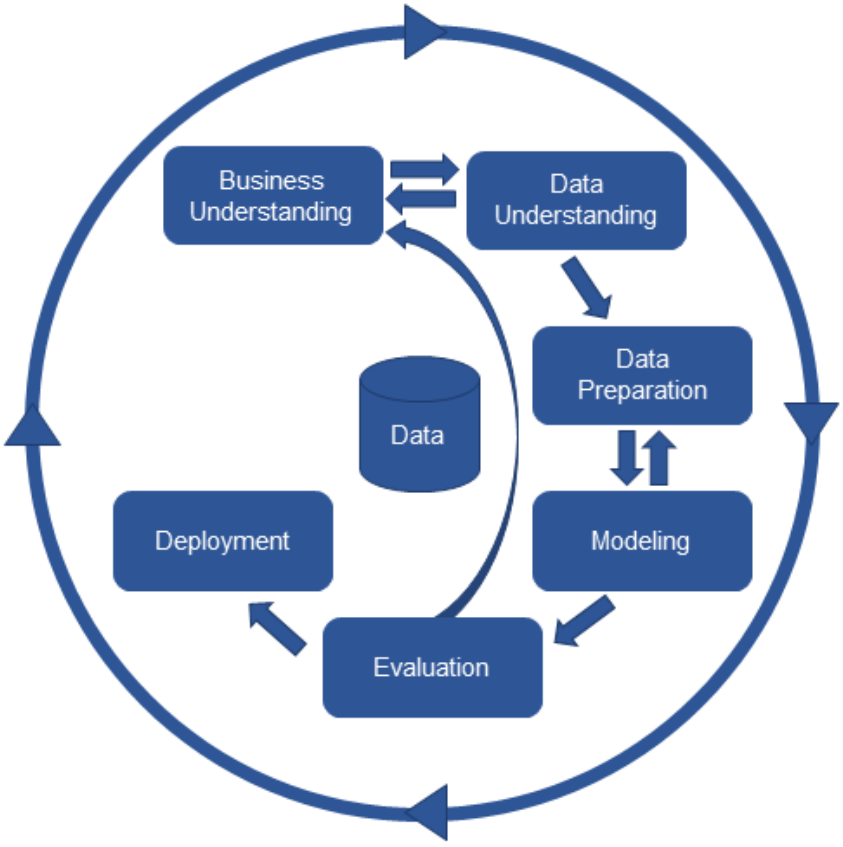
Data Scientist Application Exercise

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 problem statement and a 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 one csv file: 'intent.csv' containing the use of proceeds of 6,726 credit applications.

Each use of proceed is properly labeled (has a 1 in the respective columns/class, there are 9, check the appendix for the codebooks).

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.

Giving more context to the problem: during the evaluation process that occurs in our funnel, we ask the clients the reason they need the loan for. Understanding that, has many applications, so we considered the following approaches to classify them accordingly:

1. **Machine Learning**: worked with sentence level representations, and did a grid search over 4 different algorithms (SVC, LogReg, RandomForest, Multinomial NBayes). Used the Hamming Loss as metric to optimize.
2. **Deep Learning**: designed a simple CNN architecture to perform text classification. Considered this approach useful when our classes didn't depend on long term dependencies, in other words when our texts were short. It had the advantage that could learn to detect patterns in the text no matter where they are and also are very fast since we can parallelize the training and inference. It's hard to think CNNs applied in NLP, but could be used to find complex patterns in text and use them to generate powerful representations. Then we decided to have in the last layer as many outputs as labels we have, and treat each class as an individual probability using Sigmoid activation and for the loss function we used Sigmoid cross entropy which measures the probability error in discrete classification tasks in which each class is independent and not mutually exclusive. Finally worked with word level representations adding a fast text embedding to the preprocessed and lemmatized docs.

The best classifier was logistic regression using a Binary Relevance approach, this means we are training N classifiers where N is the count of labels, and treat each label as an individual binary classification problem. We achieve a loss score of 0.06821 and an average precision of 67% having the weakness score for the 'temp' class since is the fewer populated class.

In this experiment we see how a good representation such as Laser was capable to overpass more complex algorithms using a simple logistic regression and also add multi language support for our model.

Do you think you could do it better? Based on your interest in NLP problems and DeepLearning, we would want to see if you could design an ad-hoc architecture or use a novel approach. Not problem if you could not improve the metric this time. We would prefer to see how deep is your knowledge in this field. Please pin all the decisions you made down, so we could follow you.

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

- **Business Understanding:** understanding of the problem that should be resolved
- **Data Understanding:** Describe the dataset and perform exploratory analysis over the different classes to gather information about each of them to verify if some of them have less representation.
- **Data Preparation:** Perform cleaning, transformation (create new features), remove irrelevant characters, lemmatization, normalize cases.
- **Modeling:** You have the flexibility to experiment with various models and architectures, including cutting-edge deep learning techniques, hybrid models, or

ensemble approaches. However, use the following as a starting point: First, develop and train a deep learning model to create features from text inputs. Then, utilize these generated features to train a traditional machine learning classifier.

- **Evaluation:** Describe which metrics you are using to evaluate your model and why you choose them.
- **Deployment:** Communicate your findings and recommendations to enhance your approach in subsequent iterations.

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. Also attach the code used to generate your insights.

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. intent.csv codebook:

- motivos: use of proceeds
- cred: pay debts
- equ: buy equipment
- inic: start a business
- mkt: marketing

- no: use of proceeds not destined to working capital
- renta: rent
- sueldo: payroll
- temp: seasonal sales
- crec: grow (not an specific reason or plan was provided, but growing strategy was insinuated)