Uncover Relation of Technical and Non-technical Pilot Skills in Civil Aviation

Introduction:

CAE is a company which provides flight simulators as well as training solution for the airline companies. To improve our products and solution**, we can count on the huge amount of data that we are currently collecting through our simulators**. Eventually, it redirects us to key insights for the development and the improvement of our services. Today we would like you to help us improve our way to train pilots and make flight safer.

Problem:

A significant part of the aviation training tends to understand whether a pilot owns, or not, a specific competency. During a training session, after each maneuver, the instructor is supposed to tag the competency if the pilot does not comply with all requirements. **Unfortunately, it could happen that some maneuvers are mislabelled (the instructor simply forgot or tagged the wrong one)**.

We think it is possible to relabel correctly through machine learning algorithms. Thus, **the objective of this theme is to develop a model which, given the features of the simulation (altitude, speed, angle...), will return if the competency has to be flagged**.

We define 2 cases/labels:

- Label 0: Competency not flagged
- Label 1: Competency flagged

Challenges:

- **Time series analysis**
- **Under 500 samples**
- **Risk of mislabel data**
- **Unbalanced data (2:1)**

CAE

Description of the data:

The data that we are going to give you is in terms of a CSV file named **"*CAE_dataset.csv*"**. It contains time series of 10 different features (*velocity, altitude, different angles, etc.*) that we are calling *feature0, feature1, etc.* Another column is the Id of the pilot which will be a basic number. Finally, the last column is the label which indicates the status of the competency flagging. In general, **the maneuvers last between 1 and 2 minutes**.

index	Feature0	…	Feature9	Id	Label
0	4.56	…	2.56	0	1
1	3.55	…	4.2	0	1
…	…	…	…	…	…
202534	3,2	…	1,2	1	0
…	…	…	…	…	…
342554	2.89	…	10.4	345	1
…	…	…	…	…	…

Requirement and evaluation:

We also provide another file, named **"*CAE_test_dataset.csv*"**, which contains the test set we will use for evaluation. At the end of this challenge, we would like you to provide us a csv which will contain the label prediction on this test dataset as well as the corresponding id.

We will use the **F1 score** to distinguish teams. Also, we would like you to provide us with a quick description to explain your strategy, vision and methodologies.

Thus, we will need to receive two files:

- ***Team_Name*.csv**
- ***Team_Name*.txt**

Advice and recommendation:

Because of the nature of the problem, we advise you to combine both **supervised and unsupervised** algorithms. We have resampled the time series so that they get the **same frequency**. Finally, because of the risk of the mislabelled data, even poor results could be treated as significant results if they could be correctly explained.