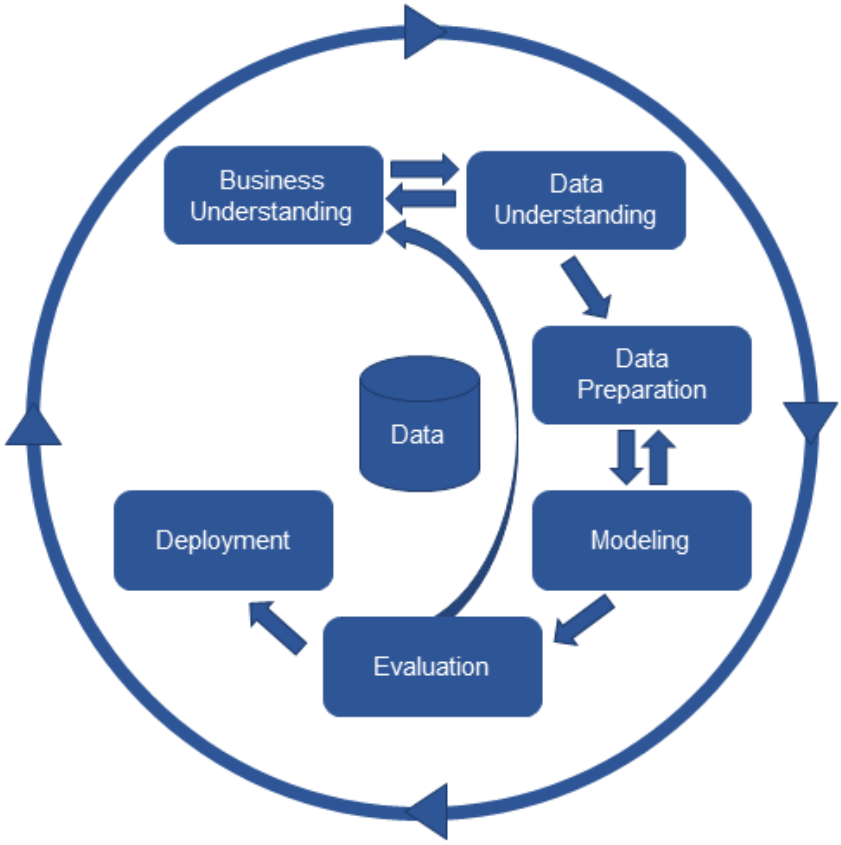
Data Scientist Application Exercise

1. Introduction:

The main purpose of this exercise is to test the applicant technical skills and familiarity with the Cross Industry Standard Process for Data Mining (CRISP-DM) which consists in 6 major phases:



1. **Business Understanding:**
   This phase focuses on understanding the project objective and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives.
2. **Data Understanding:**
   This phase starts with an initial data collection (the dataset we will provide you) and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.
3. **Data Preparation:**

This phase covers all activities to construct the final dataset (data that will be fed into the modeling tools) from the initial raw data. Tasks usually include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.

4. **Modeling:**
   In this phase, various modeling techniques and algorithms are selected and applied to the dataset, and their parameters are calibrated to optimal values.
5. **Evaluation:**
   At this stage it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives.
6. **Deployment:**
   The deployment phase can be as simple as generating a report and communicating results or as complex as implementing a model in a production environment.

2. Dataset:

The dataset attached ('dataset.zip') consists on two csv files: 'users.csv' containing the income, outcome and a label for 1000 users, and 'credit_reports.csv' containing all the accounts from the user's credit reports[1]. (Check the appendix A and B for the codebooks).

Each user is labeled as 1 if the user was a good client and 0 if the user was a bad client

3. Objectives:

In your role of Data Scientist you will be cycling around the CRISP-DM. Your objective with this exercise is to go through the data mining process using the provided dataset completing each step to finally communicate your results and work for every stage.

On the last step (deployment) you will need to answer the following 3 questions:

- Pick the best clients you will give a loan to, based on the model you created. It could be as complex as you decide (even as simpler as knock out rules), as long as the metrics support it
- Propose an amount to be lended to those clients and a term in which the loan will need to be paid back.
- Finally choose an anual interest rate the lended amount must have in order to be profitable.

---

[1] A credit report is a statement that has information about your credit activity and current credit situation such as loan paying history and the status of your credit accounts

Don’t forget to provide **justification** for you answers.
You are strongly encouraged to use whatever tools you know or are available to you. (e.g. Python, R, Excel, Weka).

For every step/phase you should include at **least** the following (Use them as a starting point):

- **Business Understanding:** By using the information of the label (column class), What characteristics differentiate between a good client and a bad client?
- **Data Understanding:** Describe the dataset and perform exploratory analysis over the different features/attributes to gather information about the data.
- **Data Preparation:** Perform cleaning, transformation (create new features), normalize and/or perform attribute selection on the data to prepare the dataset for the next step.
- **Modeling:** Choose (and explain why) a classification/regression algorithm and train it using your prepared data.
- **Evaluation:** Describe which metrics you are using to evaluate your model and why you choose them.
- **Deployment:** Communicate your findings by answering the previous three questions providing sensible justification for each one.

4. Deliverables:

Document or presentation showing the process and results of each step (Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment). Add discussion and conclusion if needed.

5. Tips:

- Include plots, diagrams and tables for easily explaining your findings.
- Keep your findings short and concise.
- Focus on generalization and the big picture.

This is not a test rather an exercise, therefore there are not right or wrong answers, we just want to see how your work and how you think, so have fun and enjoy the process!

6. Appendix:

A. users.csv codebook:

- id: User's unique identifier.
- monthly_income: User's monthly declared income.
- monthly_outcome: User's monthly declared outcome.
- class: Boolean value 1 if the client was good or 0 if bad.

B. credit_reports.csv codebook:

- user_id: User's unique identifier.
- institution: Institution granting the loan.
- account_type: Type of account for the institution.
- credit_type: Type of loan granted by the institution.
- total_credit_payments: Length of the credit (in amount of payments).
- payment_frequency: Frequency of the payments.
- amount_to_pay_next_payment: Amount to be paid on the next loan payment.
- account_opening_date: Date the account was opened.
- account_closing_date: Date the account was closed.
- maximum_credit_amount: maximum amount of credit used by
- the consumer.
- current_balance: Current balance needed to pay off the loan.
- credit_limit: Credit limit for this account.
- past_due_balance: Balance that is delinquent.
- number_of_payments_due: Number of payments that are delinquent.
- worst_delinquency: Worst delinquency (in payments) during the loan's life.
- worst_delinquency_date: Worst delinquency date.
- worst_delinquency_past_due_balance: Worst accumulated delinquent balance.