**Capstone Project**

**Document Skeleton**

# Process overview

The following diagram shows the overall end-to-end process for defining, designing and delivering the Capstone project.



Note: The following are the candidate sections of the document. They are presented here for guidance. Questions in each section could be used as possible aspects to cover. Some questions may not be applied to each project. On the other hand, additional information may be needed.

# Problem statement

* What is the problem or the opportunity that the project is investigating?
  + After a decline in revenue for the Central bank of the Portuguese Republic (Banco de Portugal)
  + , a marketing scheme needs to be implemented. This will be based off the idea that their customers seemingly aren’t depositing as frequently as they have been previously. Long-term deposits allow financial institutions to retain investment capabilities for longer.
  + In terms of business problem, the aforementioned Portuguese bank would benefit in identifying which existing customers would have a higher probability of responding positively to a long-term deposit campaign (the success rate of telemarketing phone calls aimed at selling these long-term deposits.
  + [Can a Machine Learning model predict which clients are most-likely to successfully respond to telemarketing calls aimed at selling long-term deposits?]
* Why is this problem valuable to address?
  + It distinguishes whether or not the retail bank in question is able to successfully reach clients and sell them the campaign. More specifically, which clients are most financially viable to target (reduced opportunity cost, as well as financing marketing ploys aimed at the wrong clientele.
* What is the current state (e.g. unsatisfied customers, lost revenue)?
  + Lowered business activity, revenue decline, lack of long-term client engagement.
* What is the desired state?
  + Marketing selling campaigns are a typical strategy employed to enhance business engagement. By using this targeted marketing ploy, the retail banking firm is aiming to meet a specific goal, such as increased client long-term deposits.
* Has this problem been addressed by other research projects? What were the outcomes?
  + It would only be expected that the eventuated success rate of marketing models have been researched in the past.
  + This is a key example of the practical link between business consulting and data science, and how they go hand in hand

# Industry/ domain

* What is the industry/ domain?
  + Business, banking, marketing
* What is the current state of this industry? (e.g. challenges from start-ups)
  + Facing a lot of disruption as the digital age is exponentially changes business processes and client behaviour. Specifically the shift towards mobile/ online banking, away from having to go to a physical bank to reap benefits of trade.
  + COVID-19 further stimulating predominantly online banking resources. Efficiency gains from technology are of high value to the consumers.
  + High competition from disruptive start-ups and “neobanks” that are changing service expectations
* What is the overall industry value-chain?
  + Key activities: Processing of transactions, provision of financial advice and guidance, wealth/investment management, asset/liability management, maintenance of client relationships/sales, fiscal product/service development and execution.
* What are the key concepts in the industry?
  + Wealth management
* Is the project relevant to other industries?
  + Absolutely, although specific in the banking industry in this context, the evaluation of marketing campaigns is highly adaptable (and necessary!) to every other business entity. Predicting the success of such campaigns adds huge value in terms of increased revenue, clientele retention and reduced failed marketing investments.

# Stakeholders

* Who are the stakeholders? (be as specific as possible)
  + Clients, Board of Directors, Management, Employees, Shareholders/Investors, Authorities
* Why do they care about this problem?
  + The problem is based on the success of marketing for long-term deposits. The higher the success rate, the higher the financial investment actions of the banking entity. Increasing the banks investment capabilities, increases the services available to clients. For obvious reasons, shareholders will only benefit from increased revenue/ stability of long-term deposits ect.
  + Long-term deposit benefits for clients: increased saving capabilities, decreased investment volatility, increased interest rates.
* What are the stakeholders’ expectations?
  + Assuming clientele wish to gain competitive rates on their long-term deposit and shareholders expect to benefit from increase financial flow and business activites to promote the economic success of their institutional foundations.

# Business question

* What is the main business question that needs to be answered?
  + Can it be identified (the success rate of telemarketing phone calls aimed at selling these long-term deposits.
* What is the business value of answering this question? (quantify value and make necessary assumptions)
  + Increased long-term deposits = Significant increased financial investment capabilities. The bank can invest the long-term value in various other products that are deemed a higher RoR (rate of return)
* What is the required accuracy? What are the implications of false positives or false negatives?
  + Should be high accuracy. Incorrect predictability would result in the investment of unsuccessful marketing campaign, increased opportunity cost (missing out on better opportunities/client engagement), reduced revenue and investment capabilities, lowered financial certainty/ increased volatility.

# Data question

* What is the data question that needs to be answered?
  + [Can a Machine Learning model predict which clients are most-likely to successfully respond to telemarketing calls aimed at selling long-term deposits?]
* What is the data required to answer the question?
  + Survey data of the marketing campaign, attributes of customers on the receiving end, resulting response/action (successful, unsuccessful)

# Data

* Where was the data sourced?
  + UCI Machine Learning Repository
  + Based on a 2014 study: [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
  + The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.
  + Four datasets:

1. bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014]
2. bank-additional.csv with 10% of the examples (4119), randomly selected from 1), and 20 inputs.
3. bank-full.csv with all examples and 17 inputs, ordered by date (older version of this dataset with less inputs).
4. bank.csv with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs).   
   The smallest datasets are provided to test more computationally demanding machine learning algorithms (e.g., SVM).

* What is the volume and attributes of the data?

1. bank additional full: (41188, 21)

2. bank additional: (4119, 21)

3. bank full: (45211, 17)

4. bank: (4521, 17)

* How reliable is the data?
  + Used for significant research and cited in multiple publications, the data set seems highly targeted and reliable for relevant purposed.
  + Advocated through the University Institute of Lisbon, Portugal and Research Centre of the University of Minho, Portugal
* What is the quality of the raw data?
  + Seems very thorough and cleaned to some extent. Multiple unknowns are within the attributes and denoted as “unknown”
* How was this data generated?
  + I drew the data from the UCI Machine Learning Repository
  + The original data was obtained as per the following:
    - A Portuguese retail bank was addressed, with data collected from 2008 to 2013 (note the inclusion of the financial crisis.
    - A significant set of 150 features was analysed related with bank client, product and social-economic attributes.
    - A semi-automatic feature selection was initiatied in the modelling phase within the study, resulting in a reduced set of 22 atrributes.
* Is this data available on an ongoing basis?
  + Not that I am aware of, it was a very targeted study conducted from 2008-2014.
  + However, this data could potentially be replicated in correspondence to a multitude of corresponding marketing campaign across numerous industries.

**Attribute Information:**

Input variables:  
# bank client data:  
1 - age (numeric)  
2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')  
3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)  
4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')  
5 - default: has credit in default? (categorical: 'no','yes','unknown')  
6 - housing: has housing loan? (categorical: 'no','yes','unknown')  
7 - loan: has personal loan? (categorical: 'no','yes','unknown')  
# related with the last contact of the current campaign:  
8 - contact: contact communication type (categorical: 'cellular','telephone')   
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')  
10 - day\_of\_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')  
11 - 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 attributes:  
12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)  
13 - 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)  
14 - previous: number of contacts performed before this campaign and for this client (numeric)  
15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')  
# social and economic context attributes  
16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)  
17 - cons.price.idx: consumer price index - monthly indicator (numeric)    
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)    
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)  
20 - nr.employed: number of employees - quarterly indicator (numeric)  
  
Output variable (desired target):  
21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

# Data science process

## Data analysis

* What data pipeline was to wrangle the raw data?
* What are the highlights of the Exploratory Data Analysis (EDA)?
* Is the pipeline reusable? (for example, to process future data?)
* What are the intermediary data structures used (if any)?

## Modelling

* What are the main features used?
* Did you find any interesting interactions between features?
* Is there a subset of features that would get a significant portion of your final performance? Which features?
* How did you select features?
* What feature engineering techniques are used?
* What are the models used?
* How long does it take to train your model?
* What are the tools used? (cloud platform, for example)
* What are the model performance metrics?
* Which model was selected?

## Outcomes

* What are the main findings and conclusions of the data science process?

## Implementation

* What are the considerations for implementing the model in production?

# Data answer

* Was the data question answered satisfactorily?
* What is the confidence level in the data answer?

# Business answer

* Was the business question answered satisfactorily?
* What is the confidence level in the business answer?

# Response to stakeholders

* What are the overall messages and recommendations to the stakeholders?

# End-to-end solution

* What is the overall end-to-end solution to use the model developed in the project?

# References

* Where are the data and code used in the project? (show a simplified list of main items: notebooks, datasets, exported models)
* What are the resources used in the project? (libraries, algorithms, etc)