The Problem

In this project we look at predicting the outcome of a soccer game from the five major European leagues in England, France, Germany, Italy and Spain. Our client is a betting agency who is looking into improving its predictive power to better produce odds for the games. Our benchmark is the average correctness of the major betting agencies accepting bets on European soccer. As odds are inversely proportional to the probability of a certain outcome (Home win, Draw, or Away win) lower odds correspond to higher probability of a given outcome. By selecting the lowest odds as the betting agency prediction, we will measure their prediction accuracy against the actual outcome of a game. Our goal here is to produce a prediction model that does better than the average consensus of the agencies.

The leagues selected and the computed betting agencies correctness in each are shown in the graph in the next section. The overall accuracy is 53.25%, and in the 5 major leagues is 53.02%.

The Data

The data comes from the Soccer Database at https://www.kaggle.com/hugomathien/soccer on the Kaggle website. This database is actually a collection of data from different sources:

1. http://football-data.mx-api.enetscores.com/: scores, lineup, team formation and events
2. http://www.football-data.co.uk/: betting odds
3. http://sofifa.com/: players and teams attributes from EA Sports FIFA games

The database consists of 4 different tables; *Player*, *Player_Attributes*, *Team*, and *Match*. The Player table includes name, birthdate, height and weight of the player associated with a particular player_id. The Team table contains the full and short (3 letters) names of a team. The Player_Attributes comes from the EA Sports Fifa game and includes several characteristics: overall ranking, preferred foot, goal keeping abilities, shooting abilities, offensive and defensive rankings, and others for a total of about 40. Finally, the Match team contains all details of games played starting with the 2008/09 season and up to the 2015/2016 season. It contains the teams involved, the goals scored, the team formations, the teams' starting 11, major games' events, and betting odds from several betting agencies for a total of 115 variables. Most of the information relating to the events happened during a game, such as free-kicks awarded, shots on goal, or scoring players' names, is contained within a few columns in HTML format.

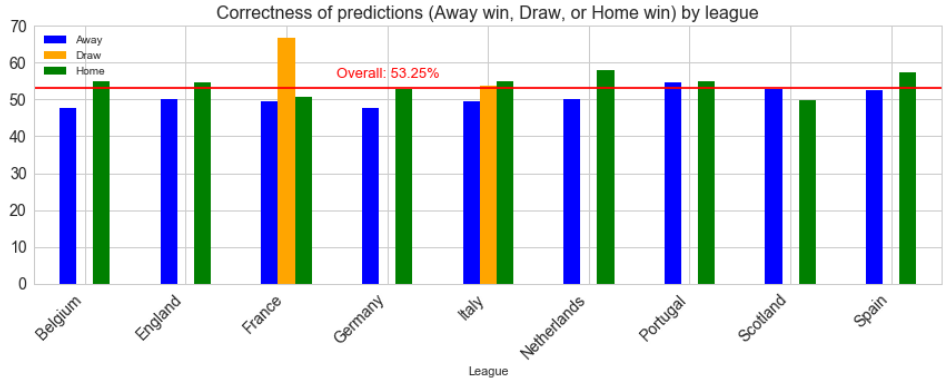
Most of the data is clean, there are only a few missing values, corresponding to a few games in the Italian Serie A during the 2010/11 season. In addition the betting odds were not collected for two of the leagues: Poland Ekstraklasa and Switzerland Super League. These leagues will be excluded from the analysis.

Data Exploration

Our preliminary analysis of the data will look into five different areas; Upsets, Home Advantage, Stage Effect, Team Performance, and Goals Scored.

Upsets

The goal here is to better understand how the betting agencies are doing in predicting outcomes for different leagues and different outcomes. The following graph shows how the betting agencies consensus is faring in each league, for which betting odds are available.

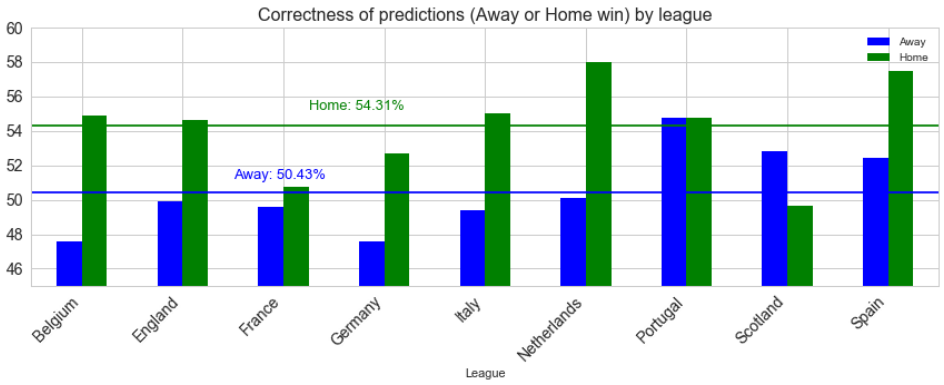


A few interesting things stand out from the above bar chart. First of all booking agencies almost never predict a draw for games. This basically means that they are ok with being wrong roughly one fourth of the times, as 25.3% of all games actually end in a draw. Of all games only 31 were *expected* to more likely be a draw, 28 in the Italian league and 3 in the French league. This is something to consider if one wants to improve on the booking agencies overall correctness rate, computed at 53.2%. It is important here to specify that betting agencies have several ways to *hedge* against such predictions which are reflected in the odds offered on the games. In other words, while they might favor a result over another their overall odds strategy will offset such problems, as the odds' difference will be small.

Another observation is that generally agencies do a better job of predicting a home win rather than an away win. The only exceptions are the Scottish league, where the opposite is true, and the Portuguese league where the difference is minimal. Comparison with predictions of draw should not be done given the really low number of such cases. Predicting a home win seems, intuitively, easier than an away win, especially when the difference between two teams is high. We will return to this point a little bit later to confirm our intuition.

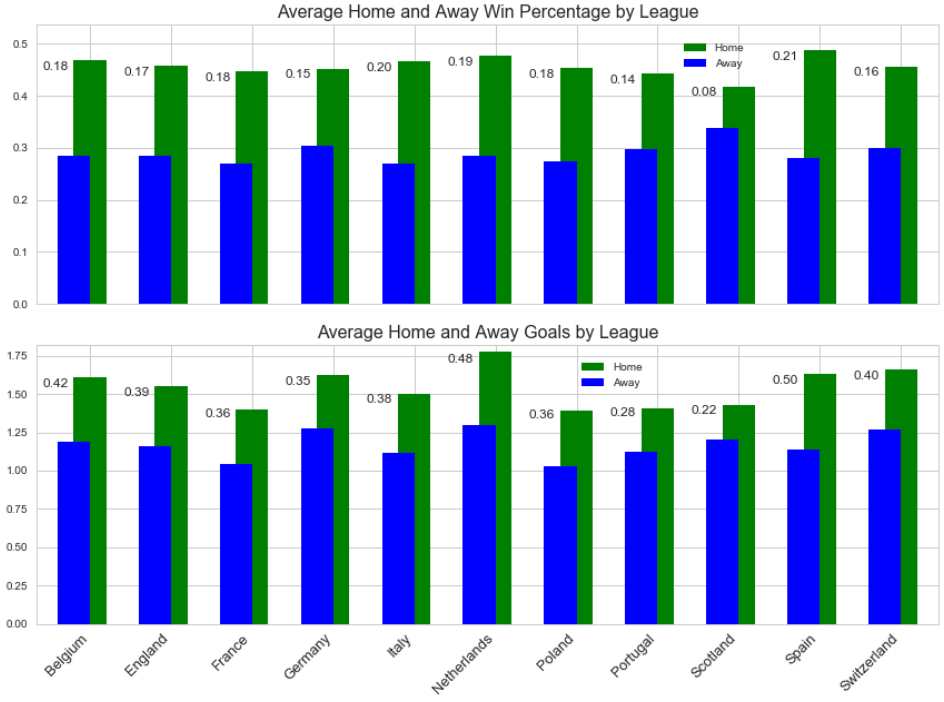
To learn more from this graph we zoom in on the results and display horizontal lines at the values of the average predictions across all leagues (by predicted outcome of a game). In this graph we drop the draw category, due to the low amount of games and reliability of related statistics.

As we can see the agencies seem to do better on certain leagues. The Spanish and the Portuguese leagues are both predicted correctly above average for both Home and Away wins. Other leagues have differing outcomes, given Home or Away. It is easier than average to predict a home win, and worse than average to predict an away win in; Belgium, England, Italy and the Netherlands, while the opposite is true for Scotland. The hardest leagues to predict are the French and German leagues where betting agencies do worse than average in both home and away wins.



Home Advantage

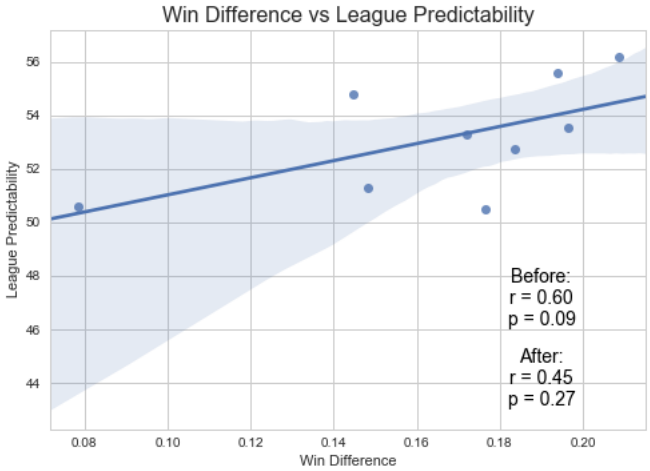
We seek here confirmation of the intuition that home teams have what is usually referred to as home advantage, that is: home teams win more often than away teams. We look at the value of the difference between home and away wins' percentages, and goals scored, for teams in all leagues.



There seems to be a clear home advantage as home teams score almost half a goal extra compared to away teams. In addition home teams win about 46% of all games while away teams only win about 29% of all games. The remaining 25%, as stated earlier, ends in a draw.

We also looked at whether there is any relation between this result and how predictable a league is. To do this we first computed a weighted average for correctness of predictions, including all predictions: Home, Draw, and Away, and then looked at whether there is any correlation when using the above *home advantage* values to explain league predictability. Dropping, once again, the Polish and Swiss leagues as we don't have betting odds for them. The Pearson test seemed to show that there is indeed

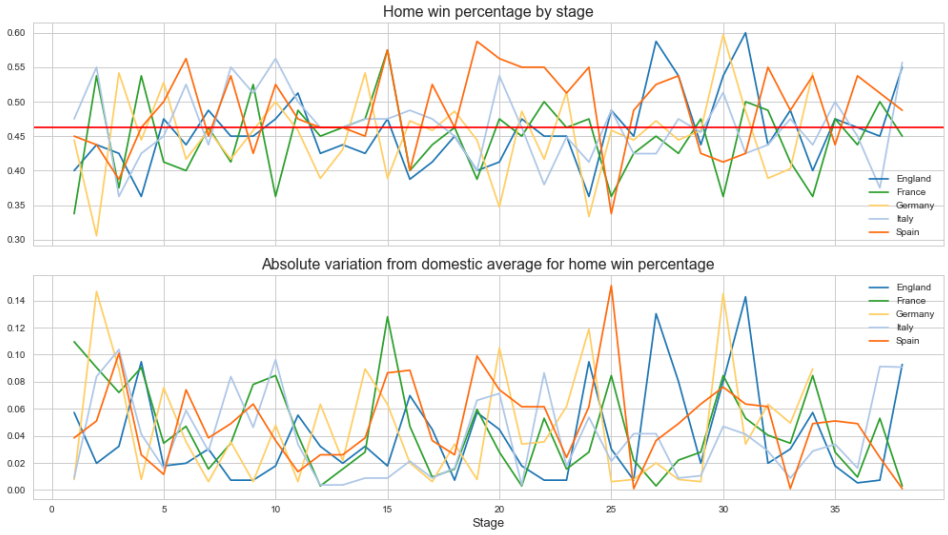
a positive linear correlation between these variables, as the Pearson's correlation coefficient is $r = .60$ and the $p - value = 0.09$. Looking at the graph below we see that the value for Scotland is somewhat of an outlier, indeed once Scotland is removed the new values are: $r = .45$ and $p - value = .27$.



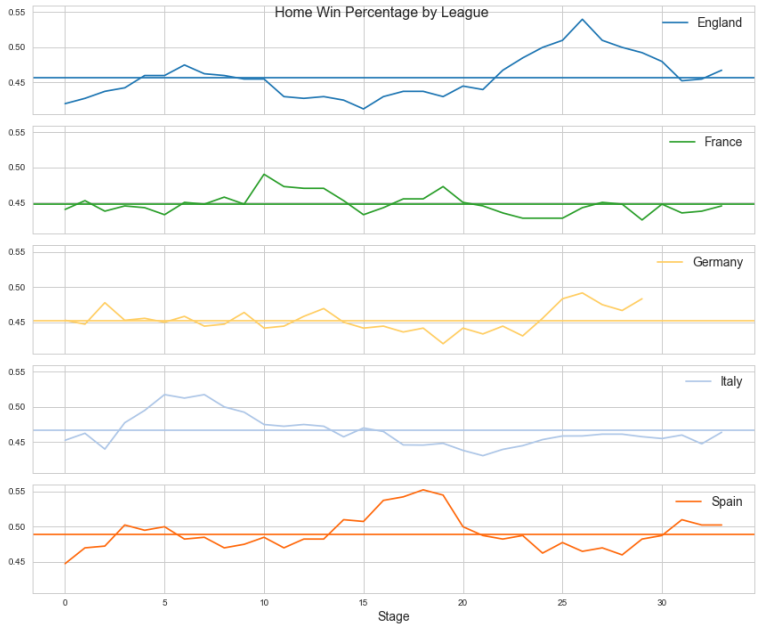
The p-value, in both tests, is too high to be able to state that the positive correlation observed is not due to randomness in the data. Nevertheless, the p-values might be high due to the few data points available. Having previously noticed that home wins are predicted correctly more than away wins we can still say that it seems to be somewhat easier to predict games when the home advantage is higher.

Stage Effect

Is there any difference in home wins percentage by stage number? That is: is it easier to win home games early, or late, in the season, or is there any difference at all? To answer such questions we limit ourselves to the 5 major European leagues: England Premiere League, France Ligue 1, Germany Bundesliga, Italy Serie A and Spain Liga BBVA. Although we group our data by stage number we still have about 80 games per stage number per league, so that our results will be somewhat *robust*.



The top graph displaying the home win percentage by stage does not reveal anything unusual, results seem to hover around the average for all stages, which for the top 5 leagues is 46.27%. The second graph looks at the (absolute) difference with the average in each league, and while showing some peaks again does not seem to show that there is any meaningful difference.
The variation among single stages is too high for us to pick up any meaningful difference, so we try to group stages and display the average of N consecutive stages, with N a parameter we set equal to 5, corresponding for most leagues to about a month of games. We also separate the plots and display the individual country average home win percentage. Results are shown in the following graph.

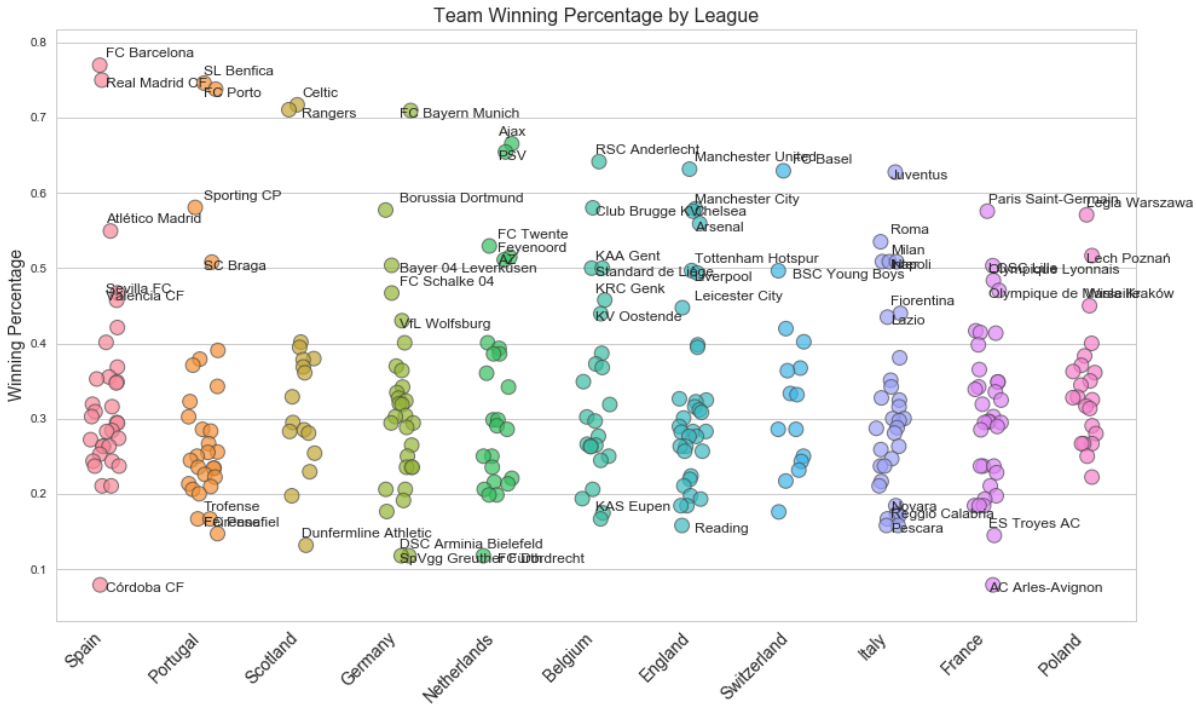


All graphs use the same axes limits so that it is easier to see any anomalies. It looks like the graph for France Ligue 1 doesn't display any pattern, while those for the other leagues do. In England and Germany home field advantage seems to be more important towards the end of the season, while in Italy the opposite is true. Finally, in Spain the best time to play at home is the middle of the season.

Another interesting observation is that generally home field advantage is less important through the first few games, as all graphs above are at or below their averages. We recall that the first point in each graph, at the 0 mark, correspond to the average of the first $N = 5$ games, while the last corresponds to the average of the last 5 games.

Team Performance

A look at individual teams performance in terms of winning percentage. The top 50 teams, and the bottom 15, in terms of winning percentage are labeled on the graph. Not surprisingly, the two big Spanish juggernauts Barcelona and Real Madrid come out on top of this special ranking. Several less known teams appear in this ranking as well, given that all the games in question are played within each country league. Thus we see close to the top; Celtic and Rangers (Scotland), RSC Anderlecht and Club Brugge KV (Belgium), FC Basel (Switzerland), and Legia Warszawa and Lech Poznan (Poland). All these teams can be considered juggernauts in their domestic leagues.



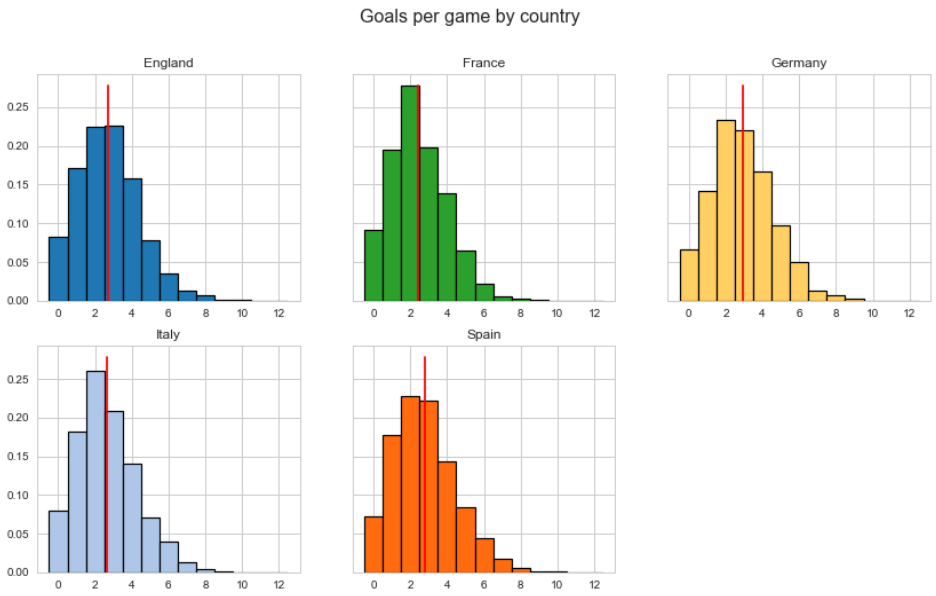
As a further consideration we now have a better idea about why the Spanish, Dutch, and Portuguese leagues turned out to be the three most predictable leagues, with respectively 56.2%, 55.8%, and 54.8% of results predicted correctly. Each of these leagues has two teams in the top 9. An anomaly seems to be Scotland again. Although Celtic and Rangers are right after the two Spanish juggernauts Barcelona and Real Madrid and the the two top Portuguese teams Benfica and Porto, the league overall is the least predictable with the exception of the French league. This is probably a consequence of the fact that the next most winning team sits well below the two top teams.

One interesting thing that stands out from this graph is how the first two teams in Spain, Portugal and Scotland are well separated from the rest, a fact that only applies to another team: Bayern Munich from Germany. Such a dominance is not seen in other leagues. In particular we see that France has seen no dominating team (during those years) and this has probably contributed to making the French league a little less predictable with only 50.51% correctly predicted outcomes.

Goals

The game of soccer is all about scoring one more goal than the opponent, so we conclude with a look at how many goals are scored per game. The following graph shows histogram of the distribution of goals scored per game for each of the 5 major leagues.

All the distribution are right skewed, but in general seem to look roughly bell-shaped. No differences are apparent from these graphs, besides that in some leagues, like Germany and Spain, slightly more high scoring games are played.



The Approach

Features

In order to improve on predicting outcomes for the games we will build onto the work done by the betting agencies and therefore use the *consensus* odds for Home Win, Draw, and Away Win for the first three features. These odds were collected exactly 48 hours before the start of the games. Not all the betting agencies offered odds on all the games, so the average is calculated on *up to* 10 betting agencies.

- **avgH**: average odds for Home win for up to 10 betting agencies
- **avgD**: average odds for Draw for up to 10 betting agencies
- **avgA**: average odds for Away win for up to 10 betting agencies

A second set of features tries to account for the overall strength of a team, in terms of average points during the current season. Points are awarded as: 3 for a win, 1 for a tie, 0 for a loss.

- **points_home**: average points by the home team up to the previous game in the current season
- **points_away**: average points by the away team up to the previous game in the current season

A third set of features accounts for whether a team is hot or cold, by looking at the average points over the last 5 games. Notice that if a team has played less than 5 games the average will be out of only the games played that far.

- **streak_home**: average points of the home team up to the previous game in the current season for up to the last 5 games
- **streak_away**: average points of the away team up to the previous game in the current season for up to the last 5 games

A fourth set of features is given by the actual formation type used by the teams. For example a team could play with a 4-3-3, that is: with 4 defenders, 3 midfielders, and 3 forwards. As the number of possible starting formations in the dataset are more than 20, this would create about 40 binary features. We, instead, summarize these findings into two features: one determining whether a team is offensive or not, and one for whether it is defensive or not. As an example: a 5-3-2 formation type is determined to be defensive as it includes 5 defenders, while a 3-4-3 formation is deemed to be offensive as it uses 3 forwards, and only 3 defenders.

- **defensive_home**: binary, whether the home team starting 11 is defensive (1) or not (0)
- **defensive_away**: binary, whether the away team starting 11 is defensive (1) or not (0)
- **offensive_home**: binary, whether the home team starting 11 is offensive (1) or not (0)
- **offensive_away**: binary, whether the away team starting 11 is offensive (1) or not (0)

More features are added by looking at the value of the attributes of the field players, given in the Player Stats table. The field players are all the players, except the goalkeeper. Their attributes' values are averaged over the home and away teams. In addition the goal keepers attributes are added as well, for both teams. Missing values in the team data are simply ignored, that is: the averages are computed only on the available players. Goal keepers missing data is replaced with the average value of that attribute among all existing data. In total this produces 30 team and 5 goal keeper features for each team, adding 70 features total. All values are integers between 0 and 100. A few examples are shown below:

- **overall_rating_home**: average overall rating of players of the home team
- **heading_accuracy_home**: average accuracy with headers of players of the home team
- **stamina_home**: average player's ability to run for the entire game for players of the home team
- **vision_home**: average player's ability to foresee the course of an action for players of the home team
- **gk_diving_home**: home team goal keeper's ability to dive to stop a ball
- **gk_reflexes_home**: home team goal keeper's reflexes evaluation
- ...

As games are often decided by the presence of a single good player, another feature introduced is the overall rating of the best player in the team.

- **best_player_home**: highest overall rating of players of the home team
- **best_player_away**: highest overall rating of players of the away team

While exploring the actual predictions of the model, we noted that the model never predicted a Draw, but only Home wins and Away wins. This is not too different from what the betting agencies do. Only in 34 instances out of more than 14,500 games under consideration the betting agencies favored a Draw. This is somewhat surprising as more than a fourth of all games end with a Draw. In order to *force* the model to predict more draws we used the class_weight option of the LogisticRegression classifier of sklearn. While this forced the model to make some Draw predictions, it did not improve on the accuracy of the model. A different approach was then taken. Adding two more features to characterize games that are more likely to end in a draw.

- **diff**: the difference between highest and lowest value of odds consensus
- **tie**: a score from 0 to 5 based on whether a match satisfies, or not, 5 criteria more likely to identify tie games

The *diff* feature should help the classifier identify games likely to end in a tie. The assumption is that those are the games where the odds difference among the various predicted results is small. The *tie* feature is based on a set of criteria most likely to identify games whose outcome was a tie. The number of criteria chosen is 5, and these are automatically identified from the given features. Games were separated based on whether the actual outcome was a tie, or not a tie. The averages for all features are then computed on both sets and compared to determine which features produce the greatest normalized difference. Games are then scored by counting for how many of these features their values are beyond a fixed number of standard deviations from the mean. In the final run of the chosen classifier these were:

- *diff, avgD, avgA, best_player_home, overall_rating_home*

The classifiers considered were scored using the median accuracy score over 10 runs of the algorithm using different training and test sets. An analysis of the correlation matrix showed that the features from the player stats table were highly correlated among each others. It was decided to use only two of them: *best_player* and *overall_rating*.

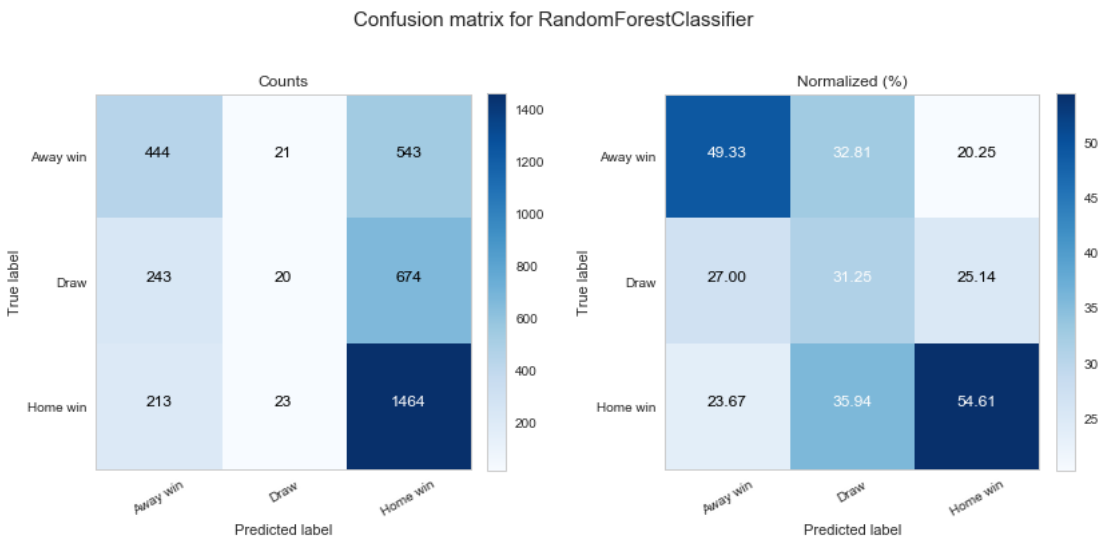
Classifiers

Various classifiers were tested: LogisticRegressionCV, LogisticRegression with $C = 0.01$, RandomForest, and ADABoost using RandomForest as a base estimator. They all produced very similar results in terms of accuracy and log loss scores, except for ADABoost, whose log loss score was significantly higher than the rest. From the table we can see that Random forest did slightly better, both in accuracy and log loss.

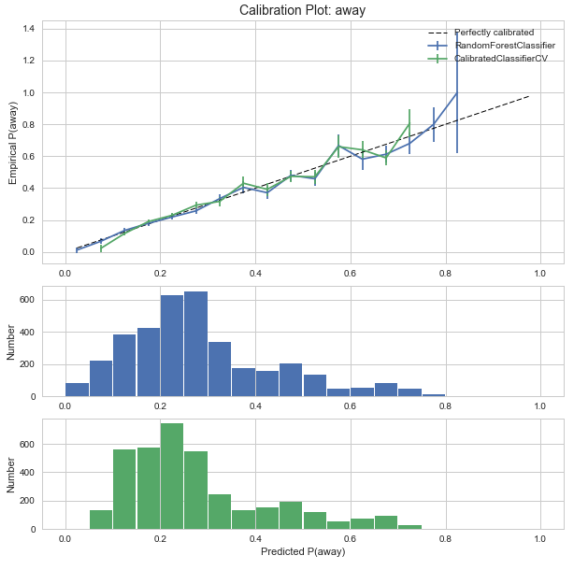
	LogisticRegressionCV	LogisticRegression	RandomForestClassifier	AdaBoostClassifier
Accuracy Test score	52.592593	52.386831	52.784636	52.524005
Accuracy Train score	53.169594	53.229052	55.374131	58.100073
Log Loss Test score	0.984087	0.984294	0.983665	1.046829
Log Loss Train score	0.975790	0.974779	0.910624	1.021158

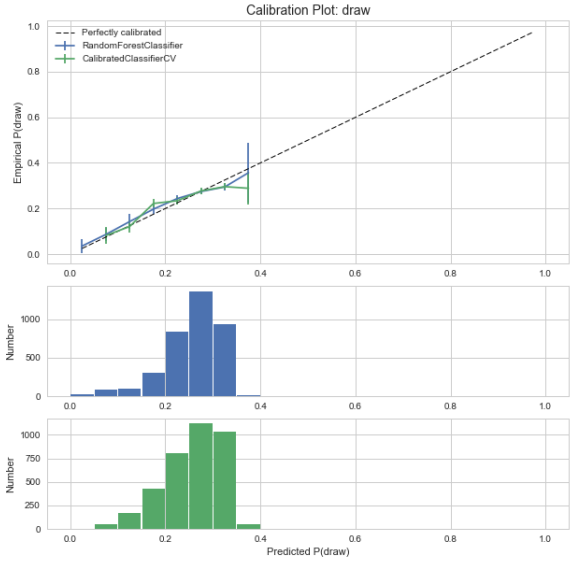
Using a run of RandomForest, with random_state=10, that obtained an accuracy score of 52.89% and a log loss score of .9832, produced the confusion matrices below. On the left, we display the counts of the predictions, on the right we have the normalized version.

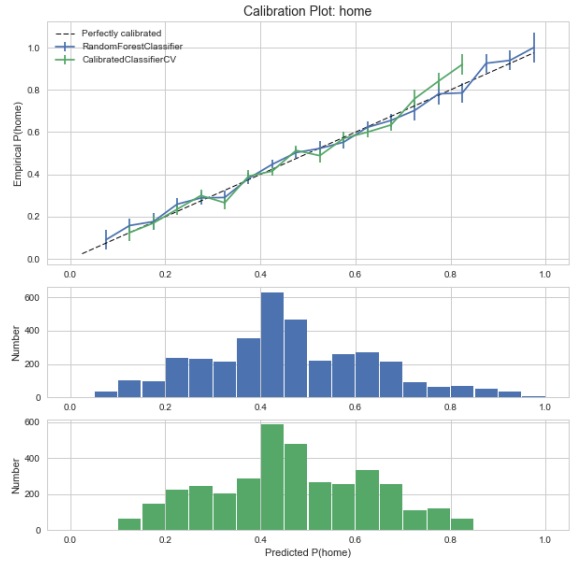
As we can see, most predictions (73.47%) are for Home Win but only a few more than half (54.61%) are correct. This is similar to what bookies do. Most of their predictions are for home win, and almost none are for a Draw. This might seem somewhat arbitrary, given that about a fourth of all games ends in a draw. Here we should consider though, that what is important for bookies is that the probabilities



assigned to each outcome are correct. To test the generated probabilities we run CalibratedClassifierCV, prefit on the RandomForest classifier selected earlier. We then compare the uncalibrated and the calibrated classifiers by means of a calibration plot. This is a plot that groups together samples that were assigned a similar probability by the classifier, say p, and tests whether they indeed happen about p percent of the times. The three plots for Away win, Draw, and Home win are below:



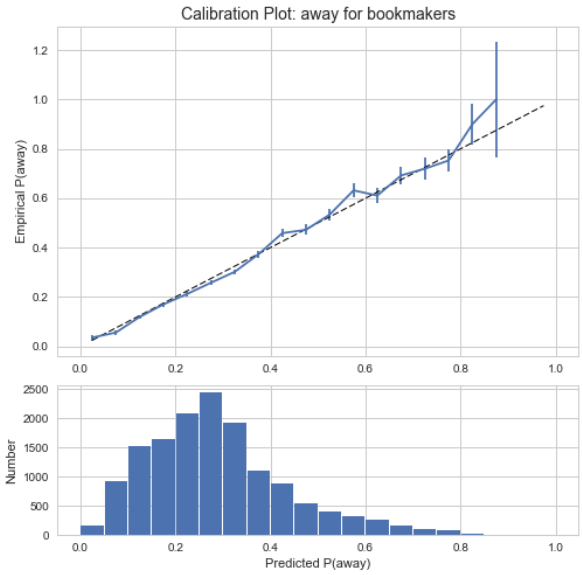


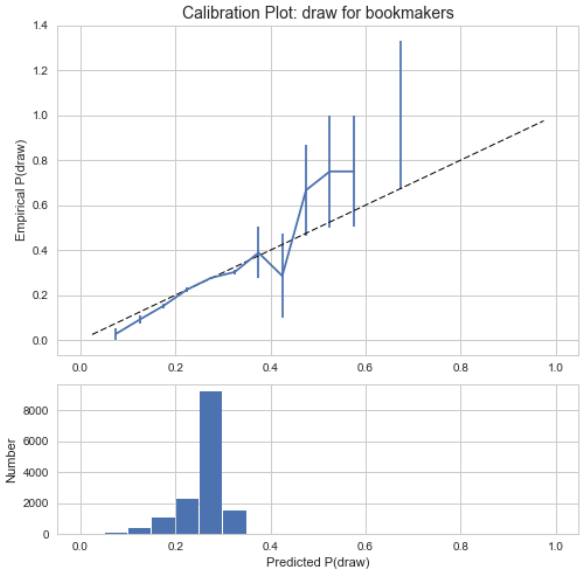


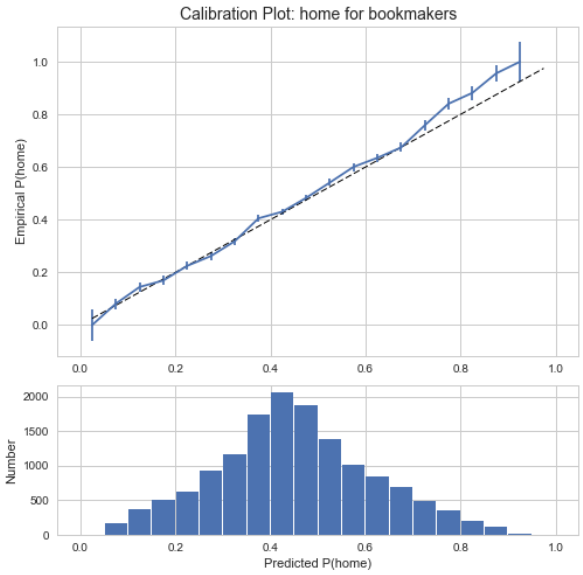
The plots show that the RandomForest classifier is well calibrated, for all outcomes. In particular it does very well with assigning probabilities for Home win, which is clearly the most important as most predictions are made for a Home win. We also notice that the CalibratedClassifierCV method tends to avoid assigning extreme values of probability, this is shown particularly in the Home win plot. Before calibration we had some probabilities between 5% and 10% on the low end, and some between 85% and 100% on the high end, instead after calibration these bins are empty. This is actually a bad thing, as predicting odds correctly for games with a heavy favorite is extremely important for bookmakers. Indeed, with high probabilities come low odds, whereas if the probabilities are thought to be

lower, odds will be higher, raising the risk of a high payout, especially since most bets are made for the team thought to be the favorite, in particular if the odds are perceived to be higher than they should.

The LogisticRegression classifier shows a similar behaviour, with a well calibrated plot, even before applying the CalibratedClassifierCV method, which actually shows some of the same issues described above. We show the calibration plots of the bookmakers below. These plots are drawn using all the data available and therefore show much smaller error bars in general.



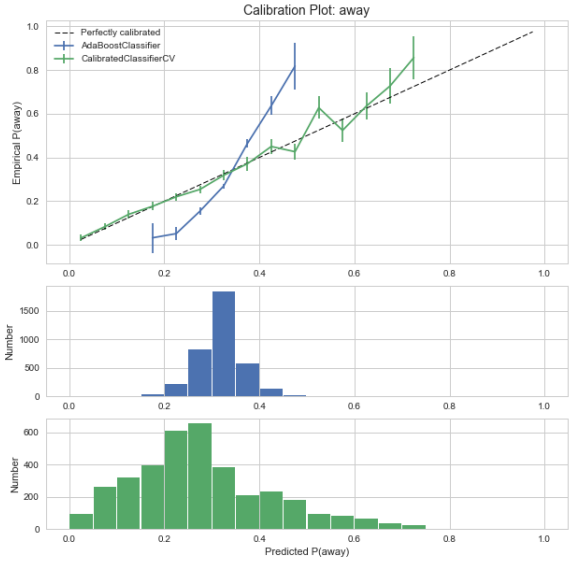


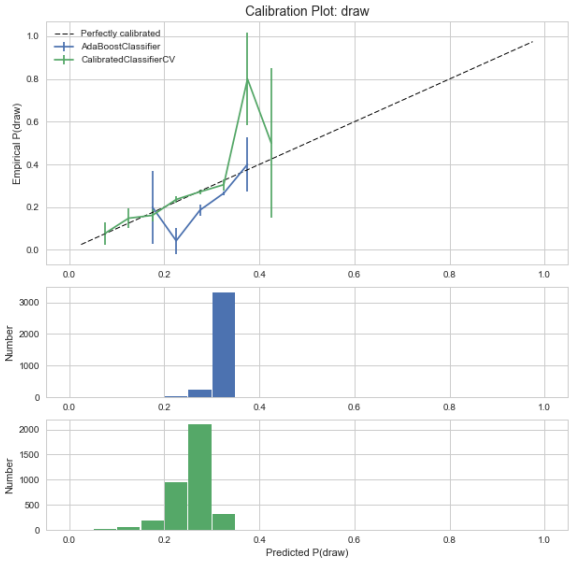


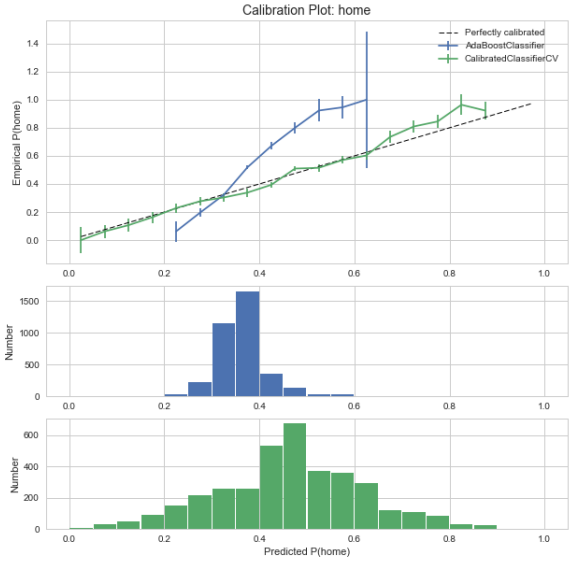
We can see that bookmakers tend to under-predict extreme outcomes as well, this is more evident once again in the Home win plot. We should point out that the bookmakers' probabilities were *estimated* from their odds. It is safe to assume that when a team is thought to have a very low probability of winning, odds are not set as the inverse of the probability, but are somehow maxed. If this were not the case, a team with a 1% chance of winning, would be offered at 100. Instead, the maximum odds from the data were found to be 36, 29, and 67 for Home win, Draw, and Away win respectively.

One more thing that these plots show is that bookmakers think that the probability of a tie for about two thirds of all games is between 25% and 30%, and that the games in this range are extremely well calibrated. One of the hardest things is to predict a Draw, and the Draw calibration plot of the bookmakers suggests why. If most games have a chance between 25% and 30%, and according to the calibration plot, these values represent well their actual probability, they will not be the most likely outcome, as at least one of the other outcome will be higher.

Next, we take a look at how the ADABoost RandomForest classifier performs, considering that its log loss score is higher than RandomForest and LogisticRegression, thus suggesting that it might not be well calibrated.


Calibration Plot: away
Perfectly calibrated
AdaBoostClassifier
CalibratedClassifierCV
Empirical P(away)
Number
Number
Predicted P(away)


Calibration Plot: draw
Perfectly calibrated
AdaBoostClassifier
CalibratedClassifierCV
Empirical P(draw)
Number
Number
Predicted P(draw)

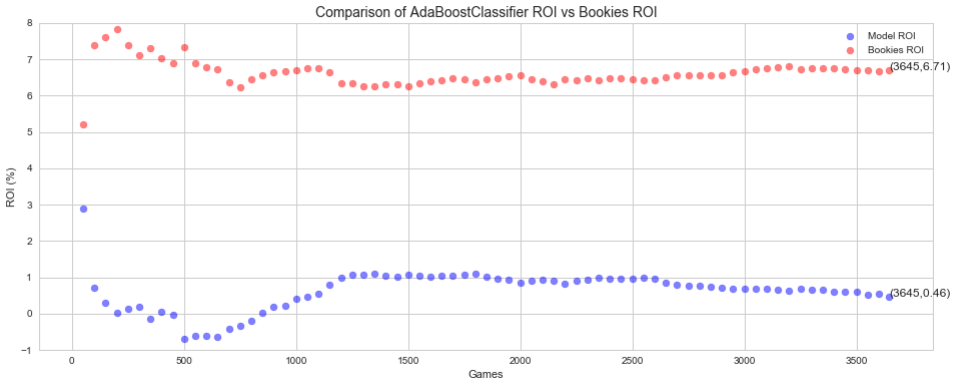


We see that this classifier does not do very well at all. It tends to under estimate high probabilities and over estimate low probabilities. This is most likely due to the fact that it does not assign extreme values at all! For example it assigns to more than 90% of the games a probability of a Draw between 30% and 35%, and no game has a probability of winning higher than 60% for the home team! The calibrated version does better, as can be inferred from the log loss score, which is reduced from 1.0471 to .9876, as well as from the calibration plots.

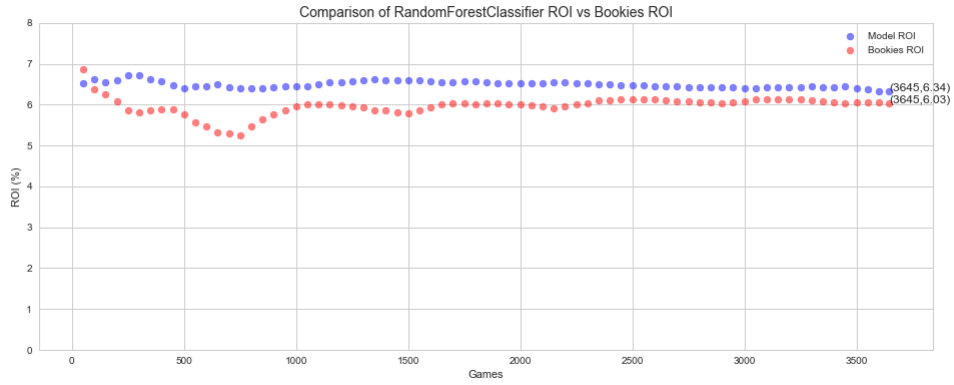
Testing the models' ROI

The calibration plots give us a good feeling about how the classifiers are doing but we can actually compute the ROI, Return On Investment, of such models for a bookmaker. In order to perform the comparison fairly we modify the odds suggested by our classifiers, to introduce the bookmaker's commission. We set this, on average, equal to the average commission computed on the bookmaker's odds, which is 6.11%, when computed on the test set. We also make the assumption that the odds, whether from the model or the bookmaker are balanced. That is: they are set, so that an equal payout, given the odds, is expected regardless of the outcome. As the odds are balanced differently for the bookmakers and our model, when we compare the odds we assume that the amount played on each outcome is given by the average of the *balanced* amount for the two set of odds. This is clearly a simplification, so that the actual ROI values might not apply to a real-life situation, where bets will be most likely somewhat unbalanced. On the other hand, while these values would not be accurate, for unbalanced odds, they can still be used for a comparison between models.

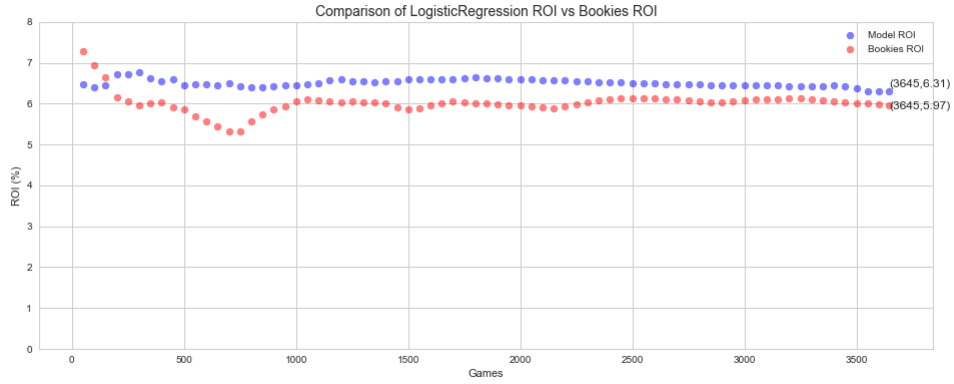
We show first how the ADABoost classifier performed, when compared to the bookmaker's odds. We can see that the bookmakers' odds are clearly better returning 6.71% versus ADABoost's 0.46%. The calibrated version returns 6.13% just barely edging the bookmakers' odds at 6.01%.



Next, we have the RandomForest classifier compared to the bookmakers. This model does better showing a 6.34% ROI versus the bookmakers' 6.03%. The calibrated version of this classifier performs worse as it brings ROI down to 6.18%, just slightly above the bookmakers' 6.12%.



Finally, we show the best performing classifier, which turns out to be LogisticRegression with an ROI of 6.31% versus the bookmakers' 5.97%. The calibrated version shifts both values slightly higher to 6.36% and 6.03% respectively.



Conclusion

In this project we were tasked with building a Machine Learning model that would improve on the accuracy of the bookmaker's odds, when it comes to predicting the final outcome of a soccer match in the 5 major European Leagues. Our most accurate classifier, RandomForest had a median accuracy of 52.78% versus the bookmaker's accuracy of 52.73% when computed only on the test set. Although RandomForest was the most accurate, LogisticRegression, a close second in terms of accuracy performed better when we looked at the probabilities assigned to the possible outcomes of a match. Indeed, in a situation of balanced odds, it will actually have an ROI 0.34% higher than the bookmakers' model.

Recommendations

1. When trying to decide on a classifier to use to predict soccer games outcomes we recommend that LogisticRegression or RandomForest be used, without further calibrating them, or trying to boost their performance.

2. Player statistics to be used as features should be computed on actual values, such as; number of crosses, penalty conversions, and similar for a team. The use of statistics from a third party, such as the FIFA game, did not improve prediction accuracy. One reason was the high correlation of many of the attributes, but perhaps also because such values are not really representative of players' quality.

3. In comparing classifiers we recommend that a separate score function be produced, similar to what shown in our ROI approach, as standard score function were not able to clearly identify a winner. In particular accuracy score does not work well in this situation where the actual assigned probabilities are so important.

4. Particular attention should be given to assigning odds for games with a heavy favorite, in particular for the team who is indeed favorite to win. Classifiers, like ADABoost, that tend to average too much will not work well especially in these cases, and should thus be avoided. Similarly, much attention should be given to the choice of the parameters in Randomforest, to make sure that the consensus sought by the classifier does not turn into a straight average. For example, the number of estimators should be kept low, and trees should not be allowed to have leafs with too few samples.