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 paper presents a proposal for a project of predicting if a client will 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. The data is available on UCI Machine Learning Repository, and will be cleaned, transformed and modeled in order to do the classification task. We will try 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). Then, we will compare the results of every model by means of ROC curve, ROC-AUC metric, and other classification metrics like precision, recall, F1-score.

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, mainly when we're talking about banking institutions, which frequently contact their clients in order to try to strengthen relationship and sell products.

In [1], it applied 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%).

In [2], it applied 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.

In [3], it proposed 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.

2. DATASETS

We'll use the dataset available in the UCI Repository [4] named as "Bank Marketing" [5].

More specifically, the one that contains all the 41,188 examples and 20 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.

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

3. METHODOLOGY

Some steps will be taken in order to accomplish reasonable performance in the prediction task, while cleaning the data properly:

1. Data Loading:
   (a) Load the columns in the correct format.
2. Data Preparation:
   (a) Encode the categorical variables;
   (b) Analyse outliers (and possibly remove them);
   (c) Impute null values.
3. Data Split into train and test datasets:
   (a) Evaluate best option: Stratify the sampling by the target variable or execute out of time split.
4. Modeling:
   (a) Apply Boruta to select the most related variables to the response variable;
   (b) Remove very correlated input variables;
   (c) Train and test every proposed modeling technique;
   (d) Evaluate with classification metrics.

We will train not only traditional machine learning models, but also Neural Networks as a novelty model compared to the references we've studied.

4. LIBRARIES

We'll use the following Python libraries for the development:

- *pandas*: to manipulate tabular data as DataFrames;
- *numpy*: to do math computations;
- *sklearn*: to train traditional models and apply other data transformations;
- *matplotlib*: to plot graphics;
- *seaborn*: to plot graphics;
- *PyTorch*: to create and train neural networks;
- *BorutaPy*: to apply the Boruta technique for selecting features.

We're going to do every development on *Jupyter Notebooks*, which are going to be made available on a GitHub repository.

5. EXPECTED RESULTS

We expect to have good F1-score for the best model, as well as good ROC-AUC metric, which indicates that the model is giving reasonable predictions regarding to the response variable. We'll present the classification metrics of Precision and Recall as well, and do a comparison with the papers we've studied on the same dataset whenever possible.

6. REFERENCES

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

[2] 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.

[3] 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.

[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.