SUBSCRIPTION FOR TERM DEPOSIT PREDICTION

Marcos Paulo Pereira Moretti
Pedro Alves Quilici Coutinho
Victor de Magalhaes Deboni

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), SVM models, Logistic Regression, until more sophisticated models, like Neural Network models (Multilayer Perceptron) to do the classification task. The best results we achieved was ... of ROC-AUC for the model ..., which compare with literature on the topic. We discovered that ... and ... variables were the most important to determine if a client would subscribe or not, which compare with literature on the topic, and that specify the profile clients 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 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 o 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, precision, recall an F1-score.

[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

Size of the dataset.

Proportion of the target (balanced, unbalanced?)

Proportion of data in each date reference (balanced, unbalanced?)

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.

2.2.3. Correlation analysis

Show the correlation table of the variables after preparing the dataset.

2.3. Pipeline of modeling

We followed the sequence of steps drawn in the figure cite figure with pipeline and length of the generated databases.

We present next each step with further details.

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

What features were created? What were removed? Why did we choose to create these new variables (any special meaning)?

2.3.4. Feature selection

Specify what techniques were used to select the most important variables. How does this technique work?

2.3.5. Models' training

Specify how the dataset was divided, what the techniques were used, how were the hyperparams chosen (GridSearch with K-Fold validation, which metric was used?), preprocessings (scaling? imputations?)

2.3.6. Models' evaluation

What metrics were used? What are their definitions?

3. RESULTS

After we trained the models, we tested the generalization power of the models, and compared them in terms of predictions, metrics and importance of variables. Here are the results we obtained.

3.1. Predictions

How were the predictions like, what was their distribution?

3.2. Metrics

Present the metrics of each model. Select the best model. Compare with literature.

3.3. Importance of variables

What did we get as best variables for each model? And for the winning model? Compare with literature.

3.4. Subscriber's profile

According to the model, find the profile of the clients with higher probability of subscribing to a term deposit. Is there any possibility of thinking about a specific campaign for this public?

4. CONCLUSION

Recall what we did briefly.

What interpretations can we get from the results?

Cite the importance of Data lifecycle for this project, in the case of it being used monthly.

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

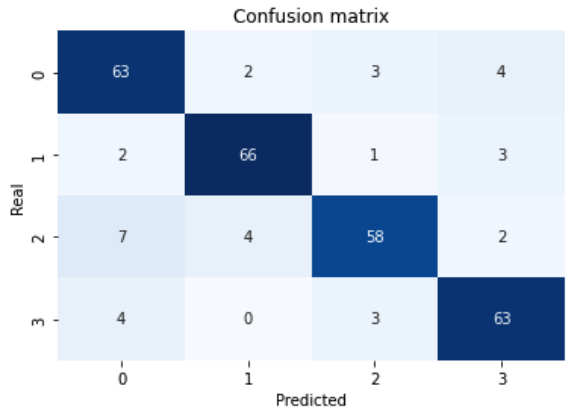

Fig. 1. Confusion matrix of the RF model.