COURSERA CAPSTONE FINAL ASSIGNEMENT

Are common betting strategies successful?

DECEMBER 19, 2020

VICTOR N.

Contents

I) Introduction

Please note that Foursquare developer API is not properly working in my geographical are, thus cannot validate my account. Has been several days/weeks and my email cannot be validated so I decided to produce another kind of analysis, using what we learnt during these data science modules.

As mentioned at the beginning of this Coursera & IBM lesson, data science is useful in various jobs and areas. The only thing that matters is us to find our passion within data science and use as much as we can what we have learn during this course.

Therefore, I hope you will understand I had to do something a bit different from what has been initially asked for this final project, as I cannot use Foursquare API.

Thank you very much for your understanding and your review!

Victor

II) Business Problem:

Most of the people trying to get money through sport betting are applying a similar strategy. We can observe 3 main types of betting strategies:

1) The crazy fan: Only betting on his favourite team, feelings are taking control of the bet. Absolutely no analysis before betting.

2) The analyst: Has a quick look at the previous games of each teams. Trying to produce a statement on who is in a good shape or not. Removes the lowest and highest odds in order to avoid either a “low return” bet or a “high risk” one.

3) The most used one: I do not want to waste my time analysing hundreds of games. I want a safe bet, not too expensive, but I need to feel like I will win 100% of my bets.

The problem here is to understand if these “quick wins” strategies, which represent what most of the people are doing on this market, are successful or not:

- Could I just repeat the same bet on each game and be successful at the end?
- Is there any correlation between odds and my earnings after hundreds of bets?
- If my strategy is working for one year, would it be relevant to use it the next one?

III) DATA

Data definition:

My scope of analysis is the English Premier League of Football with a focus on 4 years: 2015-2016, 2016-2017, 2017-2018, 2018-2019.
Four years of football represents a good sample with 1520 games. Enough to produce some statistics around sport betting.

In terms of indicators available within the data, please find below the main data definitions for you to understand the analysis (all definitions are available in the GitHub repository):

Div = League Division
Date = Match Date (dd/mm/yy)
Time = Time of match kick off
HomeTeam = Home Team
AwayTeam = Away Team
FTHG = Full Time Home Team Goals
FTAG = Full Time Away Team Goals
FTR = Full Time Result (H=Home Win, D=Draw, A=Away Win)
BbAvH = Betbrain average home win odds
BbAvD = Betbrain average draw win odds
BbAvA = Betbrain average away win odds
AvgAHH = Market average Asian handicap home team odds
AvgAHA = Market average Asian handicap away team odds

Betbrain is a betting website referencing more than 180 other betting websites in order to have a great average of odds.

Asian handicap is a bet with a game result with a specific goal difference.

Data source:

Football results and odds: https://datahub.io/sports-data/english-premier-league (csv file)

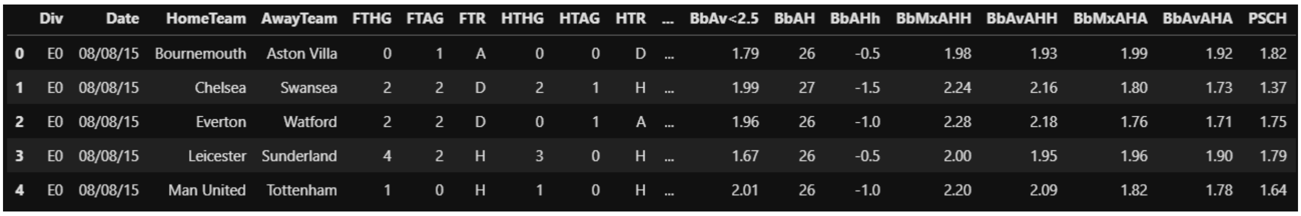
Python Libraries:

Pandas, Sickit-learn, Matplotlib, Numpy, Pylab, Plotly

IV) METHODOLOGY and RESULTS

Data preparation

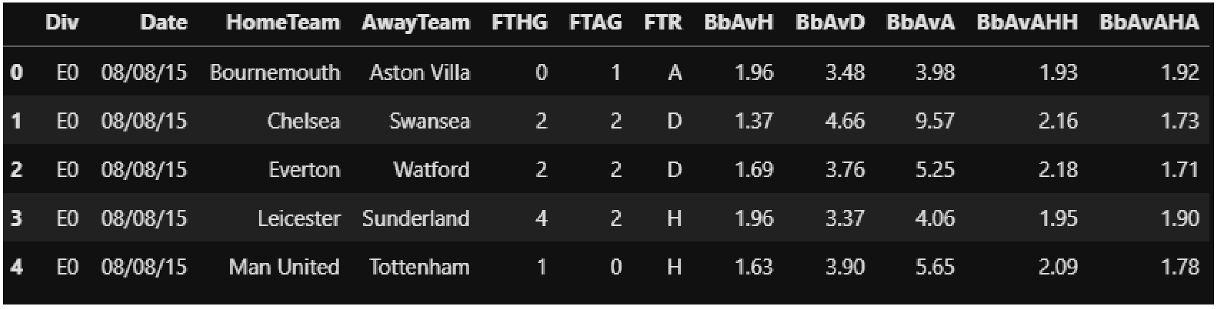
Without preparation, my data frame looks like the below (1520 rows, 65 columns):



	Div	Date	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR	...	BbAv<2.5	BbAH	BbAHh	BbMxAHH	BbAvAHH	BbMxAHA	BbAvAHA	PSCH
0	E0	08/08/15	Bournemouth	Aston Villa	0	1	A	0	0	D	...	1.79	26	-0.5	1.98	1.93	1.99	1.92	1.82
1	E0	08/08/15	Chelsea	Swansea	2	2	D	2	1	H	...	1.99	27	-1.5	2.24	2.16	1.80	1.73	1.37
2	E0	08/08/15	Everton	Watford	2	2	D	0	1	A	...	1.96	26	-1.0	2.28	2.18	1.76	1.71	1.75
3	E0	08/08/15	Leicester	Sunderland	4	2	H	3	0	H	...	1.67	26	-0.5	2.00	1.95	1.96	1.90	1.79
4	E0	08/08/15	Man United	Tottenham	1	0	H	1	0	H	...	2.01	26	-1.0	2.20	2.09	1.82	1.78	1.64

I do a first cleaning by selecting the indicators that will be useful for the analysis: Date, Team, Results, Goals and Odds.

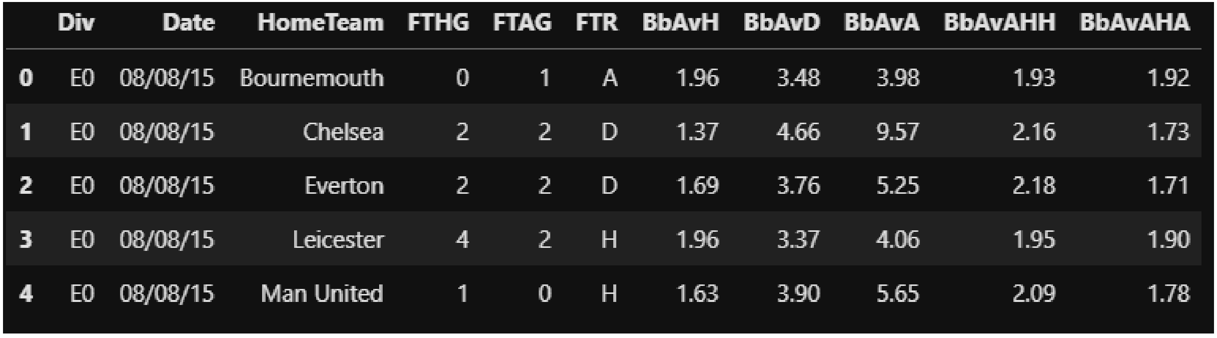
(1520 rows, 12 columns)



	Div	Date	HomeTeam	AwayTeam	FTHG	FTAG	FTR	BbAvH	BbAvD	BbAvA	BbAvAHH	BbAvAHA
0	E0	08/08/15	Bournemouth	Aston Villa	0	1	A	1.96	3.48	3.98	1.93	1.92
1	E0	08/08/15	Chelsea	Swansea	2	2	D	1.37	4.66	9.57	2.16	1.73
2	E0	08/08/15	Everton	Watford	2	2	D	1.69	3.76	5.25	2.18	1.71
3	E0	08/08/15	Leicester	Sunderland	4	2	H	1.96	3.37	4.06	1.95	1.90
4	E0	08/08/15	Man United	Tottenham	1	0	H	1.63	3.90	5.65	2.09	1.78

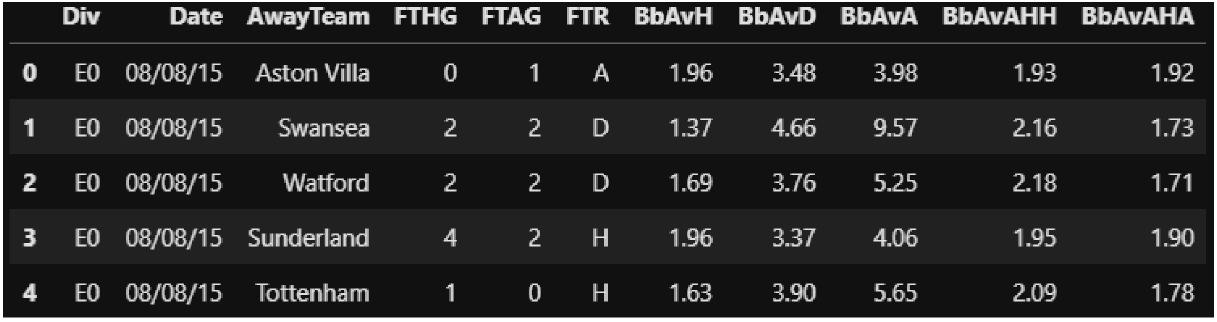
Then, I create different data frame to replicate various betting strategies:

a) Only bet on Home Team wins:



	Div	Date	HomeTeam	FTHG	FTAG	FTR	BbAvH	BbAvD	BbAvA	BbAvAHH	BbAvAHA
0	E0	08/08/15	Bournemouth	0	1	A	1.96	3.48	3.98	1.93	1.92
1	E0	08/08/15	Chelsea	2	2	D	1.37	4.66	9.57	2.16	1.73
2	E0	08/08/15	Everton	2	2	D	1.69	3.76	5.25	2.18	1.71
3	E0	08/08/15	Leicester	4	2	H	1.96	3.37	4.06	1.95	1.90
4	E0	08/08/15	Man United	1	0	H	1.63	3.90	5.65	2.09	1.78

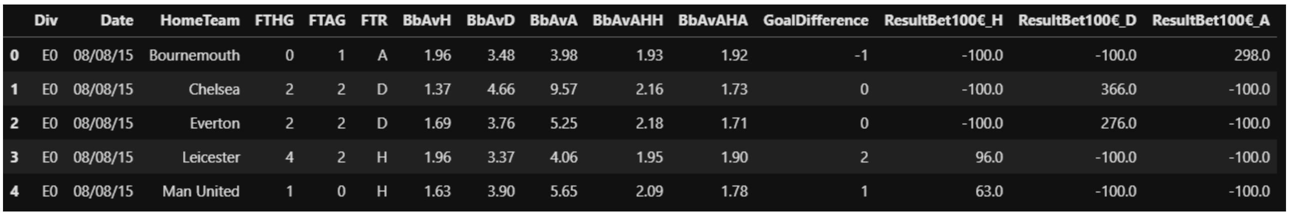
b) Only bet on Away Team wins:



	Div	Date	AwayTeam	FTHG	FTAG	FTR	BbAvH	BbAvD	BbAvA	BbAvAHH	BbAvAHA
0	E0	08/08/15	Aston Villa	0	1	A	1.96	3.48	3.98	1.93	1.92
1	E0	08/08/15	Swansea	2	2	D	1.37	4.66	9.57	2.16	1.73
2	E0	08/08/15	Watford	2	2	D	1.69	3.76	5.25	2.18	1.71
3	E0	08/08/15	Sunderland	4	2	H	1.96	3.37	4.06	1.95	1.90
4	E0	08/08/15	Tottenham	1	0	H	1.63	3.90	5.65	2.09	1.78

I add to both ‘HomeTeam’ data frame and ‘AwayTeam’ new columns showing how much you win or lose after each game if you bet 100€:

ResultBet100€_H = Bet 100€ on Home team wins
ResultBet100€_D = Bet 100€ on draw
ResultBet100€_A = Bet 100€ on Away team wins



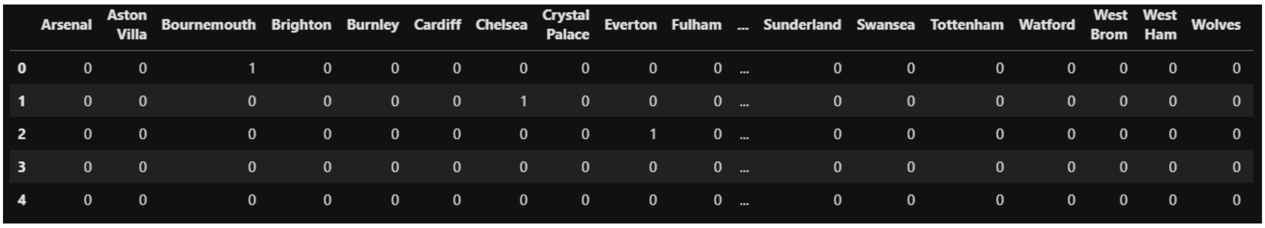
	Div	Date	HomeTeam	FTHG	FTAG	FTR	BbAvH	BbAvD	BbAvA	BbAvAHH	BbAvAHA	GoalDifference	ResultBet100€_H	ResultBet100€_D	ResultBet100€_A
0	E0	08/08/15	Bournemouth	0	1	A	1.96	3.48	3.98	1.93	1.92	-1	-100.0	-100.0	298.0
1	E0	08/08/15	Chelsea	2	2	D	1.37	4.66	9.57	2.16	1.73	0	-100.0	366.0	-100.0
2	E0	08/08/15	Everton	2	2	D	1.69	3.76	5.25	2.18	1.71	0	-100.0	276.0	-100.0
3	E0	08/08/15	Leicester	4	2	H	1.96	3.37	4.06	1.95	1.90	2	96.0	-100.0	-100.0
4	E0	08/08/15	Man United	1	0	H	1.63	3.90	5.65	2.09	1.78	1	63.0	-100.0	-100.0

I also convert the Date column to date time to then be able to sort by date.

Correlation analysis

Here I am looking for a correlation between who is playing and how much I win or lose. If there is one, I should then be able to predict if I will win my bet or not depending on who is playing.

I build a new data frame to be tested and trained through Sklearn linear regression model.



	Arsenal	Aston Villa	Bournemouth	Brighton	Burnley	Cardiff	Chelsea	Crystal Palace	Everton	Fulham	...	Sunderland	Swansea	Tottenham	Watford	West Brom	West Ham	Wolves
0	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
1	0	0	0	0	0	0	1	0	0	0	...	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	1	0	...	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0

After modifying some parameters, the best lr.score I found was : 0.022. It means that there is no direct correlation (probability close to 0%) between whi is playing and my PnL (PnL = Profit and Loss).

If the lr.score was closer to 1, the idea was to create a coefficient base on lr.coef_ to better define which team will lead to a win :

```
df_CoeffHomeResults = pd.DataFrame(data={'HomeTeam' : X.columns, 'CoeffResult' : lr.coef_})
df_CoeffHomeResults
```

However, we can conclude here that a linear regression model cannot be used to predict a result based on the teams playing.

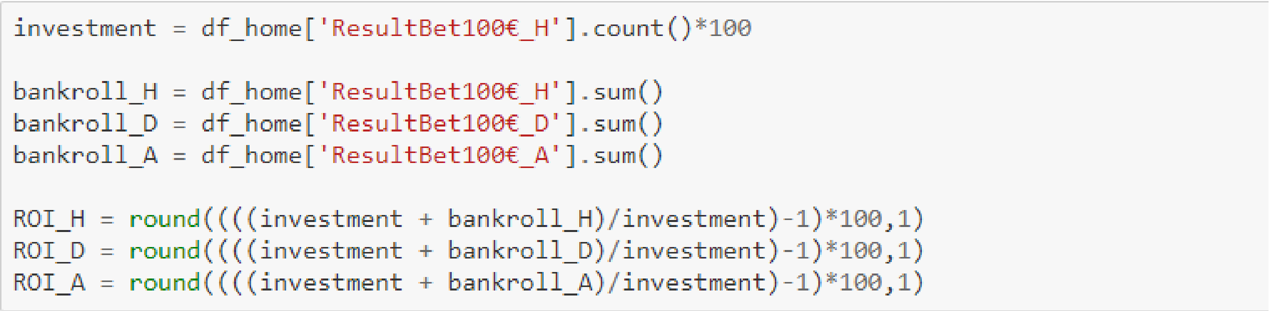
Bankroll simulation

Here I decided to analyse some straight-forward strategies: only betting home team wins, only betting away team wins, only betting on draws.

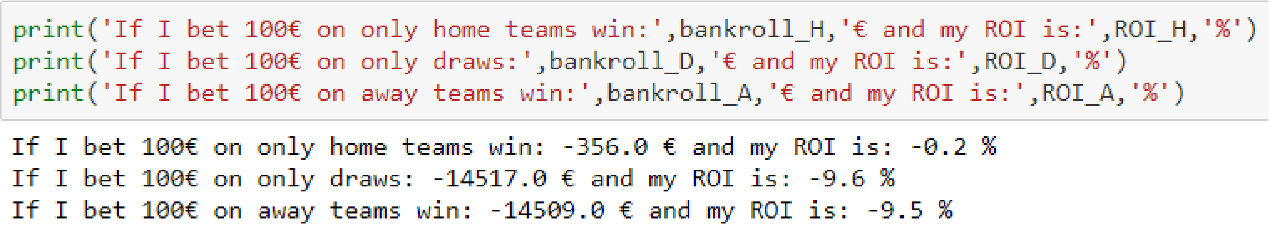
I created 3 additional variables per scenario:

Investment = How much is my total investment (1520 games x 100€ per bet = 152,000€)
Bankroll = How much do I have on my account (starting at 0€)
ROI = Return on Investment ratio [((investment + bankroll)/investment - 1) *100]

Then I print the result of this simulation on 1520 games:



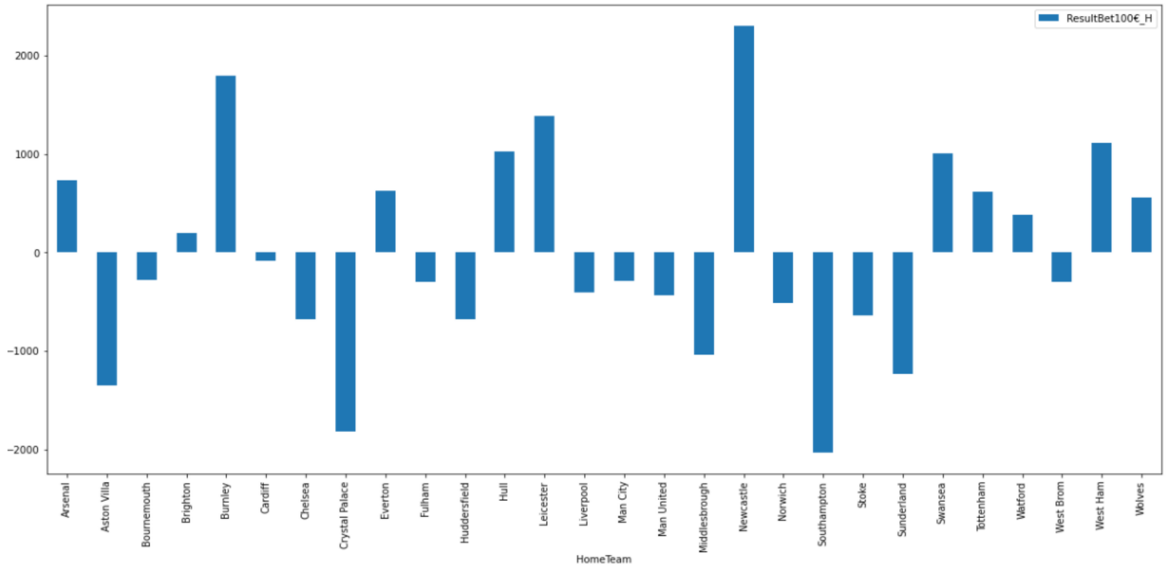
```
investment = df_home['ResultBet100€_H'].count()*100

bankroll_H = df_home['ResultBet100€_H'].sum()
bankroll_D = df_home['ResultBet100€_D'].sum()
bankroll_A = df_home['ResultBet100€_A'].sum()

ROI_H = round((((investment + bankroll_H)/investment)-1)*100,1)
ROI_D = round((((investment + bankroll_D)/investment)-1)*100,1)
ROI_A = round((((investment + bankroll_A)/investment)-1)*100,1)
```



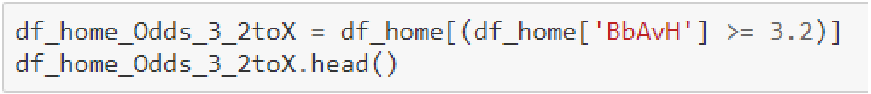
```
print('If I bet 100€ on only home teams win:',bankroll_H,'€ and my ROI is:',ROI_H,'%')
print('If I bet 100€ on only draws:',bankroll_D,'€ and my ROI is:',ROI_D,'%')
print('If I bet 100€ on away teams win:',bankroll_A,'€ and my ROI is:',ROI_A,'%')
```

```
If I bet 100€ on only home teams win: -356.0 € and my ROI is: -0.2 %
If I bet 100€ on only draws: -14517.0 € and my ROI is: -9.6 %
If I bet 100€ on away teams win: -14509.0 € and my ROI is: -9.5 %
```

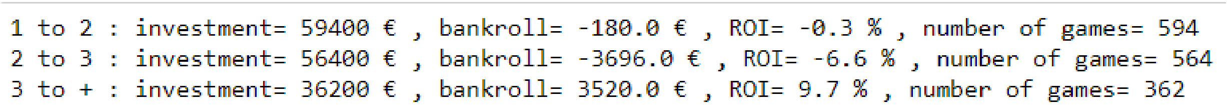
Here we can see that no scenarios are profitable, but the “only home team wins” scenario is close to 0 with only 0.2% of loss within 4 years of betting and 1520 games.



Then I wanted to go deeper into this scenario and create 3 more scenarios: Home team wins with odds between 1 and 2, with odds between 2 and 3, with odds above 3. This gave better results and after some research I optimized those scenarios to: odds between 1 and 2, odds between 2 and 3.2, odds above 3.2 (example below).



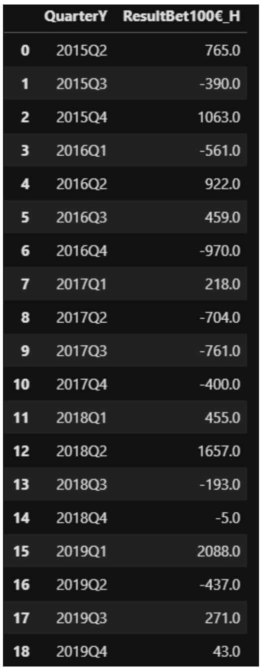
```
df_home_Odds_3_2toX = df_home[(df_home['BbAvH'] >= 3.2)]
df_home_Odds_3_2toX.head()
```



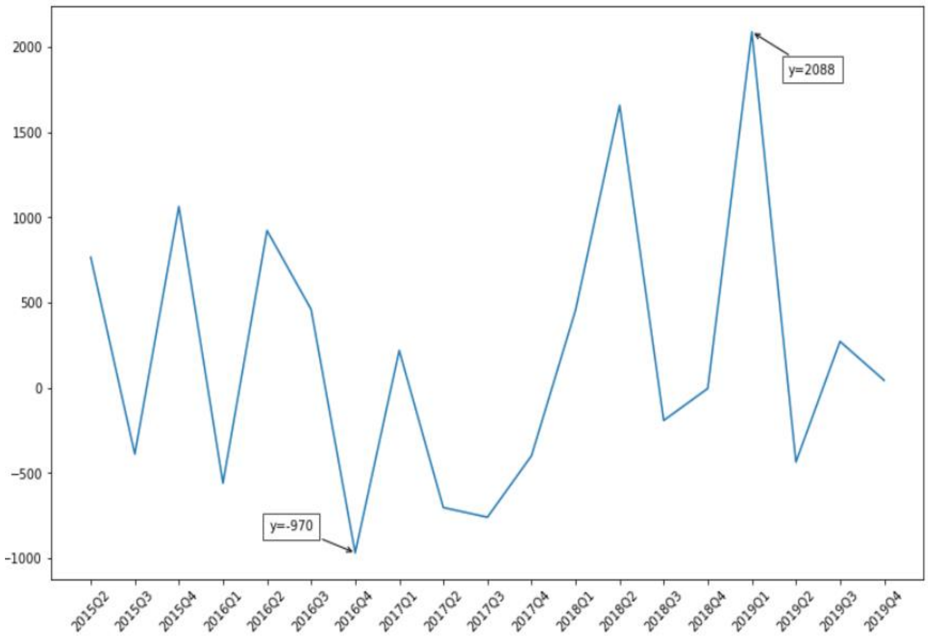
```
1 to 2 : investment= 59400 € , bankroll= -180.0 € , ROI= -0.3 % , number of games= 594
2 to 3 : investment= 56400 € , bankroll= -3696.0 € , ROI= -6.6 % , number of games= 564
3 to + : investment= 36200 € , bankroll= 3520.0 € , ROI= 9.7 % , number of games= 362
```

We can see that betting on home team wins and odds being above 3.2 is profitable with 3520€ of profits and almost 10% of ROI. Now that we can point out a profitable strategy, the questions that logically comes to us is: Is it replicable in the coming years?

The easiest way to answer this question is not to wait for the coming years results. We can just have a look at our data and analyse our cumulative PnL per quarter or per year to see if this strategy works in a smaller timeframe:



	QuarterY	ResultBet100€_H
0	2015Q2	765.0
1	2015Q3	-390.0
2	2015Q4	1063.0
3	2016Q1	-561.0
4	2016Q2	922.0
5	2016Q3	459.0
6	2016Q4	-970.0
7	2017Q1	218.0
8	2017Q2	-704.0
9	2017Q3	-761.0
10	2017Q4	-400.0
11	2018Q1	455.0
12	2018Q2	1657.0
13	2018Q3	-193.0
14	2018Q4	-5.0
15	2019Q1	2088.0
16	2019Q2	-437.0
17	2019Q3	271.0
18	2019Q4	43.0



It is not hard to notice that this strategy is not stable, and some quarters are profitable when others or loss-making. Having a look at 2017 only, it shows a net loss of around 1600€. On a single quarter you have to be able to absorb a -970€ loss while your highest quarterly profit would be about 2100€ (labels on chart).

Thus, even if this strategy seems interesting from 2015 to 2019, we observe that it would be totally unsafe to replicate it as results are erratic.

V) DISCUSSION / CONCLUSION

As we observed in the previous section, we can easily find profitable scenarios by defining specific scopes of analysis based on Teams and Odds levels. However, this statement is not enough to define a profitable strategy on the short term.

This is where data science takes place. We could and we should go deeper into the data: moving from a linear regression to a polynomial one with odds as variables instead of teams, taking a larger set of data, displaying the data on box plot to better understand PnL repartition. There is an infinite way of better understanding a data, and I will keep going through it.

These Coursera/IBM modules were more than interesting and helped me to discover what I finally really like to do.

THANK YOU FOR YOUR TIME, THANK YOU FOR REVIEWING.