Sports Analytics Challenge

- 2019 Analytics challenge organized by **Paris Saint-Germain** and **École Polytechnique**
- Data set based on French Football league matches. (Opta F24 files)
- Prizes: $100,000 funding, Top 5, Top 11, Top 50 prizes.
- Summary of challenge problem, deliverables, approach and result given below

Data

A learning database consisting of the F24 Opta files accurately describing all the ball events in all the games of the first part of the Ligue 1 season, 2016-2017.

Task

Train an algorithm (Python or R) on the first part of the season that will be tested on the second part of the season. 15 minutes of a game will be randomly selected in the second half of the season. The algorithm will have to return the identity of a player who has performed certain actions as well as predictions about the next event to take place on that game.

Test Procedure

- Randomly choose a match from the test database (F24 Opta files of all matches in the second half of the 2016-2017 Ligue 1 season).
- Randomly choose the first or second half for this match.
- Randomly pick 15 minutes of the selected halftime (all events between t and t+15 minutes with t randomly selected).
- Replace all the names of the teams by “1” (home) or “0” (away). In the files, it means replace “team_id” by “1” or “0”.
- Delete all the players IDs and write “0” instead, except for one randomly chosen player (who played more than 800 minutes on the learning dataset and did not change team in January). “1” is written for the ID of this specific player.
  - In the files, it means replace “player_id” by “1” or “0”.
  - In addition, when: Type ID=140, Type ID=141, qualifier_id=140 or qualifier_id=141 appears, the values are replaced by " ".
- Delete all position information (y, x) for all events except the last 10 Opta events within 15 minutes.
  - In the files, this means that replace “y” with “0” or “x” with “0”.
- Remove everything that is written in the F24 files before the first OPTA event. Replace the values of Event timestamp, Event id, Q id and version with " ".
- Reduce some of the information of the last 10 events. In the files it means:
  - For the last 10 events, “outcome” is replaced by " "
  - For the last 10 events, any information about qualifier_ID is gotten rid of.
    - Meaning : the “value” is replaced by " " and all the “qualifier_id” is replaced by " ".
- Sample test file with before and after test procedure provided.

Deliverable

The algorithm (in python or R) applied to the test base should allow one to: 1. Find out the identity of player 1.
2. For the next Opta event (first Opta event after t+15), find out if the team will be 1 (home) or 0 (outdoors).
3. For the next Opta event, find the associated position (y,x).

The algorithm must return (relatively quickly) a CSV file without a header whose components are four real numbers; Player ID, 1 or 0 for the team at home or away, Y and X.

The submitted code should be quick and short. Calculation time on test file should be 5 seconds' maximum, for the test base file on a laptop computer.

Evaluation

50% of the assessment will be based on the first question and 25% on each of the other two. The success score for the first 2 tasks will be the percentage of correct answers. For the last task an average error in standard L2 will be calculated. The final ranking will be based on the weighted average of the rankings for the 3 tasks.

Approach

- F24 XML files inspected and parsed to based on desirable features. Subsequent feature engineering considered eventual model training results and required multiple iterations.
- 3 different models were built for each deliverable due to nature of data and desirable model performance.
- Algorithms used were Random Forest, CART, Boosted GLM and Adaptive Mixture Discriminant Analysis.

Results

- Top 11 selection out of over 3000 submissions.
- Interview with organizers and data professional from Paris Saint-Germain and École Polytechnique.