SPORT ANALYTICS

Fall 2019 Presentation

Wednesday December 4th, 2019

Antoine Bachet, Brittany McLellan, Maxime Laharrague, Thomas Gaillard, Yiwen Sun

Context

2660
Games to analyze (38 matches per team per season, 7 seasons)

71
Predictors for each game (including fouls, corners, cards, ...)

505
Players from 20 different teams (65.2% of whom are not from the UK)

£8.39B
Total cumulated market value of PL teams for season 18/19

Objectives of the project

The project core objective is to **predict** the result of English Premier League soccer games as well as the **final rankings** at the end of a season.

We will rely on a Bayesian approach to establish **team rankings** and develop a **predictive model** on top of it to generate a win, draw and loss probability for a given game.

Benefits of the project

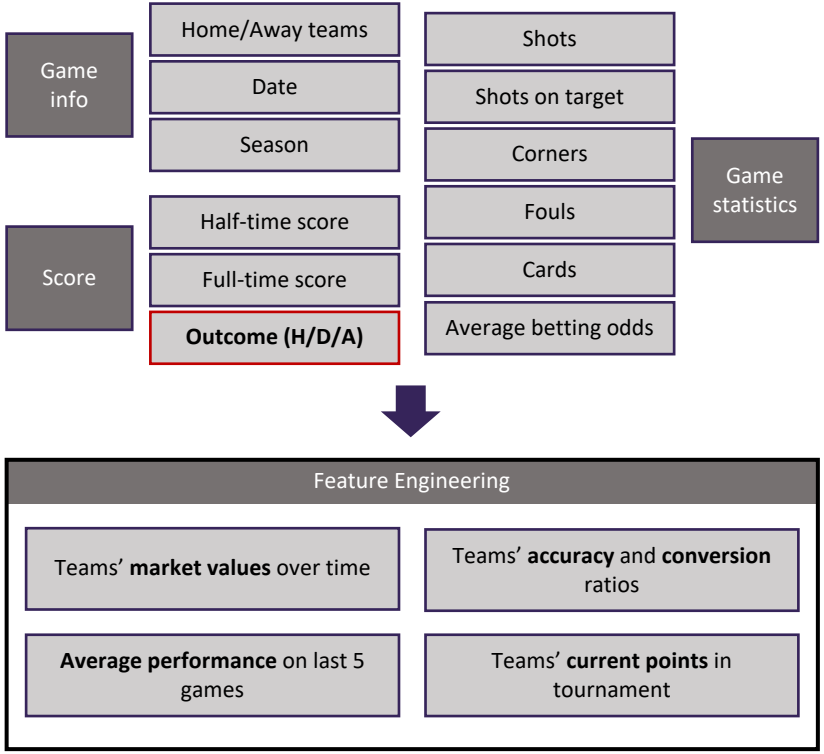
Team Management: Optimize team training program to ensure perfect fit for key decisional games and determine when to substitute the B team

Club-level decisions: Better quantify the market value of a scouted player and shape the club's willingness to pay by estimating the impact of the acquisition

Betting strategy: Provide probabilities estimation for gamblers willing to increase their chances over the teams they support

Columbia Business School

Discovering the data



Methodology

1

The first step of our approach will consist in establishing **team rankings** for each season in our data, that we can further reuse for game/season outcome prediction as we did in class.

For each season, we aggregate previous seasons' data to form a **prior** estimator and then use a **Bayesian approach** to formulate an estimation of the **posterior ranking**.

We then use this to compare the **performance of different models** with home and/or away advantage at predicting games outcomes.

2

In parallel to this, we build additional **intensive features** aiming at capturing more accurately **momentum** in one team's performance: past 5 games averaged shot accuracy, shot conversion, etc.

Once data is cleaned and normalized, we try different **supervised-learning models** to predict probabilities of win/draw/loss for each game, and choose the best performing one.

We then incorporate rankings and predictions from the Bayesian approach to the model, as well as teams' **market value** at game time.

Premier League

Sport Analytics – Predict English Premier League games outcome and final season rankings

3. Bayesian approach for team rankings

Columbia Business School

1. Simple rating model with Home advantage for each season

Year	RMSE
2013	1.486
2014	1.589
2015	1.470
2016	1.518
2017	1.558
2018	1.871

2018 Rankings			
1	Man City	11	Man United
2	Liverpool	12	West Ham
3	Chelsea	13	Bournemouth
4	Arsenal	14	Newcastle
5	Tottenham	15	Southampton
6	Wolves	16	Crystal Palace
7	Watford	17	Burnley
8	Leicester	18	Fulham
9	Brighton	19	Huddersfield
10	Everton	20	Cardiff

We try to predict the **outcome of each game** using teams' **ratings**, obtained by **minimizing the RMSE** on all game predictions.

2. Bayesian prior on past seasons to account for macro trends

Games	Test
380	304

Model test (4,3)	Model SR	Model Bayes
RMSE	47.3%	46.1%
Model SR		.3%
Model Bayes	99.7%	

We add a term for **Bayesian mean reversion** based on all previous seasons of data.

3. Simulation over many games to estimate current season ratings

Win league: **rank 1** in points

Champions league qualify: **top 4** in points

Relegated: rank **>= 18** in points

Team	Win League	Champions	Relegation
Man City	52.5%	99.1%	0.0%
Liverpool	37.0%	98.4%	0.0%
Chelsea	7.3%	84.4%	0.0%
Arsenal	2.1%	59.7%	0.0%
Tottenham	1.1%	45.1%	0.0%

We perform a **simulation** of 1000 trials for 3 probabilities: **winning** the league, **qualifying** for UCL and being **relegated**.

4. Predict season final rankings & create a feature from this rating

2019 Rankings			
1	Liverpool	11	Burnley
2	Leicester	12	Crystal Palace
3	Man City	13	Bournemouth
4	Chelsea	14	Brighton
5	Tottenham	15	Newcastle
6	Man United	16	Everton
7	Sheffield United	17	West Ham
8	Wolves	18	Southampton
9	Arsenal	19	Norwich
10	Aston Villa	20	Watford

We use the ratings obtained with this training phase to **predict outcomes** for the current season's **games** & **final rankings**.

Premier League

Sport Analytics – Predict English Premier League games outcome and final season rankings

4. Feature engineering & selection

Columbia Business School

Web scrapping

Cut-off date: Nov 15, 2019 Show

#	Club	Current value
1	Manchester City	£1.15bn
2	Liverpool FC	£965.48m
3	Chelsea FC	£726.30m
4	Tottenham Hotspur	£886.50m
5	Arsenal FC	£621.68m
6	Manchester United	£677.93m

Teams' market value on 11/15/19

Feature selection

- We want to keep only **intensive features** that will be available to predict future outcomes.
- We get rid of features related to **fair-play** (ie fouls, cards, etc.) and **game facts** (corners, free-kicks, etc.).
- We also keep the **average odds provided by betting websites** at the time of a game.

Feature Engineering

1. Filling in for teams without market value

For each season, some teams do not have market value because they just got promoted/relegated.

We tried filling the missing values with
- Average
- Median
- **Minimum value** (worked best)

2. Adding conversion & accuracy features

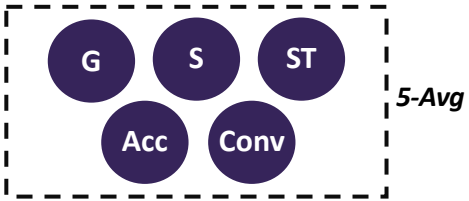
Creating **ratio features** better fitted and scaled to **comparing performance between teams**: goal conversion and shot accuracy.

$$\text{Accuracy} = \frac{\#\text{ Shots on target}}{\#\text{ Shots}}$$

$$\text{Conversion} = \frac{\#\text{ Goals}}{\#\text{ Shots}}$$

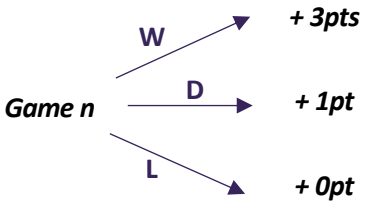
3. Creating momentum features

We build features aiming at capturing the team's **momentum** based on its **last 5 games** (with a minimum of 2 games played in the season).



4. Championship points counter

To determine **final season** rankings, we implement a **point counter** to track the number of points after every game played by a team.



Premier League

Columbia Business School

Model

1 Logistic Regression → **2 Random Forest**

Description

The **logistic regression** is the equivalent of a linear regression for when the variable to predict falls into **categorical buckets** (ie is not continuous).

Here, we use all previous features to predict the *RESULT* dependent variable which has **multiple classes**: Win / Draw / Loss.

Logistic regression is a **simple and interpretable model** and its results are easy to convert into applicable **business decisions**.

The random forest is a more complex classifier made of **decision trees** where **the features are randomly sampled,** and which outputs the **most frequently predicted class** among all trees.

This model is more likely to accurately extend its predictions to **new teams** in the league and will work better if one of the 3 **classes** is less represented.

Random forests **combine** multiple hypothesis to form a better one and **generalize** better to new data.

Results

73%	Of game outcomes well predicted on **training** seasons (2012/13 to 2017-18)	**77%**	On **training** seasons	**80%** On **training** seasons
52%	Of game outcomes well predicted on the **test** season (2018/19)	**55%**	On **test** season	**60%** On **test** season

With Bayesian ratings

Premier League

Columbia Business School

1 Naive Strategy

- For **each match** of the season bet $1 on the **predicted outcome.**
- The money made is based on the **true odds** of the match.

We bet **$1** on **all 360 games**

Expected return: **$232**

ROI: +64.4%

2 Odd-based Strategy

- Hypothesis that the **inverse of the odd** is their probability.
- If our max probability is **higher than the odd's** max probability, we bet $1 on our prediction.

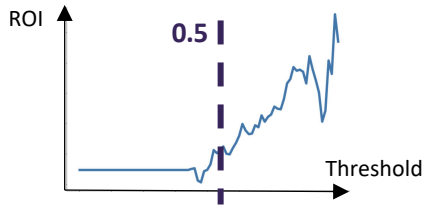
We bet **$1** on **only 142 games**

Expected return: **$95**

ROI: +66.9%

3 Threshold-based strategy

- If the **highest prediction probability is above a threshold,** we place a bet.
- We bet on **each outcome** by placing a regular amount (eg $1 per betting)



ROI: +75.9%

Premier League

Thank you for your attention!

Any questions?

Antoine Bachet, Brittany McLellan, Maxime Laharrague, Thomas Gaillard, Yiwen Sun