14TH ANNUAL
MIT SLOAN
SPORTS ANALYTICS CONFERENCE
MARCH 6 - 7, 2020 | BOSTON, MA
presented by ESPN

A Poisson Betting Model with a Kelly Criterion Element for European Soccer

Soccer
7WA5RO9BD

1. Introduction

Sports betting has experienced a rapid rise in popularity as accessibility and commercialization of daily fantasy and live betting has increased. As sports betting is legalized in different countries and states, we are presented with a new opportunity to create statistical models that we can utilize to predict outcomes of different sporting events that can then be used to find inefficiencies in sportsbooks. In this paper, we attempt to create a model for European Soccer and measure its performance against betting markets to understand if this model can be used to generate profits. The paper shows how we predicted the number of goals by each team in a game and utilized these predictions to create a Poisson distribution that determined the probability of each event (Win, Lose, Draw). To add another dimension to our model, we used an optimization technique known as Kelly Criterion to determine the optimal amount of money that should be bet on each match. This technique generates bet amounts while also creating a value (KCO Value) that acts as an accurate estimator of the risk associated with each match. By exploring the characteristics of this value, we were able to maximize the success of the model. After running the model for the 2018 and 2019 seasons across the five major European soccer leagues, we can safely say that our model was not only successful in predicting outcomes, but also in generating significant profit yields for a user. A profit percentage of 119.48% can be yielded using this model, which implies that a user would pretty much double their money using this model. We also evaluate how the model performs in different leagues to understand which league characteristics benefit the model. The highest profit percentage was seen in the 2019 Bundesliga season with a profit percentage of 226.68%. The success of the model can not only help users generate significant profits, but it can also expose certain inefficiencies in the market.

2. Generating λ Values for Our Poisson Distributions

The first step to use our Poisson Distribution is determining the λ to be used. In a Poisson Distribution, λ value represents the expected rate of occurrence within a given time interval. In this case, the λ value will represent the number of goals we expect a team to score in a game. To ensure we have the most accurate value of λ, we used four different models to create different λ values and tested our models with each value of λ. An important aspect of any betting model is accounting for the context surrounding the game. Specifically, if a team is playing on home or away and the strength of their opponent. Each model we ensure that our λ value we generate are adjusted for the strength of the opponent and if the team is home or away. In addition to this, we ensure that we prevented data leakage within our model. Data leakage is when models use data for predictions that at the time of the prediction would not be available. The λ values we calculated for each model were calculated as the season progressed, for example we would use data available through 12 weeks into the season to predict outcomes for the matches in game week 13. This also denotes that our λ values are being updated and improved upon as the season progresses. To summarize, for each we ensured each of our λ values factors in 3 major characteristics:

1. The λ value accounts for the opponent
2. The λ value accounts for whether the team is playing home or away
3. The λ value updates as the season progresses, becoming more accurate

2.1. λ Values based on Goals

The first model we created used goals scored and goals allowed to create attacking and defending strengths for each team. The formula we use in the model to develop the aforementioned strengths are:

$$Attacking\ Strength = \frac{Goals\ Per\ Game\ For\ The\ Team}{Goals\ Per\ Game\ Across\ The\ League} \tag{1}$$

$$Defending\ Strength = \frac{Goals\ Allowed\ Per\ Game\ For\ The\ Team}{Goals\ Allowed\ Per\ Game\ Across\ The\ League} \tag{2}$$

These strengths were calculated with home and away splits. This resulted in every team having 4 different strengths:

1. Home Attacking Strength
2. Home Defending Strength
3. Away Attacking Strength
4. Away Defending Strength

We use these strengths to calculate each team's λ value for a game. The formula below is used:

$$HomeGoals\lambda = HomeAttackingStrength \times AwayDefendingStrength \times AverageLeagueHomeGoals \tag{3}$$

$$AwayGoals\lambda = AwayAttackingStrength \times HomeDefendingStrength \times AverageLeagueAwayGoals \tag{4}$$

To provide more clarity, we will use the Crystal Palace (H) vs West Brom (A) game as an example during the last game week of the 2018 Premier League season. Below is a calculation of both team's attacking and defending strengths along with our final λ value for each team:

Crystal Palace Average Goals Scored and Goals Allowed at Home Respectively: 1.5 and 1.5
West Brom Average Goals Scored and Goals Allowed when Away Respectively: 0.56 and 1.39
Average League Goals Scored and Goals Allowed at Home Respectively: 1.52 and 1.15

$$CRYHomeAttStrength = \frac{1.5}{1.52} = 0.99 \quad WBAAwayAttStrength = \frac{0.56}{1.15} = 0.48$$

$$CRYHomeDefStrength = \frac{1.5}{1.15} = 1.31 \quad WBAAwayDefStrength = \frac{1.39}{1.52} = 0.91$$

$$CRY\lambda = 0.99 \times 0.91 \ \times 1.52$$

$$WBA\lambda = 0.48 \times 1.31 \times 1.15$$

$$\text{Crystal Palace } \lambda \text{ Value} \approx 1.37$$

$$\text{West Brom } \lambda \text{ Value} \approx 0.73$$

2.2. λ Values based on Expected Goals (xG)

The second mode we used to generate λ values is exactly like the first model, however rather than goals we used expected goals (xG). Expected goals is a popular metric in the world of soccer today that represent the number of goals that are expected to be scored based on the location and way a shot is taken. By substituting xG in place of goals, are formulas now look like this:

$$Attacking\ Strength = \frac{xG\ Per\ Game\ For\ The\ Team}{xG\ Per\ Game\ Across\ The\ League} \tag{5}$$

$$Defending\ Strength = \frac{xG\ Allowed\ Per\ Game\ For\ The\ Team}{xG\ Allowed\ Per\ Game\ Across\ The\ League} \tag{6}$$

The game used as an example earlier (CRY vs WBA) would have the following calculation for the λ values for this model:

Crystal Palace Average xG Scored and xG Allowed at Home Respectively: 1.73 and 1.2
West Brom Average xG Scored and xG Allowed when Away Respectively: 0.77 and 1.37
Average League Goals Scored and Allowed at Home Respectively: 1.39 and 1.08

$$CRYHomeAttStrength = \frac{1.73}{1.39} = 1.24 \quad WBAAwayAttStrength = \frac{0.77}{1.08} = 0.71$$

$$CRYHomeDefStrength = \frac{1.2}{1.39} = 1.11 \quad WBAAwayDefStrength = \frac{1.37}{1.08} = 0.98$$

$$CRY\lambda = 1.24 \times 0.98 \times 1.39$$

$$WBA\lambda = 0.71 \times 1.11 \times 1.08$$

$$\text{Crystal Palace } \lambda \text{ Value} \approx 1.70$$

$$\text{West Brom } \lambda \text{ Value} \approx 0.85$$

2.3. λ Values based on Linear Regression

The third model we used to predict λ was linear regression. We used the cumulative statistics of a team and the team's opponent to develop the model. The linear regression predicts the expected number of home and away goals using the metrics of both teams. These metrics include:

1. Goals
2. Shots
3. Shots on Target
4. Fouls
5. Corners
6. Yellow Cards
7. Red Cards

The λ value generated from this model for the Crystal Palace West Brom game we used as an example earlier would be:

$$\text{Crystal Palace } \lambda \text{ Value} \approx 1.54$$
$$\text{West Brom } \lambda \text{ Value} \approx 0.42$$

2.4. λ Values based on a Random Forest Regression

This model predicts λ values in the same manner as the third (Linear Regression) model. Rather than a linear regression, this model makes use of a Random Forest regression instead. The metrics used by the model to make predictions are the same used by the Linea Regression model.

The λ value generated from this model for the Crystal Palace West Brom game we used as an example earlier would be:

$$\text{Crystal Palace } \lambda \text{ Value} \approx 1.70$$
$$\text{West Brom } \lambda \text{ Value} \approx 0.38$$

3. Generating Probabilities for Every Event

3.1. Using λ Values to Generate Probabilities for a Home Win, Away Win, and Draw

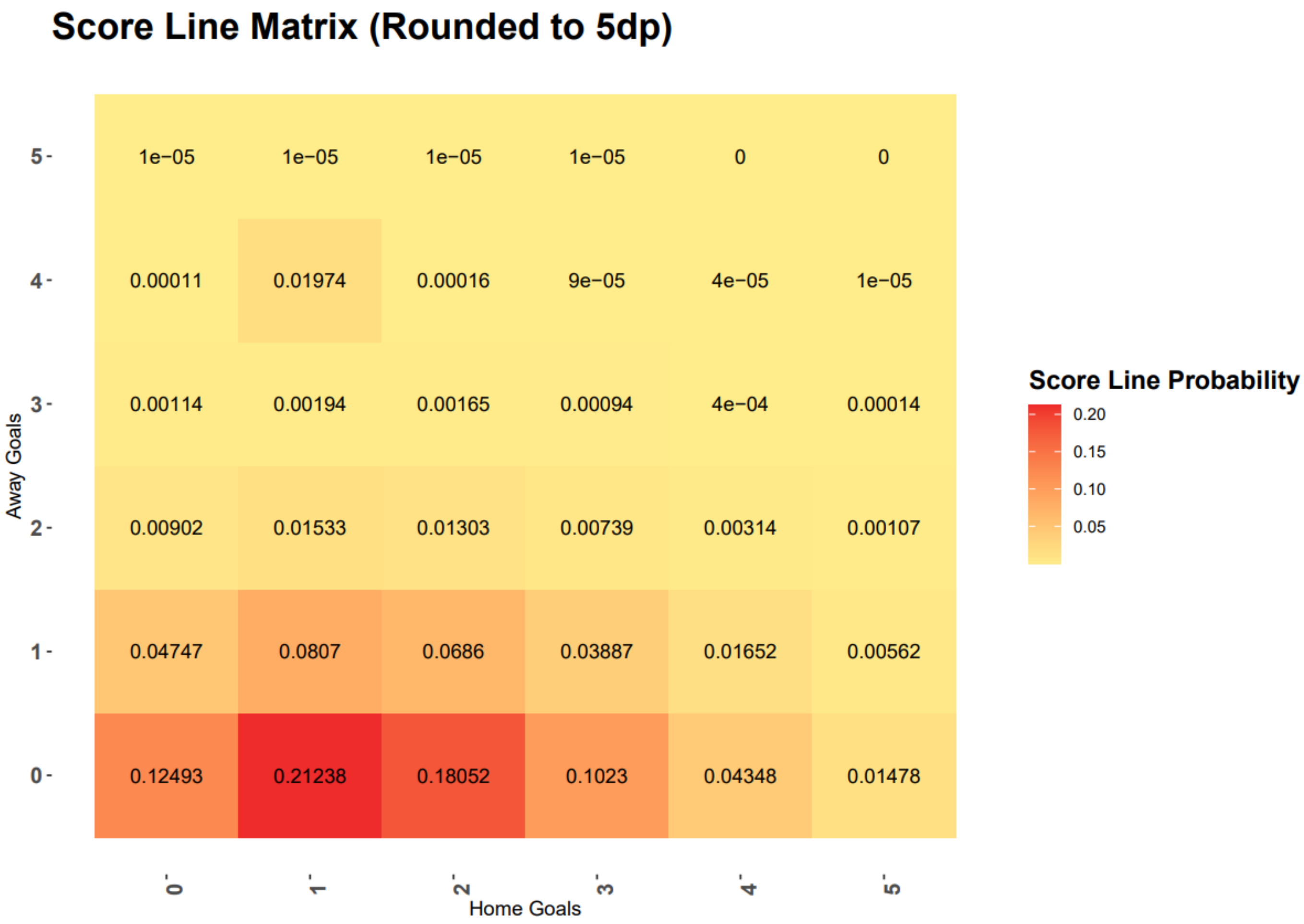
Now that we have our λ values, we can use the Poisson Distribution to calculate our probabilities. Below is the formula the Poisson distribution uses to predict x number of events:

$$P(x) = \frac{e^{-\lambda} \times \lambda^{x}}{x!} \qquad (7)$$

For example, if we would want to predict the probability of a team scoring 3 goals, x would take the value of 3. Essentially, to calculate each outcome's total probability we used the above formula to calculate every possible score line from 0 to 0 to 5 to 5. Using simple probability rules, the probability of each score line can be calculated in the following manner:

$$Prob\ of\ Scoreline = P(HomeGoals) \times P(AwayGoals) \tag{8}$$

So, the probability of a 3-2 score line would be the probability of the home team scoring three goals times the probability of the away team scoring two goals. We can show the probability of each score line as “Score Line Matrix” as seen below:



Score Line Matrix (Rounded to 5dp)

Away Goals \ Home Goals	0	1	2	3	4	5
5	1e−05	1e−05	1e−05	1e−05	0	0
4	0.00011	0.01974	0.00016	9e−05	4e−05	1e−05
3	0.00114	0.00194	0.00165	0.00094	4e−04	0.00014
2	0.00902	0.01533	0.01303	0.00739	0.00314	0.00107
1	0.04747	0.0807	0.0686	0.03887	0.01652	0.00562
0	0.12493	0.21238	0.18052	0.1023	0.04348	0.01478

The rows correspond to the number of goals scored by West Brom (Away) and the columns correspond to the number of goals scored by Crustal Palace (Home). The number present within each cell is the probability of that score line happening. As discussed earlier the likelihood of each score line is calculated by multiplying the probabilities. Based on the matrix, we can see that the model believes the most likely score line is 1 – 0 which has a 0.21238 (0.31056 × 0.68386) chance of occurring.

The total probability for every event is the sum of all the score lines that correlate with that outcome. Below are the formulas we used for this:

$$HomeWin = \sum Probabilities\ of\ Scorelines\ with\ Home\ Team\ Winning \tag{9}$$

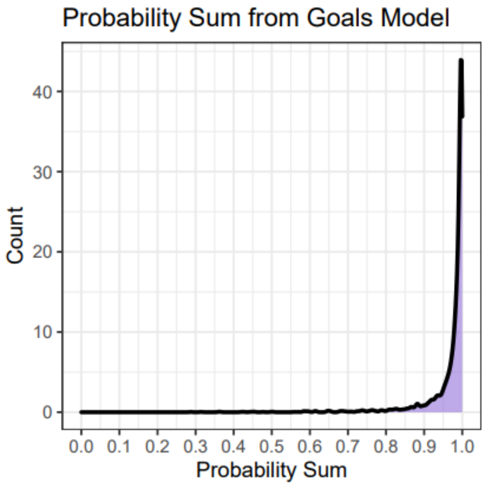
$$AwayWin = \sum Probabilities\ of\ Scorelines\ with\ Away\ Team\ Winning \tag{10}$$

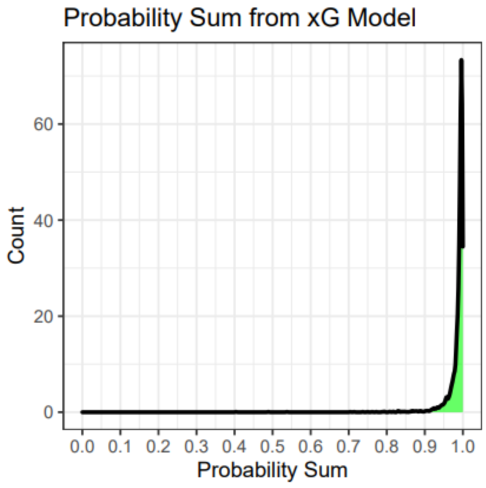
$$Draw = \sum Probabilities\ of\ Scorelines\ ending\ with\ a\ Draw \tag{11}$$

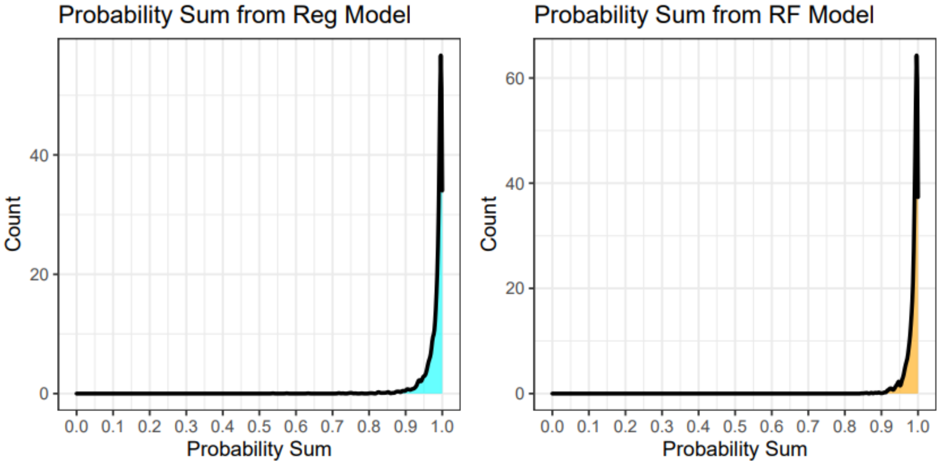
For each of our models, here were the probabilities generated for the Crystal Palace West Brom game:

Model	*Home Win Probability (Crystal Palace Win)*	*Away Win Probability (West Brom Win)*	*Draw Probability*
Goals	0.52	0.198	0.297
xG	0.567	0.186	0.239
Linear Regression	0.649	0.098	0.248
Random Forest	0.696	0.077	0.22

We considered the score lines from 0-0 to 5-5 to cover enough outcomes to generate an accurate probability for each event in the game. Below is a distribution of the total probability for every game for each model. Ideally, we want the probability to be equal to 1 as that would entail that all outcomes are covered.







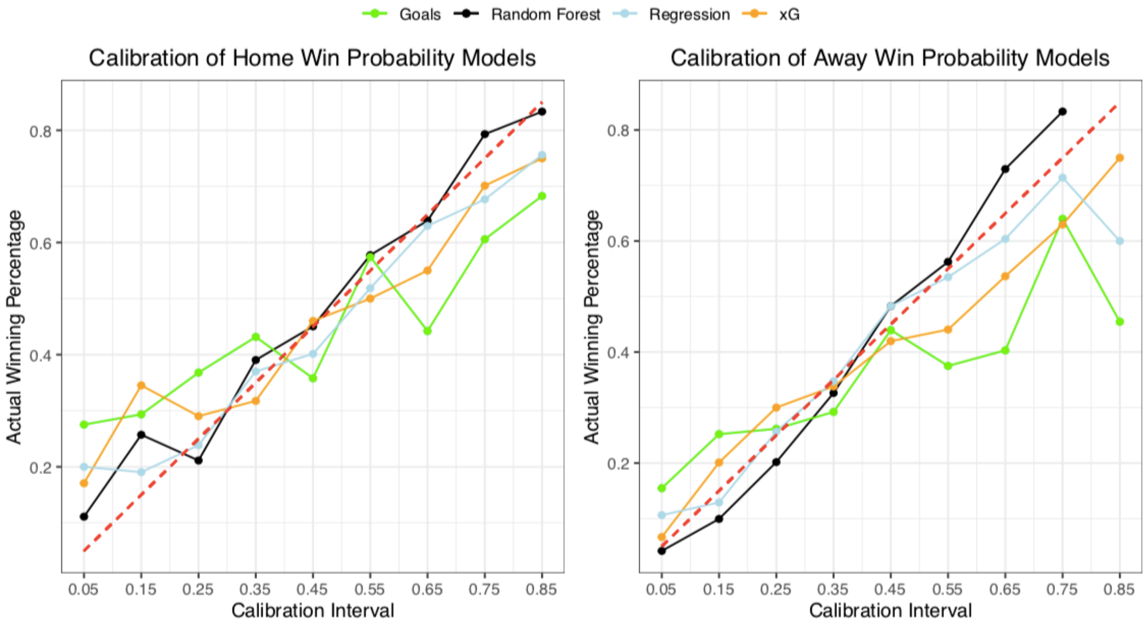
If we look at the distributions of the total probability for each model, we can see an overwhelming majority have a sum over 0.9 with most being equal to at least 0.99. These are also from games beginning in match week 5, hence as the season progresses, we anticipate the distribution to move even closer to 1. Hence, the Poisson Distribution we used with each λ value we generated accurately accounts for each outcome in the game.

4. Model Calibration (Brier Score)

We calibrated the model using a Brier Score to determine which model generated the most accurate λ value. The Brier Score is a score function that helps determine the accuracy of any probabilistic model. The Brier score for predictions is given by the formula below:

$$BS = \frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2 \tag{12}$$

Essentially, we are measuring if the outcomes we predict with a certain probability are occurring in that proportion. The best performing model would have probability values corresponding to the respective interval. For example, among all the events for the rest of the season that we predicted to occur with a probability of 0.6 and 0.7, we would want to be correct around 65% (0.65) of the time. Theoretically, calibration results after each game week would improve as the models would have more data. We wanted to identify which model and at which game week the marginal improvement is minimal to be able to understand the ideal time to begin to use the best performing model. Based on our calibration results, we determined that the best model to estimate λ was the **Random Forest Model after game week 13.** This is the model we use when evaluating our model's success. The calibration results were calculated after each game week, with the best calibration result being shown in the graphs below. The graphs showcase all 4 models and how well they do in predicting the home team and the away team to win. These are graphs for the calibration results after game week 13. The dashed red line is what perfect calibration would look like and provides a reference for comparison. The model closest to the red line would be the best model in terms of predictions. In the case of both home wins and away wins the Random Forest is the best model.



5. Bankroll Management System (Kelly Criterion)

The Kelly Criterion model is a form of probability theory often used by investors that we plan to incorporate into this model. The goal of the model is to maximize profit while accounting for the risk associated with a lost bet. This is done by maximizing the logarithm of the potential ending bankrolls after the bet is placed. Many online sources choose to simplify the math behind the Kelly Criterion Theory into a generic formula that looks like this:

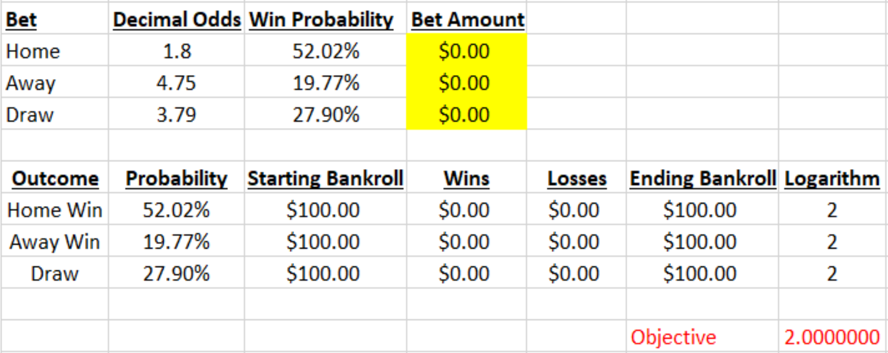
$$\frac{(O \times P_w) - P_l}{O} = B \tag{13}$$

$$\mathrm{O} = Odds, P_w = Prob\ of\ Bet\ Win, P_l = Prob\ of\ Bet\ Loss, and\ \ B = Percentage\ of\ Bankroll$$

While this formula does involve some principles of the Kelly Criterion Theory, it does not reflect the theory itself. Also, a soccer match has three outcomes, slightly complicating the math of this simplified two outcome formula. Our goal is to simply maximize the logarithm of our potential bankroll according to the Kelly Criterion Theory. Below is our optimization for the Crystal Palace vs. West Brom example discussed earlier. The inputs are the odds and probabilities of each possible result as well as the starting bankroll. Using this information, we can calculate our ending bankroll for each possible event and then take the logarithm of that. Lastly, the "objective" (KCO Value) is a weighted average of the logarithms and their associated probabilities.

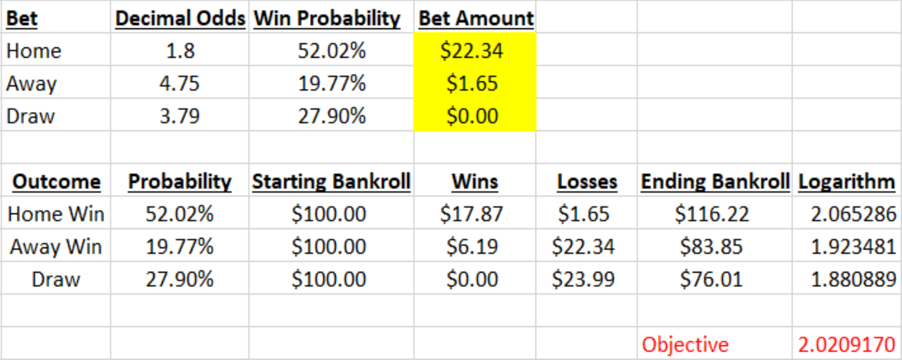
$$KCO = (HomeWinProb \times log(HW)) + \ (AwayWinProb \times log(AW)) + \ (DrawProb \times log(DW)) \tag{13}$$

$$HW = Ending\ Bankroll\ Home\ Win,\ AW = Ending\ Bankroll\ Away\ Win,\ DW = Ending\ Bankroll\ Draw$$



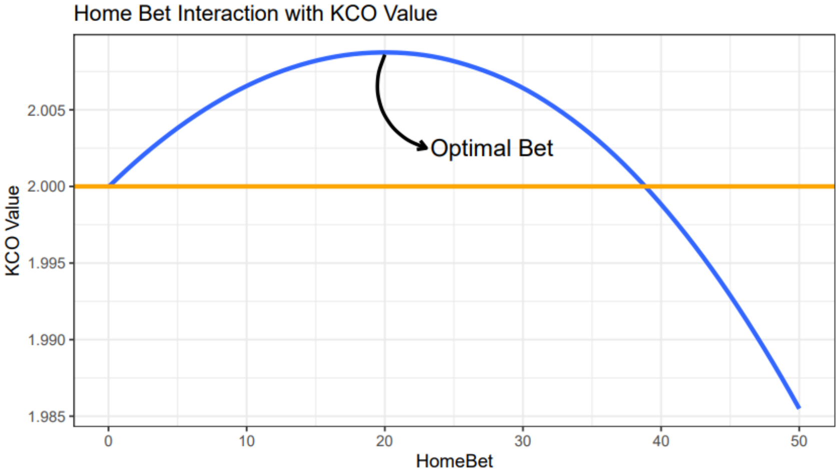
Bet	Decimal Odds	Win Probability	Bet Amount			
Home	1.8	52.02%	$0.00			
Away	4.75	19.77%	$0.00			
Draw	3.79	27.90%	$0.00			
Outcome	**Probability**	**Starting Bankroll**	**Wins**	**Losses**	**Ending Bankroll**	**Logarithm**
Home Win	52.02%	$100.00	$0.00	$0.00	$100.00	2
Away Win	19.77%	$100.00	$0.00	$0.00	$100.00	2
Draw	27.90%	$100.00	$0.00	$0.00	$100.00	2
					Objective	2.0000000

Above is an image consisting of the inputs and using this information and R, we maximize the objective cell by changing the “Bet Amount” cells. The image below is what a typical output would look like. Based on these odds and probabilities, the model suggests a bet of $22.34 on Crystal Palace to win and a bet of $1.65 on West Brom to win are the optimal bets to place that maximize long term profitability. The result of the game was 2 – 0 in favor of Crystal Palace, hence we would have profited $16.22 for this bet. We can see here that the model predicts a form of hedging, to minimize loss to some degree



Bet	Decimal Odds	Win Probability	Bet Amount			
Home	1.8	52.02%	$22.34			
Away	4.75	19.77%	$1.65			
Draw	3.79	27.90%	$0.00			
Outcome	**Probability**	**Starting Bankroll**	**Wins**	**Losses**	**Ending Bankroll**	**Logarithm**
Home Win	52.02%	$100.00	$17.87	$1.65	$116.22	2.065286
Away Win	19.77%	$100.00	$6.19	$22.34	$83.85	1.923481
Draw	27.90%	$100.00	$0.00	$23.99	$76.01	1.880889
					Objective	2.0209170

To further understand how the objective cell interacts with different bet amounts we can use the graph below:



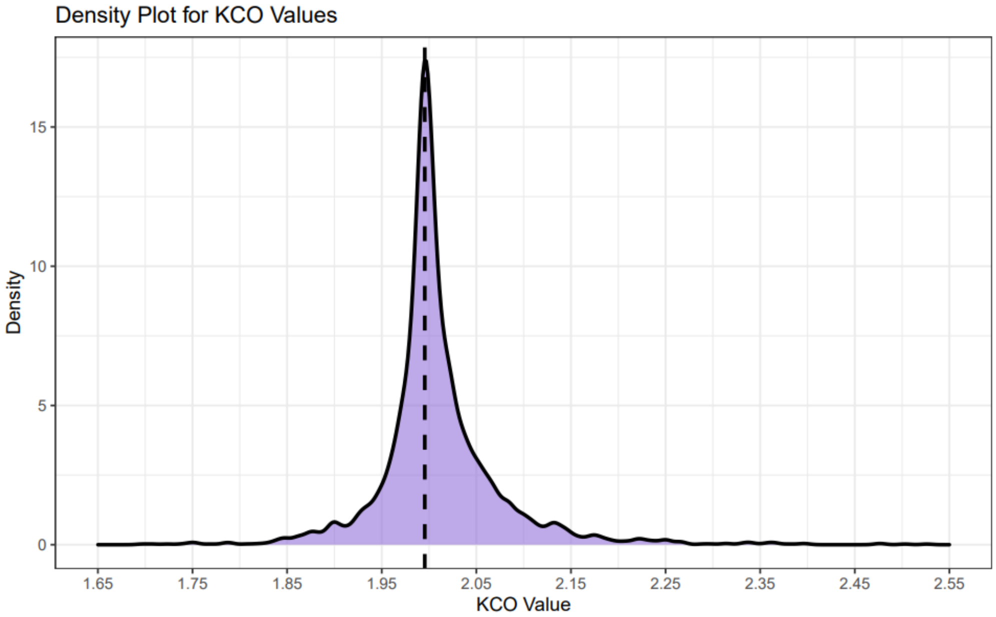
The above graph shows how the objective cell interacts with different bet amounts for a sample game (blue) in which the team being bet on has +100 betting odds and a 60% chance to win the game. It follows somewhat of a quadratic form that varies in shape depending on the betting odds and probability of winning. The left side of the curve indicates bet amounts that are profitable, but not as profitable as higher bet amounts. The right side of the curve indicates higher bet amounts would be considered higher risk bets that would prohibit long term success. The orange line indicates the objective number if no bet is placed. In other words, bets in which the blue line is below the orange line should not be placed. As mentioned before, our model essentially looks to find the maximum of the blue line. This method of risk management and profit maximization has been proven very effective in a study conducted by Victor Haghani and Richard Dewey.

Before moving on to our results, it is important to point out two more aspects of this methodology. The first is how to interpret the "objective" number produced by the optimization. As mentioned earlier, the objective number is the product of the logarithm of all possible ending bankrolls. Since we assumed a starting bankroll of $100, the objective cell when placing no bet is 2 since there is a 100% chance of ending with $100 and log(2) = 100. After we find the optimal bet, the objective cell becomes a representation of our confidence in profitability with a KCO value over 2 having a lot of confidence relative to a value below 2. Lastly, there are some games in which a bet might be placed on a team to win as well as a smaller bet on one of the other outcomes. This occurs as a form of hedging that reduces the potential loss of our bet not hitting.

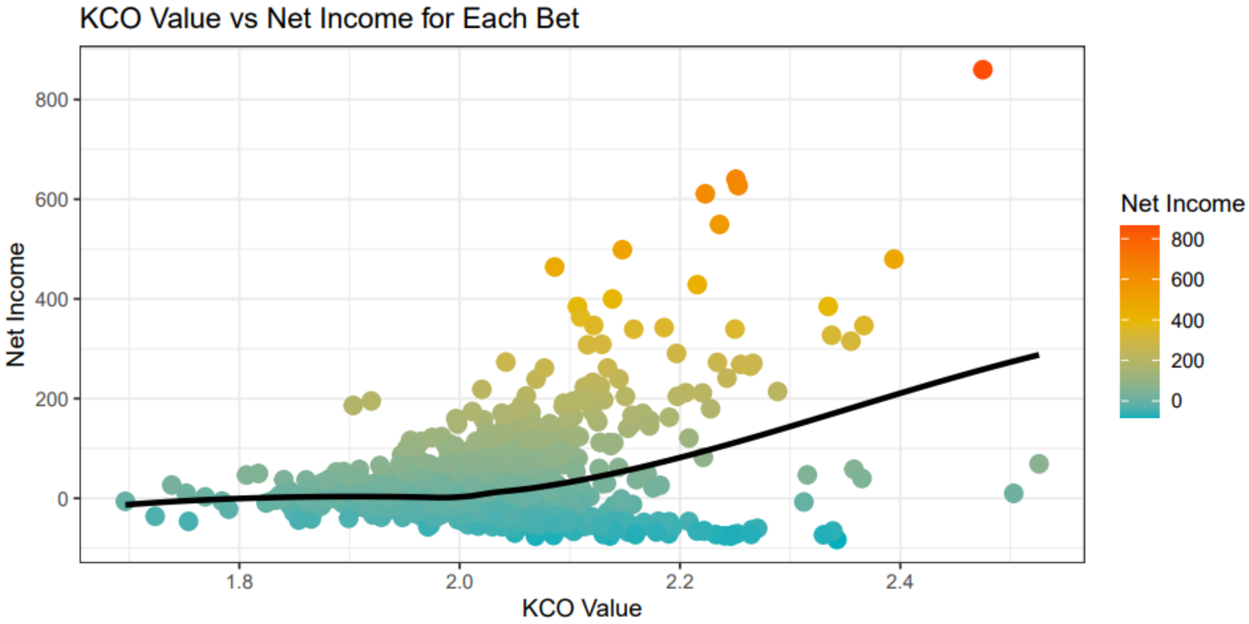
6. Model Evaluation

6.1. Interpreting KCO Values as Risk Estimators

Before analyzing our model's results, it is important to under the KCO value associated with each game (Game Week 5 onwards). Below is a density plot for the KCO values associated with each bet:



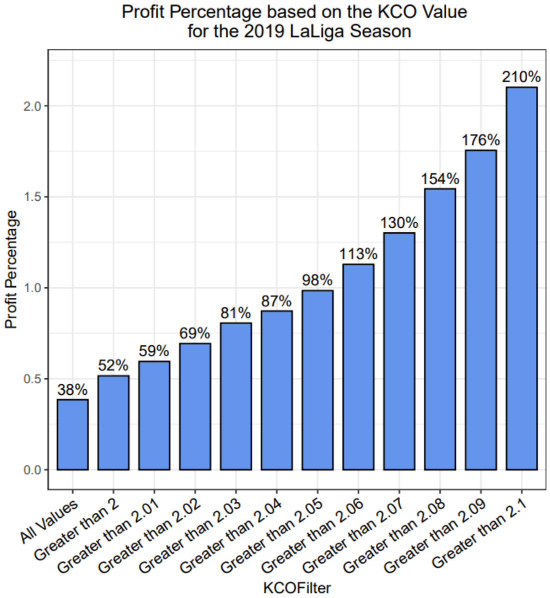
As mentioned earlier, a KCO value of 2 is associated when no bet is placed. We can see that the peak of the distribution is at a value extremely close to 2. This makes sense as oddsmakers are extremely effective at setting odds and hence a majority of the model's suggestions are to not bet anything. The values on the left side of the distribution are bets that are deemed "risky" with bets to the right of distribution being a bet with less risk. We can also see a weak but positive relationship between the success of a bet and the bet's KCO value. We can see this relationship below, with Net Income being the profit or loss from a bet:

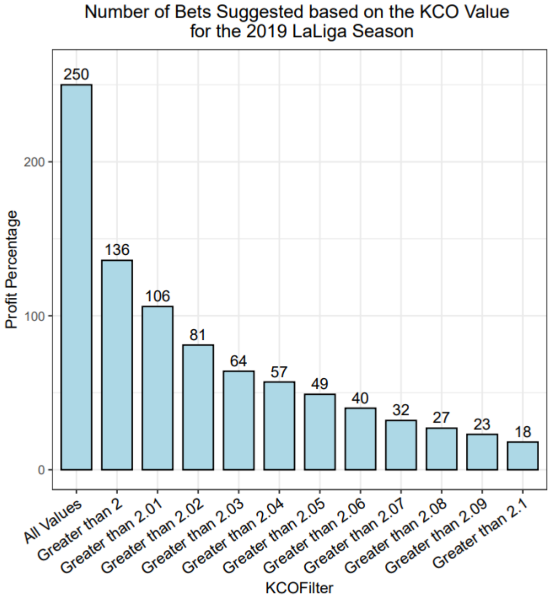


We can see as the KCO values increases, the net income of a bet also tends to increase. We can also see that the proportion of bets that make a profit are greater when the KCO value is greater than 2. This weak positive correlation between the Net Income and KCO Value shows us that the KCO Value can be used as a risk estimator. This also entails, that if we filter the model based on the KCO Value, we can increase our profit percentage. Theoretically, if we only place bets with a KCO value greater than 2 or 2.1, our profit percentage will improve. We further explore this hypothesis in the next section.

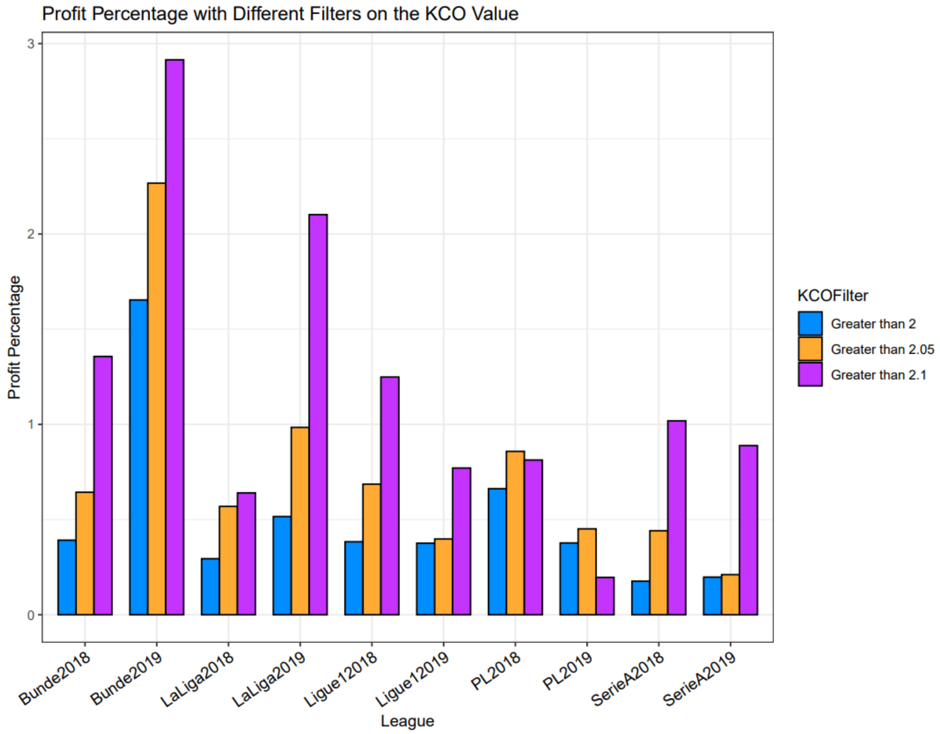
6.2. Success of the Model in terms of Profit

In this section, we will observe how successful the model was in terms of overall profit. It is important to note that the Kelly Criterion suggests different amounts to bet for each game based on the risk, hence we decided to use profit percentage, as an indicator for the model's success in addition to simply profit. As mentioned earlier, we used a bankroll of $100 for each game so that each bet could be made independent of the success of another bet. The first aspect we looked at was the relationship between profit percentage and the KCO value. If we look at the 2019 La Liga Season, we can see that as we increase our filter for the KCO value with increments of 0.01, the profit percentage also steadily increases. In addition to this, the number of bets suggested also decreases steadily in this case. This is highlighted below:


Profit Percentage based on the KCO Value for the 2019 LaLiga Season

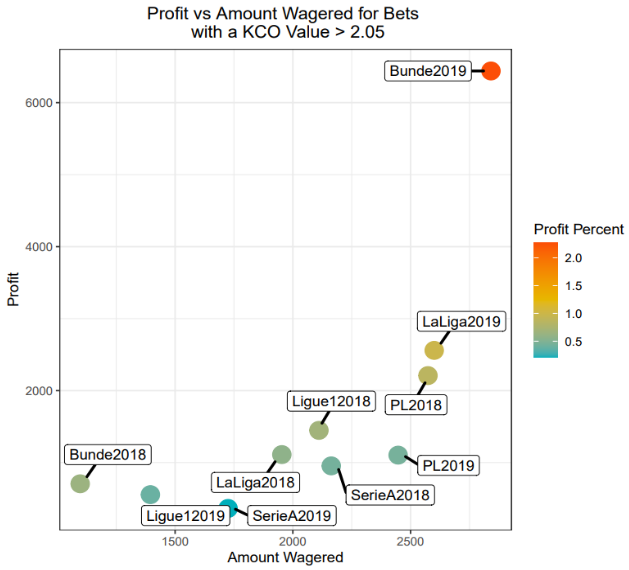

Number of Bets Suggested based on the KCO Value for the 2019 LaLiga Season

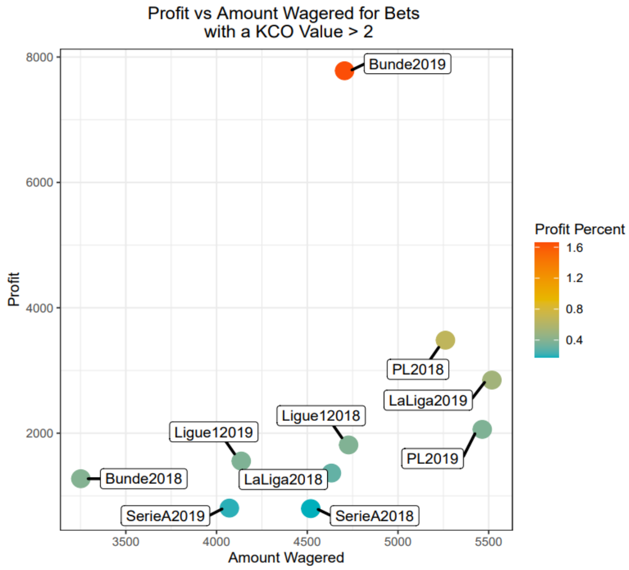
We then apply to this to every league, looking at profit percentage with 3 categories for the bets based on their KCO Value. If the KCO Value is greater than 2, greater than 2.05, and greater than 2.1.


Profit Percentage with Different Filters on the KCO Value

We can clearly see for every season, the overall profit percentage drastically improves as we select bets with higher KCO values. Every orange bar is higher than the league's respective blue bar and every purple bar is higher than the league's respective orange and blue bar outside the Premier League. While there is a massive amount of success based on profit percentage for bets with a KCO value greater than 2.1, the number of bets being placed is minimal. The Bundesliga in 2019 had only

26 bets with a KCO value greater than 2.1 which was the most we saw in a season. While the profit percentage is higher, the overall profit is higher when placing bets with a KCO value greater than 2 or 2.05. Hence, if the goal is profit maximization a used could place bets with a KCO value greater than 2 or 2.05. Below is a graph highlighting the amounts risked and won for these bets:





The KCO value provides a unique aspect of the model as it can be tailored to an individual's preference of how risky they would like to be with their bets. To provide a holistic explanation, below is a table highlighting the results for the model:

League	*KCO Value Greater than 2.05*		*KCO Value Greater than 2.1*	
	Amount Risked	Amount Profited	Amount Risked	Amount Profited
Bundesliga 2018	$3251.95	$1273.59	$454.73	616.80
Bundesliga 2019	$4705.58	$7780.69	$1720.42	$5014.22
La Liga 2018	$4633.20	$1363.87	$1098.77	$703.33
La Liga 2019	$5517.66	$2847.05	$1215.85	$2555.30
Ligue 1 2018	$4727.76	$1812.51	$894.89	$1117.37
Ligue 1 2019	$4136.08	$1555.22	$674.37	$519.70
Premier League 2018	$5261.52	$3483.21	$1148.28	$933.22
Premier League 2019	$5464.42	$2060.52	$897.53	$176.17
Serie A 2018	$4519.20	$796.81	$1178.63	$1200.40
Serie A 2019	$4070.81	$804.77	$704.14	$625.49

If we were to use our model for the past 2 years, placing bets that had a KCO value of greater than 2 we would profit $23,778.24 from betting on 1164 games. If we only placed bets with a KCO value of greater than 2.05 we would profit $17,450.94 from betting on 387 games. Lastly, if we only placed bets with a KCO value of 2.1 we would profit $9,987.61 from betting on 161 games. In order to fully understand the importance of the KCO value, we can look at the amount of money we would win or lose without implementing the KCO aspect.

If we were to simply bet on the outcome with the highest probability for each game, over the 10 seasons we would profit $62,340. While this seems like a large amount, we would be risking over $237,800. This yields a profit percentage of 26.22% which is not very good compared to the model with the KCO element. There is a significantly larger return on your investment when using the KCO aspect within model and hence we believe it is extremely important to use. Moreover, as discussed earlier the different KCO intervals allow for user to tailor the model to their preference in how risky they would like to be. This unique aspect is a also a big advantage provided by the KCO aspect of the model.

7. Future Research

7.1. Bivariate and Zero Inflated Poisson Distributions

It would be extremely interesting to repeat this process with a Bivariate or Zero Inflated Poisson Distribution. A bivariate distribution provides probabilities when you have two independent variables, allowing for each combination to be accounted for. Hence by using Home Goals and Away Goals as the independent variables, the distribution would provide probabilities for each combination of home goals and away goal. In a similar manner as we do in this paper, we can sum the probabilities for each corresponding event.

A zero inflated Poisson distribution is similar to a Poisson Distribution, however it is primarily used when an excess of 0 is expected within the data. Considering that outcomes with 0 goals being scored by a team tend to be higher in nature, a zero inflated Poisson Distribution may be more suited to predicting probabilities.

7.2. Factoring Lineup Selections Within Our Predictions

A major limitation for our model is that our model does not account for injuries and lineup changes. If we can factor in the strength of lineups using a holistic statistic, our predictions can drastically improve in terms of its accuracy. An example would be using the recently created Advanced RPM metric by the Syracuse University Soccer Analytics Club. By using each player's Advanced RPM value within our model, each lineup's strength is being accounted for. This factors in which players are playing for each team, allowing us to create even more accurate predictions.

7.3. Adjusting for Different Leagues

We can see a decent variation in success between the leagues we looked at. While all of our models are positive and yield a great amount of profit, there is possible room for improvement by adjusting for the talent distribution among the leagues. For example, the Premier League is considered one of the more competitive leagues, if we can factor that within our model, the suggested bets and overall profit from these bets will significantly improve.

8. Conclusion

When appropriately using past data to predict game outcome probabilities and then allocating funds efficiently to minimize risk it is clear that there is money to be made betting in European soccer. Utilizing our predictive model and with the Kelly Criterion, we found a large amount of success. The filtering of our KCO values in each respective league help us maximize our return on investment when betting. We found the most success and greatest profit percentage for all games where the KCO value was greater than 2.1 with a profit percentage increasing up to 291.45% in the Bundesliga in 2019. As discussed above, there are always improvements that can be made with our model. We hope to expand our model and continue to grow those profit percentages, finding the most accurate and profitable way to predict European soccer match outcomes.