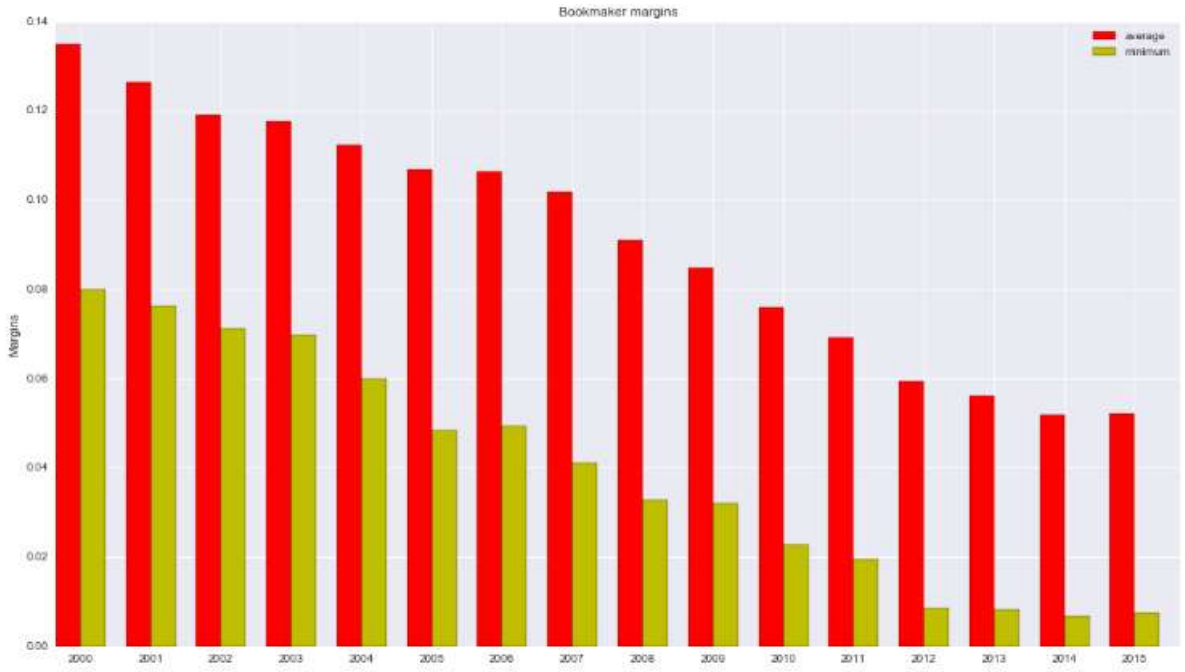
Soccer betting – on top European leagues

Jacky Yeung

Objective

Sports gain significantly more popularity with betting involved. This especially true in European soccer. With historical data for each game available including different bookmaker odds, it will be interesting to see if data science techniques can provide an edge in this market, given already a significant bookmaker margin of 5%+ and even higher at 13%+ in early years. Shopping around from the 13 bookmakers from this dataset does reduce margins by a significant of 4-5%. However, credibility and max acceptable bet size of each individual bookmaker should be taken into account when we start betting real money.

As with recent years, the minimum margins are reduced to less than 1%, likely due to the elevated bookmaker market competition. For this project, for each game each of the W/D/L odds are chosen from the bookmaker which offers the highest.



Arbitrage

Sometimes odds discrepancies between bookmakers happened and created arbitrage opportunities according to my dataset. Odds do change over time while progressing towards game start, so extra scrutiny is needed. Even so, one needs to have significant working capital in all of the bookmakers to place bets if such arbitrage opportunities arise; depositing and withdrawing at last minute likely will not work. Risks on the credibility on the bookmakers, sizing, and execution cannot be neglected. Nevertheless, the returns are indeed very appealing in recent years given the substantial availability of arbitrage games.

	per game return	number of games	total return
season_yr			
2000	0.83%	3	2.48%
2001	2.20%	3	6.60%
2002	2.48%	4	9.92%
2003	4.04%	1	4.04%
2004	0.77%	2	1.54%
2005	2.24%	3	6.71%
2006	2.05%	6	12.31%
2007	0.53%	7	3.72%
2008	1.64%	14	22.94%
2009	0.42%	12	5.06%
2010	0.70%	54	37.98%
2011	0.49%	64	31.52%
2012	0.75%	315	235.91%
2013	0.76%	299	228.34%
2014	0.85%	358	304.85%
2015	0.89%	336	299.81%

Data

Raw data

The datasets were obtained from http://www.football-data.co.uk/ in multiple files specified for each league from each country.

Five countries and 12 leagues data are provided:
-English (4 leagues), Spanish (2 leagues), German (2 leagues), Italian (2 leagues), French (2 leagues)

Data is mostly available from 1993 onwards but some second tier leagues history started later than 1993. For this report, we will just focus on the top league for each country, due to impracticality reasons such as higher margins and sizing in betting in less prominent leagues.

Raw data fields:

Div = League Division
Date = Match Date (dd/mm/yy)
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)
HTHG = Half Time Home Team Goals
HTAG = Half Time Away Team Goals
HTR = Half Time Result (H=Home Win, D=Draw, A=Away Win)
BookmakerX_H = Bookmaker X home win odds
BookmakerX_D = Bookmaker X draw win odds
BookmakerX_A = Bookmaker X away win odds

The game here is maximizing profits, so predicting right does not necessarily mean high return if odds are low in the long run. As a result, the probability of our prediction vs the bookmaker odds have to be taken into account. The dataset includes 13 bookmakers but some of them are not always in business. For usability and fairness, median of the odds are used for analysis.

Match statistics such as attendance, referee, home/away team shots, shots on target are available only for late years. For robustness argument, those are sacrificed for longer backtest duration.

Features

As this is a moving window problem, we track the past N games for modelling. Which N is a big question, after multiple trial and errors, past six games data are used. Another big problem arises for training set in which number of past years should be used. I finally settled with trailing past 5 years.

Each data row is structured so that every game is a home game to avoid duplication. Features tested are the following:

Current fields

'**team**' = home team
'**against_team**' = away team
'**Div**' = league division
~~'favour'~~ = indicator of which game result outcome has the lowest odds (W/D/L)
~~'win_odds'~~ = median win odds for all available bookmakers
~~'draw_odds'~~ = median draw odds for all available bookmakers
~~'lose_odds'~~ = median lose odds for all available bookmakers
~~'week'~~ = which week in the year
~~'month'~~ = which month in the year
~~'weekday'~~ = which weekday in the week

Trailing fields (H_ for the history of the home team / A_ for away team; X ranges from 1 to 6)

'**H_Result_-X**' / '**A_Result_-X**' = win/draw/lose
'**H_is_Home_-X**' / '**A_is_Home_-X**' = whether the past X game is home or not for Home/Away team
~~'H_Goal_-X' / 'A_Goal_-X'~~ = number of goals
~~'H_Concede_-X' / 'A_Concede_-X'~~ = goals conceded

'**H_against_team_-X**' / '**A_against_team_-X**' = the past X game whom is the opponent for the current Home/Away team
'**H_Div_-X**' / '**A_Div_-X**'
~~'H_favour_-X' / 'A_favour_-X'~~
'H_win_odds_-X' / 'A_win_odds_-X'
'H_draw_odds_-X' / 'A_draw_odds_-X'
'H_lose_odds_-X' / 'A_lose_odds_-X'
~~'H_week_-X' / 'A_week_-X'~~
~~'H_month_-X' / 'A_month_-X'~~
~~'H_weekday_-X' / 'A_weekday_-X'~~

The bolded ones are the final used fields. Initially, I refrained from using bookmaker odds fields but results were less than ideal. Then, using the odds greatly enhance predictive power. Even greater result was achieved by not using historical number of goals and conceded. Eliminating current game odds surprisingly brings another improvement, and great for practicability as current game odds do change before game starts.

Favour field is derived from current game odds with much reliability, since while current game odds do not always stay constant they do not fluctuate drastically; however, it is not a useful field. Perhaps, there is too much information for the current game or even match fixing possibilities. The reason why historical odds in conjunction with the indicators of which teams playing and game end result work amazingly well versus others, maybe because the odds already have lots of implicit information and the end result of the game provide a pattern for machine learning model to detect.

Modelling

Many models were tried. The list includes Logistic Regression, KNN, Random Forest, Extra Tree Classifier. Finally, XGboost was used. Some grid searches were done but not a lot. I change the learning_rate from 1 to 0.1. Higher levels of max_depth and increasing number of estimators create overfit.

SelectFromModel using either ExtraTreeClassifier, RandomForest, or LogisticRegression were tried. Either the randomness is too unstable or just there is just worse predictive power. PCA and LDA were tested but not used at the end.

As previously mentioned, we use a window of 5 years for train set and then use such model to predict next 1 year. Since I am predicting 2005 to 2015, 11 models were created. To avoid overfitting, same type of model and parameters were used.

There was an initial approach of having one model for each league and then aggregate to get the total return. Due to time constrain and seemly endless iterations for trials and improvements, aggregating all 5 leagues for modelling is the final decision.

How to use the model?

-**pred_return**
- bet on what the model predicts regardless of the odds

-**pred proba**
- get the model predicted probability for each of W,D,L outcomes and compare with the implied probability from bookmakers odds, bet on the result in which my model has highest probability vs bookmakers implied probability

-**pred_return** & **proba_return**
- bet only if the predicted choice equals the predict proba choice

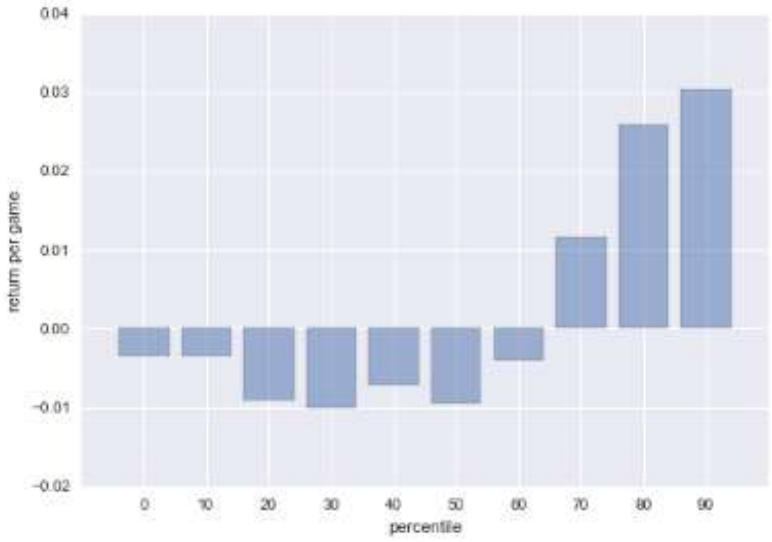
-**v_proba_return**
- there is another quick non-machine learning strategy, which does not care what the exact team is. Each data row has a Home team and Away team, but the team is classified by total goal difference for past N games, and let's fix N as six for now, ie it does not matter if it is team Manchester United or Watford as long as their past six games goal-diff is the same. The game result of different combination of Home goal-diff and Away goal-diff is recorded to form a distribution for prediction.

-**pred_return** & **v_proba_return**

-**pred_return** & **pred proba** & **v_proba_return**

Percentile selection

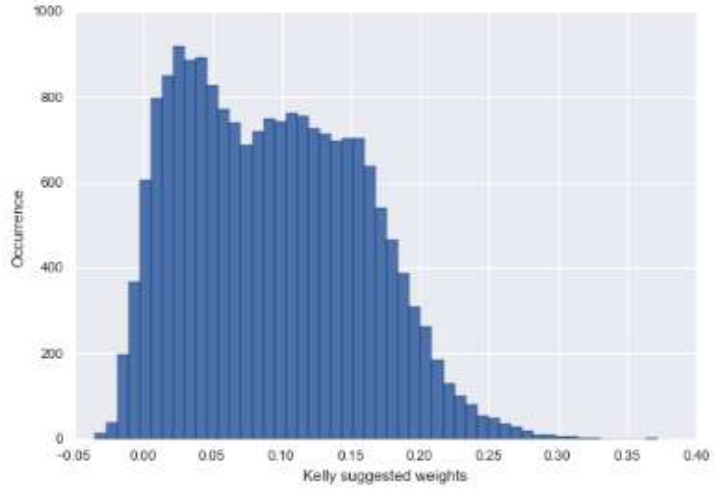
Since my model will give a probability for each of the predicted outcomes (W,D,L), comparing with the bookmakers odds implied probability gives a numeric difference. I can then sort the difference by percentile and choose the ones which have highest difference, equivalent of higher winning edge. The downside of this is reducing the number of games for betting. To avoid overly cherry picking, only top 50% or 20% should be considered. The tail end such as top 10% can be volatile and too few games available. Top 50% can screen out half but yet end result can still be negative. Given the fitted model and the theoretical rationale mentioned, top 20% is used.



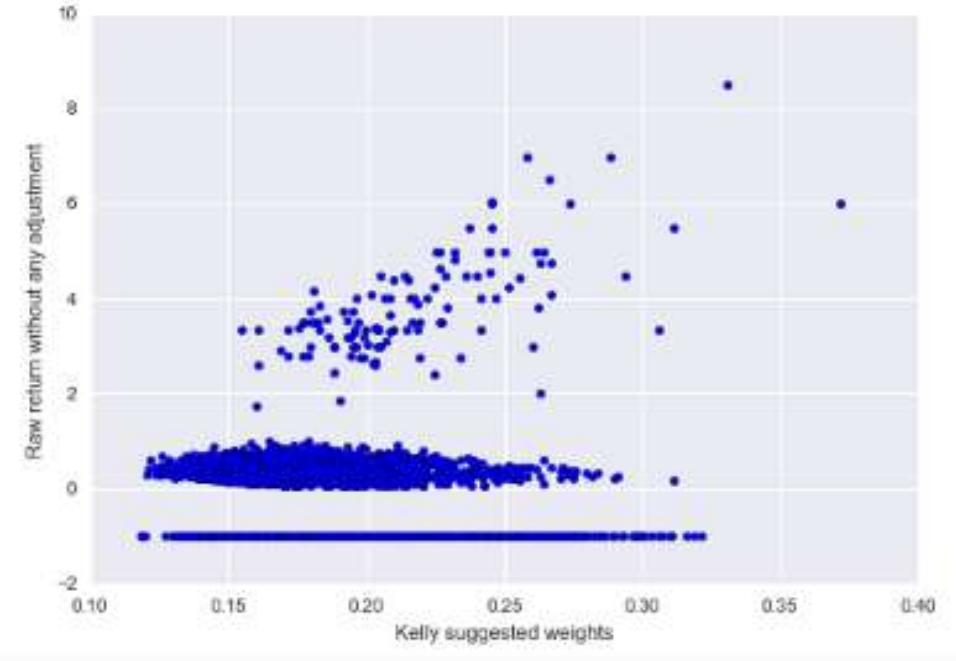
Sizing

With both the probability and target payoff available, we can use the famous Kelly formula for sizing

$$f^* = \frac{\text{expected net winnings}}{\text{net winnings if you win}}$$



Bulk of weights are around at 10% level without adjustment.



Kelly seems to work as higher weight yields higher return per game.

I also multiply by a factor 0.10 so each individual bet will not be too big at around 1-3% of total asset per bet assuming in an asset management setting where there will not be any reset in asset. Without such 0.10 factor scale down, a continuous streak of loss can easily lose more than 90% of asset and to make up such loss requires a 10 times return. Another strategy to consider can be an annual reset in asset level.

Trend momentum modification

Basically this means that if I lose on my last bet I will wait until my model wins and then restarts betting on next game so to avoid losing streaks. This only works if the real returns have a positive autocorrelation with lag 1. Some models show that this works. Statistically unreliable, I decide to drop this.

Final decision

Below is the total aggregate return of all games for the 5 strategies and the percentiles. The 3 far right ones which involves the non-machine learning value bet strategy have horrible results plus super unstable when adjusting with different past windows of historical games.

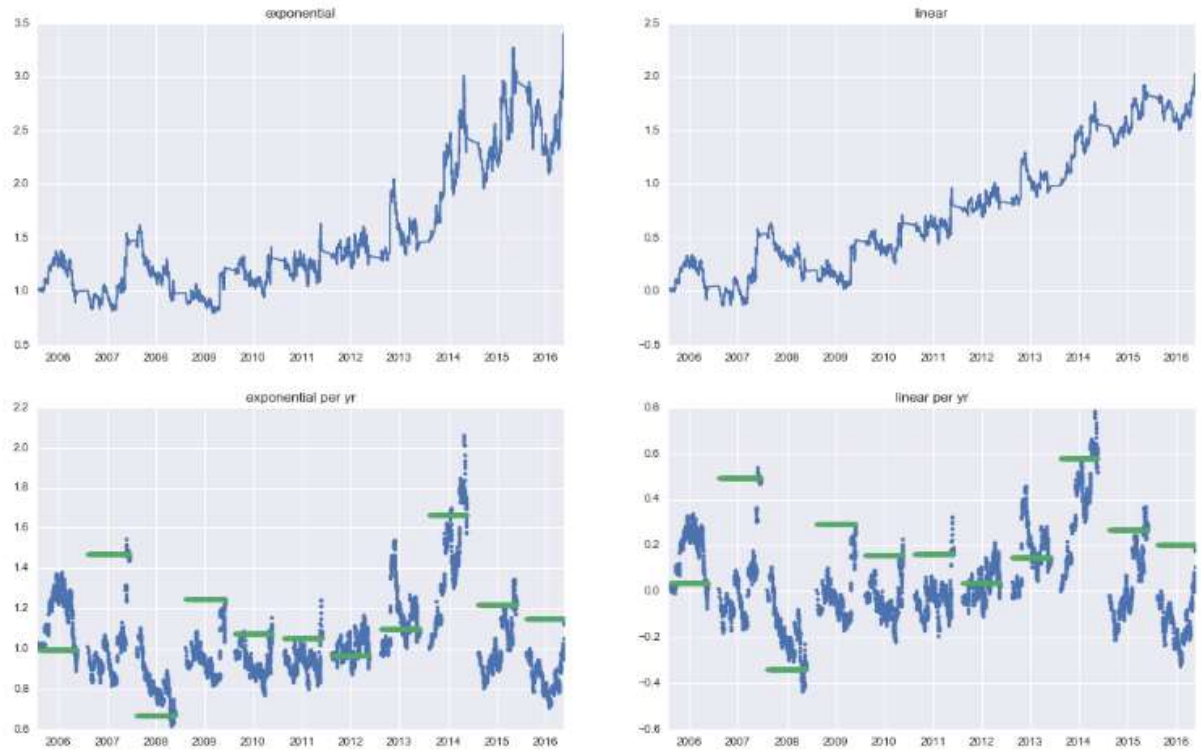
	pred_return	proba_return	proba_ret_select	v_proba_return	v_proba_ret_select	v_pred_proba_ret_select
Total						
0	2.10%	-874.05%	44.97%	-1169.60%	-726.52%	47.14%
10	4.02%	-884.28%	42.42%	-1168.93%	-762.89%	53.87%
20	-3.97%	-873.70%	32.25%	-1172.90%	-706.09%	60.28%
30	-31.95%	-841.40%	22.48%	-1074.70%	-612.02%	37.64%
40	-10.34%	-811.11%	33.68%	-1095.83%	-616.38%	56.20%
50	-5.75%	-631.15%	90.12%	-1075.84%	-528.67%	96.27%
60	55.02%	-630.87%	101.98%	-1026.56%	-488.38%	102.62%
70	169.02%	-583.13%	127.19%	-906.42%	-375.77%	117.93%
80	202.58%	-685.54%	122.50%	-463.52%	-136.19%	128.83%
90	135.59%	-347.39%	62.02%	-287.70%	55.92%	95.93%

I will focus on the left 3.

'**pred_return**' bets on what the model suggests regardless of bookmaker odds. The percentile here is derived from the differences of scikit learn's predict_proba and bookmaker implied probabilities of the outcomes (W/D/L). Hence, the higher top percentile eliminates low return high risk bets.
'**proba_return**' bets on the highest difference in probabilities vs bookmakers implied probabilities as explained above. Unfortunately it does not work well here and is usually less stable.
'**proba_ret_select**' only chooses to bet when the choices of '**pred_return**' and '**proba_return**' are the same. This is like the hyped up more stringent version of '**pred_return**' when using percentiles.

'**pred_return**' at top 20% is used. Below is the average game return on asset by each soccer season. All are positive except 2007.

```
Total return: 202.58%
Avg per game return: 0.05%
           pred_return avg ret  pred_return num games
season_yr
2005                      0.01%                    350
2006                      0.14%                    347
2007                     -0.10%                    349
2008                      0.08%                    343
2009                      0.05%                    348
2010                      0.05%                    347
2011                      0.01%                    353
2012                      0.04%                    352
2013                      0.17%                    348
2014                      0.08%                    351
2015                      0.06%                    349
```



The top left chart is the asset curve over compounded exponentially. The top right is linear assuming asset level is always the same, ie if I start with 100 dollars and after 1 game this becomes 103, exponential will choose bet size according to 103 and linear will always use 100.

The bottom charts show the performance for each individual season assuming starting from scratch. The green lines are the performance of where the season ends. Notice for season 2007 which starts in calendar year mid of 2007 and ends mid of 2008, the return is a negative 30%+. If the 0.10 scale down factor is not used, this is almost a complete ruin of asset.

Conclusion

There seem to be endless iterations in many areas: feature selections, years of window train set, model selection, model parameters. Not only much time is required but also a detail clear method of recording down the results of each iteration. Also, there can be possible predictive power improvement with more data such as player stats and historical game play details, but this is another huge step both in analysis and computer run time. Nevertheless, this model has some usefulness and provides much insights for real money soccer betting.