Trying to beat the odds

Jarle Kvile

Abstract

In trying to beat the odds, I am using the Football events dataset to try to beat the bookmakers' odds which teams wins a soccer match. The project first explores some of the aspects of the dataset and determines that there is indeed a statistical home advantage but does not determine the underlying cause. The project proceeds into finding what rate the odds predict the correct outcome of a match, before trying to beat it with 5 different ML-algorithms, commonly used for categorical values. Although it does not manage to beat the odds, some of the results are encouraging for future research.

Motivation

Predicting the correct outcome of a soccer match has been a concern for many a fans, gambler and pundit for as long as we have had fans. My project wants to try different machine learning algorithms, to see if we can outscore the bookmakers' odds.

The audience here is wide, the ability to predict outcomes when there is a lot of unknown factors, would be useful for any kind of research. This is strongly linked to other sports markets, but also the stock exchange (will the price go up or down). It has therefore been important to include cross-validation, to ensure there is no overfitting of the model.

Datasets

- I have used the “Football Events” database,open available to anyone here: https://www.kaggle.com/datasets/secareanualin/football-events
- The dataset contains a dictionary, containing the text description of integer coded variables.
- It also contains two csv files. The “events” file contains detailed game events, such as information on how a goal is scored, etc. The “ginf” set, or “game information” contains information on a more aggregated level, including the pre-game odds for each game.

Data Preparation and Cleaning

- Basically, there are several steps or checks you need when preparing and cleaning the data.You need to:
  - Remove duplicates
  - Fix structrual errors (typos etc)
  - Fix unwanted outliers (improper data entrys)
  - Handle missing data (what do do with NA?)
  - Validate. Check the data makes sense.
- In the data I have used for this project, no genuine data cleaning was needed. The data was of very high quality, and I just had to ensure I had control of all the steps, including proper filtering to make a good analysis.

Research Questions

- The project tries to answer a few different questions, at all once.
- First, at what aspects of the game does “home advantage” seem to be true? Is there a statistical advantage to playing at home? It is true that away teams score more on the counter-attack than others?
- Secondly, using the odds to predict the correct outcome, it wants to know, can we use any of the following 5 machine learning algorithms to beat the odds?

1) Logistic Regression (Lasso)
2) Support Vector Machine
3) K-Nearest Neighbours
4) Decision Tree Classifier
5) Random Forest Classifier

Methods

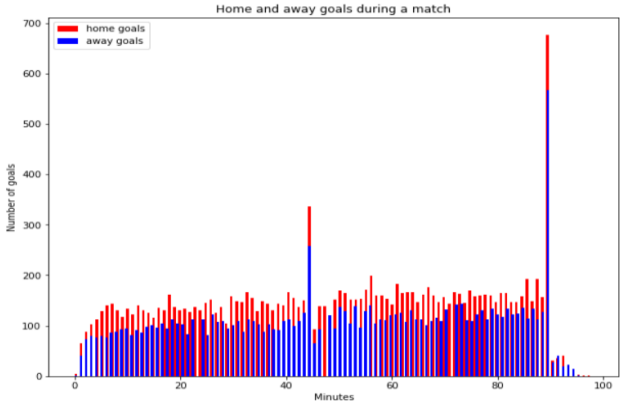
- This projects classifies trough an algorithm whether a game has ended in home win, away win or a draw. Each of outcomes have a certain set of pre-game odds assigned to them, and I firstly predict whether the odds predict the correct outcome of the game. The odds have a 53 % success rate of predicting the correct score.

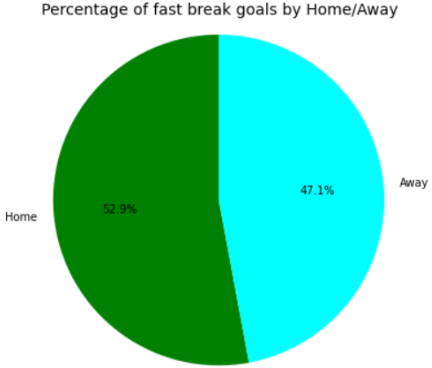
- I then move forward and using 5 different machine learning algorithms in order to beat that success rate of the pre-match odds. I have the five most common classifiers and have added a layer of cross validation to ensure I do not overfit the models, a common problem in classification algorithms. The ones I have used are:

  - Logistic Regression (Lasso)
  - A Support Vector Model Classifier
  - K-Nearest Neighbors Classifier
  - Decision Tree Classifier
  - Random Forest Classifier

Findings – Home advantage is real

- During a game, a lot of the goals is scored during the end of the halves. And at the end of those, the most is scored by the home side.
- It is however not true that away teams score more on the break than home sides.
- However, we do see that away sides score **47.1 per cent of all fast-break goals**, and only *42.6 per cent of all goals scored*.



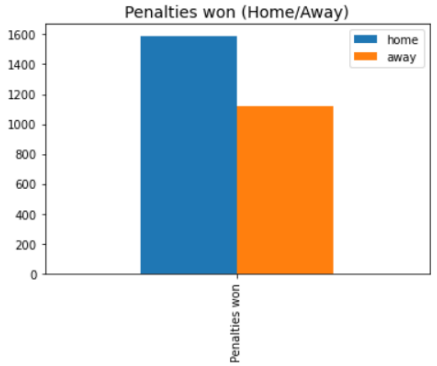


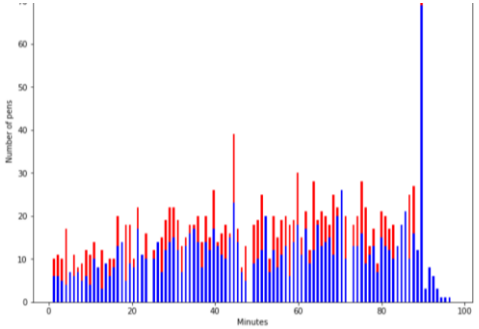
Findings – Home advantage is real

- The home side does get more penalties than the away side!

- They also get them more at the end of the halves, but it is not as a clear advantage as people might have thought. We see the away side gets a few at the end of the games too.

	home	away
Penalties won	1589	1117





Before iteration

```
In-sample accuracy: 0.46
Test accuracy: 0.48
In-sample Precision Score: 0.97
Test Precision Score: 0.43
In-sample F1 Score: 0.97
Test F1 Score: 0.43
```

After iteration

```
In-sample accuracy: 0.53
Test accuracy: 0.55
In-sample Precision Score: 0.54
Test Precision Score: 0.48
In-sample F1 Score: 0.44
Test F1 Score: 0.43
```

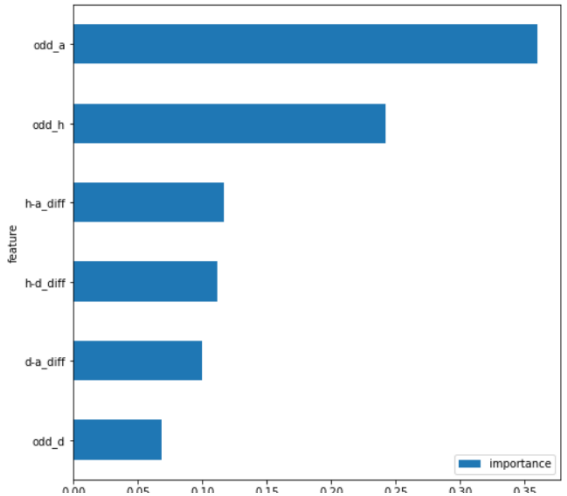
Findings – Trying to beat the odds!

- In order to know if we beat the odds, we defined how well the odds preform as a predictor of the result. The odds predict the correct outcome of the game **53 per cent** of the time.
- My highest score was also 53 per cent, and the Random Forest Classifier is the winner, after a much-needed iteration of the model that improved the accuracy from 46 % to 53 %.

Findings – Trying to beat the odds!

- We can see that these models yield about the same (or even worse) accuracy than by just guessing based on the odds 53% accuracy. So, I would have liked to have better predictions.
- When we look at the factors, the Random Forest is the winner* as it has the highest accuracy, although the K-nearest neighbor had a higher Test set F1 score.

	Classifier	Training Accuracy	Test Accuracy	Training Precision	Test Precision	Training F1 Score	Test F1 Score
0	Logistic Regression (Lasso)	0.53	0.55	0.35	0.37	0.38	0.40
1	Support Vector Machine(Linear)	0.53	0.55	0.35	0.38	0.37	0.40
2	K-Nearest Neighbours	0.50	0.52	0.53	0.47	0.50	0.45
3	Decision Tree	0.52	0.54	0.53	0.46	0.42	0.42
4	Random Forest	0.53	0.55	0.54	0.48	0.44	0.43



	Actual Home Win	Actual Draw	Actual Away Win
Predicted Home Win	3141	1361	1084
Predicted Draw	71	137	55
Predicted Away Win	512	556	1172

Limitations

- This plot is a good summary of the limitations. The winning odds are the most important for the random forest classifier, as well as their difference.
- The lack of hit rates on the draws that actually happened, is also a problem, as the confusion matrix for the RF-classifier shows.
- This, along with the model that have been selected, shows that in order to actually being able to beat the odds, we need even more information. However, that is outside of the scope of this paper.
- Further analysis should use the form of the sides, the styles of the sides, the previous results, the factors of the different leagues, etc. All of these metrix are "baked into" the odds, so it is hard to beat it, without being able to split it completely.

Conclusions

The project set out to answer two questions. Does the home side have an advantage in the game? Using common talking points as metrics, we were able to show that yes, indeed, they do have an advantage, but it is not as huge as the conversation shows. A difference between perception and the data, might be because humans tend to remember the main events, such as when the home team turns the tide around, to the great excitement of the home fans.

The project then tries to envestigate whether or not we can beat the odds by using ML-techniques. The results are not conclusive here, but they are promising. However, more data, further analysis to include different information is needed. The project did find that the over-fitting problem of the decision tree and random forest classifers are indeed very true for this type of problem indeed. Future research, bewarned!

Acknowledgements

I collected the data myself from the kaggle website. I wanted to see how to begin, and to determine a good function for the odds and the predicted score, I spent a dozen hours on stack overflow, even asking for help a few times.

I pitched the research question to my friends, some of them soccer fans and some of them couldn't care less, and all of them were very helpful in framing my research question to the scale it needed to be. Thank you especially Mikael Gursli and Sara Sølberg for your patient efforts.

In addition, spending several hours of my year listening to different football podcasts have helped me.

The rest of the work is my own.

References

Kuhn, Max; Johnson, Kjell (2013). *Applied Predictive Modeling*. New York, NY: Springer New York. doi:10.1007/978-1-4614-6849-3. ISBN 9781461468486.

https://www.vantage-ai.com/en/blog/beating-the-bookies-with-machine-learning

https://medium.com/analytics-vidhya/beating-soccer-odds-using-machine-learning-project-walkthrough-a1c3445b285a