Machine Learning Engineer Nanodegree

Using Supervised Classification Algorithms to Predict Bank Term Deposit Subscription

Fabiano Shoji Yoschitaki
June 28th, 2018

Proposal

Domain Background

The domain background of this project is about analyzing marketing campaigns of banking institutions. In order to promote products and services, financial institutions like banks generally run marketing campaigns using two approaches [1]: 1) mass campaigns, which targets general indiscriminate public, broadcasting the same message to different customers and 2) directed marketing, which targets specific contacts, creating a directing relationship to customers.

Banks which run marketing campaigns following the first approach have had their campaigns' performance reduced over time, having less than 1% of positive responses [2]. On the other hand, marketing campaigns which follow the second approach have shown better results compared to the first [3]. For this reason, banks are more likely to spend their budget on directed marketing campaigns than on inefficient mass campaigns.

The personal reason to work on this domain background comes from the fact that I've worked on a project related to a bank company with the goal to offer the most coherent products to its customers based on their characteristics. I believe that having applied machine learning techniques could have helped us to get better response rates.

Problem Statement

Given the Bank Marketing data set [4], which is related to direct marketing campaigns of a Portuguese bank institution, a supervised binary classification model has to be created and trained with the objective of predicting whether or not a client will subscribe to a term deposit.

Datasets and Inputs

The data set chosen for this project is related to direct marketing campaigns based on phone calls of a Portuguese banking institution. It was obtained by exploring the University of California Irvine's Machine Learning Repository [5]. The data set file which will be used is the bank-full.csv and it contains 45211 instances with 17 columns each. The last column is the target label: whether or not the person subscribed a term deposit. The probability for the label 'yes' (did a term deposit) is approximately 12% and for 'no' (didn't do a term deposit) is approximately 88%. The description of the columns follow:

- Bank client features:
  - **age**: the age of the client (numeric).
  - **job**: the type of job of the client (categorical). Possible values: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown'.

  - **marital**: the marital status of the client (categorical). Possible values: 'divorced', 'married', single', 'unknown'; note: 'divorced' means divorced or widowed.
  - **education**: the education level of the client (categorical). Possible values: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown'.
  - **default**: whether or not the client has credit in default (categorical). Possible values: 'no', 'yes', 'unknown'.
  - **housing**: whether or not the client has housing loan (categorical). Possible values: 'no', 'yes', 'unknown'.
  - **loan**: whether or not the client has personal loan (categorical). Possible values: 'no', 'yes', 'unknown'.

- Features related with the last contact of the current campaign:
  - **contact**: contact communication type (categorical). Possible values: 'cellular','telephone'.
  - **month**: last contact month of year (categorical). Possible values: 'jan', 'feb', 'mar', 'apr', 'may', 'jun', 'jul', 'aug', 'sep', 'oct', 'nov', 'dec'.
  - **day_of_week**: last contact day of the week (categorical). Possible values: 'mon','tue','wed','thu','fri'.
  - **duration**: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

- Other features:
  - **campaign**: number of contacts performed during this campaign and for this client (numeric, includes last contact).
  - **pdays**: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted).
  - **previous**: number of contacts performed before this campaign and for this client (numeric).
  - **poutcome**: outcome of the previous marketing campaign (categorical). Possible values: 'failure','nonexistent','success'.

- Target label:
  - **y**: whether or not the client subscribed to a term deposit (categorical). Possible values: 'yes', 'no'.

Solution Statement

The main goal of this project is to apply different machine learning techniques, including supervised classification algorithms such as Gaussian Naive Bayes, Linear Discriminant Analysis, K-Nearest Neighbors, Logistic Regression, Decision Tree, Extreme Gradient Boosting, SVM and Random Forest, to predict from the Bank Marketing data set whether or not a client will subscribe to a term deposit (so it is a binary classification task). After applying the chosen supervised algorithms to the dataset, their prediction scores will be evaluated by accuracy metric and the best one (maybe more than one if we have similar high scores) will be selected so that its hyper-parameters are tuned using Grid Search technique and eventually apply the newly tuned model on the same training and testing data to compare before and after scores.

Benchmark Model

The Bank Marketint data set is a supervised learning classification problem, so I propose to apply - without tuning - all the supervised classification algorithms described in the Solution Statement section and choose the one with the best accuracy score as the benchmark model for this project. Then, after we tune the hyper-parameters of the model, it is expected that the newly tuned model may overcome the benchmark model accuracy score.

Evaluation Metrics

The evaluation metric that will be used in this project is accuracy. It is an appropriate metric for supervised classification problems. Accuracy is a metric that takes considers both correct and incorrect classifications in its formula:

(TP + TN)/(TP + TN + FP + FN)

Where:

- **TP**: True Positive, in our case a person that makes >50k a year and is correctly classified as >50k a year.
- **TN**: True Negative, in our case a person that makes <=50k a year and is correctly classified as <=50k a year.
- **FP**: False Positive, in our case a person that makes <=50k a year and is incorrectly classified as >50k a year.
- **FN**: False Negative, in our case a person that makes >50k a year and is incorrectly classified as <=50k a year.

Project Design

The project is designed in the following steps:

- **Data and Library Loading:** the first step is to load the Bank Marketing data set in the CSV format from the UCI's Machine Learning Repository and all the libraries needed for the project.
- **Data Exploration:** in this step, we'll do some tasks like: visualize the data, print some samples, check its dimensions, check the most relevant features, show its statistical summary.
- **Data Preparation:** after exploring the data, pre-processing tasks will be done: data cleaning, remove null values, convert categorical features into dummy/indicator variables and split the data into training and testing datasets.
- **Model Selection:** with the prepared data, various supervised classification algorithms will be experimented in order to find compare their results and choose the best one (taking into account the accuracy score) for model tuning.
- **Model Tuning:** after we choose the best model, grid search cross validation will be applied with the objective to tune the hyper-parameters of the model.
- **Final Evaluation:** in this step, the accuracy score of the tuned model will be evaluated by applying it to the testing dataset.

References

[1] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimaraes, Portugal, October, 2011. EUROSIS.

[2] Ling, X. and Li, C., 1998. Data Mining for Direct Marketing: Problems and Solutions. In Proceedings of the 4th KDD conference, AAAI Press, 73–79.

[3] Ou, C., Liu, C., Huang, J. and Zhong, N. 2003. On Data Mining for Direct Marketing. In Proceedings of the 9th RSFDGrC conference, 2639, 491–498.

[4] https://archive.ics.uci.edu/ml/datasets/bank+marketing

[5] Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml].