PREDICTING SOCCER TEAM PERFORMANCE TO BEAT BOOKMAKERS' ODDS

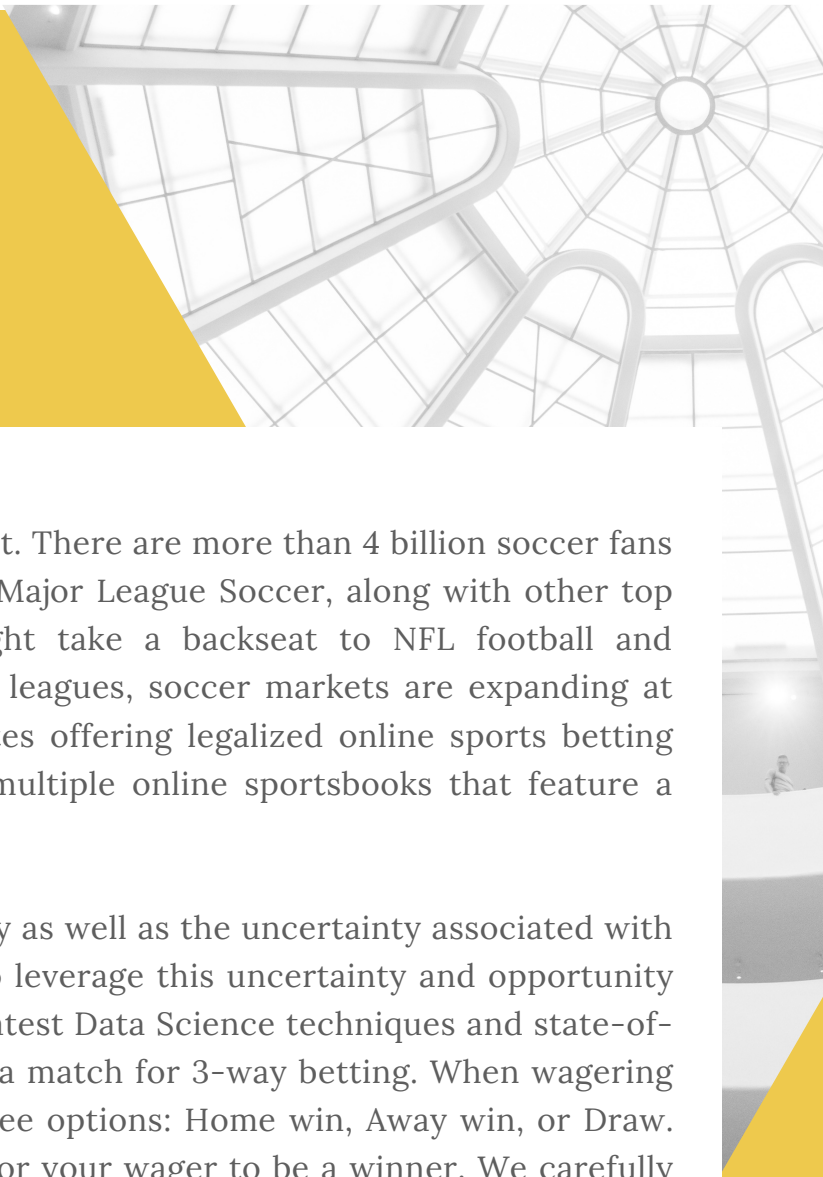
PREPARED BY

Team 23

East Coast Regional Datathon 2021

TABLE OF CONTENTS

EXECUTIVE SUMMARY

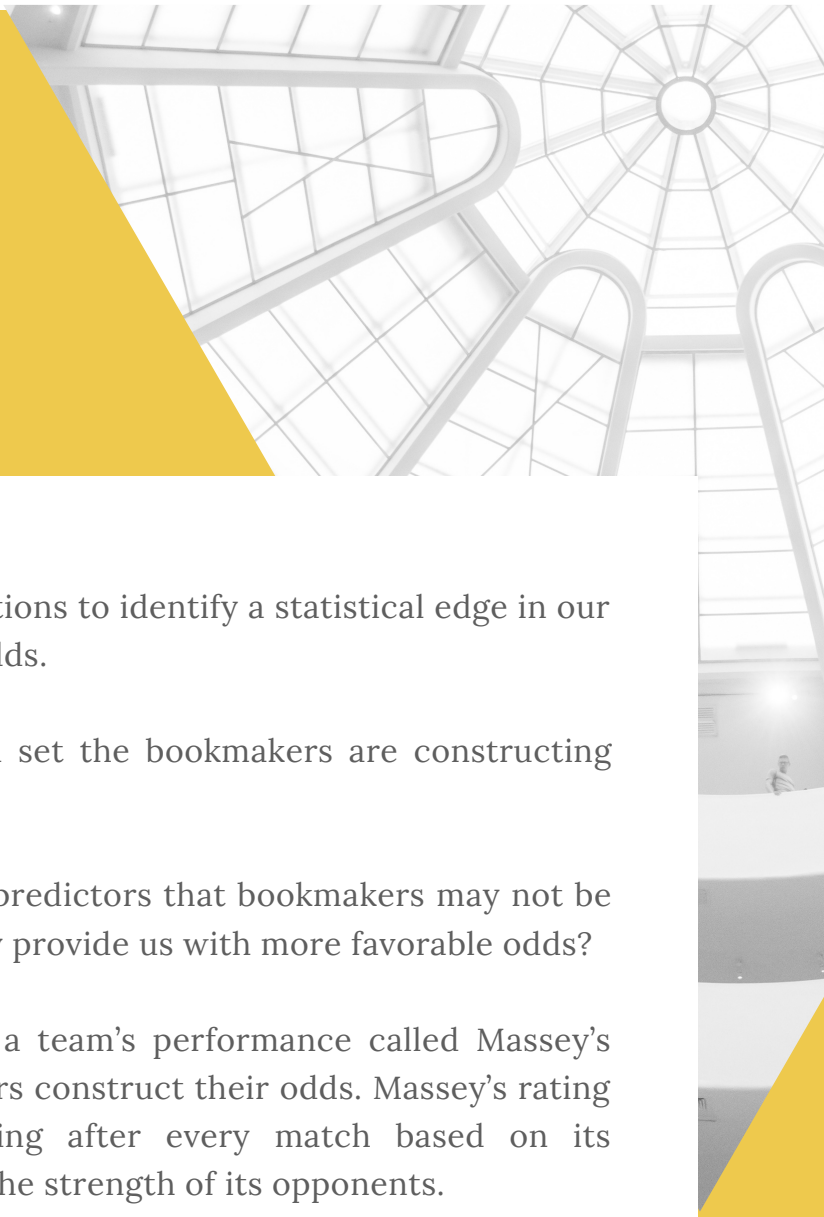


Soccer is the world's most popular sport. There are more than 4 billion soccer fans worldwide. In the US, many wager on Major League Soccer, along with other top leagues in the world. The sport might take a backseat to NFL football and basketball, but with so many different leagues, soccer markets are expanding at American sportsbooks. With more states offering legalized online sports betting than ever before, most are home to multiple online sportsbooks that feature a variety of soccer betting odds.

With these huge markets and popularity as well as the uncertainty associated with this relatively recent sport, we want to leverage this uncertainty and opportunity to beat the market returns. Using the latest Data Science techniques and state-of-the-art models we predict the odds of a match for 3-way betting. When wagering on the 3-way money line there are three options: Home win, Away win, or Draw. Your selected option must be correct for your wager to be a winner. We carefully curate a strategy based on the results and further data analyses of the model and data on when to bet and how much to bet. In the following pages, we will explain comprehensively our thought process, data analysis techniques, and the methods and actions taken to reach our results. We look at various aspects of the game from team composition, ratings data, players data, and past soccer match data to reach our conclusions.

Our testing shows that we can identify a betting strategy that can outperform the bookmakers' odds. Our development is based on understanding how the bookmakers construct those odds and identifying additional information bookmakers may not be considering.

RESEARCH QUESTIONS



We investigated the following questions to identify a statistical edge in our strategy to beat the bookmakers' odds.

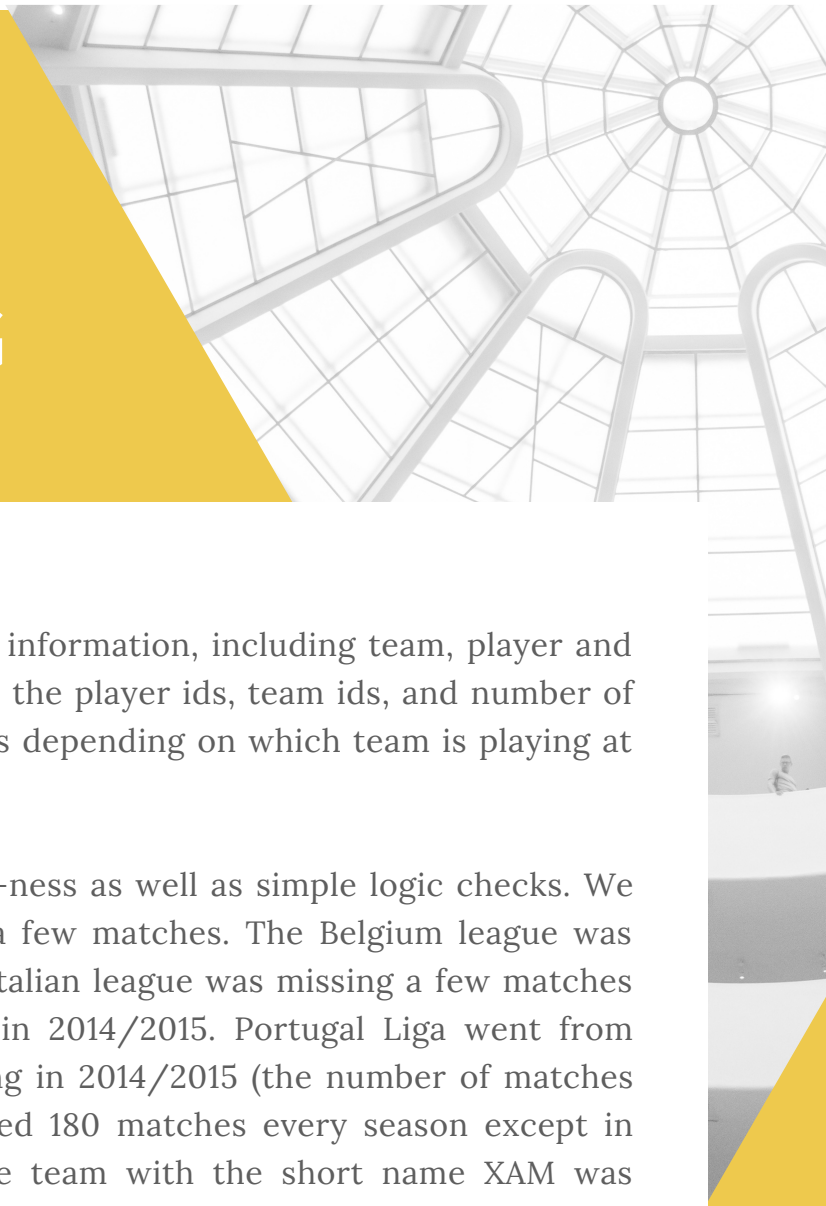
1. Can we identify the information set the bookmakers are constructing their betting odds with?

2. If so, can we identify additional predictors that bookmakers may not be incorporating, that can consistently provide us with more favorable odds?

First, we explore a way of rating a team's performance called Massey's rating as a proxy of how bookmakers construct their odds. Massey's rating is able to capture a team's rating after every match based on its performance from that match and the strength of its opponents.

Once we know how bookmakers make their decisions, we seek to find additional information they might not be incorporating. We investigate whether a team's winning or losing streak can continue after taking into account the strengths of its current and previous opponents.

Next, we test whether there are attributes or playing styles that consistently drive results. Given that teams' player composition and play styles change from season-to-season or even matches-to-matches, being able to identify when some attributes or play styles beat another will help us identify scenarios we may have an edge in.



DATA WRANGLING

Match Data

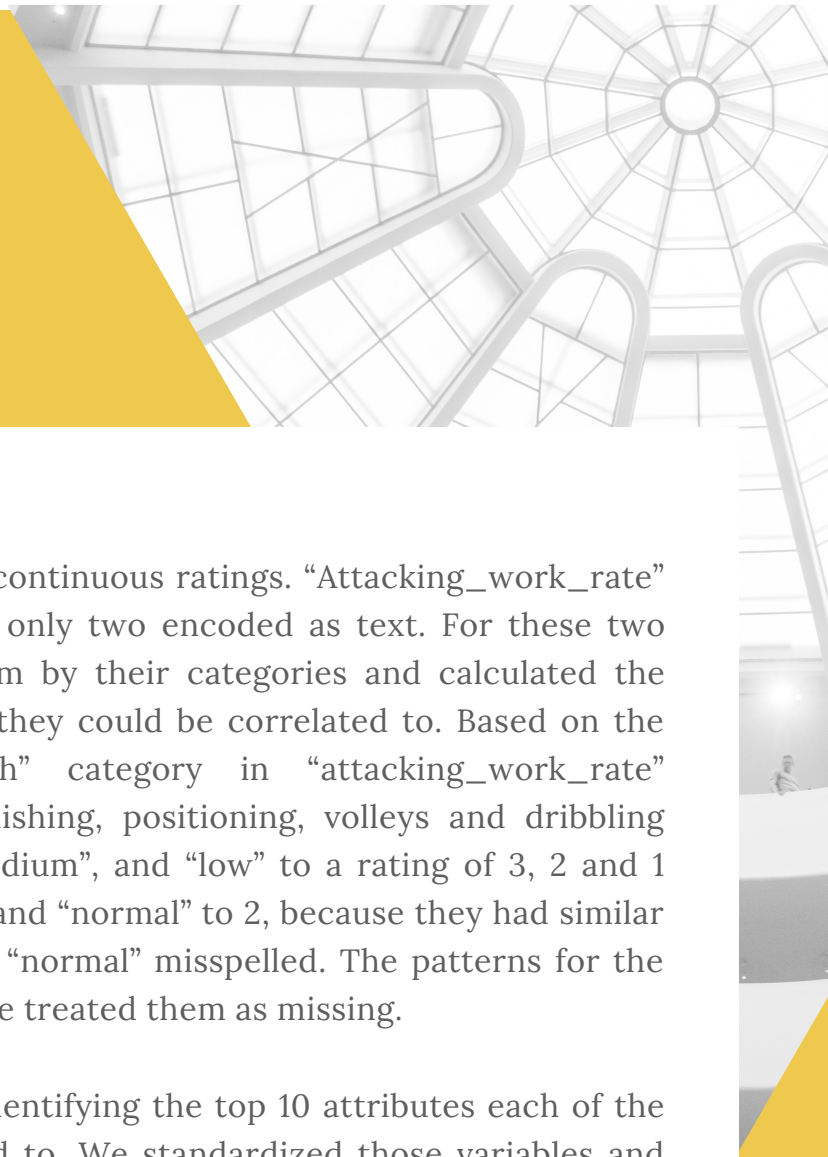
The match data contains the match information, including team, player and odds data, in a wide format. That is, the player ids, team ids, and number of goals are split into separate columns depending on which team is playing at home.

We checked each variable's missing-ness as well as simple logic checks. We found that we were missing quite a few matches. The Belgium league was missing the 2013/2014 season, the Italian league was missing a few matches in 2011/2012 and one later match in 2014/2015. Portugal Liga went from playing 16 teams to 18 teams starting in 2014/2015 (the number of matches from 240 to 306). Switzerland played 180 matches every season except in 2011/2012, where it played 162; the team with the short name XAM was disqualified halfway through. We were able to replace the missing data with data from Sports Statistics. The downloaded data was nearly identical (some of the teams were named slightly different). We also verified their data against ESPN, Scorespro, and Europsport's websites.

We later separated this dataset based on the home team and away team, and stacked them so each team's information is by row (a home indicator column was created). This made it easier to merge in the cleaned version of the player and team attributes.

DATA WRANGLING

Player Attributes Data



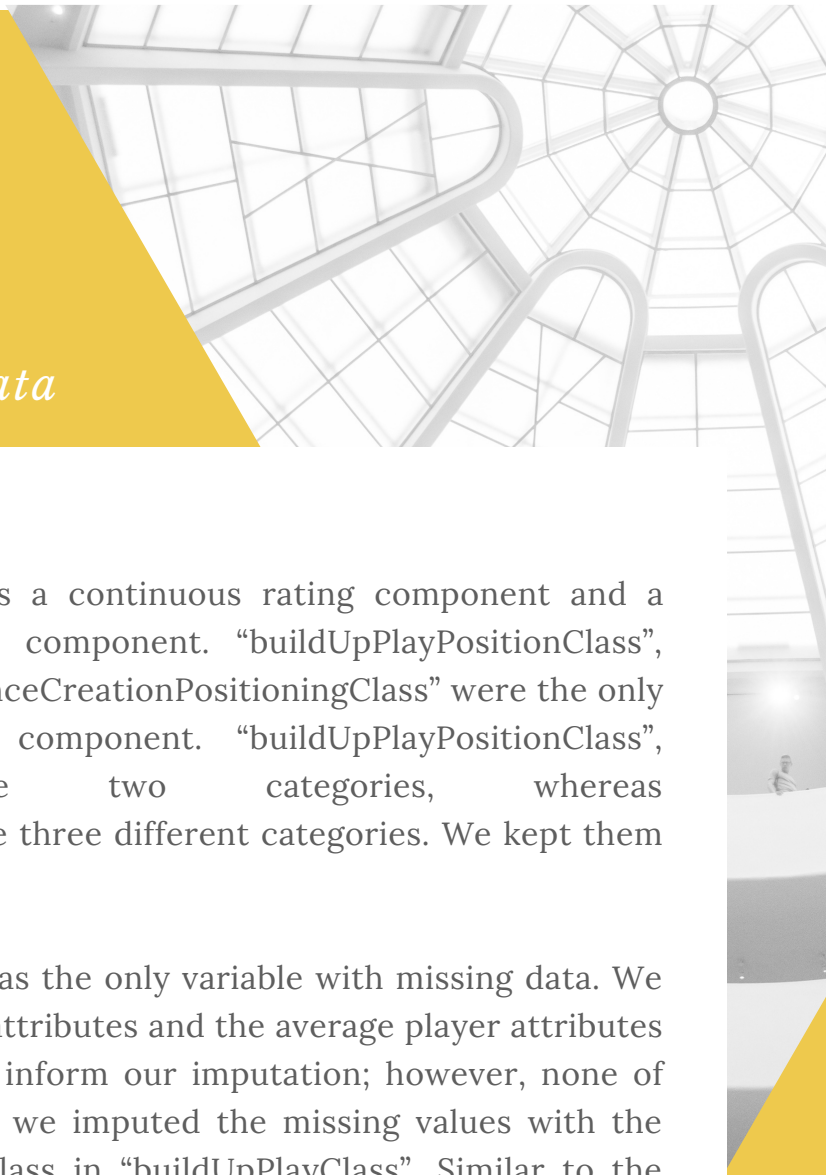
Most of the player attribute data are continuous ratings. “Attacking_work_rate” and “defensive_work_rate” were the only two encoded as text. For these two categorial variables, we grouped them by their categories and calculated the mean rating of variables we thought they could be correlated to. Based on the patterns we saw (e.g. the “high” category in “attacking_work_rate” corresponding to a high average finishing, positioning, volleys and dribbling ratings), we then mapped “high”, “medium”, and “low” to a rating of 3, 2 and 1 respectively. We also mapped “norm” and “normal” to 2, because they had similar ratings to medium and appears to be “normal” misspelled. The patterns for the other categories weren’t as clear, so we treated them as missing.

We then dealt with missing data by identifying the top 10 attributes each of the missing attribute was most correlated to. We standardized those variables and used 30 nearest neighbors to impute the missing data. For example, the attacking work rate attribute was highly correlated to the attributes long shots, volleys, crossing, curve, finishing, positioning, dribbling, agility, sprint speed and acceleration; these attributes made sense logically. The other attributes we applied the same methodology to include defensive work rate, volleys, curve, agility, balance, jumping, vision and sliding tackle.

Since player data was updated at different times and not by match, we also had to tie the appropriate data in time to each match. If a player played a match yesterday, we can only use that player’s attributes that were updated prior to yesterday and not attributes that were updated tomorrow (as to avoid information leakage). We then created a mapping of player attributes to the appropriate match id, where we only kept the attribute data that was updated most recently prior to the match date.

DATA WRANGLING

Team Attributes and Odds Data



Team Attributes

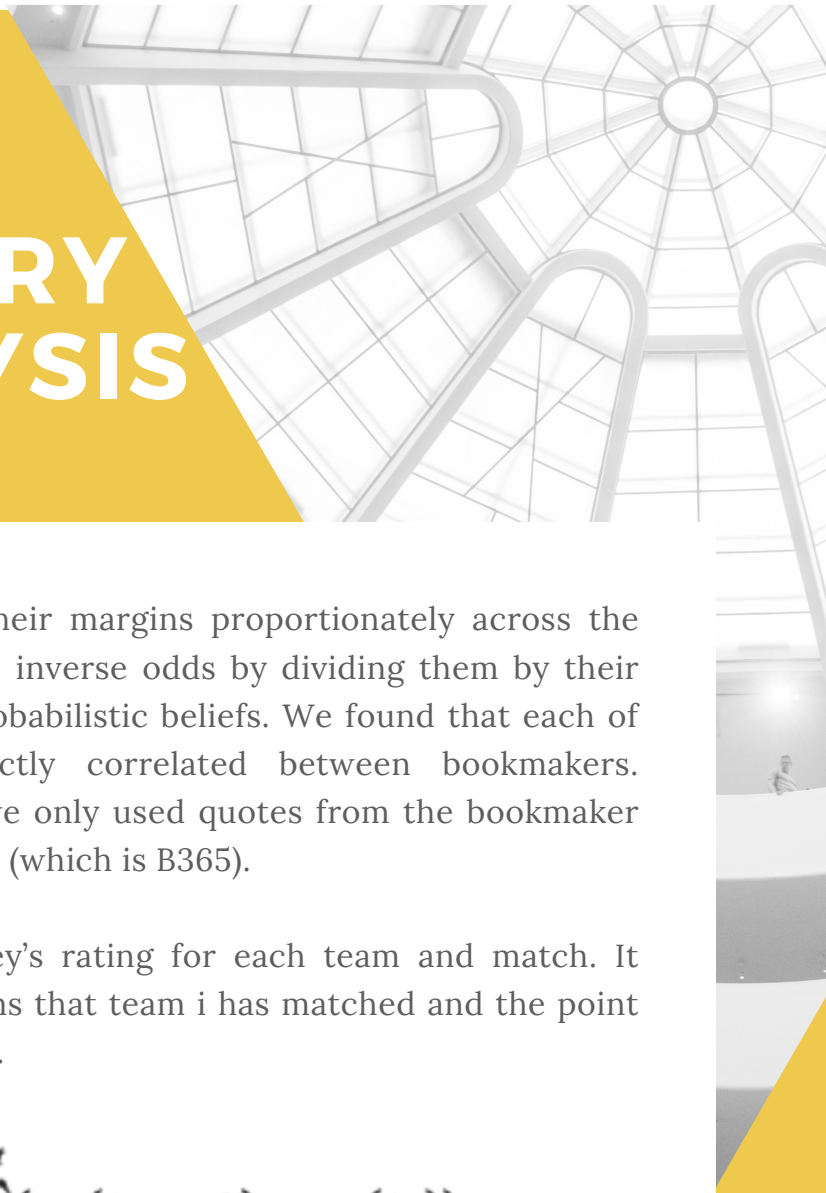
Most of the team attribute data has a continuous rating component and a categorial encoding (named “class”) component. “buildUpPlayPositionClass”, “defenceDefenderLineClass” and “chanceCreationPositioningClass” were the only three variables without a rating component. “buildUpPlayPositionClass”, “defenceDefenderLineClass” have two categories, whereas “chanceCreationPositioningClass” have three different categories. We kept them as categorical variables.

The variable “buildUpPlayDribbling” was the only variable with missing data. We first tried to identify correlated team attributes and the average player attributes with the hope of using them to help inform our imputation; however, none of them had a strong correlation. Thus, we imputed the missing values with the average value of its corresponding class in “buildUpPlayClass”. Similar to the previous section, we matched the team attributes to the appropriate match in time (we used the most recent data prior to the match date).

Odds Data

We noticed that significant portions of betting odds data Switzerland Super League and Poland Ekstraklasa were missing. Therefore, we gathered the odds data for those leagues from a database of historical European betting odds. However, the database only had records dating from 2012, so we were only able to gather half of the odds data needed for Switzerland and Poland.

In order to merge the datasets, we had to reformat the teams' names and dates. The team names in our data had a lot of special characters in Polish and German that could not be displayed. The dates had inconsistent ordering of month and day. For Switzerland, we eventually had 720 new entries. For Poland, we got 825 new entries.



EXPLORATORY DATA ANALYSIS

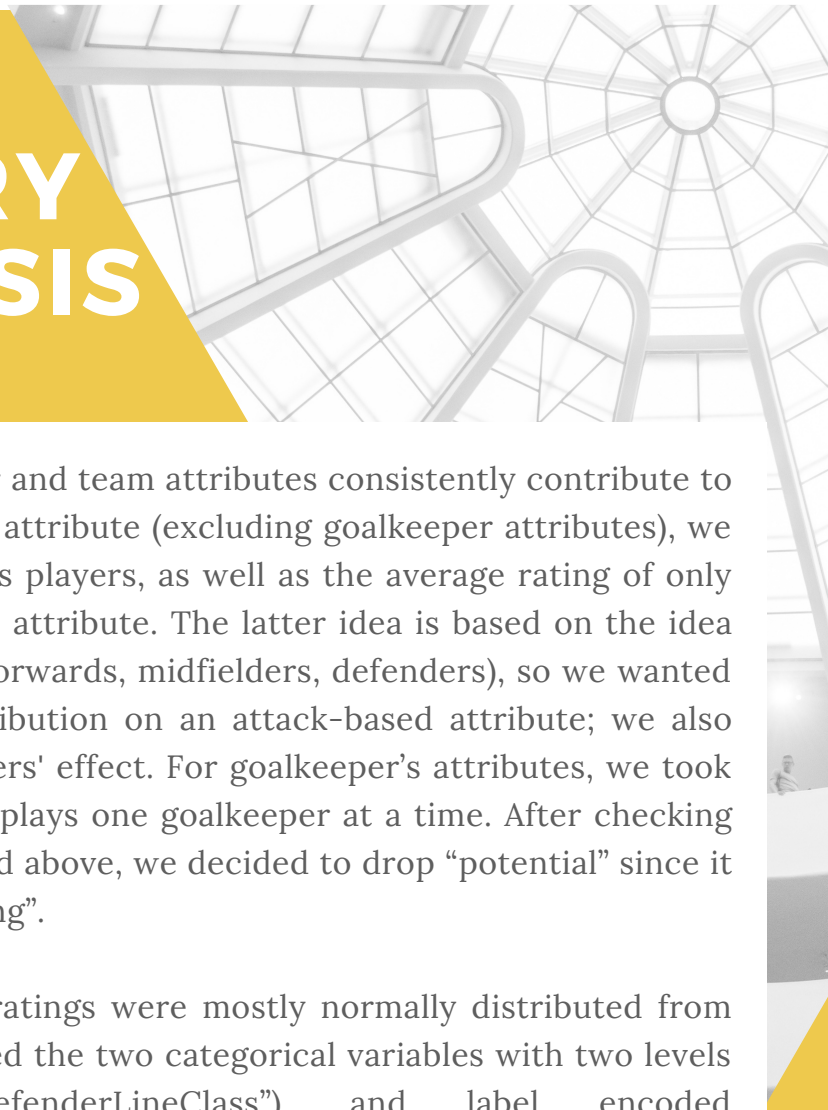
Bookmakers' Odds

We assume that bookmakers add their margins proportionately across the three outcomes, and normalized the inverse odds by dividing them by their sum to arrive at the bookmakers' probabilistic beliefs. We found that each of these probabilities is near perfectly correlated between bookmakers. Therefore, throughout this report, we only used quotes from the bookmaker with the least amount of missing data (which is B365).

We then calculated temporal Massey's rating for each team and match. It captures the historical rating of teams that team i has matched and the point spread (outcome) of team i's matches.

$$r_i(t) = \frac{1}{m_{i,t}} \sum_{k=1}^{m_{i,t}} (r_{j_k}(t_k - 1) + s_i(t_k)).$$

For each match, we calculated the difference between the two team's Massey's ratings; this had a 0.805 correlation with the spread quoted by the bookmakers after stage 4 (and 0.872 after 10 stages). By identifying that bookmakers's underlying probabilistic beliefs are closely related to this rating, we can test which predictors contain additional/orthogonal information.



EXPLORATORY DATA ANALYSIS

Player and Team Attributes

We analyzed which combinations of player and team attributes consistently contribute to a win across seasons. For each match and attribute (excluding goalkeeper attributes), we calculate the average rating of each team's players, as well as the average rating of only the 3 players with the top ratings for that attribute. The latter idea is based on the idea that players play different positions (e.g. forwards, midfielders, defenders), so we wanted to give less weight to a defender's contribution on an attack-based attribute; we also wanted to capture any potential star players' effect. For goalkeeper's attributes, we took the maximum value since each team only plays one goalkeeper at a time. After checking the correlation between features generated above, we decided to drop "potential" since it was over 90% correlated with "overall rating".

For the team attributes, the continuous ratings were mostly normally distributed from visual inspection. We then one hot encoded the two categorical variables with two levels ("buildUpPlayPositionClass", "defenceDefenderLineClass"), and label encoded "chanceCreationPositioningClass". We decided not to bin the variables given that almost all of our features are

To create the training data, we calculated the difference between each of the two team's attributes with respect to the home team (e.g. avg overall rating of home team's players - avg overall rating of away team player's). We then standardized the differences by league and season, because we saw distinct differences between leagues and seasons; this allows us to more closely analyze the effects of attributes across leagues and seasons. We then labeled our target variable as 1 if the home team won, 0 if it's a draw, and -1 if the home team lost. Finally, we only kept matches where we had both complete player attribute (about 82% of the matches each season had the complete 11 player data for both teams) and team attribute data.

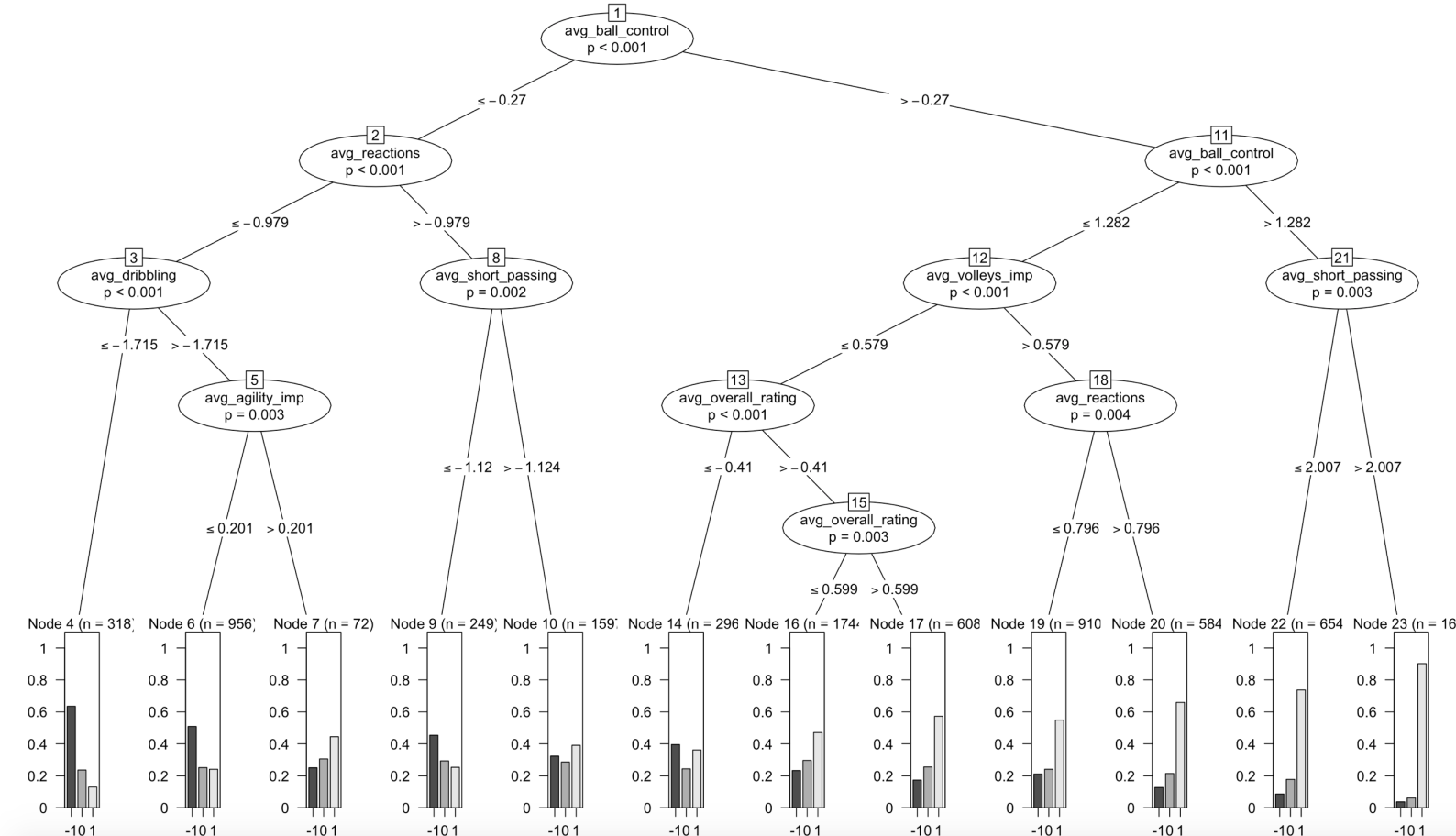
We then applied time-series cross-validation for hyperparameter tuning and identifying features that were statistically significant across seasons. We tested different training periods looking back one season, two seasons, and three seasons then evaluated on the next season. We applied Conditional Inference Trees (Ctree), which recursively partition variables based on its correlation to the target variable.

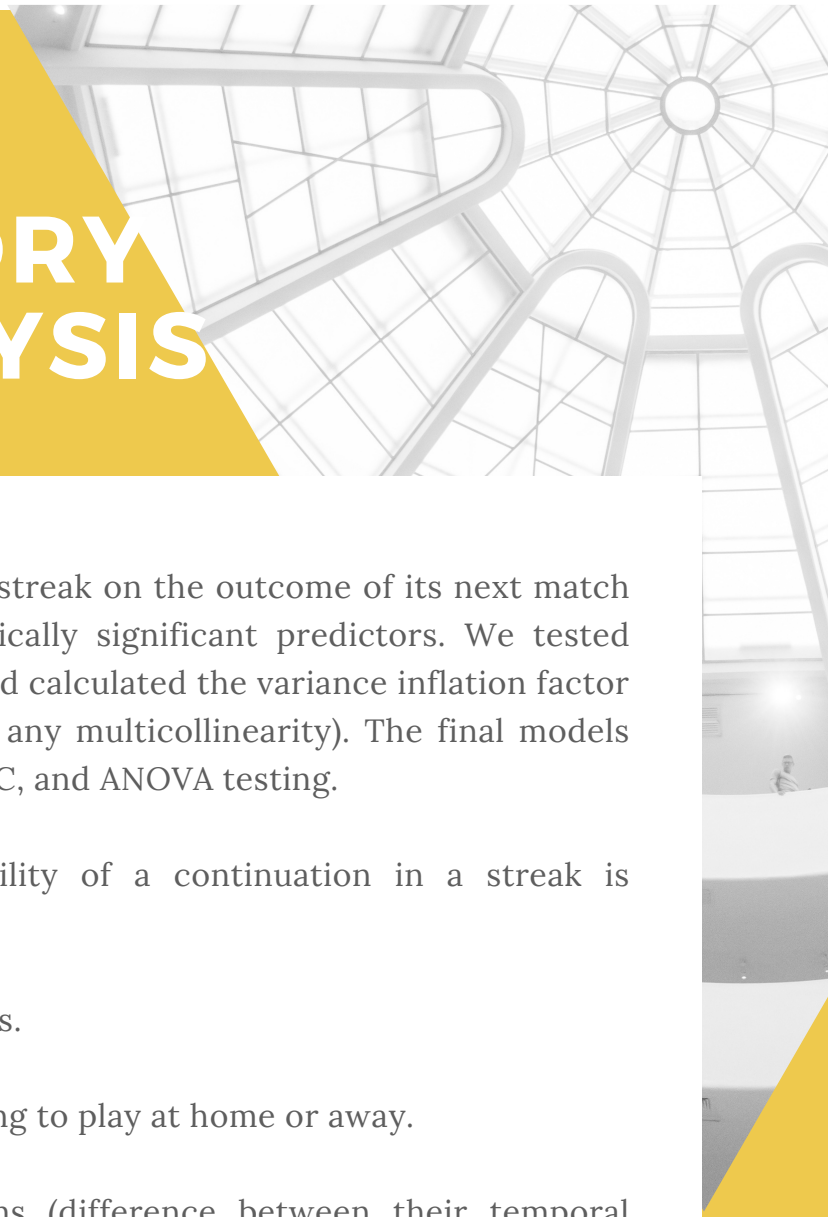
EXPLORATORY DATA ANALYSIS

Player and Team Attributes

Ctree algorithm tests the null hypothesis of independence between the differences in attributes and outcome and selects the difference in attributes with the highest p-value after applying Bonferroni correction. We chose Ctree because we hypothesized that higher-order interactions of the ways (attributes) teams are different contribute to a higher chance of a particular outcome. Moreover, we can perform significant testing to identify the degree/level of interaction.

The best hyperparameter set was chosen with the smallest average cross-entropy loss on the test sets. Ctree generates the classification by nodes and provides the most significant features like in the graph below. We then selected the features that were consistently predictive across time. Those features are avg overall rating, avg volleys, avg long passing, avg finishing, avg ball control, avg dribbling, avg short passing, avg reactions, avg vision, avg penalties, avg agility. It turns out that the difference in team attributes wasn't as predictive after accounting for player attributes.





EXPLORATORY DATA ANALYSIS

Wining/Losing Streak

We analyzed the impact of a team's streak on the outcome of its next match with the goal of identifying statistically significant predictors. We tested several logistic regression models and calculated the variance inflation factor of each variable (we did not detect any multicollinearity). The final models were chosen based on parsimony, AIC, and ANOVA testing.

We hypothesized that the probability of a continuation in a streak is determined by:

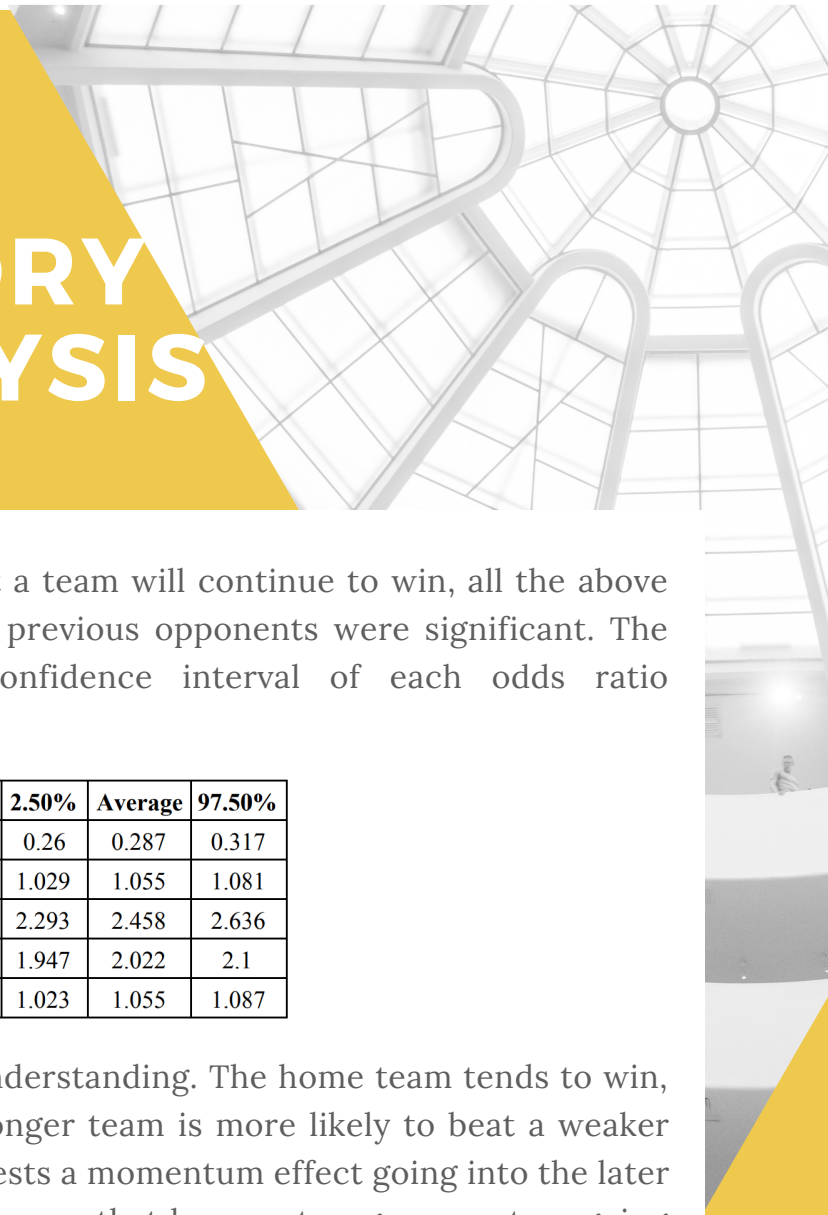
1. The number of cumulative win/loss.

2. Whether the streaking team is going to play at home or away.

3. The strengths of the two teams (difference between their temporal Massey's ratings).

4. The strengths of the previous teams the streaker played against (the average historical temporal Massey's ratings of the opposing teams).

5. Which stage of the league is the match going to be played (e.g. first match, semifinal, final). We first transformed the stage variable to each team's match number to put every team in the same league on the same starting point. We then binned the match numbers into four equal-sized bins so that each league is put on the same reference point (since some leagues play more matches than others).



EXPLORATORY DATA ANALYSIS

Wining/Losing Streak

We found that for the probability that a team will continue to win, all the above variables except the strengths of the previous opponents were significant. The table below shows is the 95% confidence interval of each odds ratio (exponentiated log odds).

	2.50%	Average	97.50%
Intercept	0.26	0.287	0.317
Cumulative Win	1.029	1.055	1.081
Home	2.293	2.458	2.636
Rating Difference	1.947	2.022	2.1
Match # Binned	1.023	1.055	1.087

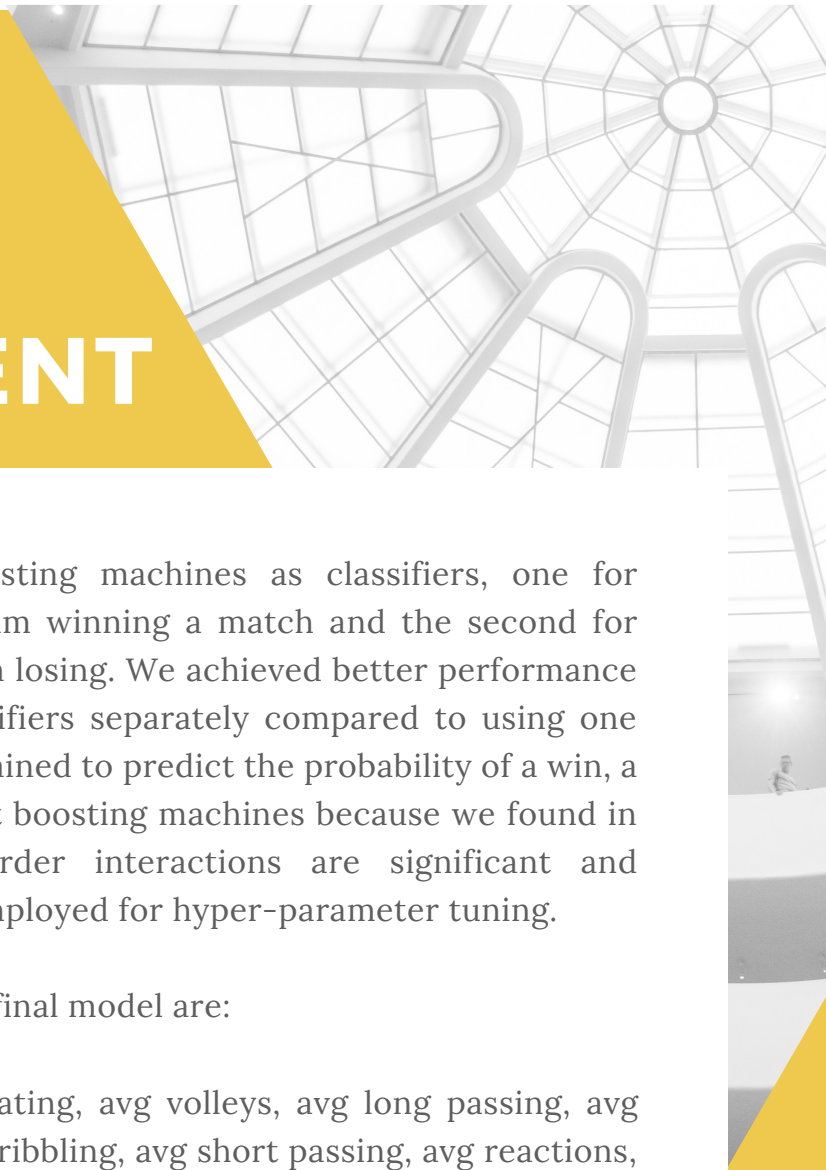
These findings are aligned with our understanding. The home team tends to win, streakers tend to continue, and a stronger team is more likely to beat a weaker team. The match number's effect suggests a momentum effect going into the later stages. Looking at data for the finals, teams that have a strong momentum going into the game are more likely to win.

We found that for the probability that a team will continue to lose, all the hypothesized variables except the strengths of the previous opponents were significant. A similar table is provided below.

	2.50%	Average	97.50%
Intercept	0.732	0.812	0.901
Cumulative Loss	0.937	0.966	0.995
Home	0.41	0.438	0.469
Rating Difference	0.51	0.53	0.551
Match # Binned	1.02	1.05	1.082

These findings show that a team on a losing streak is less likely to continue losing. This could be a "backs against the wall" effect. The home team and the stronger teams are less likely to lose. Finally, the match number suggests a momentum effect going into the later stages. A team is more likely to continue losing going into the finals because of the negative mentality.

MODEL DEVELOPMENT



We developed two gradient boosting machines as classifiers, one for predicting the probability of a team winning a match and the second for predicting the probability of a team losing. We achieved better performance optimizing the two different classifiers separately compared to using one multi-class classifier (which was trained to predict the probability of a win, a draw, or a loss). We chose gradient boosting machines because we found in previous sections that higher-order interactions are significant and predictive. Cross-validation was employed for hyper-parameter tuning.

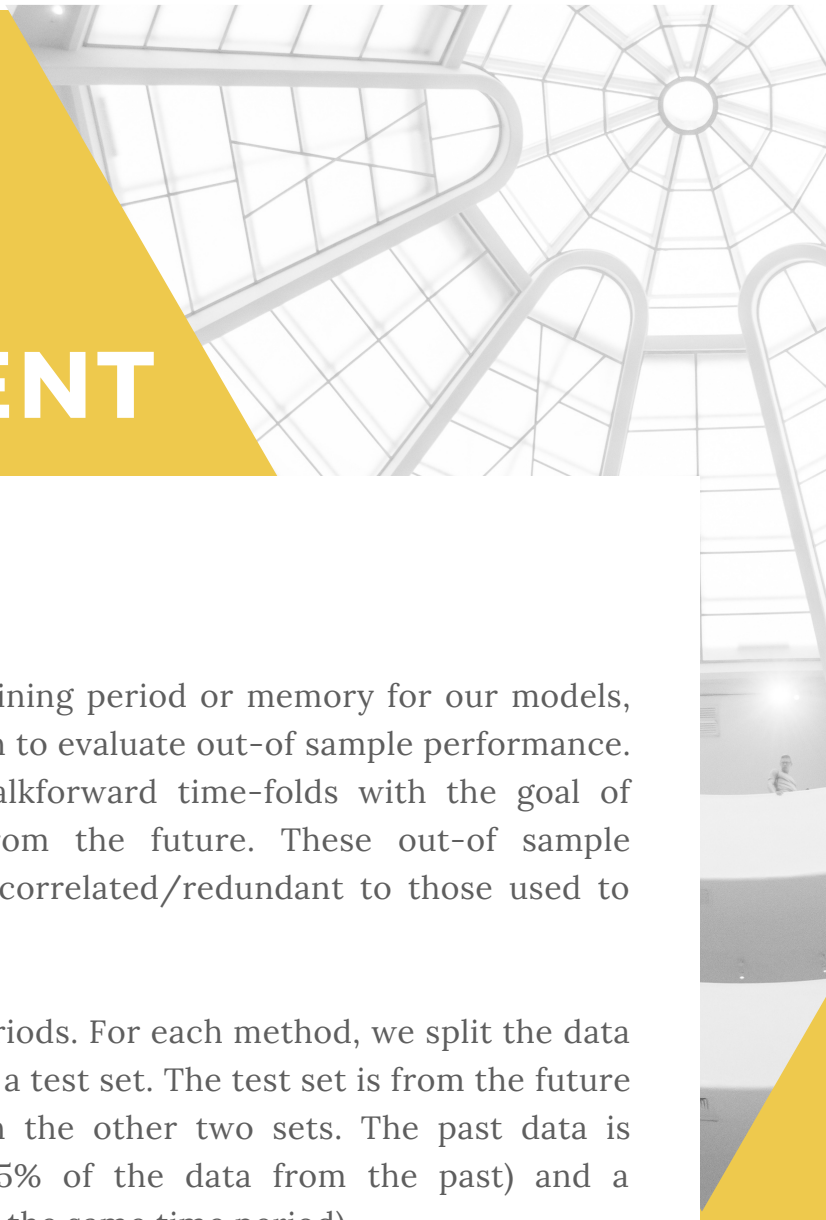
The features we've included in the final model are:

- Player attributes: avg overall rating, avg volleys, avg long passing, avg finishing, avg ball control, avg dribbling, avg short passing, avg reactions, avg vision, avg penalties, avg agility,
- Other: league id, rating difference, home, match # binned, cumulative win/loss

Probability Calibration

Since our strategy is based on predicting the probability of a win or loss, it is crucial that we have accurate probability estimates. We used a sigmoid regressor based on Platt's logistic model for calibration, and cross-validation to ensure unbiased data is fed to the calibrator.

MODEL DEVELOPMENT



Time-Series Cross-Validation

To determine the most relevant training period or memory for our models, we used time-series cross-validation to evaluate out-of sample performance. That is, we split the data into walkforward time-folds with the goal of preventing information leakage from the future. These out-of sample observations are least likely to be correlated/redundant to those used to train in the model.

We tested four different training periods. For each method, we split the data into a train set, a calibration set and a test set. The test set is from the future and does not overlap in time with the other two sets. The past data is randomly split into a train set (75% of the data from the past) and a calibration set (25% of the data from the same time period).

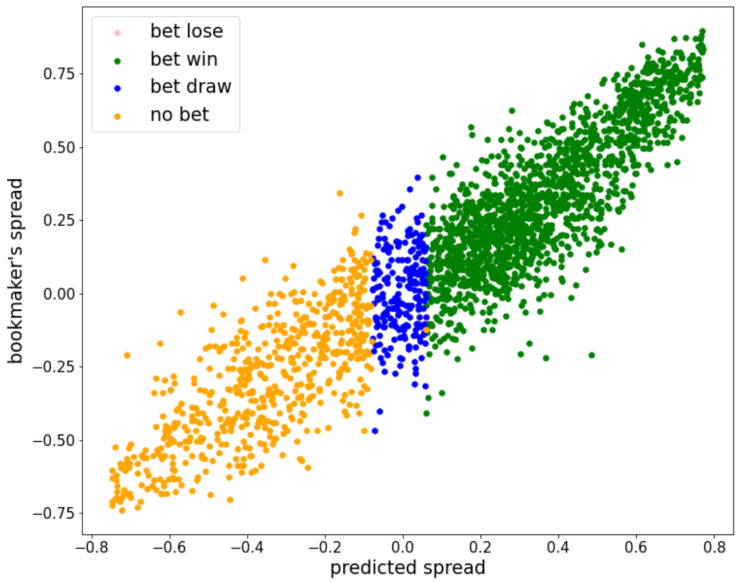
The test sets are the following seasons: 2010/2011, 2011/2012, 2012/2013, 2013/2014, 2014/2015 and 2015/2016. For each of the test seasons, the train and calibration sets are constructed from the previous one, two and three seasons.

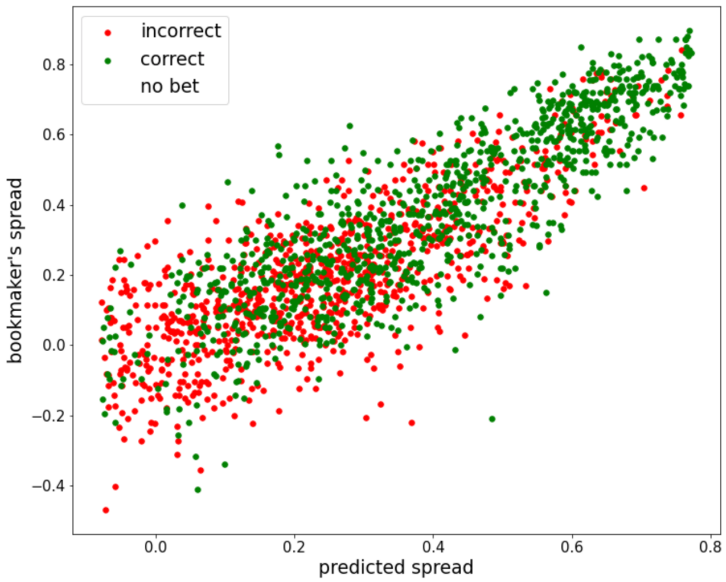
The fourth method constructs the train and calibration sets from the previous 3 seasons and the first half of the current season, then evaluates the calibrated model on the second half of the current season.

MODEL DEVELOPMENT

Results

We used the two calibrated classifiers to predict on future test data, normalized the probabilities (so that they sum up to one) then calculated the spread. Throughout this analysis, we will present results with respect to the home team. The graphs below show our bets for the 2013/2014 season (trained on 2010/2011, 2011/2012, and 2012/2013).





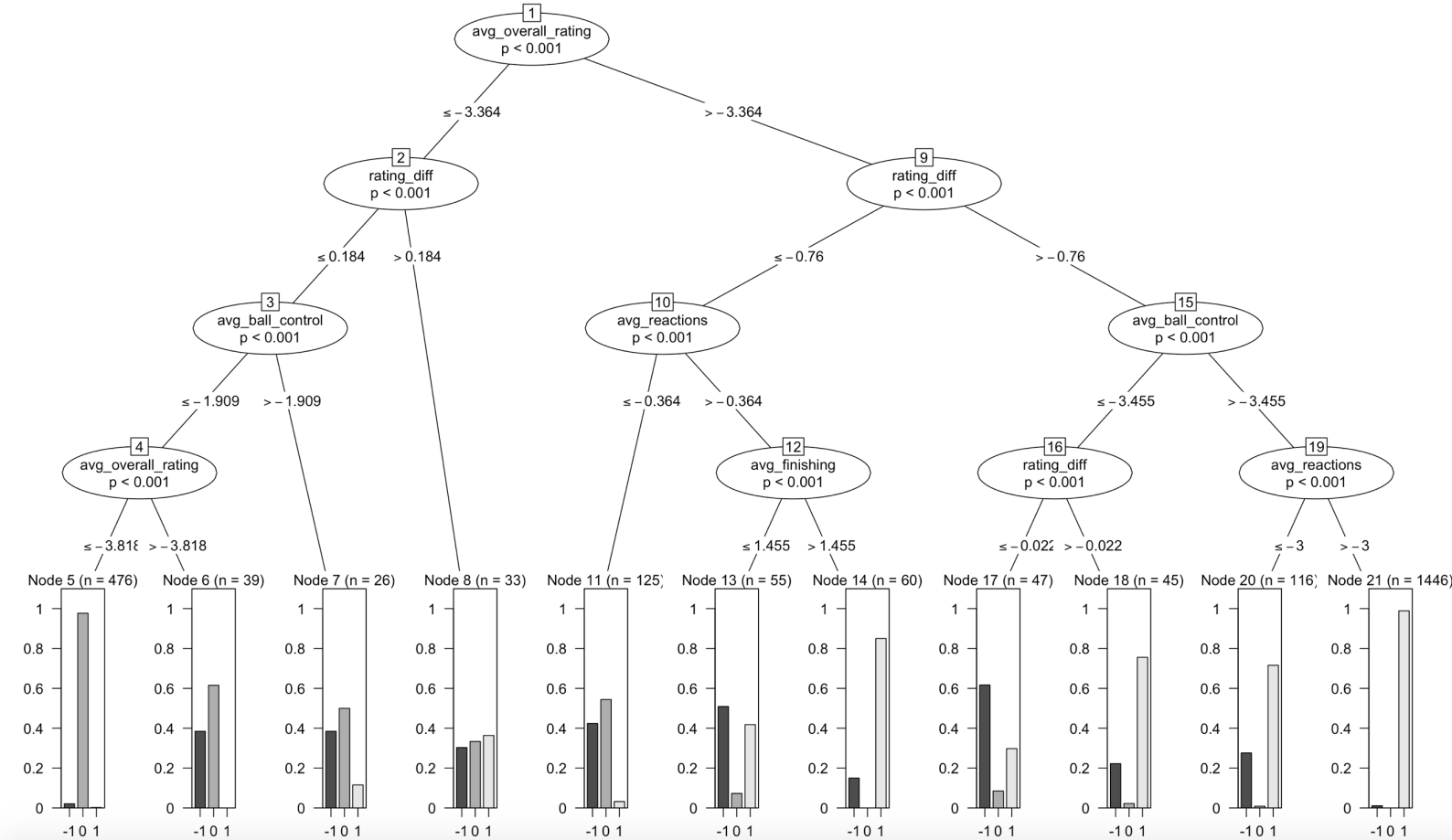
The graph on the left shows our predicted spread against the bookmaker's spread. The green dots are when we bet on the home team winning, the blue dots are when we bet on a tie and the orange dots are when we take no positions. The graph on the right shows whether the bets we took had a positive return. The right graph shows that our models are more accurate when both the spreads are high (when our model and the bookmaker are more confident in a win).

MODEL DEVELOPMENT

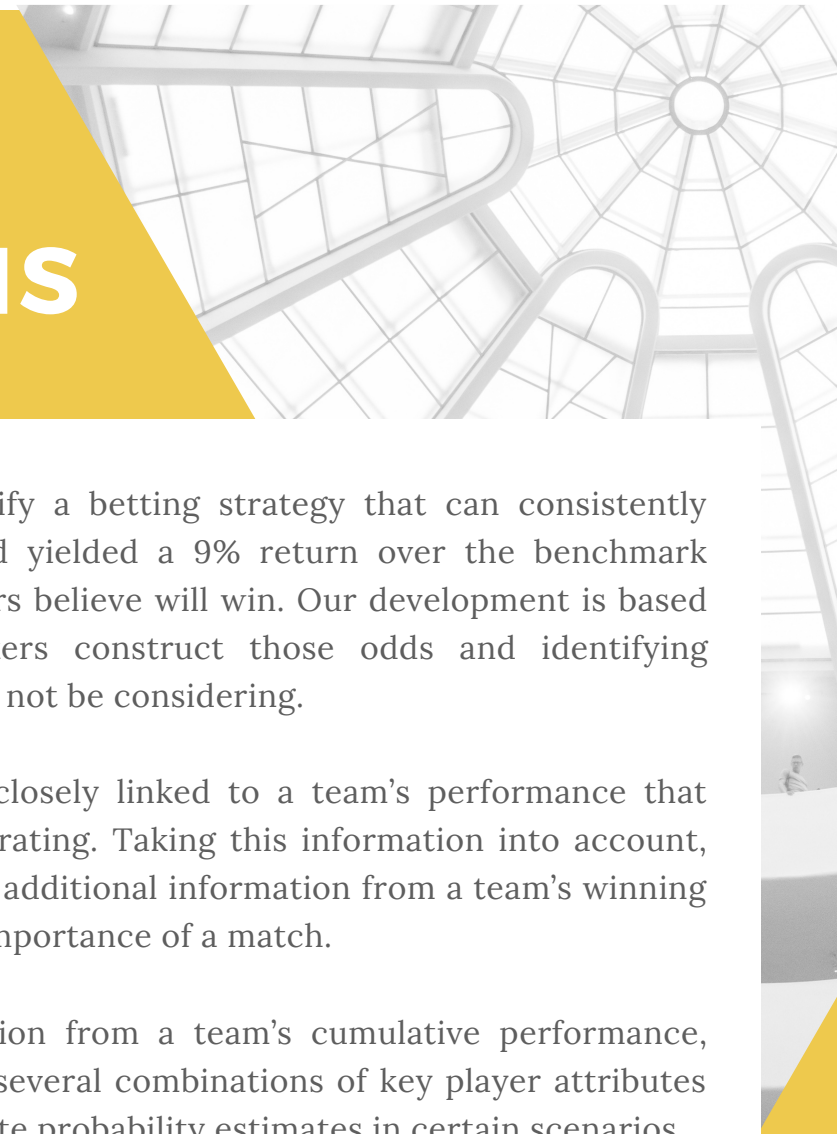
Results

Our betting strategy is to bet that the team will win if the spread is greater than 0.06 and bet that the teams will draw if the spread is between -0.08 and 0.06. The thresholds were determined by binning the differences in spreads into 10 different bins using k-means; then starting from bin 10, which corresponds to the largest positive differences in the spread, we added one bin at a time until it was no longer profitable to do so. The thresholds were determined using the calibration set (2012/2013) and then we evaluated the strategy on the other test sets (2013/2014, 2014/2015, 2015/2016).

The evaluation is done by betting one dollar using the above rules against betting one dollar on what the bookmaker believes is most likely to happen (the strategy with the lowest payoff). This strategy was able to return $132, $185, and $211.81 more than the bookmaker's strategy over 1877, 1956, and 2011 bets; this is equivalent to a 9% return over the bookmaker's strategy.



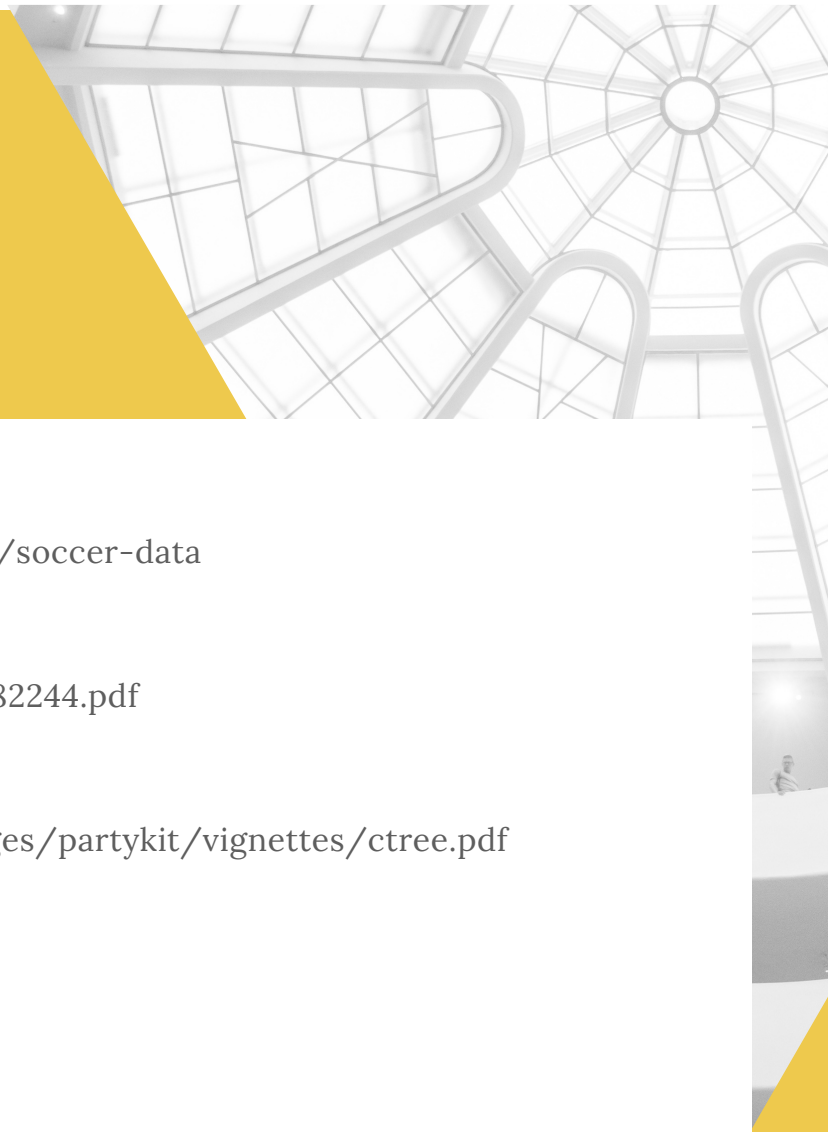
CONCLUSIONS



Our testing shows that we can identify a betting strategy that can consistently outperform the bookmakers' odds and yielded a 9% return over the benchmark strategy of betting what the bookmakers believe will win. Our development is based on understanding how the bookmakers construct those odds and identifying additional information bookmakers may not be considering.

We saw that bookmakers beliefs are closely linked to a team's performance that season as measured through Massey's rating. Taking this information into account, we were able to test the significance of additional information from a team's winning or losing streak, home advantage, and importance of a match.

Our betting strategy utilizes information from a team's cumulative performance, home advantage and the difference in several combinations of key player attributes between teams to arrive at more accurate probability estimates in certain scenarios.



APPENDIX

Sports Statistics
https://sports-statistics.com/database/soccer-data

Massey's rating
https://core.ac.uk/download/pdf/84482244.pdf

Conditional Inference tree
https://cran.r-project.org/web/packages/partykit/vignettes/ctree.pdf