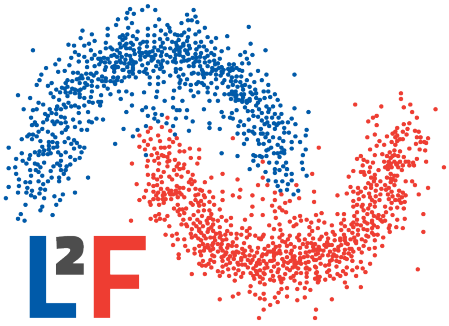

Modeling expert challenge

1 Description of the challenge

1.1 Introduction

Pokémon are small creatures that fight in competitions. All Pokémon have different numerical characteristics (strength of attack, defense, etc.) and belong to one or two so-called classes (water, fire, etc.). Professor Oak is the inventor of the Pokedex, a useful portable device that keeps information about all the Pokémon available. As his lead data scientist, you just received a request from him asking you to update the software on all Pokedex devices. In this challenge, you will work toward improving the Pokedex by

1. Exploring a dataset of Pokémon and the battles they fight
2. Developing a model for predicting the outcome of battles between Pokémon
3. Finding a method for ranking Pokémon by their overall strength.

1.2 Dataset description

Professor Oak has dumped the memory of one Pokedex device, resulting in the dataset you'll work with in this test. In the file pokemon.csv, each row represents the features of one Pokémon.

- pid: Numeric - ID of the Pokémon
- HP: Numeric - Health points
- Attack: Numeric - Strength of the regular attack
- Defense: Numeric - Strength of the regular defense
- Sp. Atk: Numeric - Strength of the special attack
- Sp. Def: Numeric - Strength of the special defense
- Speed: Numeric - Moving speed
- Legendary: Boolean - 'True' if the Pokémon is rare
- Class 1: Categorical - Pokémon class
- Class 2: Categorical - Pokémon class

Please note that a Pokémon can have either one or two classes. If a Pokémon has two classes, they are both considered to have the same importance.

In the file combats.csv, each row represents the outcome of one battle between two Pokémon.

- First-pokemon: Numeric - ID (match with pid)
- Second-pokemon: Numeric - ID (match with pid)
- Winner: Numeric - ID of the winner

2 Challenge

2.1 Part 1: Exploring the data (30 min)

1. Compare the probability distribution of the "regular attack" feature with that of the "regular defense" feature. In particular:
   (a) visualize the relation between these two variables using an violin plot;
   (b) list the names of the 3 Pokémon with highest attack-over-defense ratio;
   (c) list the names of the 10 Pokémon with the largest number of victories.
2. Professor Oak suspects that Pokémon in the **grass** class have a stronger regular attack than those in the **rock** class. Check if he is right and convince him of your conclusion with **Statistical** arguments. Note: If there are Pokémon that are in both classes, discard them.

2.2 Part 2: Prediction (40 min)

The model should take as input the features of two Pokémon and generate a binary value to predict who of the two will win.

1. Generate the feature vectors and the labels to train your model.
2. Train several models (at least a linear and non linear based model) to predict the winner of a match. And evaluate their performance using appropriate validation techniques and metrics.
3. Summarize and describe the results obtained with different models, highlighting the differences.
4. Find your best model(s). Motivate your choice.

2.3 Part 3: Ranking (20 minutes)

Pokémon tournaments can be represented as a so-called **dominance graph**. The dominance graph is a directed graph in which each Pokémon is represented by one vertex, and directed edges point from stronger to weaker Pokémons: in particular, if Pokémon i has won a strictly larger number of times against Pokémon j than j has won against i(that is, if i dominates j), there is a directed edge pointing from node i to node j. The dominance graph can be represented as an adjacency matrix G where entry G_{ij} is 1 if there is an edge pointing from i to j, and 0 otherwise. Summing up all elements of a row of G results in a **dominance score** for the respective Pokémon.

- Compute G and extract the 10 Pokémon with the highest dominance score.
- If a pokemon i dominates j and j dominated k, does i dominates k?

Good Luck !