Building a machine learning model for sport betting

Syllabus

[1 Business Understanding 2](#_Toc497229708)

[1.1 Objectives 2](#_Toc497229709)

[1.2 General concepts in Sport betting 2](#_Toc497229710)

[1.2.1 Odds calculation 2](#_Toc497229711)

[1.2.2 Betting strategies 2](#_Toc497229712)

[1.3 Why Football? 3](#_Toc497229713)

[1.4 The bet type 3](#_Toc497229714)

[1.5 Selecting the competition 3](#_Toc497229715)

[1.6 The Model 4](#_Toc497229716)

[1.6.1 Data Sources 4](#_Toc497229717)

[1.6.2 Outcome 5](#_Toc497229718)

[2 Resources 6](#_Toc497229719)

# Business Understanding

## Objectives

The main goal of this project is to build a simple and efficient model that will be able to predict the outcome of future football matches. In a second phase, this model will have to be accurate enough to be used in sport betting, with a positive expectancy.

From a personnal point of view, the goal of this project is to build knowledge on all the aspects of a data mining project, including machine learning and project methodology

The business perimeter is sport betting. Let’s first introduce general concepts and then focus on our strategy and the data.

## General concepts in Sport betting

Sport betting is the activity of predicting sport result and placing a wager on the outcome. The vast majority of sport betting is done online. The main betting websites are: Unibet, Betclic, Winamax, Bwin and pmu.fr .

### Odds calculation

The odds provided by the bookmakers are calculated in two steps:

1. **Statistical analysis**

* Last matches outcomes
* Sportsmen health
* The venue
* The news coverage
* Other specific criterias according to the sport

1. **Refining the odds**

The odds are then refined taking into account:

* The bookmaker margin
* The bets made by the customers

In average, the bookmakers distribute between 80% and 90% of the total amount of bets. The odds are adjusted to make sure that the bookmaker will always make a benefit in the process.

### Betting strategies

An obvious point that all the good strategies have in common is to have good knowledge in the field in which one is going to bet. The main strategies are, from the simplest to the most complex:

**Baseline strategies**

* Always bet on the favorite: Safe, but with very little gains overall
* Always bet on the outsider: Risky, with potentially high gain or loss, high variance,

**Following best bettors**

* Some people are known for their skills at betting. A simple strategy is to replicate their bets**Arbitrage betting**

**Betting arbitrage**

Betting arbitrage, miraclebets, surebets, sports arbitraging is a particular case of [arbitrage](https://en.wikipedia.org/wiki/Arbitrage) arising on betting markets due to either [bookmakers'](https://en.wikipedia.org/wiki/Bookmaker) different opinions on event outcomes or plain errors. When conditions allow, by placing one bet per each outcome with different betting companies, the bettor can make a profit regardless of the outcome. Arbitrage is an extremely fast-paced process and its successful performance requires lots of time, experience, dedication and discipline, and especially [liquidity](https://en.wikipedia.org/wiki/Liquidity" \o "Liquidity).

**The statistical method**

From the least important to the most important

1. Current performance of the team: analyzing the 6 last matches played by the two teams
2. Follow the news: Transfers, rankings, last matches, missing players, etc.
3. Statistical analysis of the competition
4. Experience/Feeling

The statistical method provides good indicators that can be used for building a machine learning model.

## Why Football?

Football is by far the most popular sport in Europe, totaling $24.6 billions of revenue for the 2015/2016 seasons. This is the same situation in sport betting industry: In France, football betting accounts for two thirds of online betting.

Moreover, there are regular matches and competitions all over the year: some teams plays up to 2 matches a week. All these conditions will guarantee an interesting amount of historical data, statistical data, and news coverage.

## The bet type

The main bet types are:

* Match: Bet on the outcome: Team1, Draw, or Team2
* Total Goal: Bet on the number of goals, for example “over 2.5” or “under 2.5”
* Both Team To Score: “Yes” or “No”
* Half-Time: Bet on the half-time: Team1, Draw, or Team2
* Total Goal – First half

The “Match” bet seems to be the easiest to understand and predict. Other interesting bets are “Total Goal” and “Both Team Score”, as they are “yes” or “no” bets. The outcome is simple, but the explanation seems to be much harder than just predicting the match outcome. Thus, these are bet types where machine learning could potentially bring a huge value.

At first, we will focus on the “Match” bet. Then, according to the feedback, we might consider studying the other interesting bets: “Both Team Score” and “Total Goal”. They will likely be based on the same dataset.

## Selecting the competition

The selection criterion is to have the maximum amount of data available, so the questions we are going to ask ourselves are:

* What are the biggest competitions? Focusing on the biggest competitions will guarantee more data available, more news coverage, etc.
* What is the average match frequency? More observations to build our model. The competitions where matches take place once a week are better that once every 3 month. It enables us to test our model in the short term, and adjust it on a regular basis. Thus, the future calendar of the competitions is important.
* Are the teams going to play the whole competition, or are there eliminations? Competitions without elimination (like national championships) allow us to follow teams throughout the competition.

Considering the above criteria, we have two possibilities:

**1-National football teams:**

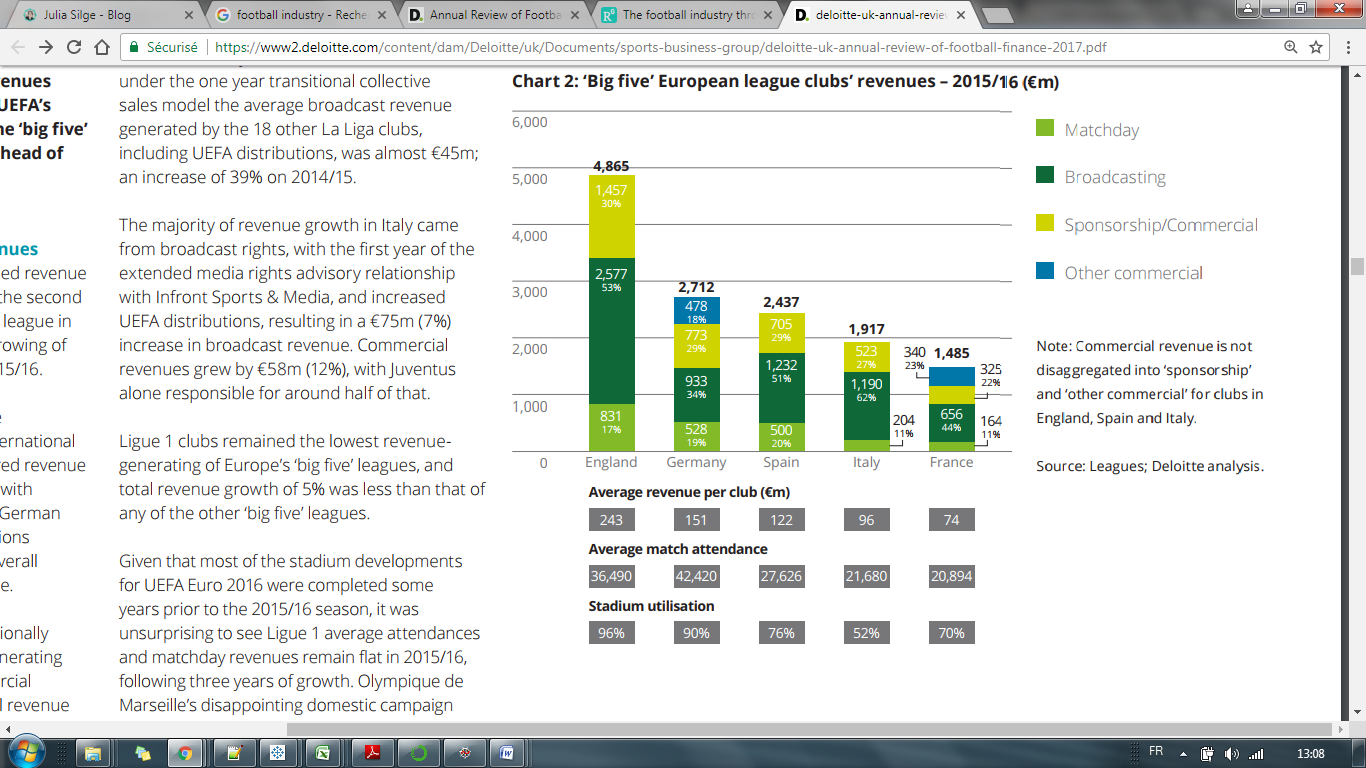
There is a faire amount of data if we take into account all the competitions and the friendly matches.

The interesting point is that it is a field that seems to have very little variance in the results compared to national championships. For example, when a national team is in good shape, it is generally over a long period of time, like several years.

Another point is that the national teams have a huge amount of players that can be selected to play a match. This will minimize the impact of wounded or unavailable players and smoothen the performance over time.

**2-European National Championships**

The main national championships in Europe are called the “Big Five” European Leagues. Let’s assume that the data available for each championship is positively correlated with the amount of money they generate. In the chart below, we see that the English “premier league” is by far the biggest market. Thus, it is the most interesting.



## The Model

### Data Sources

* Google Cache for historical data

#### FIFA Ranking

* FIFA Ranking, Age, Value, Confederation History by Date : [www.transfermarkt.com](http://www.transfermarkt.com)
* FIFA Ranking History by Team : [www.fifa.com](http://www.fifa.com/fifa-world-ranking/ranking-tools/compare-teams.html#dataform=true&t1=GER&t2=BRA&t3=FRA&t4=ARG&timeRange=fifarankingcreation&fromDate=1993-08-01&toDate=2017-10-16)
* FIFA Ranking History by Date : [FIFA](http://www.fifa.com/fifa-world-ranking/ranking-table/men/rank=273/index.html)
* FIFA Ranking History by Date: [wildstats](http://wildstat.com/p/7001/ddate/1999-11-17)

#### ELO Ranking

* ELO Ratings : [eloratings](http://www.eloratings.net)
* ELO Ranking by Date (Club) : [clubelo](http://clubelo.com/2017-10-01/Ranking)
* ELO Ranking API (Club): [clubelo](http://clubelo.com/API)
* ELO Ratings before each World Championship : [Wikipedia](https://en.wikipedia.org/wiki/World_Football_Elo_Ratings#Elo_all-time_records)

#### UEFA Ranking

* UEFA Ranking History by Date (club coefficient) : [wildstats](http://wildstat.com/p/7003)

#### Matches

* Match History by Team (National) : [RSSSF](http://www.rsssf.com/tablesl/landen-intres.html#euro)
* Match History All Team & Date (National): [soccer-db](http://www.soccer-db.info/index.php?option=com_php&Itemid=224)
* Match History All Team by Year (National): [theroonba](http://www.theroonba.com/football/men/2017.html)
* Match History All Team by Year (National) : [wildstats](http://wildstat.com/p/6)

#### Championships

* National Championship results by year (Club) : [wildstats](http://wildstat.com/p/26)
* Club Tournaments Results by Year (Club) : [wildstats](http://wildstat.com/p/5)
* National Championship Results by Year : [football-data.co](http://www.football-data.co.uk/englandm.php)

#### Stadium

List of European Stadium : [RSSSF](http://www.rsssf.com/miscellaneous/eur-stadium.html)

#### News

* Twitter API
* Google News

### Outcome

The model outcome will be 3 classes:

* Team1 :Team 1 wins
* Draw
* Team2 : Team 2 wins

# Resources

<http://www.sv-europe.com/crisp-dm-methodology/#businessunderstanding>

<http://www.dummies.com/programming/big-data/phase-1-of-the-crisp-dm-process-model-business-understanding/>

<http://www.pariezmieux.com/methode_statistique.html>

<http://www.pariezmieux.com/aide-aux-pronostics.html>

<http://www.abcargent.com/paris-sportifs/>

<http://www.soft-concept.com/surveymagazine/les-statistiques-et-les-paris-sportifs/>

<https://www.kelbet.com/comment-devenir-trader-paris-sportifs/>

<https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/sports-business-group/deloitte-uk-annual-review-of-football-finance-2017.pdf>

<http://lasek.rexamine.com/football_rankings.pdf>

<http://www.football-rankings.info/p/elo-ratings.html>