Acumen 2022 Case Document

EXL

December 2022

EXL

Predict the winner of soccer matches

With the world in the grips of football fever as the FIFA World Cup hits its stride, it's only fair that we too play ball and get with it. This year's ACUMEN problem statement provides participants with an opportunity to gaze into the crystal ball and try their hands at predicting winners for football games.

Background

EXL's client ABC is a fantasy sports platform. They organize multiple forecasting-based games for different sports where users can predict winners, win points & redeem prizes.

Problem Statement

ABC is planning to roll out a 'Predict & Win' competition for PFL 2023 (Premier Football League). They require EXL's help to build a forecasting model that can predict how the upcoming season might unfold. This will help them decide what odds to offer for each game so that the users are rewarded points appropriately for their predictions. The task for EXL is to utilize 10 years of historical data (2012-2021) to predict PFL game winners for the 11th season (2022).

PFL is a football competition involving 20 top clubs from Europe. There is no relegation/promotion, and the set of teams remains same across seasons. The league is round-robin styled with no knockout stages. A PFL season typically runs from February to November with each team playing 76 games (playing all 19 other teams 4 times each in total: twice at home, twice away). All away (visitor) teams travel to home team's stadium for the games and return to their home stadium/city to prepare for the next game.

Each game has a winner. There are no draws. If scores are tied at end of 90 minutes, golden goal-based overtimes are played until one team scores and wins.

The Clubs have strict 'Cantera' policies meaning that they can only recruit players that are local / home-grown. Foreign players / player trades are not allowed. All this ensures that team squad strength (player quality) doesn't change dramatically overnight / across seasons.

Foreign managers are allowed but clubs are not allowed to sack managers or modify squad during the season.

Financial Fair Play rules forbid extravagant investments ensuring that team financials also don't change drastically overnight / across seasons.

EXL

Guidelines

Round 1

The participants are provided with 10 (2012-2021) seasons worth of data. Each game is identified by a unique game_id. The 'Events Data' files contains results of games from 2012-2021. The 'Game Metadata' file contains metadata for the training period (2012-2021) as well as test period (2022). File 'Test Data' contains the fixtures from 2022 season that need to be predicted for and shared back with EXL.

Participants will need to forecast the win probability for home team for all of 760 fixtures from the 2022 season. They should share the output / forecasts as an excel with two columns: 1) game_id and 2) win probability for home team (Refer file: 'Submission Template").

- **Events Data:** Game level characteristics for all games from 2012 to 2021
- **Game Metadata:** File containing details about each team, manager, manager history & referee.
- **Test Data:** Game level characteristics for all games of 2022. Participating teams have to predict the winning probability of the home team for each game of 2022.
- **Data Dictionary:** Attributes of all data elements
- **Submission Template:** Format of the final submission template

Along with the output, participants should also share their code-files (used for data ingestion, cleaning, transformation, feature creation, modeling, scoring & other steps involved)

Participants should only use the data shared by EXL for solving this problem statement. The data is decoupled from real-world context. We encourage not making assumptions outside of the data provided. (For eg. England clubs will beat Greek clubs because in the real world, that has been the case generally – is not a valid assumption. Similarly, COVID did not impact anything in this case study since the case study universe is decoupled from real world).

Submissions will be evaluated to select top X teams (undecided) for the next round.

Round 2

The shortlisted teams are required to submit a PowerPoint presentation (PDF) of maximum 8 slides covering:

- Data Exploration / Analytics
- Variable Concepts Explored / Feature Creation
- ML Pipeline / Algorithms
- Summary of Results
- Important Factors affecting Win Probability
- Unexplored Topics / Other Interesting Findings

Round 3: National Finals – Stage Round

Teams selected for National Finals will be invited to present their solutions to the panel of judges.