Company's Data Scientist Challenge

Introduction

Measurement of performance is vital to business success and most performance studies are related to the so-called key performance indicators (KPI). A KPI measures how well the organization is performing on operational, tactical or strategic activities. This is critical for the current and future success of the organization.

In the Oil and Gas industry, a KPI can be used to monitor and measure operational quality and crew performance. In a *Drilling* operation, the process where a hole is bored using a drill bit to create a well for oil and natural gas production, an example of KPI is the **Slip to Slip connection time**. This KPI can offer significant improvements in identifying sources of Non-Productive Time.

The **Slips** is a device used to grip and hold the upper part of a drill pipe to the drill floor, which is the area where the pipe begins its trip into the earth. The Slips are used when making a **connection**: the pipes are joined in order to advance further into the hole. Therefore, each pipe is picked up by a hook, temporarily gripped by the slips and then joined to another pipe. After the joint, the slips are removed and the entire pipe is carefully lowered into the hole, resuming the drilling. A skilled rig crew can physically accomplish all of those steps in a minute or two.

In the following video, you can have a better idea on how a drill pipe connection works. Video link: https://www.youtube.com/watch?v=QrPg8sMgRWk.

Note that, in the video, the **Slips** is the device added at 1:13 by the operator in a white T-shirt (image below). Also, at 1:48 the two pipes are joined together and the Slips is taken off at 4:32.

The challenge

In this challenge, you should do an **Exploratory Data Analysis** in order to extract useful information for the development of the KPI mentioned in the Introduction (Slip to Slip connection time).

We expect you to (not necessarily in this order):

1. Describe the data:
    a. Identify relations between variables;
    b. Identify the most important variables for detecting the when the slips is on or off;
2. Preprocess the data:
    a. Define and apply a strategy to treat missing values, if there are any;
    b. Define and apply a strategy to treat for duplicated values, if there are any;
    c. Usually sensor data has embedded some amount of noise. Analysis of such raw data may often fail to give accurate information. Define a strategy to deal with noise and explain your choice;
3. Machine learning methods are usually based on the assumption that the data generation mechanism does not change over time. However, some series in the data set do not present this characteristic, presenting change in statistical properties over time. Define a strategy to deal with this problem;
4. Build additional features that can help the detection of when the slips is on or off;
5. Propose and implement a model that can point out when the drill string is placed in slips and when it is taken off. Explain your proposition. What are the strengths and flaws of your model? It is important to emphasize that the model must be able to handle online inference;
6. Present the conclusions of your work.

Data

The data represents a multivariate time series collected by sensors during a **tripping** operation and is sampled at a rate of 2 samples per second. In this operation, the pipe connections happen in the same manner as explained in the Introduction section.

In the `data/` folder, you'll find the following files:

- `challenge_data.csv` - The data for the challenge.
- `challenge_annotation.csv` - Annotated data.
- `metadata.txt` - Some information on the data.

Solve the Challenge

For the development of the challenge, you must:

- Implement a **Jupyter Notebook** using **Python** as the programming language;
- Elaborate a presentation (**slides**) with your methodology and findings (**.pdf format**);
- Justify your choices.

All tasks should be done in **English**.