Mini-project 1

Direct marketing campaigns of a bank

Objective

The classification goal is to predict if the client will subscribe to a term deposit (variable y).

Number of Instances: 45211 for bank-full.csv. The marketing campaigns were based on phone calls. Often, more than one contact with the same client was required, in order to access if the product (bank term deposit) would be (or not) subscribed.

Input variables:

Customer data

a) Age (*numeric*)

b) Job : type of job (*categorical: "admin.", "unknown", "unemployed", "management", "housemaid", "entrepreneur", "student", "blue -collar", "self-employed", "retired","technician", "services"*)

c) Marital : marital status (*categorical: "married","divorced","single"; note: "divorced" means divorced or widowed*)

d) Education (*categorical: "unknown","secondary","primary","tertiary"*)

e) Default: has credit in default? (*binary: "yes","no"*)

f) Balance: *average yearly balance,in euros. (numeric)*

g) Housing: has housing loan?(*binary: "yes","no"*)

h) Loan: has personal loan? (*binary: "yes","no"*)

Related with the last contact of the current campaign:

i) Contact: contact communication type (*categorical: "unknown","telephone","cellular"*)

j) Day: last contact day of the month (*numeric*)

k) Month: last contact month of year (*categorical: "jan", "feb", "mar", …, "nov", "dec"*)

l) Duration: last contact duration, in seconds (*numeric*)

Other attributes:

m) Campaign: number of contacts performed during this campaign and for this client (*numeric, includes last contact*)

n) pdays: number of days that passed by after the client was last contacted from a previous campaign (*numeric, -1 means client was not previously contacted*)

o) previous: number of contacts performed before this campaign and for this client (*numeric*)

p) poutcome: outcome of the previous marketing campaign (*categorical: "unknown","other","failure","success"*)

Output variable (desired target):
y - has the client subscribed a term deposit? (*binary: "yes","no"*)

Planning

A four-step planning is proposed to develop a complete modelling project. Each phase will emphasize an important aspect of conducting AI consulting projects.

Proposed development phases for the project

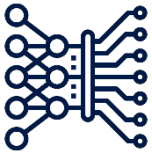
Think strategies

- Ensure the clear understanding of the objectives
- Design of the strategy for its resolution
- Identification of IT resources to be used
- Detail working plan for ensuring the delivery at the end of the week

Research options

- Downloading and exploratory analysis of the database
- Understanding the data
- Preprocessing of data
- Research of the best models and metrics to use.

Find solutions



- Selection of the best approach
- Development of the models
- Validation and final adjustments
- Understanding the output

Communicate results

- Development of the presentation for the communication of the results
- Preparation of data and graphic support
- Thinking of possible improvements and next steps

Thinking the strategy

Organization of the workflow and workload is paramount to solve the problems in a timely manner. It is necessary to define a clear strategy to achieve success in the realization of the project.

Main questions to ask

What is the objective?

- Select the best classification algorithm for the provided database. Identify the best predictors.

Where is the data?

- The data is the full_bank dataset.
- You can access it throughout the Git Hub datasets:
  https://github.com/Fabiancaru/Machine_Learning

How to solve it?

- Design a strategy to solve the problem. It should include:
  - Your initial approaches.
  - A schedule with deadlines for the data analysis, model selection, model development…
  - The resources you will need, such as libraries or more data.

Recommendations

- Create a new environment for the project.
- **Document your code**.
- Check the official documentation of functions and libraries.

Researching the options

Predict under the best predictors, if the product (bank term deposit) would be (or not) subscribed.

Classification algorithms

Tasks

- Which algorithms and goodness of fit metrics could be used?
- What conclusions could we obtain from this preliminary models?
- Preprocessing the data (for example, applying PCA) makes the model better?
- Are they useful for the objective?
- Compare the main characteristics of the models used

Finding the solutions

Choose the models that work best for you and work on them to improve the results. Try to modify the input hyperparameters of the model and to work more on the preprocessing.

Selecting models

Tasks

- Explain why you choose these models and compare results.
- Explain the preprocessing you are going to do and why you think it is the best.
- Explain the benefits and drawbacks of applying PCA.
- What does PCA make to dataset?
- What other models serve to solve the problem?

Communicating results

Submit a report explaining your results. Use images, data, text and anything else you think will be helpful to understand your results. No assumptions: everything you say must be based in facts. Attach to your delivery the code of the final models worked on.

Understanding the results

Tasks

- Explain the models you used to explore the data and the preprocessing you have done.
- Think about how you would explain your result to someone who doesn't know coding.
- Explain what you would need to improve the accuracy of your model and how.
- Explain what is overfitting and underfitting and how you have avoided it in your model.

Evaluation parameters.

A) Word document (report) not exceeding 6 pages.

Front page. The author's data is detailed, as well as the title of the report.

Index. The content of the report is listed, indicating the page it is on.

Introduction. The parts of the report and its total length are detailed.

Development. The subject addressed is captured and developed (it is the body of the report).
Some expected items:
1.EDA
2. Dimensionality reduction (appropriate or not)
3. One-hot coding is the process by which categorical data are converted into numerical data for use in machine learning.
4. Normalization (standardization) as appropriate.
5. Comparison of models.

Conclusion. The most important results are reviewed.

Bibliography. The material consulted for the preparation of the report is arranged alphabetically.

B) Slides and presentation.

C) Notebook