|  |  |  |
| --- | --- | --- |
|  |  |  |
| Ramses Meza  April 2021  [https://github.com/lordtable/UTDSGC\_C2T](https://github.com/lordtable/UTDSGC_C2T2)4 |  | Data Analytics Course 2: CAPSTONE Project proposal |

Business Acumen

An undisclosed Portuguese banking institution has been conducting phone calls as part of a direct telemarketing campaign, aimed at reaching and persuading potential customers that could (or could not) subscribe to the bank’s term deposit product. Customers were contacted more than once, and several features were collected related to such calls. The bank’s goal is to identify the main drivers during the calls that lead to a new client to subscribe the product, and use the derived predictive model to optimize the bank’s marketing campaign.

Raw Data Description

The dataset is located on the University of California-Irvine Center for Machine Learning and Intelligent Systems web repository: (https://archive.ics.uci.edu/ml/datasets/Bank+Marketing)

This dataset has been locally stored as a .csv file. It consists of 41,188 phone calls, which represent the instances/samples. It is structured in a table format: each row represents a phone call to an undisclosed customer. Each column represents an attribute or feature collected for each phone call instance, such as:

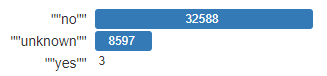
1. **Age**: Customer’s age (numeric)
2. **Job**: Customer’s type of job (categorical)
3. **Marital**: Customer’s marital status (categorical)
4. **Education**: Customer’s highest education (categorical)
5. **Default:** Customer’s general credit default status (binary categorical)
6. **Housing**: Customer has housing loan? (binary categorical)
7. **Loan**: Customer has personal loan? (binary categorical)
8. **Contact**: Contact communication type (categorical)
9. **Month**: last contact month (categorical)
10. **Day\_of\_week**: last contact day (numeric)
11. **Duration**: last contact phone call duration, in seconds (numeric)
12. **Campaign**: number of contacts performed during the current campaign and for the same client (numeric)
13. **Pdays**: number of days that passed by after the client was last contacted from a previous campaign (numeric)
14. **Previous**: number of contacts performed before the current campaign and for the same client (numeric)
15. **Poutcome**: outcome of the previous marketing campaign (categorical)
16. **Emp.var.rate**: National Economic Index: employment variation rate (numeric)
17. **Cons.price.idx**: National Economic Index: consumer price index (numeric)
18. **Cons.conf.idx**: National Economic Index: consumer confidence index (numeric)
19. **Euribor3m**: National Economic Index: Euribor 3-month rate (numeric)
20. **nr.employed**: National Economic Index: Number of employees (numeric)
21. **y**: Has the customer subscribed to a term deposit? (binary: “yes”,”no”)

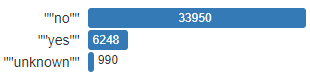
Based on the Business Acumen, the target feature is ***y***, which being a binary variable makes it suitable for a classification modelling approach.

This raw data exploration was managed as a Python Pandas’ DataFrame, with local backups in .csv format at major checkpoints, if needed.

Preliminary Exploratory Data Analysis (EDA)

The Pandas Profile Report of the raw data allowed performing a preliminary EDA of the dataset, in order to start gaining familiarity with it and start envisioning the data processing/cleaning strategy and challenges. Gladly, there are zero missing cells, and very few duplicated instances/rows (12). Twelve out of the total 21 features are categorical, which will require a systematic feature formatting, labelling and one-hot encoding. Some of these categorical variables are highly imbalanced, such as ***default***, ***loan***, ***poutcome***, and ***y***. Their histograms can be seen below on Figure 1. These imbalances would need to be mitigated for an enhanced predictive model-building.





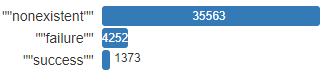




Figure 1: Histograms (Pandas Profile Report) of highly-imbalanced categorical variables, from top to bottom: ***default***, ***loan***, ***poutcome*** and ***y***.

Regarding the numeric variables (9 features), there is a varying degree of inter-correlation amongst each other, as deducted from the Spearman’s r Feature Correlation Matrix shown on Figure 2. For instance, there is a very high positive correlation between ***euribor3m***, ***nr.employe***d and ***emp.var.rate***, while ***previous*** has high negative correlation with ***pdays***, ***emp.var.rate***, ***euriborn3m*** and ***nr.employed***. Scatter plots were built only for the highly-correlated numerical variables (as indicated by Figure 2) and shown on Figure 3; although it seems that these plots do not honor such high level of correlation. There are other less pronounced feature mutual correlations, but the observations above indicate that a systematic and robust feature selection/engineering scheme may be needed after processing/cleaning the data but before building a predictive model, in order to guarantee as many mutually-independent features calibrating the model as possible.

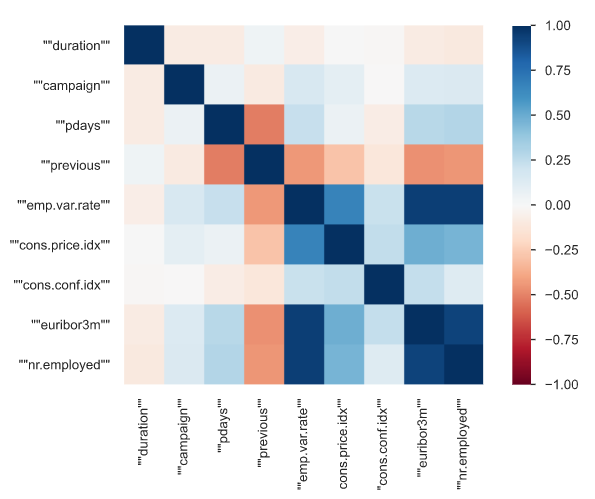


Figure 2: Feature Correlation matrix (Sperman’s r)

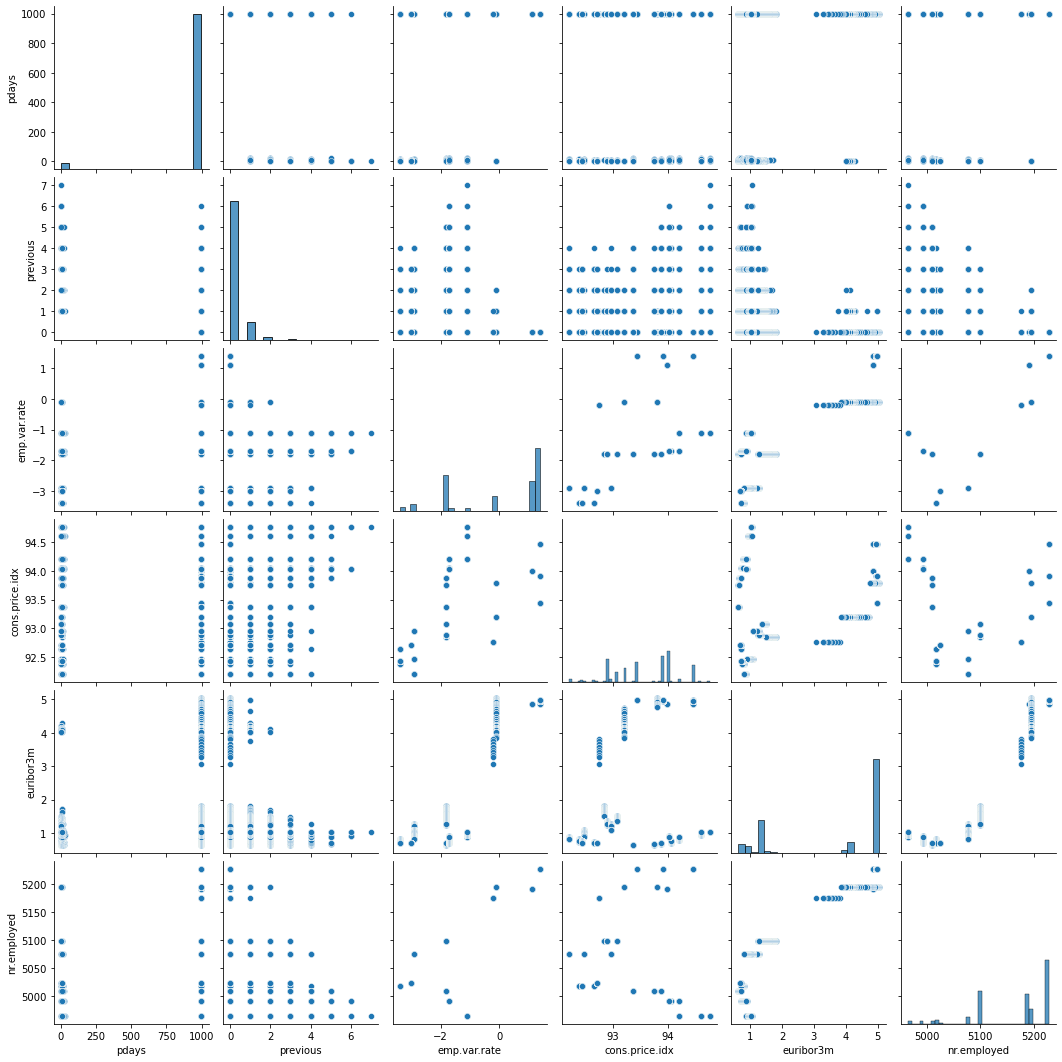


Figure 3: High-Correlation Numerical Features pairplot (larger image on accompanying Jupyter Notebook)

Toolbox(es) to use

The main tool to use in order to perform data processing, cleaning, feature selection, model building, model evaluation and assessment will be Python, via Jupyter Notebooks. I am envisioning to use the following Python libraries (not an exclusive list):

* Pandas
* Pandas Profiling
* Dora
* MatPlotlib
* Seaborn
* SciKit-learn

I will also use R/RStudio as a complementary tool, whenever is suitable.

Data Science Framework

This Capstone project proposal is already implicitly supported on the following Data Science Framework, to be followed throughout the project timeline:

* Definition of the goal & mission statement
* Data Gathering, processing & Management
* Predictive Model-building
* Model Evaluation
* Report results
* Model deployment & Maintenance

I am proposing this framework because of its intuitive sequence that can have well-defined milestones and checkpoints, while still allowing for iterative reviews at almost every step when needed.

Deliverables

I am committing to deliver the following items at the end of the Capstone project:

* One (1) annotated Jupyter Notebook containing all the tasks and graphical outcomes, outlined under the proposed Data Science Framework.
* One (1) executive summary/report in .doc format
* One (1) PowerPoint presentation, ready for oral delivery, if required.
* One (1) Pandas Profile Report of the raw data
* One (1) Pandas Profile Report of the final processed data.
* Raw and final processed data on .csv file format.

All these items will be made available via direct submission, and through a dedicated repository on my GitHub account, to be announced later: <https://github.com/lordtable/>

Data Copyrights

Citation Request:

***“This dataset is publicly available for research. The details are described in [Moro et al., 2014]. Please include this citation if you plan to use this database:***

***[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014.”***