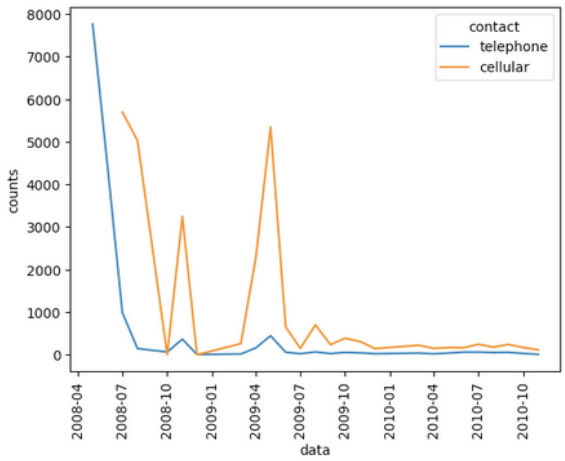

Fig. 11. Distribution of clients contacted by each contact type in time.

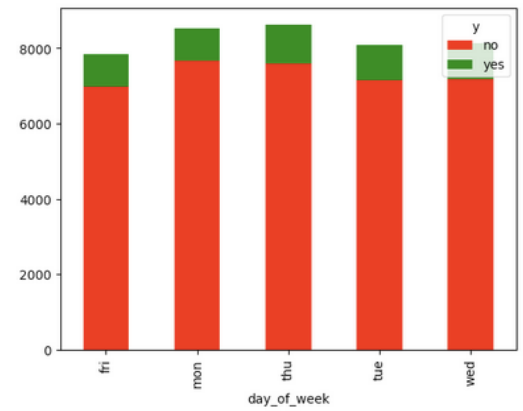

Fig. 12. Distribution of day of week of the contacts.

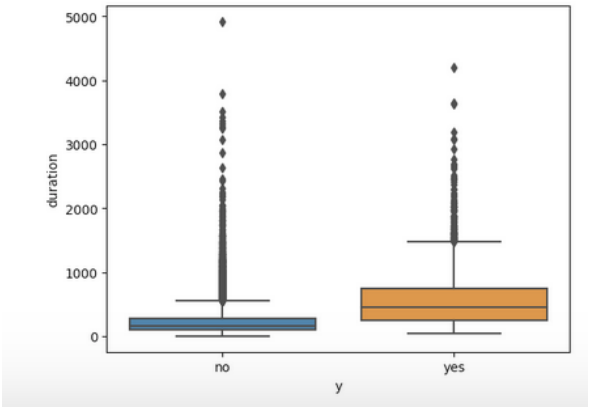

Fig. 13. Distribution of the duration of the call.

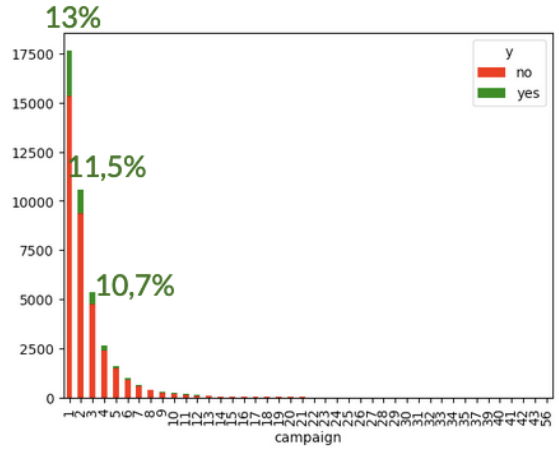

Fig. 14. Distribution of the recontacts in the current campaign.

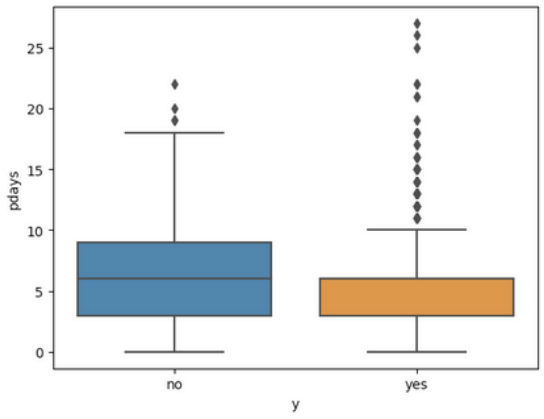

Fig. 15. Distribution of days passed since the last call in this campaign.

previous	no	yes	%yes
5	5.000	13.000	0.722
6	2.000	3.000	0.600
3	88.000	128.000	0.593
4	32.000	38.000	0.543
2	404.000	350.000	0.464
1	3594.000	967.000	0.212
0	32422.000	3141.000	0.088
7	1.000	0.000	0.000

Fig. 16. Distribution of previous contacts and subscription.

poutcome	no	yes	%yes
success	479	894	0.651
failure	3647	605	0.142
nonexistent	32422	3141	0.088

Fig. 17. Distribution of previous campaign success and subscription in the current campaign.

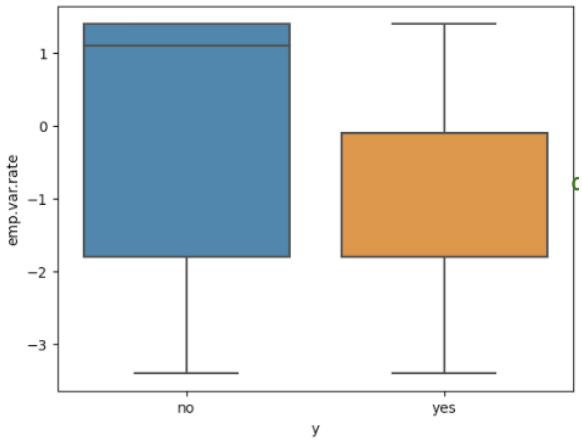

Fig. 18. Distribution of the quarterly index of variation of employment.

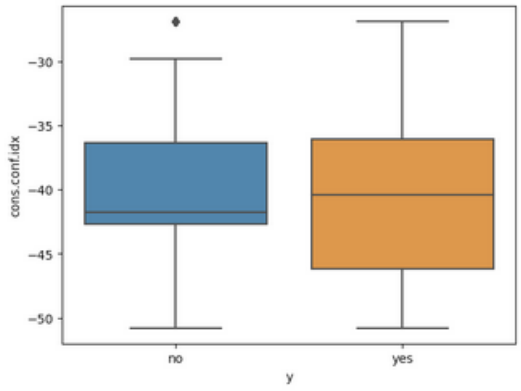

Fig. 19. Distribution of the Consumer confidence index (CCI).

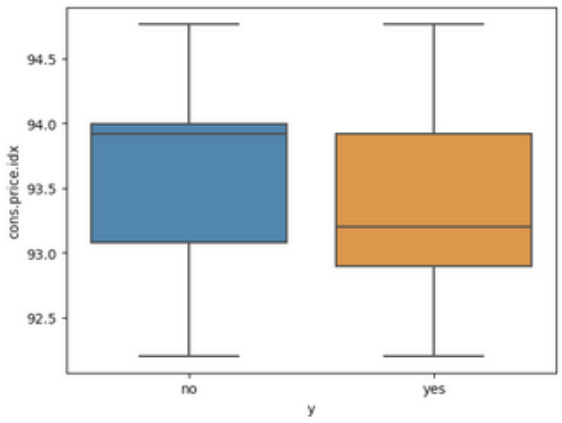

Fig. 20. Distribution of the Consumer price index (CPI).

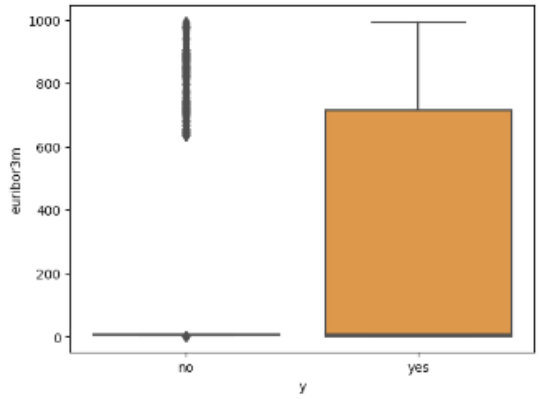

Fig. 21. Distribution of the Euribor quarterly index.

duration	0.405	loan_unknown	−0.002
anomes	0.352	default_yes	−0.003
year	0.348	loan_yes	−0.004
euribor3m	0.319	job_self-employed	−0.005
poutcome_success	0.316	job_technician	−0.006
previous	0.230	job_housemaid	−0.007
contact_cellular	0.145	day_of_week_fri	−0.007
month_mar	0.144	month_aug	−0.009
month_oct	0.137	month_jun	−0.009
month_sep	0.126	marital_divorced	−0.011
default_no	0.099	housing_no	−0.011
job_student	0.094	month_nov	−0.012
job_retired	0.092	job_entrepreneur	−0.017
month_dec	0.079	day_of_week_mon	−0.021
month_apr	0.076	month_jul	−0.032
cons.conf.idx	0.055	job_services	−0.032
marital_single	0.054	marital_married	−0.043
education	0.044	campaign	−0.066
month_nb	0.037	job_blue-collar	−0.074
poutcome_failure	0.032	default_unknown	−0.099
job_admin.	0.031	month_may	−0.108
age	0.030	cons.price.idx	−0.136
job_unemployed	0.015	contact_telephone	−0.145
day_of_week_thu	0.014	poutcome_nonexistent	−0.194
housing_yes	0.012	emp.var.rate	−0.298
day_of_week_tue	0.008	pdays	−0.325
day_of_week_wed	0.006	nr.employed	−0.355
marital_unknown	0.005		
loan_no	0.005		

Fig. 22. Correlation of the input variables related to the target variable.

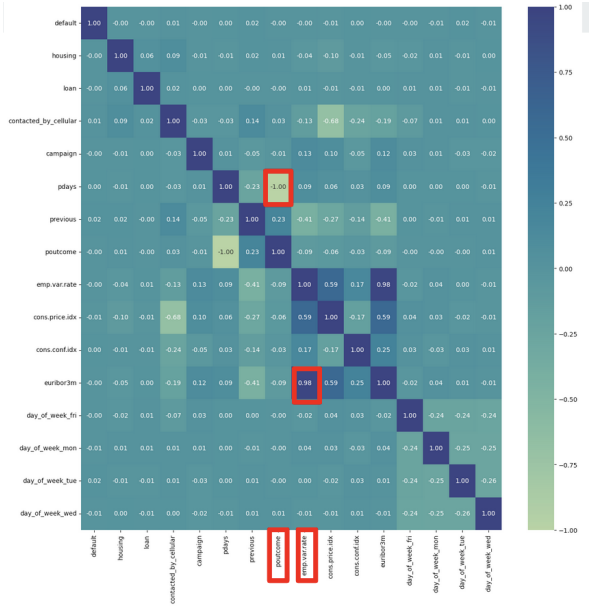

Fig. 23. Correlation of the input variables with each other, pointing in red the removed features.

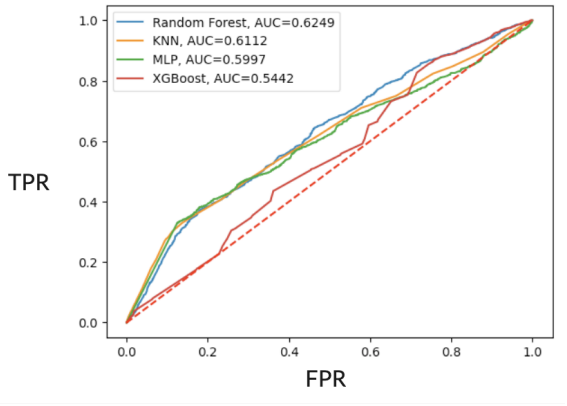

Fig. 24. ROC curves of every tested model, showing the value of the AUC as well.

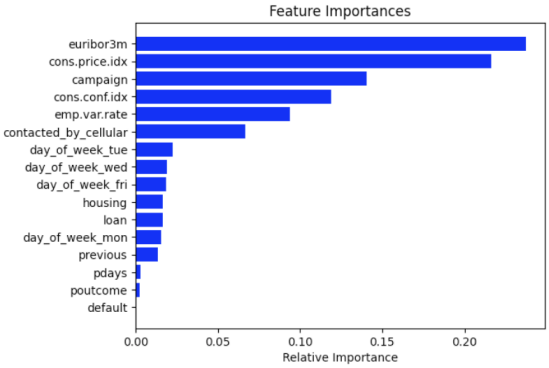

Fig. 25. Feature importances we got from the model Random Forest.

score_classification	Qtt. of calls	age	job	marital	education	default	housing	loan	contact	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m
3.00	2	58.50	admin.	divorced	university.degree	no	yes	no	cellular	70	20.50	NaN	0.00	-1.80	92.84	-50.00	1.66
4.00	91	38.33	blue-collar	married	high.school	no	yes	no	cellular	7	6.80	2.64	0.95	-1.85	92.90	-45.90	1.30
5.00	3263	38.31	blue-collar	married	high.school	no	yes	no	cellular	11	3.01	4.82	0.65	-1.86	92.92	-45.89	1.57
6.00	7180	39.22	admin.	married	university.degree	no	yes	no	cellular	104	1.71	5.57	0.25	-2.13	92.85	-43.13	133.07
7.00	844	42.93	admin.	married	university.degree	no	no	no	cellular	207	1.76	6.47	0.15	-3.08	92.54	-33.11	583.84
8.00	70	45.16	admin.	married	university.degree	no	yes	no	telephone	120	1.30	13.00	0.01	-3.17	92.40	-30.00	788.50

Fig. 26. Client's average profile for each score group assigned by the Random Forest model.