**INTRODUCTION**

Hello, this is my first experience putting my previous knowledge to practice. Since my former boss, told me that I need to find methodology in what I do, I took my time to find a very useful methodology for doing Data Science. It is called "Foundational Methodology for Data Science" developed by IBM, and the link to download it is here:

<http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=IMW14824USEN>

This methodology involves 10 steps. It was developed in a top-to-bottom form, as it is the general approach on the art of Data Sciences. First I describe the methodology my own words, and then I will go through a step-by-step walkthrough on our problem.

**0) Understanding the business:** This starts under the premise that the business sponsors need an analytical solution to a problem. In this process it is very important to keep in contact throughout the project. This part plays an essential role, hence the problem and the goal of it must be well posted and well understood.

**1) Analytic Approach:** After the problem is understood, we as data scientists must be able to find an analytical approach. We need to identify an analytical approach so as to have a top to bottom process. Assuming we are in the realm of supervised learning, we first need to identify whether the model is a classification or a regression model. Then we could choose one or more different models to rely on.

**2) Data requirements:** Step 1 determines this step. Here we identify which data do we need in order to address the analytical approach

**3) Data Collection:** Initially data scientists have to identify and collect data from the available resources. Normally we require additional investment to gather more specific data. However, it is advisable to postpone the investment decision until there is more knowledge about the data and the model. If we incorporate additional data, the predictive models can better represent events such as an earthquake or a disease.

**4) Data Understanding:** This is the prequel of Exploratory Data Analysis. This starts by doing some descriptive statistics and visualization to get the first insights about the data content. Be careful because additional data may be required.

**5) Data Preparation:** This stage covers everything that involves preparing the data for the modeling stage. Its activities involve data cleaning (deal with missing values, duplicates or formatting), combining data from multiple resources and transforming the data into the useful variables. This is the most time consuming step in a Data Science project. Up to 80% of the time could be consumed if the data is messy.

**6) Modeling:** This stage focuses on developing predictive or descriptive models, following the analytical approach. When we use a predictive model, we normally split the data to leave a minority for training the model, and a majority for testing the model. Here we get intermediate insight that leads into refinement. This leads into testing multiple algorithms aiming to find the best model. Hence this stage is highly iterative.

**7) Evaluation:** Between the development of the model and its deployment, we as data scientists must evaluate how well does our model performs. We have to understand its quality and ensure it properly addresses the business problem. Here we compute various diagnostic measures and outputs such as graphs and tables. In this part we begin to understand and interpret the model's quality and efficacy. Predictive models work on a previously trained algorithm with a known dataset (in the case of supervised machine learning). The testing part uses either created data that follows the same probabilistic distribution, or part of the original data that is left exclusively for testing and tuning. Furtherly we may need to assess our model with statistical significance tests for further quality proof. Practically this proof can be used to justify the model implementation, or taking actions when the stakes are high.

**8) Deployment:** Once a model is developed and approved by business sponsors, the model is tested in the business environment or a similar test environment which is enclosed or limited for further evaluations, until it is fully assessed. It may be as simple as making a report with recommendations, or as involved as in a complex workflow scenario. This usually involves groups with additional skillsets from within the enterprise.

**9) Feedback:** After getting results from the implemented model, the organization gets feedback on the model's performance and its impact on the environment where it was deployed. This stage also has some refining to improve its accuracy and usefulness. We can automate some of the feedback gathering and the evaluation, and redeployment to respond to the dynamism from the environment.

An important aspect of this methodology is to consider that this is a highly iterative process. We as data scientists usually have to go back to previous processes to make adjustments.