Machine learning informed betting strategy applied to Football matches

Citadel Data Open 2021

Marcus Foo, Leo Zhang, Liew You Sheng, Arush Tagade

28 March 2021

1 Executive Summary

Football is a sport that commands worldwide attention from hardcore fans and casual viewers. This widespread appeal for the sport has given rise to a large betting market where the main interest is in guessing which team will win.

This report presents a methodology for predicting the outcomes of a given football match and use this prediction to make a decision to bet or not to bet. Through this we attempt to the answer two key questions of:

1. Which set of team and player attributes are the most influential in predicting match winners?
2. Can a profitable betting strategy be developed that outperforms random betting?

In addition, we hope to employ quantitative finance methodologies, specifically from The Advances in Financial Machine Learning[1] textbook authored by Dr Lopez de Prado, borrowing quantitative finance concepts to bolster our strategy's performance.

We utilized machine learning techniques to create a model that predicts the outcome of football matches based on a large range of team attributes along with individual player attributes. By using aggregated betting odds from 10 betting websites and our model prediction, we make decisions on betting with confidence in the model prediction as a parameter. For certain ranges of confidence we found that our predictions led to a net increase of 127.55% over a period of one year and 3 months.

There are a few key insights that we derived from our experiments that can be used for future analysis of deeper problems:

- The individual skill of players turned out to be much more important in our model predictions when compared to team-based skills. This might arise from the fact that skilled players would naturally exhibit better team skills.
- When looking at only just team characteristics, it was found that the centre back player, and chance creation features of the team are the most influential in determining the winner of a match.
- A profitable betting strategy was able to be developed giving statistically significant returns when compared to random betting.

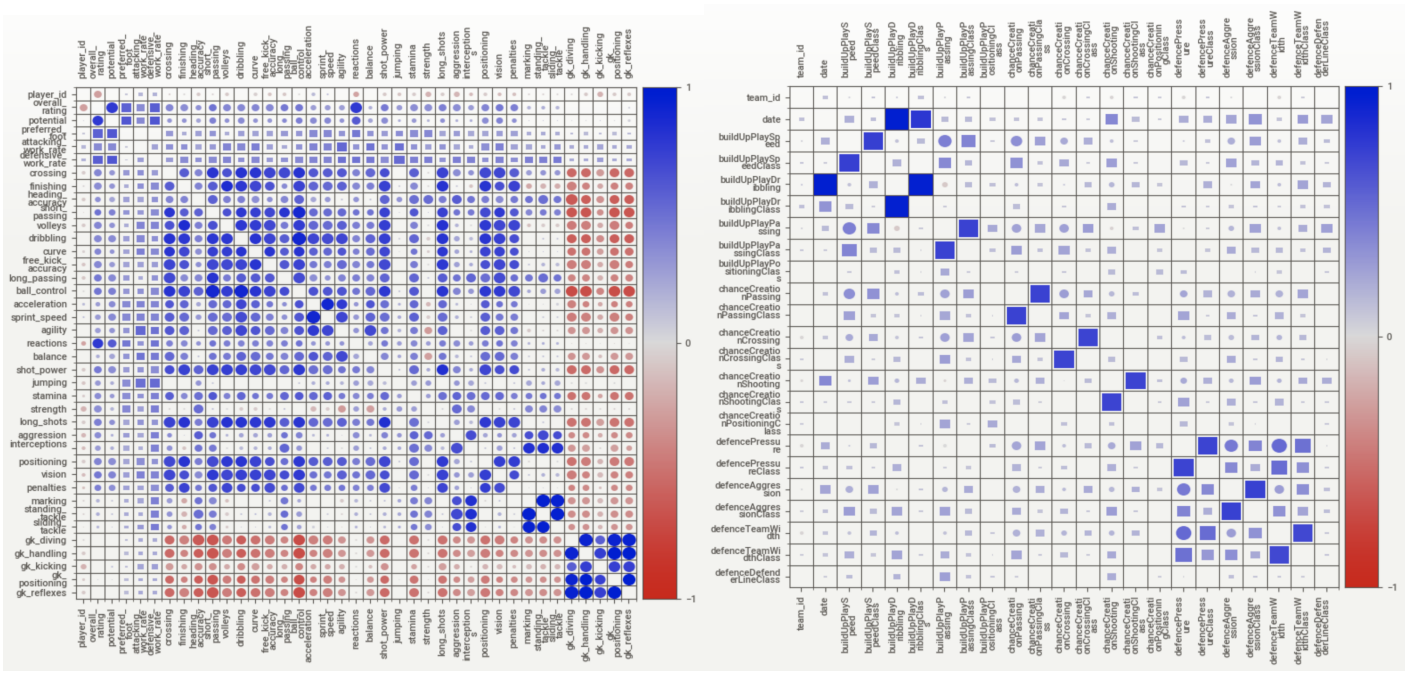
2 Background Information

Our report seeks to build upon the work of Stubinger et al.[2] whereby a machine learning framework for forecasting future football matches and achieving profitable returns through appropriate betting was developed on a similar dataset. Within their paper, an ensemble model was developed to predict the goal difference of a match given the team and player characteristics of that match. They applied a betting strategy whereby if a clear victory (defined as a goal difference of greater than two) was predicted, a predefined monetary unit would be bet on the favoured side.

Though they achieved profitable returns, we believe significant improvements can be made to their methodology, both within the construction of their predictive model and the way in which they determined how bets should placed. This served as inspiration for the formation of our second key question.

3 Exploratory Data Analysis

An initial overview of all the files within the dataset revealed that the features that would be required to build a predictive model on matches (*match.csv*) reside solely within the *player_att.csv* and *team_att.csv* files. EDA on these was conducted using the SweetVIZ[3] python library. A visual representation of the correlation matrices created for both team and player attributes are shown in Figure 1 below.



(a) Player attributes

(b) Team attributes

Figure 1: Correlation matrices of player and team attributes

3.1 Player Attributes

An initial EDA of the player_att dataset reveals that player attributes spanned from 2007 to mid 2016 (Figure 2). It is interesting to note that while regular updates to player attributes are given at around half a year intervals, in the period following 2013, frequent updates are given on a weekly basis. To prevent lookahead bias, we ensure that the most recent player data available before every match's date is used for our predictive models.

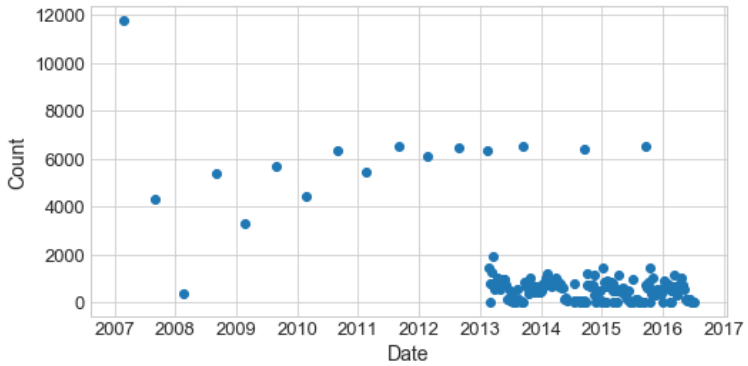


Figure 2: Number of player updates over time-frame of player attribute dataset

Figure 1 reveals that player rating and potential score are positively correlated with the majority of the other features in the dataset. As a result, we chose to select only rating and potential as the player features.

3.2 Team Attributes

Unlike player data, team attribute data is updated on a more regular and predictable yearly basis (Figure 3). Nonetheless, the we apply the same process of selecting the most recently updated team attributes for each match. We chose to

incorporate both categorical and continuous aspects of team attributes into our predictive model. BuildUpPlayDribbling was the only feature to be dropped following our EDA, due to its values being comprised of mostly nulls.

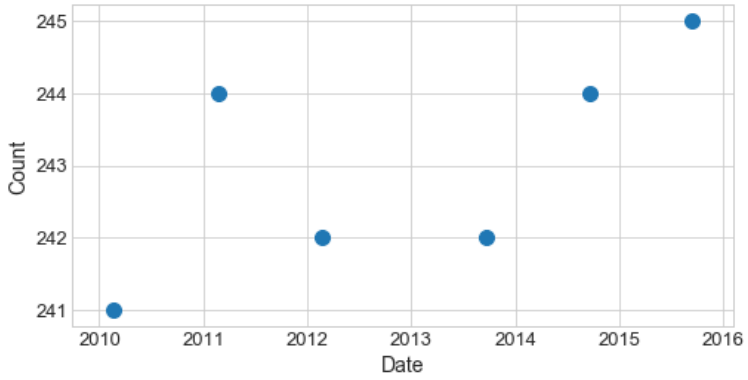


Figure 3: Number of team attribute updates over time-frame of team attribute dataset

3.3 Matches

The match data set comprised of 25,979 matches spanning from 2008 to 2016. 5% of these contained null player data and were consequently excluded from our analysis. While many betting websites contained a large proportion of missing data, this is largely irrelevant to our analysis as we always only require the single most favourable odds given for our prediction.

We compared the number of goals for each match by each team, and there is an advantage by the home team as compared to the away team. The mean is much higher on the home team, and the away team has scored more zeros as well.

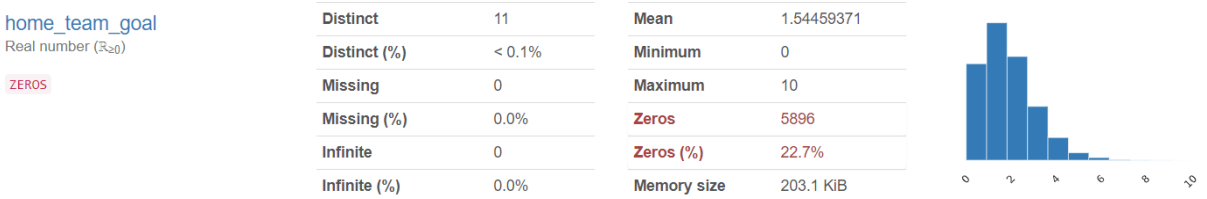


Figure 4: Number of goals by home team

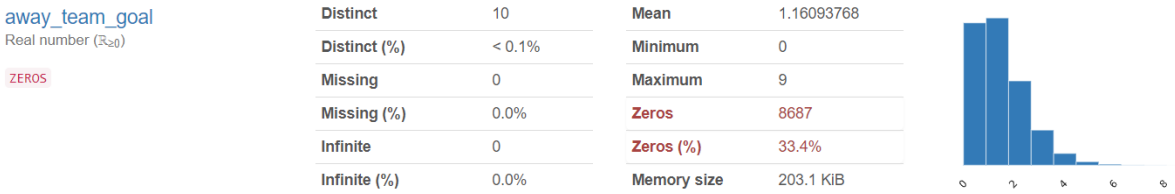


Figure 5: Number of goals by away team

We also compared the number of match outcomes and the home team advantage is clearly shown, a total of 46% games end in wins for the home team. Surprisingly the number of away team wins and draws are quite close to each other with 29% wins and 25% wins respectively.

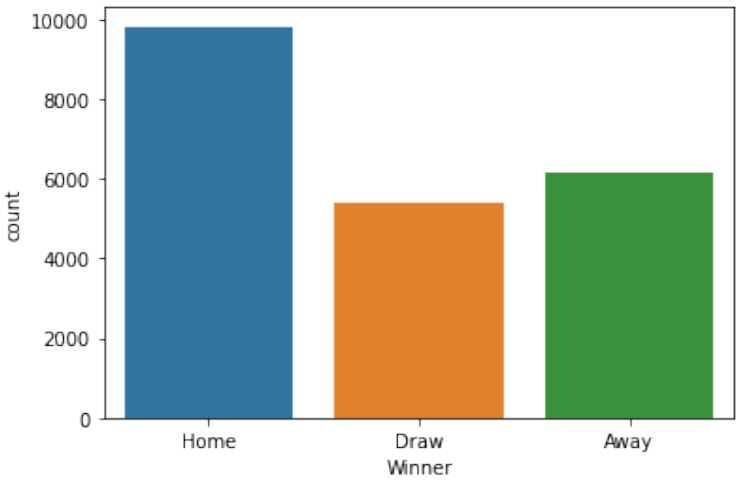


Figure 6: Comparison of number of matches ending in home team wins, away team wins and draws

4 Data Preparation

4.1 Data Cleaning and Wrangling

Before aggregation with the player and team attribute datasets, we removed all rows in the match dataset which did not include player id values. Intuitively, we felt that the individual skill level and characteristics of all players are vital features for determining a logical guess for which team would win. We confirm our hypothesis later on within the Feature Selection subsection. For each match in the remaining set of matches, we retrieved the most recent team and player attributes (set to NaN if not found) for both the home and away team, and added them as separate columns for the match. The cleaning/wrangling process and final dataset is summarised in the entity relationship diagram below (Figure 7).

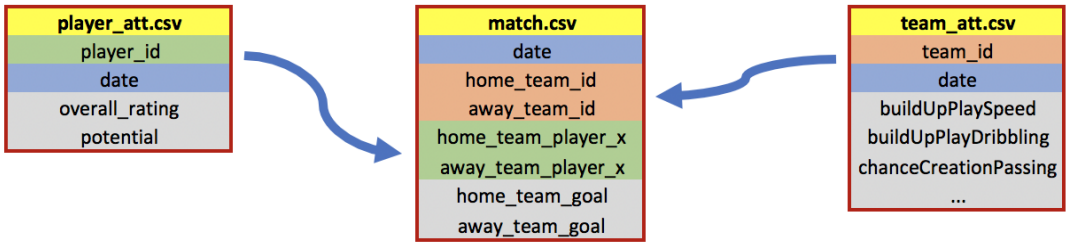


Figure 7: Entity relationship diagram of our final dataset

We also removed features which we felt were inaccurate to include in our model, which includes date, season, player id and multiple others which are mentioned within our pipeline notebook. Our main concern was to ensure that our model was robust across time, whereby the addition of newer players, arenas, as well as environmental factors like seasons and datetime should not breakdown our underlying model.

4.2 Feature Engineering

The match dataset contains information about betting odds from 10 websites, which we aggregated into three columns by taking the average odds from all 10 brokers for home team win, draw and away team win. This gave us a very strong feature that informs the model's prediction heavily. It is important to note that these betting odds do not correspond to the actual odds of the competing teams but are in fact slightly modified to be profitable for the betting website that

publishes them. In addition, we also created a feature called home-away spread which is done by taking the difference between the average of home and away odds. Similar to the concept of market makers offering a bid-ask spread, the spread could contain valuable information for conveying the broker's confidence of their estimation of the match's true odds. Regardless of this fact, we experimented with this feature and found that it leads to a model that performs very well for our problem.

4.3 Training, Validation and Test Set Split

We performed a 60-20-20 split for training, validation and test sets. The first 80 % of the data was split randomly by 60-20 to provide a better chance for the model to generalise and predict match outcomes. To realistically test the viability of our strategy, we opted to take the most recent 20% of data chronologically to simulate the use of our model to inform our betting strategy as a backtest on out of sample data.

5 Implementation Details

A brief summary of our approach is as follows:

1. Aggregate data sources into a singular dataframe
2. Do a 60-20 random split to form our train-val set, and leave the latest 20% data as our test-set
3. Elect a classification model as our primary model for predicting bet direction (home/draw/away)
4. Tune hyperparameters of primary model on our val-set
5. Calculate profit-and-loss (PnL) of our primary model on our val-set to form metalabels (0-not profitable, 1-profitable)
6. Train a secondary model with the same initial features on our val-set, but with the newly generated metalabels as target variables
7. Tune hyperparameters of secondary model on val-set
8. Conduct a backtest and generate PnL curve across our test-set

5.1 Model Selection

Many cases of Machine Learning applications in quantitative finance suffer from data-snooping bias, resulting in the overfitting of parameters to the train-set[4]. In the same vein, we leaned towards bagging and boosting models.

Our criteria for model selection were stability, great performance given a moderate sized dataset and flexibility of utilising both continuous and categorical variables effectively. We chose to perform classification initially using a Random Forests[5] model. Random Forests are particularly robust to imbalanced classes and any other inconsistencies in the data that would have to be pre-processed for other models. The model theoretically provides great performance with a comparatively low amount of time being spent to process data that is used by it.

We chose to perform a classification task rather than a regression task to simplify our model's task to predict the outcome of the match rather than the actual score at the end of the match. Since the scope of our problem is to bet on the outcomes of matches, an effective classifier would serve us well to perform our experiments. Moreover, we observed a high number (25%) of draws within the dataset (Figure 6) - a simple regression would not be able to predict this set of draws.

On evaluating the Random Forests classifier we found the results to be satisfactory, but a better performing model would be necessary to facilitate an impactful betting strategy. We decided to go with a LightGBM[6] model to push

performance to meet our goals. Light Gradient Boosting Machine (LightGBM) is a highly efficient gradient boosting method based on decision trees. This is the primary model, which will be used to predict the direction of our bets.

5.2 Metalabelling

In the case that our primary model is able to perform reasonably better than random betting, inappropriate bet-sizing can still lead to the loss of capital, which introduces the problem of bet-sizing. We refer to the metalabelling technique published by Dr Lopez De Prado in [1], which generates the binary labels,

$$y_t = \begin{pmatrix} 1, & \text{if } P_{t+1} - P_t \geq 0 \\ 0, & \text{otherwise} \end{pmatrix} \tag{1}$$

where P_{t+p} and P_t represents our capital at time step t and $t+1$, where profitable bets are labelled as 1.

As is illustrated in the model architecture below, we used a LightGBM model as our risk-management layer, which we term as our secondary model. We extracted the binary classification probabilities, which will be used to help structure our bet sizes for our model. We view the probabilities as the secondary model's confidence in the primary model's bets, where we can apply a confidence threshold to further filter our unnecessary bets which reduces the chances of loss of capital.

Figure 8 is an illustration of our overall model architecture:

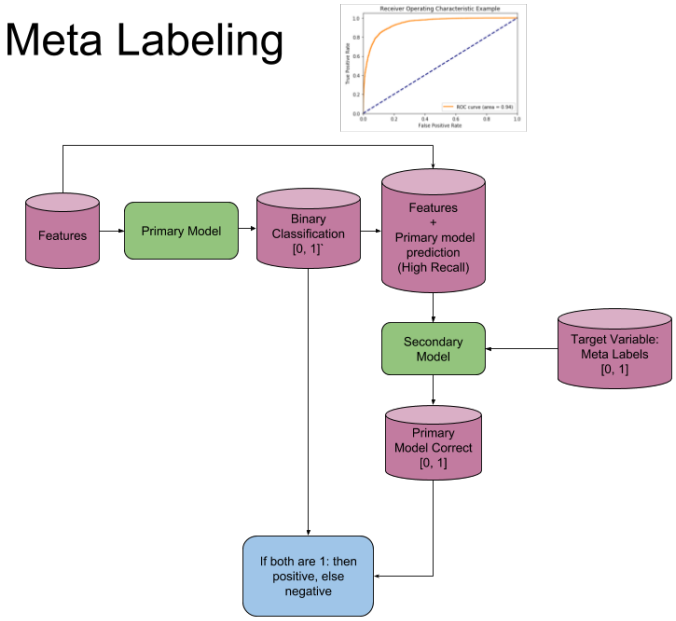


Figure 8: Overall model architecture

5.3 Hyperparameter Tuning

We used Optuna[7], a hyperparameter optimisation library that efficiently finds the best parameters using efficient halving and sampling techniques. The optimised LightGBM parameters helped to increase accuracy and prevent overfitting by adjusting the parameters according to validation accuracy.

5.4 Feature Selection

After utilising LightGBM Classifier, we used SHAP[8] (SHapley Additive exPlanations), a game theoretic approach to analyse the impact of each variable to the model. This is a holistic approach that is more consistent with human intuition.

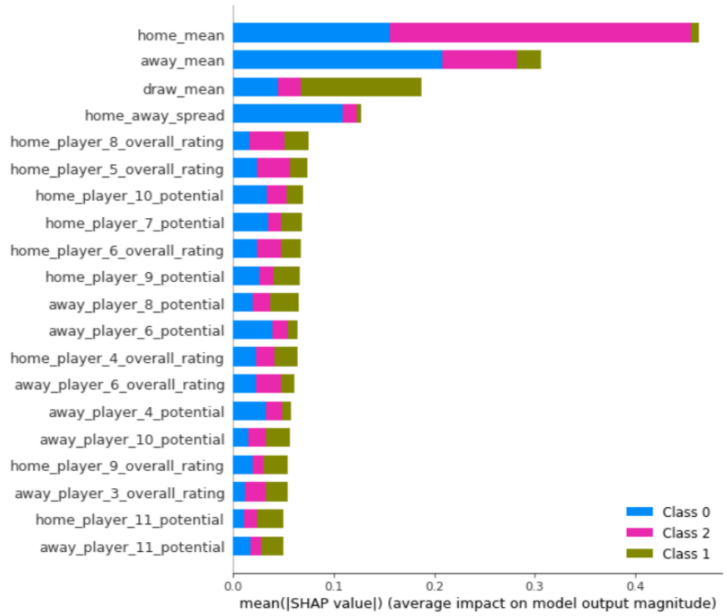


Figure 9: SHAP Plots from our model's analysis, showing the importance and priority of each variable

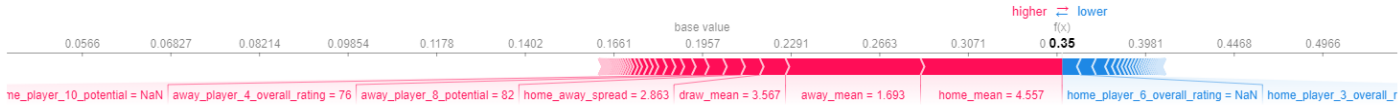


Figure 10: SHAP Explainer, after aggregating brokers

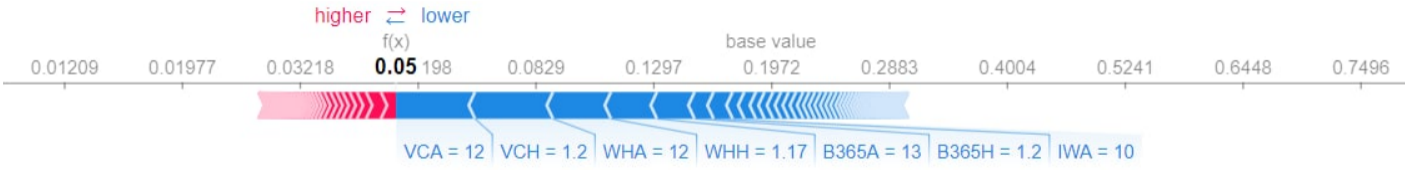


Figure 11: SHAP Explainer, before aggregating brokers

As seen from the before and after plots for aggregating broker odds, aggregating the brokers to condense broker information serves as a vital step to reduce multicollinearity effects within our data. We also confirmed our earlier hypothesis, that individual player stats were important. Looking at our SHAP plots, ratings for players 5,8 and 10 were amongst the top most important features. Interpreted in context, the Centre half back, Inside right and Inside left are critical players which could be further studied and researched upon to improve our primary model predictions.

5.5 Betting Strategy

We utilise the Kelly Criterion[9] to inform our betting strategy which focuses on wealth generation through proper asset allocation.

The Kelly Criterion is part of a mainstream betting method to maximise the expected value E log X of the logarithm of the random variable X, representing wealth. This formula most likely lead to a higher wealth compared to any other strategies in the long run, as the number of bet approaches infinity.

The practical use of Kelly Criterion is more suited for gambling, but have also been adapted into investment management and analysis over the last few decades. For our use-case, which is betting on match odds, the Kelly Criterion is arguably the best suited strategy to maximise wealth

We calculated our bet ratio as follows:

$$K\% = W - \frac{1-W}{R} \tag{2}$$

$$\beta = \gamma * K\% \tag{3}$$

where $K\%$ is the Kelly percentage, W is the metalabel classification probability for a profitable prediction (Winning probability), and R are the broker odds. β is the size of our bet, penalised by a parameter γ. The tangency portfolio in Markowitz Mean-Variance theory, is the same as the Kelly optimal portfolio which has the same objective as maximising our strategy's Sharpe Ratio[10]. In reality, the Kelly is too levered, and should be penalised according to the inherent risk of the bet. We have tested for 10%, 50%, and 100% Kelly, and have gotten significantly better performance results from using the 10% Kelly bet sizing.

6 Results

6.1 Model performance on validation and test sets

We find the model performs equally well on both validation and test sets showcasing the generalisation properties of the model. We achieve an accuracy of 51% on both sets with high precision and recall for home team wins across both sets.

TestSet Classification Report

	precision	recall	f1-score	support
-1	0.49	0.47	0.48	1301
0	0.32	0.08	0.13	1075
1	0.53	0.77	0.63	1899
accuracy			0.51	4275
macro avg	0.45	0.44	0.41	4275
weighted avg	0.46	0.51	0.46	4275

(a) Test set

ValSet Classification Report

	precision	recall	f1-score	support
-1	0.48	0.46	0.47	1221
0	0.30	0.08	0.12	1127
1	0.54	0.78	0.64	1927
accuracy			0.51	4275
macro avg	0.44	0.44	0.41	4275
weighted avg	0.46	0.51	0.45	4275

(b) Validation set

Figure 12: Classification results on Test set and Validation set

6.2 Identification of Favourable Team Characteristics

From the SHAP plot in Figure 9 we observed that the odds provided by betting sites and individual player skill exhibits the largest impact on our model. We thereby pose the theory that skilled players would naturally comprise a team with favourable characteristics, hence the reliance of our model on player potential and rating. Nonetheless, we are able to employ MDA[11] on a subset of features only comprising of only team characteristics (Figure 13). We decided to do a feature selection plot using Mean Decrease Accuracy (MDA) values as shown in Figure 13 to provide an alternative perspective. However we decided to prioritise the insights from SHAP as they have shown to be more stable to random seed permutations. Nonetheless, both plots are insightful.

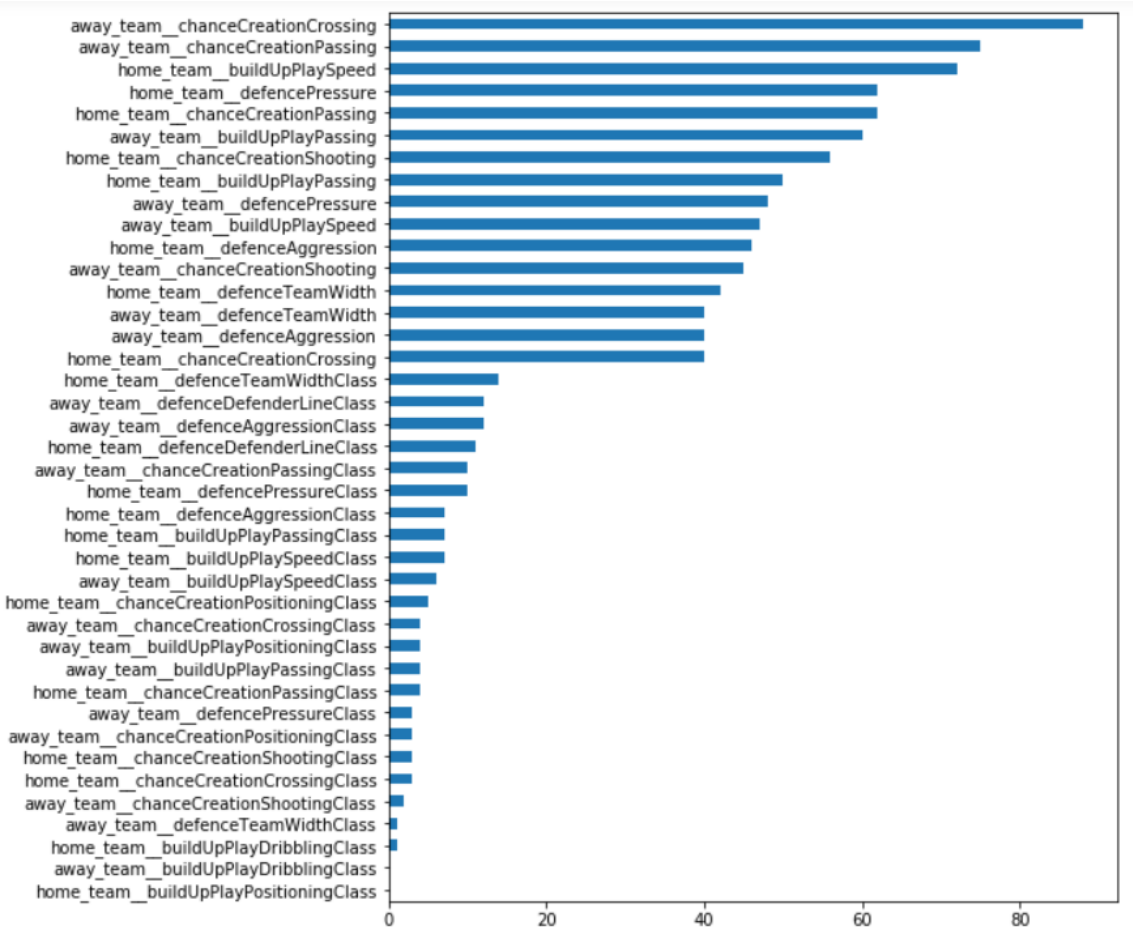


Figure 13: Team characteristics in descending order of importance

We observe that the chance creation passing, crossing and shooting variables features are the most important team characteristics to our prediction for both the home and away team sides. Team attributes with lesser but still relatively significant influence include build up play speed, and passing, as well as defence width and aggression.

6.3 Development of a Profitable Betting Strategy

6.3.1 Bet Sizing and Confidence

By betting only if we are 80% confident with 1% capital bets on our strategy and constant $1 bets for random betting, where it randomly picks out of the 3 options. The figure below shows that our strategy is outperforming the random simulation consistently through the backtested dates. We also added the comparison with S&P500 Index, adjusted, to show the comparison between our betting strategies and the index.

6.3.2 Comparison of Returns

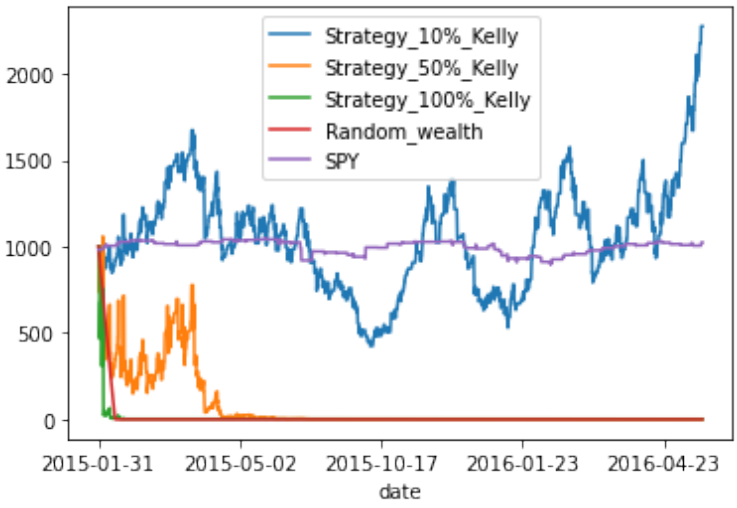


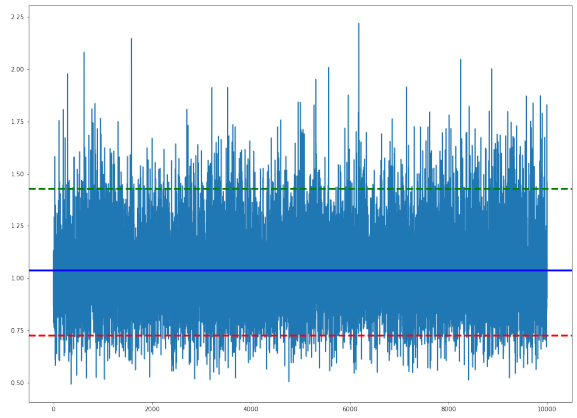
Figure 14: Comparison of Kelly Criterion against random betting and SPY

It can be clearly observed that the 10% Kelly Strategy is optimal, providing returns of 127.55%. Higher bet sizings effectively run out of capital, as does the random betting strategy. Included is the SPY index as a point of comparison to demonstrate the profitability of our strategy over the long run. Despite high drawdowns, we exhibit a significant percentage return, which is tested for statistical significance within the next section.

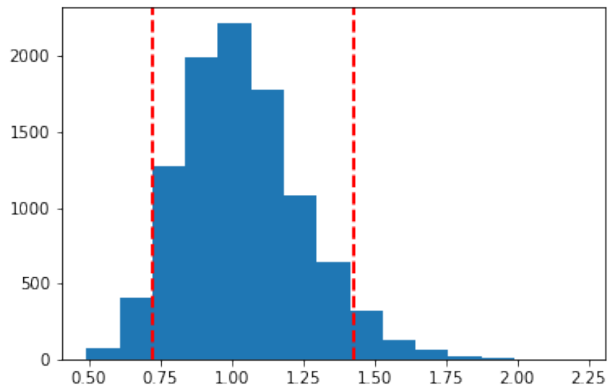
6.4 Significance Testing

A non-parametric Wilcoxon rank-sum test was performed on the returns of our strategy compared in relation to random betting to check for statistical significance in the difference of the two samples. This was chosen over a t-test for robustness, as we cannot validate the underlying normality and homoscedasticity assumptions of the t-test. The null hypothesis that the two samples possess the same mean return was clearly rejected with a p-value of <0.0001.

The bets made by our strategy yield a mean percentage return of 3.84% and variance of 4.27% per bet, as compared to a mean of -13% and variance 5.4% for random betting with constant bets. On average, our bets have a positive expected return as compared to random betting. To put this to the test, we ran 10000 simulations of 1000 consecutive bets with samples drawn from a normal distribution,



(a) Mean returns for 10000 simulations



(b) Distribution of returns for 10000 simulations

Figure 15: Monte Carlo Simulation for 10000 simulations of 1000 consecutive bets

We observe interesting results above. The red lines indicate the 5% and 95% confidence points for a sample portfolio terminal value, meaning the ending values of the portfolio after 1000 consecutive bets. The maximum loss from our simulation was 51.8%, with an expected return of 101.73%. At the 1% p-value, the expected loss is 38.96%.

Conforming to our expectations, sports-betting is a highly risky alternative to investing, which is duly compensated with a higher chance of extreme returns as compared to a buy-and-hold investing strategy in the S&P index.

7 Future Research

Due to time constraints, we were unable to expand upon our strategy as deeply as we wanted. We've listed below a few points that could be considered in future research of this area:

- Portfolio optimisation of betting-strategy to hedge risks against equity investments
- Optimise against other known ratios like Information, Calmar and Sortino ratio in addition to the Sharpe ratio
- Use SHAP highlighted features on our val-set, to include a manual intervention layer in the primary model. For example reject all bets as long as there is a player with ¡60 rating

8 Conclusion

Our report demonstrates a machine learning informed betting strategy that produces a return of 127.55% over a period of 1 year and 3 months of football matches in our test-set. The accuracy of predictions by our machine learning algorithm were significantly higher than that of random betting, achieving economically and statistically significant returns. It was found that the most profitable strategy was determined by a relatively conservative approach using a bet sizing of 10% of the Kelly criterion on a prediction confidence of 85% dictated by metalabelling.

We also showcase the features that football team management might want to focus on while building a team it to maximise chances of winning. It was found that the centre back player, and chance creation features of the team were the most influential in determining the winner of a match.

References

[1] Marcos Lopez de Prado. *Advances in Financial Machine Learning*. Wiley Publishing, 1st edition, 2018.

[2] Johannes Stübinger, Benedikt Mangold, and Julian Knoll. Machine learning in football betting: Prediction of match results based on player characteristics. *Applied Sciences*, 10(1), 2020.

[3] Francois Bertrand. Sweetviz.

[4] Ernest Chan and Ray Ng. Optimizing trading strategies without overfitting, 2017.

[5] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[6] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[7] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[8] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874, 2017.

[9] J. L. Kelly. A new interpretation of information rate. *The Bell System Technical Journal*, 35(4):917–926, 1956.

[10] William F. Sharpe. The sharpe ratio. *The Journal of Portfolio Management*, 21(1):49–58, 1994.

[11] Xin Man and Ernest P. Chan. The best way to select features? comparing mda, lime, and shap. *The Journal of Financial Data Science*, 2020.