Soccer Guru:
a football match prediction pipeline

Fernando Miñaur

Marco Ramos

Miguel Calvo

Julio Lezama (*missing in action*)

Motivation & Business Problem

Create a platform that is:

- Free
- Transparent
- Accessible
- Well documented
- Open source (Github)

So that passionate people of this sport could bet on the results of football matches.

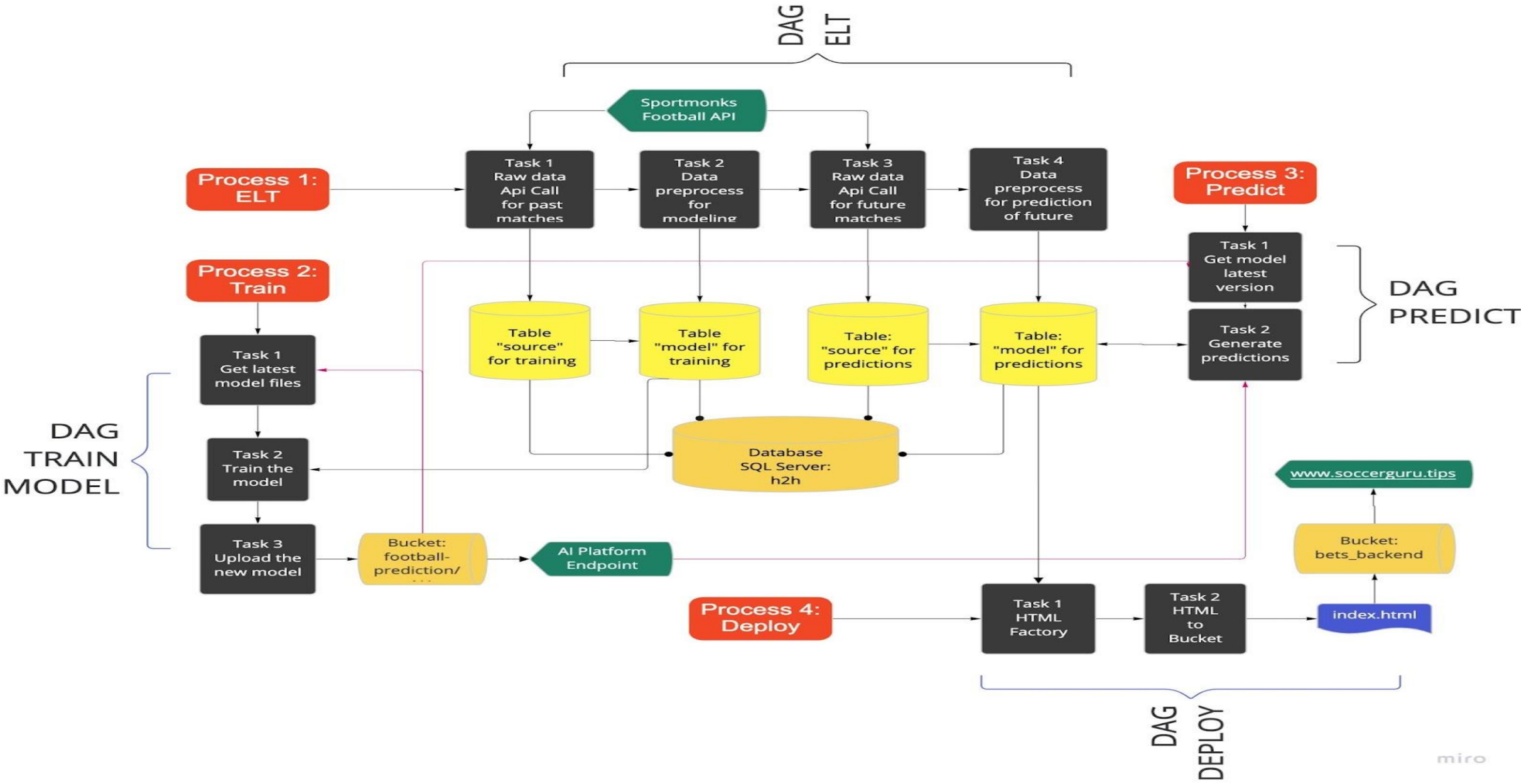
The problem that we wanted to solve throughout the project was to predict the result of soccer matches with accurate information using historical data of past confrontations.

Machine Learning

Historical data in the database can be used as an historical developement of a football team overall.

Supervised Machine Learning: known target variable (Winning team id)

System Design



Data Ingestion

sportmonks

Premier League

BUNDESLIGA

SERIE A

LaLiga

Selected features

Features

Categorical

league_id	Id of the league f.e. Copa del Rey
season_id	Id of the season f.e. 2022
venue_id	Id of the venue f.e. Santiago Bernabéu
referee_id	Id of the referee
localteam_id	Id of the local team
visitorteam_id	Id of the visitor team

Numerical

localteam_position	Current position of the local team in the table
visitorteam_position	Current position of the visitor team in the table

Feature Engineering

Four new variables:

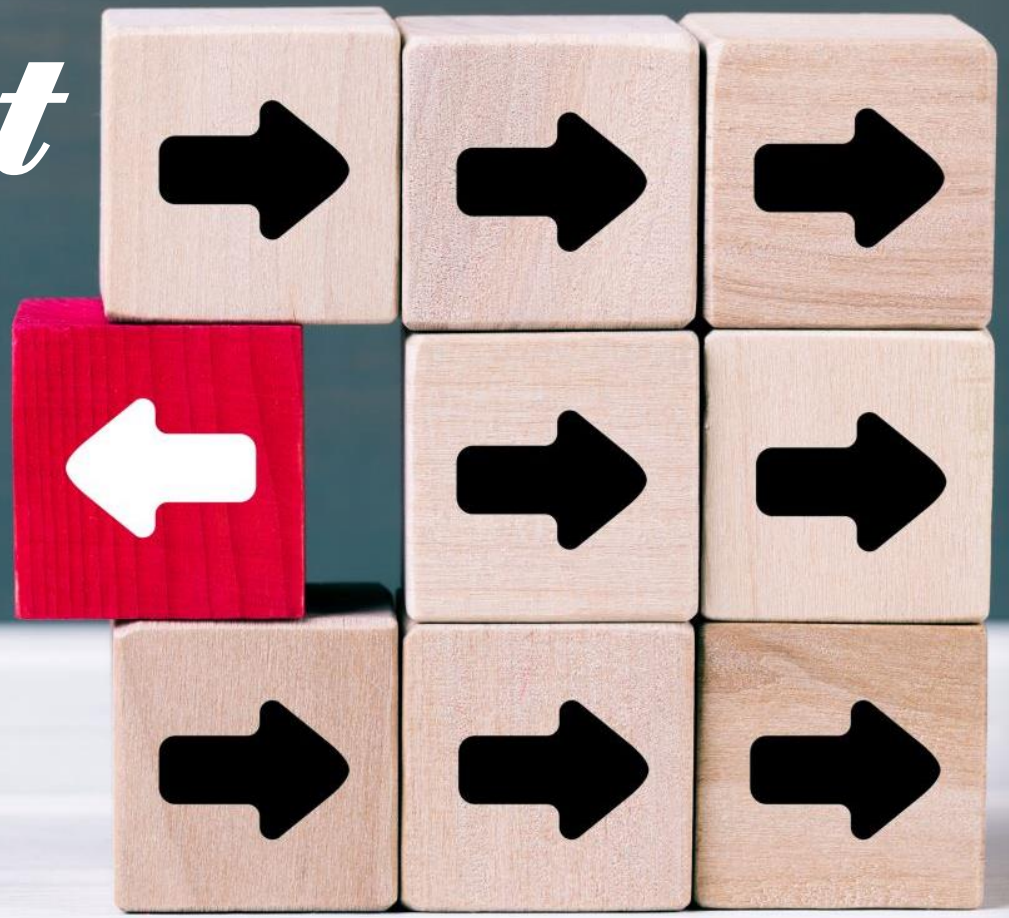
Weighted goals scored by local team

Weighted goals scored by visitor team

Current winning streak local team

Current winning streak visitor team

Model Development and Evaluation



First experiments

- Logistic Regression

Train Accuracy	Test Accuracy
63.58%	61.28%

- L1 Logistic Regression

Train Accuracy	Test Accuracy
62.14%	61.14%

- Random Forest

Train Accuracy	Test Accuracy
65.71%	66.2%

Best model XGBoost

Train Accuracy	Test Accuracy
77.34%	74.29%

Best Local Model: Random Forest with feature engineering

Train Accuracy	Test Accuracy
84.83%	85.65%

Final product

- We developed a webpage since the product is meant to be for a final user rather than a machine.
- The webpage can be found in: https://www.soccerguru.tips/index.html and it will show the next games with its respective probabilities.

	localteam_win_p	localteam_flag	localteam_name	visitor_tie_or_win_p	visitorteam_flag	visitorteam_name	match_day
0	0.614		Newcastle United	0.386		Arsenal	2022-05-16
1	0.681		Juventus	0.319		Lazio	2022-05-16
2	0.681		Sampdoria	0.319		Fiorentina	2022-05-16

References:

- -Sze Yeung. *Football Prediction by XGBoost*. Kaggle https://www.kaggle.com/code/szeyeung/football-prediction-by-xgboost/notebook?scriptVersionId=94437712
- -Utikal, N. (26 December 2019). *Predicting Football Results with Random Forest*. Obtenido de Medium: https://medium.com/@nicholasutikal/predict-football-results-with-random-forest-c3e6f6e2ee58
- -Calvo, M., Miñaur, F., Ramos, M., & Lezama, J. (3 de February de 2022). *Github*. Obtenido de Ubiquitous-Goggles: https://github.com/JulioLezamaAmastalli/ubiquitous-goggles/
- -Densmore, J. (10 February 2021). *Data Pipelines Pocket Reference: Moving and Processing Data for Analytics*
- -SportsMonk. From SportsMonk: https://docs.sportmonks.com/football/