Independent Work Report Fall, 2014

Author: Cole McCracken

Advisors: Mark Braverman, Christopher Moretti

Beating the Odds in Sports Betting Markets: English Premier League Betting and the Importance of In-game Statistics

Abstract

The English Premier League betting market is huge and continues to grow, making it worthwhile to study. I find evidence that including in-games statistics, such as shots, shots on goal, fouls committed, and corners taken, improve prediction models of match outcomes over models that just include win ratios. However, I do not find evidence of a market inefficiency documented by a recent study based on earlier data and I fail to produce a betting strategy that can earn consistent profits. Overall, the evidence presented suggests that if there was an inefficiency in the English Premier League betting market, it no longer exists.

1. Introduction

Sports gambling has become a very large market over the past several decades. Some believe that the industry could be worth up to 1 trillion dollars, 70 % of which can be credited to English football (soccer). In this one sport, over 500,000 people place bets each week and over 61 million have received a payout. [1] These statistics make the study of betting markets, and soccer in particular, a topic that warrants academic attention.

This paper focuses on the English Premier League (EPL) betting market. I chose the EPL for several reasons. First, many argue that it is the best league and it certainly is the most followed. The total television audience for the 2012-2013 season has been estimated at 4.7 billion people. [2] Second, detailed historical data for this league including betting odds is available online. [3] Finally, there have been a number of academic papers about sports betting markets and the EPL in particular since betting markets are important for testing market efficiency. [4] Many of these papers use financial analyses to determine whether the EPL betting market is consistent with the efficient market hypothesis (EMH).

According to financial theory, there are several types of efficiency. We are interested in the semi-strong form of the EMH, in which a market is considered efficient if the price of a good offered reflects all available public information. [5] This form of market efficiency leads to the implication that no one can earn positive investment returns consistently, or colloquially known as "beating the market." Turning to the studies that test the efficiency of the EPL betting market, the results are mixed and recent work suggests more research is warranted. Studies based on data from earlier periods generate betting strategies that earn positive returns. In contrast, later studies suggest that this is not feasible and the EPL betting market is generally efficient. Most notably, in a 2013 study written by active contributors on the topic (Buraimo, Peel and Simmons), the authors report "striking evidence of semi-strong inefficiency in the UK fixed-odds football betting market". [6] They construct a strategy that places bets on home wins and generate positive returns. So, an interesting question is: what explains these different results?

One possible reason for different results is that the studies estimate different models in predicting match outcomes. All models include the recent winning percentages of the teams and the home or away status of each team as independent variables. Some models also include additional variables, for example, whether the game was considered important, the number of spectators in attendance, and the distance between the stadiums of the two teams. [7] One key difference of the Buraimo, et al. study with prior research is that the authors use an online newspaper tipster, the Fink Tank Predictor, to predict game outcomes. [6] In contrast to other prediction models, the Fink Tank Predictor includes in-game statistics, in particular, the number of shots taken in recent games in its prediction model. [8]

Why should shots data in recent games matter? Soccer games are very low scoring and the team who wins is not always the team that played better. The difference between a shot going in and going wide is a matter of inches. I argue that the score-line of recent games does not fully capture the underlying skills of the teams. For example, if a team had a lot more shots than their opponent but was unlucky and did not win, the final score would fail to accurately represent the team's performance in that game. Therefore, models that include shots in recent games are better able to capture a team's strength and more accurately predict that team's chances of winning the next game. In addition to shots, other statistics, such as shots on goal, fouls committed, and corners taken (from now on referred to as 'in-game statistics') could improve the precision of a prediction model. In fact, the importance of in-game statistics in predicting outcomes has been demonstrated in other sports betting markets. For example, Zuber et al. test the market efficiency of American Football and include in their model passing yards, running yards, fumbles, and interceptions, among others, in recent games. They find that there are profitable betting strategies using in-game statistics and conclude that inefficiencies exist in the American Football market. [9]

My paper addresses two related questions: does the inclusion of in-game statistics – shots, shots on goal, fouls, and corners – improve the prediction of match outcomes in the EPL betting market over models that do not include those statistics? Can I devise a betting strategy using in-game statistics that leads to positive investment returns? If the answers to the above questions are yes, then this may partially explain the reason why Buraimo et al. find evidence of a market inefficiency while other papers find the opposite.

To explore these questions and test my hypothesis, I conduct several analyses. First, I compare models that include in-game statistics to models that just use goals scored to test whether the additional variables are important and lead to more precise predictions. The results suggest that including in-game statistics improves the model's accuracy in predicting match outcomes: the model containing in-game statistics predicts 2.9 % more games correctly than the model lacking in-game statistics. Hence, the evidence I present in this paper suggests that researchers should consider in-game statistics when constructing models of fixed-odds soccer betting markets.

Second, I test the models against the betting odds offered for each game to determine if it is possible to generate a profit. I also test a strategy that bets on only home games in an attempt to replicate the procedure Buraimo et al. use. The first model does not result in a positive investment return, losing 7.66% on average. Placing bets only on home wins does better, but also fails to support the inefficiency, losing 2.4% on average. Hence, the results presented in the paper are at odds with the inefficiency in the betting market as reported by Buraimo et al. using the Fink Tank predictions. Overall, the evidence suggests that the inefficiency in the EPL betting market no longer exists, if it once did.

The paper proceeds as follows. In the Background section, I detail how the EPL works, what odds mean and how to decipher them. I then review previous work in the Related Work section. In the Data and Algorithms section, I describe the data and algorithms I use to perform my analyses. In Part 1, I describe a model that evaluates the predictive power of in-game statistics and discuss the results. In Part 2, I test the model against the bookmakers to see if it results in profits. Finally, I conclude by discussing limitations and making suggestions for further research.

2. Background

2.1 English Premier League

The English Premier League (EPL) contains the best teams in England and is arguably the best league in the world. 20 teams participate in it annually and play each other team twice, once at home and once on the road, totaling 38 games per team. The team with the most points at the end of the season wins the title. Teams receive 3 points for every win, 1 point for each tie, and nothing for a loss. There are several important attributes that are key to understanding the league and related research on the EPL betting market.

First, individual matches can differ in terms of their importance to a team. The three teams with the least points at the end of each season are relegated to England's second division, meaning that next year they will not participate in the EPL. The three teams from the second division are promoted to play in the EPL the following year. This leads to games that could have different levels of importance for the two teams involved. Imagine, for instance, it is the last game of the season and one team will win the league if they win that game. However, the other team knows they will be relegated regardless of their result. This situation results in a different level of match importance for each team. This characteristic is relevant because some previous work uses the importance of a match in statistical modeling.

Second, there is an entirely different competition in English soccer, called the FA cup, which takes place during the season and has been used in models of existing research. The FA cup is a single-elimination tournament that every team in England competes in (including teams from the EPL and other secondary divisions). One might think that a team eliminated from the FA cup will have fewer games to play and can focus on EPL games, arguing that the team will perform better in their remaining matches. Alternatively, a team could receive a moral boost from winning games and thus remaining in the FA cup strengthens a team. Previous research includes whether the team is in the FA cup because it is unclear which argument is stronger.

Perhaps the most intriguing part about professional soccer is that ties (draws) are allowed and occur frequently, contrary to many other sports. This complicates betting because now the bettor has to consider the possibility of a draw. Nevertheless, it certainly does not deter people from placing bets.

2.2 Betting Markets

There are several ways to determine the odds of an event occurring. Examples include pari-mutuel, variable-odds and fixed-odds systems. Pari-mutuel odds are most common in United States horse racing. In this system, odds are determined from the amounts wagered on each competitor, and thus only solidified after all bets have been placed. [10] Companies that provide the odds (bookmakers) guarantee making a profit because they take a percent of the total money wagered (called the bookmaker's margin). They then define the odds to be proportional to the number of bets placed on each competitor. The variable-odds system is the other common betting scheme. Essentially bookmakers declare their odds and as people place bets, they adjust the odds based on the demand. [10] This system generally gives bookmakers assurance that they will make a profit because they adjust in order to mitigate risk. The final form of a betting market is fixed-odds betting, and is used in the EPL betting market. This system is much less prevalent. The bookmakers declare the odds and those exact odds are available up until the start of the match. All large bookmakers allow a bet to be placed on a home win, draw, or away win for any league game. [10] This system is riskier for the bookmakers as they could very well lose money if they publish poor odds. Bookmakers do reserve the right to change their odds, however, it is rarely seen in practice. [7] I will now detail what the odds mean and how they relate to probability of outcomes.

Odds are the payoffs given that the bet was successful. For instance, if you were given 2 to 1 odds that the away team wins the match, you placed that bet, and the away team wins, you would end up with twice as much money. Consider the following example, Chelsea is playing Manchester United and Chelsea is the home team for the contest. The bookmakers have given odds of (2.1, 3.2, 3.75). This notation suggests that a $1 bet placed on Chelsea will pay $2.1 if Chelsea wins; if the bet was placed on a draw, the payoff would be $3.2 if the teams tied; and finally a successful bet placed on Manchester United would yield $3.75. We can determine the probabilities for each event occurring (according to the bookmakers) by finding the reciprocal of each odd, giving us (.48, .31, .27). Notice how these numbers sum up to 1.06, not 1. This difference (i.e. 6%) is the bookmaker's margin. On average, the bookmaker will make a profit of 6 cents on the dollar for every bet placed. Note that this difference is not always .06, and it varies between games and bookmakers. We can normalize the result above by dividing each fraction by the sum, yielding (.45, .29, .26). [11] In the above calculation we assume that the bookmakers believe the generated probabilities are the true probabilities of each event occurring. However, this is not necessarily the case.

We must remember that the bookmakers are trying to maximize their own profits. If everyone bets rationally, bookmakers would maximize their profits by trying to offer markups off the true probabilities of the events occurring. However, much of the public does not bet rationally, and the bookmakers know this. Two major aspects of betting practices are perhaps easiest to understand. First, a lot of people view betting as a way to further support the team they root for. [10] Therefore, the bookmakers may give slightly worse odds than the estimated probability to a team that has a strong fan base. Secondly, people who are betting typically exhibit risk-loving behavior. This suggests that more people will place bets on teams that are underdogs because of the chance of a high payout. [12] This is known as the favorite-longshot bias and has been well studied.

Before I move on to describing the work that has been previously done on this topic, I introduce arbitrage. There is an opportunity for arbitrage when someone can take advantage of two different prices for the same good in order to generate a risk-free profit. For instance, let’s say one bookmaker offers 4/1 odds for the home take winning, another bookmaker offers 4/1 odds for a draw, and a final bookmaker offers 4/1 odds for the away team winning. If you place a $1 bet on each of those odds, spending 3 dollars, you would guarantee a $4 payout and a profit, regardless of the outcome of the match. The concept of arbitrage is useful for understanding related research.

3. Related Work

There are number of studies that investigate the question of whether the EPL betting market is efficient and if there are opportunities for consistent profits. Most of the newer studies suggest that bookmakers are getting better at accurately predicting games and there are no longer profitable strategies. However, as I have mentioned, Buraimo et al. disagree.

I will start with those who find ways to exploit the market. In 1996, Dixon and Coles find a statistical model based on a team's recent results capable of generating positive returns. [13] In 2000, Cain and others find that the same favorite-longshot bias found in horse racing also appears in the EPL betting market. [12] In their analysis in 2004, Goddard and Asimakopoulos discover that a strategy placing end of season bets can generate positive returns given bookmakers' odds. They believe that the current season's results are the most important and by the end of the season, their model has enough information to beat the bookmakers. [14]

However, each of these papers is over ten years old. It is quite possible that the market was inefficient and exploitable then, but now bookmakers are predicting outcomes well enough to prevent inefficiencies. Looking at data from 2000 to 2006 and across multiple leagues (EPL included), Stumbelj and Sikonja find that bookmakers' odds are getting better. [11] Forrest et al. conduct an analysis using data spanning EPL seasons from 1998-99 to 2002-03 and find that at the beginning of the period, statistical models outperformed bookmakers' odds. However, by the end of the period studied, the reverse is true. The authors' statistical models include the following independent variables: long-term win ratios, recent results, the importance of a match for each team, whether the team was still in the FA cup, the distance between the two team's home stadiums, and the attendance of the match relative to their league position. They attribute the failure of their model in beating the bookmakers to 'subjective adjustments'. The authors believe that the bookmakers got better at slightly tuning their odds by incorporating additional information such as injuries and suspensions, but also by using subjective analysis of how teams have played recently despite their results. [7]

Despite those findings, recent studies find inefficiencies in the market. Analyzing data from 2002-06, Deschamps and Gergaud find a positive favorite-longshot bias for home and away odds, a negative favorite-longshot bias for draws, and an overall draw bias (meaning that betting on draws results in higher returns). However, none of the strategies in this study generate returns to overcome bookmaker margins and turn a profit. [15]

This brings us to the paper I mention in the Introduction that uses a newspaper tipster, the Fink Tank Predictor, to predict game outcomes. The authors claim that betting on home wins using the predictions from the tipster result in statistically significant positive returns for each of the seasons from 2006-07 to 2011-12. [6] However, they do not investigate how these predictions are generated. On the Fink Tank website, the model is briefly described: "Our statistical model uses time-weighted shots and goals data to generate an attack and defense ranking for each club". [8] The newspaper further describes the model, but this is the only mention of what data the Fink Tank Predictor uses as independent variables.

Now we arrive at the point of my analyses. In 2005, Forrest et al. conclude that the market is efficient by showing bookmakers' odds outperform statistical models that use several different independent variables, but they do not include in-game statistics beyond the number of goals scored. [7] In contrast, in 2013, Buraimo et al. find evidence of a market inefficiency using a newspaper tipster that uses shots data in its model. [6] Is the reason for this discrepancy due to the inclusion of shots data? Do more in-game statistics, such as shots on goal, fouls, and corners, yield an even better model? And can that model outperform the bookmakers enough to generate a profit and further support Buraimo et al.'s conclusions that the EPL betting market is inefficient? In the remaining sections I detail the models I develop, describe the results, and discuss the potential explanations for those results.

4. Data and Algorithms

4.1 Data

I use data taken from a company called football-data. [3] It is in CSV format and contains match results for all EPL games, including goals scored, shots, shots on goal, fouls, and corners, for both home and away teams. I use the data from the 6 seasons between 2008 and 2014. I chose these 6 seasons because they are the most recent and because 6 seasons gives me a large enough sample to show that my results are not random. Each season contains 380 observations (the number of games played over the course of the season). An example of the data is given in Figure 1.

Home, Away Team	Home, Away Goals Scored	Full Time Result (H, D, A)	Home, Away Shots	Home, Away Shots on Goal	Home, Away Fouls	Home, Away Corners
Arsenal, Aston Villa	1, 3	A	16, 9	4, 4	15, 18	4, 3
Liverpool, Stoke	1, 0	H	26, 10	11, 4	11, 11	12, 6
…	…	…	…	…	…	…

Figure 1: Shows the dataset used for my analysis. The data included is from the first two games played in the 2013-14 season. The full time result is H (home win), D (draw), or A (away win). Dates were excluded for clarity.

The dataset also contains the fixed odds given out by several bookmakers for each of the matches. Figure 2 gives an example of that data.

Home, Away Team	Bet365 [Home Win, Draw, Away Win Odds]	Blue Square [Home Win, Draw, Away Win Odds]	Bet &Win [Home Win, Draw, Away Win Odds]	…
Arsenal, Aston Villa	1.44, 4.75, 8	1.36, 5, 7.75	1.37, 4.6, 7.5	…
Liverpool, Stoke	1.4, 5, 9.5	1.4, 4.33, 8.25	1.4, 4.4, 7.3	…
…	…	…	…	…

Figure 2: Shows the betting odds data used for my analysis for three bookmakers. The data included is from the first two games of the 2013-14 season. Dates were excluded for clarity.

In total, the data included 14 bookmakers. A glance at the values show that most of the odds are the same, and when they differ, they only do slightly. Some research shows that there have been cases in the past where arbitrage was possible. Buraimo et al. find that pure arbitrage opportunities occur perhaps every day in the EPL betting market. However, in practice, they state that arbitrage opportunities are not so large to pose a serious risk to bookmakers. [6] Nevertheless, this is not my research question, so I use a single company's (Bet365) odds for my analysis without a significant change to my results.

4.2 Algorithms

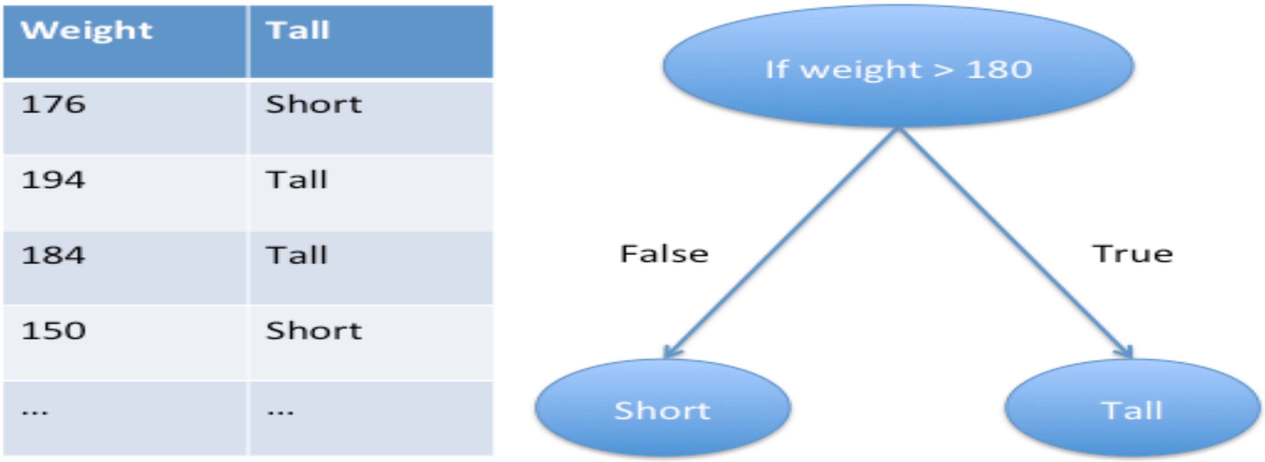
To run my analyses I use Weka, a machine learning library written in Java. I chose Weka because many advice columns considered Weka the best Java machine learning library and because there were good online tutorials and documentation. Weka's ideal input formats are ARFF files, so I used an online tool to easily convert from CSV. [16]

Much of the previous research on this topic has used statistical techniques to estimate models. To justify the choice of machine learning techniques, I refer to a couple other studies. In their book on machine learning, Witten and Frank argue that the statistical and machine learning domains have learned from each other and the gap between their approaches is small. [17] In the domain of EPL match predicting, Joseph et al. build several different Bayesian learners (another class of machine learning algorithm). The authors find that some do better than others, but call for further research on the topic. [18] I discuss the specific machine learning algorithm I chose next.

Among the many machine learning algorithms, I experimented with several to see which handled my data the best. There were a few that did better than others, but I chose random forests because it performed just as well as any other algorithm, but ran much quicker than most others.

It is important to note that the point of this paper is not to compare and contrast different machine learning algorithms. Below I give a brief introduction to random forests.

In order to describe how random forests work, I first introduce a classifier (a machine learning algorithm that labels input data into an output value) called a decision tree. Essentially, a decision tree is a structure that gets built during the training phase of a classifier. Each node has an if-else statement that then brings you to a node one level down in the graph. The nodes are based on the inputs and the decision of which node to follow down is made based on the value of those inputs. Once you are at a leaf node, you are classified into one of the output possibilities. [19] Let's consider a very simple example. Imagine you have 100 observations of 'weight' values (integers) and a boolean 'tall' output variable (tall if true, short if false). You then want to classify 100 more observations of weight values whose tall variable is unknown. A decision tree learner might construct the tree depicted in Figure 3 below.



Weight	Tall
176	Short
194	Tall
184	Tall
150	Short
...	...

Figure 3: An example of a simple decision tree using example data. In practice these are often much more complicated when there are more input variables.

A random forest is built from decision trees. There are a couple different ways to build them, but the underlying principle is that many trees are built from the training data and random vectors. Each tree then classifies the dependent variable and the most popular choice is taken. In general, random forests have been shown to perform better than decision trees on their own. [20]

5. Part 1: Importance of In-game Statistics In Predicting Match Outcomes

5.1 Preliminaries

In this section, I test the importance of adding in-game statistics – shots, shots on goal, fouls, and corners data – in predicting match outcomes. To predict the outcome of the current match, models use the home team, away team, and some information taken from each team's past matches. In order to test the importance of each of my in-game statistics, I construct models both with and without these additional variables, and use these models to predict the outcomes of future matches. The dependent (or outcome) variable captures the full time result of the match and can take on the values home win (H), draw (D), or away win (A). I hypothesize that the success rate in predicting outcomes is higher for models that include in-game statistics, therefore suggesting that these statistics improve model precision. Alternatively, in-game statistics may not provide any additional useful information in determining the strength of a team, and thus the ability to predict outcomes.

When analyzing the strength of a team, previous studies use the team's prior results as independent variables in modeling. I now introduce the term 'performance', which refers to the team's statistics over the course of its previous games. Note that the exact definition of performance differs slightly for each model, depending on which independent variables are used. In previous literature, performance includes: the full time results, the goals scored, and the goals allowed. In my models, I include these measures of performance as independent variables plus additional in-game statistics. I now distinguish between 'recent performance' and 'historical performance’. The former is the performance of a team in recent games (select games from the current season). The latter is the performance of a team over time (many games across multiple seasons).

Most of the previous work makes this distinction as well, and authors use both recent and historical performance as independent variables in their models. Forrest et al. measure historical performance as win ratios calculated over the past 24 months. [7] I argue that including in-game statistics from prior seasons as a measure of historical performance would not benefit the model’s ability to predict games. As I have mentioned earlier, in-game statistics should benefit a model because games are often very low scoring, and the team that wins is not necessarily the better team. In the course of the recent N games, a team could achieve poor results even if they have played reasonably well. Models that weight recent performance heavily, as most do, would theoretically benefit from including in-game statistics because recent random results would unduly affect the models. However, over the course of seasons, the chances of a team having a win ratio that does not represent the team’s strength are small, due to the quantity of games played. Thus, the information gleaned from historical in-games statistics beyond that of the win ratio would not add new information to the team’s strength and thus a model that uses in-game statistics would not perform better than one lacking such information. Therefore, I do not incorporate historical performance in my prediction models. Before I precisely define recent performance, I want to note a potential limitation with this approach. It is possible that historical win ratios and recent in-game statistics are highly correlated, implying that adding recent in-game statistics to previous studies would not substantially improve prediction models. While they are likely correlated, it is doubtful that historical win ratios encompass all information from recent in-game statistics.

The important question that remains is how many games are considered ‘recent’. Previous literature defines ‘recent’ somewhat arbitrarily. I now introduce the variable N, which represents the number of previous games in a season that my models use to define recent performance. Forrest et al. use N=9 for home games and N=4 for away games. [7] Goddard and Asimakopoulos vary their choice to see which ones yield significant results. [14] I choose a similar approach. I use 4, 6, 8, 10, and 12 for different values of N.

There are tradeoffs associated with using smaller vs. larger choices of N. On the one hand, a model using a very small number of games can be improperly influenced by lucky or unlucky runs. For example, a model that uses a single game, would classify a team that has one bad game as a “bad” team. Using a larger number of recent games would mitigate this effect. On the other hand, a model using a larger number of recent games may not capture an important change in a team’s strength. For example, if the model includes 20 games, but the team’s strength has declined recently due to injuries, suspensions, etc., then including the earlier games and giving less weight to very recent performance may negatively affect the model’s accuracy. Using a smaller N is somewhat similar to giving more weight to recent games. I use several values of N in order to capture the optimal value that incorporates the effect of these tradeoffs.

5.2 Model

Now I describe the models in this paper in detail and which independent variables are included. The Baseline Model includes many of the independent variables used in the previous literature, specifically the full time results, the home goals scored, and the away goals scored in each of the previous N games, in addition to the current home and away teams. I include 5 models, testing the importance of each in-game statistic – shots, shots on goal, fouls, and corners – individually and in combination. The Shots Model includes the same independent variables as the Baseline Model, but adds home shots and away shots from the previous N games. The Shots on Goal Model adds home shots on goal and away shots on goal to the Baseline Model. The Fouls Model adds home fouls and away fouls to the baseline. The Corners Model adds home corners and away corners. The Full Model adds all in-game statistics – shots, shots on goal, fouls, and corners – to the Baseline Model, testing the effect of the combination. I run this model in each of the 6 seasons. To evaluate the models, I compare each predicted result to the actual full time result and then compare the aggregate percentage correct (prediction rate) in each season. I also average across all seasons for each model in order to compare the prediction rates for each model. Since I am also varying N using 5 discrete values, I will have 5 (1 baseline model * 5 values of N) baseline prediction rates and 25 (5 testing models * 5 values of N) total testing prediction rates.

I want to note that the sample size will be slightly different for each N. In order to predict the current game, you need information for each of the N previous games, but the number of games in the season is fixed at 38 per team. So for N=4, there are 34 total predictions per team, for N=6, 32 predictions, etc.

5.3 Results

The Full Model, including all in-game statistics, improves the model's prediction rate by 2.9% ((45.07-43.78)/43.78) over the Baseline Model. I also find that including each in-game statistic individually also increases the prediction rate. The Fouls Model performs the best, doing 1.69% better than the baseline on average. The next most important variable is shots on goal, which improves accuracy by 1.25% over the baseline. The corners variable does the next best, improving the accuracy by .78%. Finally, the shots statistic barely increases its model's prediction rate by .15%. Figure 4 (below) details the full results and is useful for the discussion that follows.

	N=4 (2040)	N=6 (1920)	N=8 (1800)	N=10 (1680)	N=12 (1560)	Average
Baseline Model	41.13%	44.53%	43.67%	44.7%	44.87%	43.78%
Shots Model	39.66%	44.95%	44.33%	44.58%	45.71%	43.846%
Shots on Goal Model	41.18%	44.17%	45.72%	44.35%	46.22%	44.328%
Fouls Model	41.37%	45.26%	44.17%	45.77%	46.03%	44.52%
Corners Model	40.98%	45.05%	44.28%	45.18%	45.13%	44.124%
Full Model	43.28%	45.42%	44.78%	45.77%	46.09%	45.068%

Figure 4: The percent of games predicted correctly over the course of the 6 seasons tested using the N previous games of the current season as input into the classifier. The number of observations over the 6 seasons is in parentheses.

5.4 Discussion

Most of the predictive power in the models is captured by the full time result and goal differential variables. However, it is important to note, in-game statistics – shots, shots on goal, fouls, and corners – add some additional predictive power. It is surprising to see that of the individual models, the Fouls Model performs the best and the Shots Model the least. Perhaps, the information in shots data is mostly represented by goals scored. However, a team that moves the ball well may draw more fouls, but these fouls do not often lead to goals. Therefore the Fouls Model can identify teams that move the ball well and better predict their chances of winning the next match.

From Figure 4, we can also see that as N gets larger, the ability to predict the game improves slightly. The Full Model does better using N=12 and N=10 than using any of the other three possibilities. This suggests that it is better to use information from a greater number of recent games to offset the effect of a few random outlier performances rather than a smaller number to capture recent changes in a team's strength.

Finally, model precision varies across seasons. Figure 5 (below) details the performance of the Full Model by season.

	N=4 (340)	N=6 (320)	N=8 (300)	N=10 (280)	N=12 (260)	Average
2008-09	46.47%	47.5%	47.33%	50.71%	47.31%	47.864%
2009-10	45.29%	47.5%	44.33%	44.64%	48.46%	46.044%
2010-11	40.88%	47.5%	45.67%	41.43%	45.77%	44.25%
2011-12	43.24%	46.56%	42.67%	43.21%	40.38%	43.212%
2012-13	40.29%	38.44%	39%	44.64%	43.85%	41.244%
2013-14	43.53%	45%	49.67%	50%	50.77%	47.794%
Average across all seasons	43.28%	45.42%	44.78%	45.77%	46.09%	45.068%

Figure 5: The percent of games predicted correctly in each season tested using the N previous games as input into the classifier. These results are based off of the Full Model (including all in-game statistics). The number of observations over each season is in parentheses.

Comparing the 2012-13 and 2013-14 seasons, the number of games predicted correctly is over 6 % higher in the latter season (almost 7% in N=12, but ranges from 3+ % for N=4 to 10+ % for N=8). Since this is a fairly large discrepancy, I conduct additional analyses to determine what might be the cause. Analyzing the final standings for each of those seasons, I find a greater number of draws in the 2012-13 season than the later one. The average number of draws per team is 3 games more (10.8 vs. 7.8). [21] More draws suggest that the teams are more evenly balanced. Perhaps match outcomes were simply easier to predict in the 2013-14 season because the teams were less equal.

6. Part 2: Betting Strategies and Investment Returns

6.1 Model

Next, we attempt to devise a betting strategy capable of consistent positive returns. I use a similar approach as I do in predicting games. I let N take on values of 4,6,8,10, and 12 again. I would expect N=12 to yield the best results because it had the best prediction rate in the above experiment. For thoroughness, however, I estimate all models because the methodology for this section is slightly different.

In order to beat the bookmakers, I use the following procedure. I start with the model from the previous section by constructing the classifier based on the last N games and generate probabilities for the outcomes of the next match. I then compare these probabilities to the odds given by the bookmakers and place a bet on that game if my calculated probabilities would generate a profit in expectation. I compare my bet to the actual full time result to see if I earned money. I then sum up the money made over the course of each season and divide by the money wagered in order to determine the return on investment. I call this the Standard Model. I include all in-game statistics tested in the previous section as independent variables to this model because, while each of them performs slightly better than the baseline, the combination performs the best.

Let's look at a hypothetical example to illustrate more clearly how my model places bets. Liverpool (home) plays Arsenal (away). The odds given are 2:1, 3.5:1, 3.5:1 (home win, draw, away win). My model generates the probability vector (.52, .24, .24). The model would the place a bet on the home win, because in expectation I would make a 4% return (2*.52 = 1.04). Note that there are two other possibilities for each game. There is a chance that two of the probabilities generated would earn a profit in expectation. In this case, I bet on the higher expected return. It is also possible that none of the probabilities would yield a return in expectation because the bookmakers are taking a margin. In this case, I do not place a bet.

In theory, any time I generate a probability that would earn money in expectation, I should place that bet, which is what the above model does. However, Buraimo et al. find that betting on home games only using the Fink Tank Predictor generates a profitable betting strategy, not all games. [6] The Fink Tank Predictor includes shots data, like my model, so I also test a strategy that performs the same procedure as the Standard Model, but only bets if the the home team odds result in expected profit. Let this be called the Home Model. I detail the results from these two models next.

6.2 Results

The results show that neither strategy generates positive returns. First, I discuss the Standard Model that places bets on any type of outcome. Averaging returns across all seasons and models, I lose 7.66%. The best model, N=4, loses 4.71% on average across all seasons. The details of the returns on investments are given in Figure 6.

	N=4 (2027)	N=6 (1907)	N=8 (1786)	N=10 (1674)	N=12 (1551)	Average
2008-09	-16.62%	-10.86%	-10.09%	-11.93%	-19.7%	-13.84%
2009-10	-7.74%	-11.83%	-20.88%	-22.26%	-13.53%	-15.248%
2010-11	11.97%	5.35%	-11.75%	-0.85%	4.86%	1.916%
2011-12	7.6%	12.57%	13.97%	11.91%	13.61%	11.932%
2012-13	-7.67%	-12.55%	-23.02%	-33.01%	-24.71%	-20.192%
2013-14	-15.93%	-12.98%	-14.45%	0.56%	-9.78%	-10.516%
Average across all seasons	-4.71%	-5.05%	-11.07%	-9.28%	-8.17%	-7.656%

Figure 6: The entries are returns on investment. The values in parentheses at the head of the columns are the total number of games bet on across all seasons for each model. This table represents the results of the Standard Model.

The Home Model that places bets only on home wins (as in Buraimo et al.) does better, but also fails to generate positive profit on average. The best model, N=6, earns a 1% profit. However, averaging across all choices of N, 2.4% was lost. Due to the high variance across seasons and lack of a yearly trend among the results, I do not believe the N=6 positive return to be statistically significant. These results are detailed in Figure 7 below.

	N=4 (970)	N=6 (919)	N=8 (835)	N=10 (789)	N=12 (732)	Average
2008-09	-4.6%	1.66%	0.1%	-10.94%	-7.83%	-4.322%
2009-10	19.72%	7.24%	-3.28%	1.52%	3.41%	5.722%
2010-11	-9.16%	-1.19%	16.41%	8.01%	10.25%	4.864%
2011-12	4.89%	12.22%	10.14%	-3.39%	0.25%	4.822%
2012-13	-17.88%	-14.16%	-34.21%	-27.78%	-30.51%	-24.908%
2013-14	-0.51%	-0.89%	5.04%	-1.96%	-9.65%	-1.594%
Average across all seasons	-0.89%	0.99%	-1.01%	-5.42%	-5.71%	-2.408%

Figure 7: The entries are returns on investment. The values in parentheses at the head of the columns are the total number of games bet on across all seasons for each model. This table illustrates the results of the Home Model.

6.3 Discussion

While this paper does not provide evidence of an inefficiency in the betting market, the results are informative and relevant to the literature on this topic. The discussion section proceeds as follows. I highlight patterns in the results presented and give potential explanations for their occurrences. I then hypothesize that my failure to reproduce the inefficiency could be explained by the use of different data, specifically the inclusion of the 2012-13 and 2013-14 seasons. Finally, I discuss how the presented results may suggest that the betting market has become more efficient over time.

6.3.1 Select Insights from Results

First, notice that betting on only home wins leads to a higher return than betting on any outcome. Buraimo et al. offer a potential explanation: it may be due to the favorite-longshot bias, because home teams win almost 50% of games. [6] A bettor influenced by the bias will be more likely to bet on either draws or away wins. Bookmakers realize this and adjust for it, making the home odds slightly better.

Next, notice that using a smaller number of recent games improves the performance of the betting strategy. Specifically, models using N=4 and N=6 significantly outperform those using N=10 and N=12. This is especially interesting because the opposite is true in the prediction models of Part 1, where using larger N values improves the model's prediction rate. These findings are counter-intuitive. Clearly a model that predicts more games correctly should also perform better when bets are placed. However, the results presented demonstrate the opposite. Here is a possible explanation. The models using smaller N values do a better job of predicting the outcome probabilities for the less likely possibilities, which correspond to worse teams, than the models with larger N values. Since the model in Part 2 relies on odds given by a bookmaker, it is not always betting on the team that has the largest probability of winning, but places bets when it finds a profit in expectation. This could suggest the match outcomes of worse teams are better modeled using more recent history than the match outcomes of better teams.

Finally, we see the same variance across years as we did in Part 1. Using the Home Model, the return is almost -24.9% in 2012-13. That number increases 23% in the 2013-14 season to -1.9%, demonstrating the same pattern that we see in Part 1. It seems that a greater number of draws during a season has the same negative effect on the betting strategy as it does on predicting outcomes.

6.3.2 Evidence for the Market Becoming More Efficient

In this section I provide evidence that suggests the betting market has become more efficient since the study conducted that used the Fink Tank Predictor (Buraimo, et al.). Their analysis uses data from the 2006-07 season to the 2011-2012 season. My models use more current data and perform much better in the early period (2008-12 seasons) in comparison to the later period (2012-14 seasons). Specifically, the Standard Model performs substantially better in the first four seasons than the last two, with an average of -3.81% return compared to -15.35%. The Home Model drops from 2.77% to -13.21%. This alone demonstrates it was harder to make money over the last two years. If the latter two seasons were included in the Buraimo et al. study, it may have nullified their systematic profits and altered their conclusion about market inefficiency.

I offer another piece of evidence to further support this point. I introduce Figure 8 (below), which shows the average bookmaker margins in each year.

	Bookmaker margins
2008-09	5.31%
2009-10	5.44%
2010-11	5.45%
2011-12	5.46%
2012-13	4.11%
2013-14	2.62%
Average across all seasons	4.73%

Figure 8: The entries are the percentage margin that the bookmaker (Bet365) made on each game averaged across the whole season

It is interesting to note that the average bookmaker margins have decreased over the course of the last two years. There are many possible explanations for this trend, and it is likely due to some combination. Perhaps, the bookmakers have gotten better at generating odds. Consider the following: bookmakers develop more precise methods of generating probabilities for match outcomes. This leaves fewer inefficiencies for bettors to exploit and thus the bookmakers would be making more money on average. If there is competitive pressure, the bookmakers can reduce their markups, resulting in the data seen in Figure 8. This is one possible explanation that would suggest bookmaker odds are improving, and support the notion that the market is becoming more efficient.

7. Limitations and Opportunities for Further Research

In this section I discuss a few potential limitations of my study and I outline a number of avenues for further research.

One limitation of my study is due to data constraints. I am not able to include all independent variables used by previous literature, for example: match attendance, FA cup involvement, geographical distance, and the importance of the match for both teams. While I have shown that the use of in-game statistics improves the predictive power of statistical models, further research could determine if the effect of in-game statistics is independent of the other independent variables used in previous literature. Also, there are also other in-game statistics that are not available in my dataset, but I believe would improve a statistical model. One such example is time of possession. It is not hard to imagine that the time of possession in recent games would help contribute to representing a team's strength. Another avenue for future research would be to test if a betting strategy using a statistical model with the additional in-game statistics is capable of "beating the market."

The paragraph above discusses a number of analyses to conduct using additional data. As mentioned earlier, this paper's findings come from a model that is based on a machine-learning algorithm (specifically, random forests), whereas many of the recent studies on the topic that are referenced here estimate econometric statistical models (i.e., ordered probits). An interesting research topic is to compare these two types of models. As a first step, it would be useful to determine if adding in-game statistics as independent variables in an ordered probit specification would generate a model with greater precision in prediction rates and lead to positive returns in betting strategies.

Finally, I now turn to my year-by-year analysis. Buraimo et al. use data up to the 2011-12 season and find evidence of an inefficiency in the market. My results based on more recent data suggest the market has become more efficient. An informative and useful study would test the Buraimo et al. betting strategy in more recent seasons to determine if the positive returns of their investment strategy are robust in the later period. If not, it may suggest that bookmakers have recently gotten better, removing an inefficiency that could have been exploited by using in-game statistics.

8. Conclusion

In their paper published in 2005, Forrest et al. conclude that the EPL betting market has become more efficient over time. In contrast, in 2013, Buraimo et al. show that using the predictions of a newspaper tipster (the Fink Tank Predictions) yields systematic positive returns, suggesting an inefficiency in the EPL betting market. However, the authors do not investigate the prediction model, stating it is beyond the scope of their study.

In this paper, I present two main findings that contribute to this debate. First, I show that the inclusion of in-game statistics – shots, shots on goal, fouls, and corners – to a model that includes only prior game results and goal differentials improves the prediction of match outcomes. The use of in-game statistics increases the prediction rate by 2.9%. I then test the model against a set of bookmaker odds to see if I can replicate the inefficiency documented in the paper using the Fink Tank Predictions (which also include select in-game statistics). I am unable to generate consistent positive returns. I discuss possible explanations for this discrepancy including potential limitations of my data model and different time periods of the data analyzed in the papers. Overall, the evidence presented suggests that if there was an inefficiency in the English Premier League betting market, it no longer exists.

The evidence I present in this paper suggests that researchers should consider in-game statistics when constructing models of fixed-odds soccer betting markets. Models containing these statistics outperform those that do not. This paper's findings of betting strategies with only negative investment returns are at odds with those found by Buraimo et al. This discrepancy calls for further research on the topic – including incorporation of more data, estimation of a more robust model, and the application of the Fink Tank Predictor to more recent data – in order to convincingly evaluate whether the EPL betting market is efficient.

9. Honor Code

I pledge my honor that this paper represents my own work in accordance with University regulations.

Cole McCracken

10. Acknowledgements

I would like to thank my initial advisor, Mark Braverman, for helping me find a topic and guiding my analysis. I would also like to thank Christopher Moretti for answering late questions. I would finally like to thank my mom, Julie Wulf, for guidance in writing academic papers and in the editing process.

11. References

1. Frank Keogh, G.R. *Football betting - the global gambling industry worth billions*. 2013 10/3/13 12/17/14]; Available from: http://www.bbc.com/sport/0/football/24354124.
2. Arshad, S. 2013 [cited 2014 12/22/14]; Available from: http://www.tsmplug.com/records/most-watched-football-league-in-the-world/.
3. *Historical Football Results and Betting Odds Data*. 2014 [cited 2014 11/14/14]; Available from: http://www.football-data.co.uk/data.php.
4. Ziemba, R.H.T.a.W.T., *Anomalies Parimutuel Betting Markets: Racetracks and Lotteries.* Journal of Economic Perspectives, 1988. **2**(2): p. 161-174.
5. Fama, E.F., *Efficient Capital Markets: A Review of Theory and Empirical Work.* The Journal of Finance, 1970. **25**(2): p. 383-417.
6. Babatunde Buraimo, D.P., Rob Simmons, *Systematic Positive Expected Returns in the UK Fixed Odds Betting Market: An Analysis of the Fink Tank Predictions.* International Journal of Financial Studies, 2013. **1**(4): p. 168-182.
7. David Forrest, J.G., Robert Simmons, *Odds-setters as forecasters: The case of English Football.* International Journal of Forecasting, 2005. **21**(3): p. 551-564.
8. *The Fink Tank Predictor Help: Info*. 2014 [cited 2014 12/23/14]; Available from: http://www.dectech.co.uk/football_sites/football/help_info.php.
9. Richard A. Zuber, J.M.G., Benny D. Bowers, *Beating the Spread: Testing the Efficiency of Gambling Market for National Football League Games.* Journal of Political Economy, 1985. **93**(4): p. 800-806.
10. David Forrest, R.S. *Globalisation and Efficiency in the Fixed-odds Soccer Betting Market.* 2001.
11. E. Strumbelj, M.R.S., *Online bookmaker's odds as forecasts: The case of European soccer leagues.* International Journal of Forecasting, 2009. **26**(3): p. 482-488.
12. Michael Cain, D.L., David Peel, *The Favourite-Longshot Bias And Market Efficiency in UK Football Betting.* Scottish Journal of Political Economy, 2000. **47**(1): p. 25-36.
13. Mark J. Dixon, S.G.C., *Modelling Association Fooball Scores and Inefficiencies in the Football Betting Market.* Journal of the Royal Statistical Society: Series C (Applied Statistics), 1997. **46**(2): p. 265-280.
14. John Goddard, I.A., *Forecasting Football Results and the Efficiency of Fixed-odds Betting.* Journal of Forecasting, 2004. **23**(1): p. 51-66.

15. Bruno Deschamps, O.G., *Efficiency in Betting Markets: Evidence from English Football.* The Journal of Prediction Markets, 2007. **1**: p. 61-73.
16. Tkalcic, M. *csv2arff: Online CSV to ARFF conversion tool*. Available from: http://slavnik.fe.uni-lj.si/markot/csv2arff/csv2arff.php.
17. Ian H. Witten, E.F., *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. 2 ed. 2005.
18. A. Joseph, N.E.F., M. Neil, *Prediction Football Results using Bayesian Nets and other Machine Learning Techniques.* Elsevier B. V., 2006. **19**(7): p. 544-553.
19. Rokach, L., *Data Mining With Decision Trees: Theory and Applications*. 2007: World Scientific.
20. Breiman, L., *Random Forests.* Machine Learning, 2001(45): p. 5-32.
21. *League Table*. 2015 [cited 2015 1/6/15]; Available from: http://www.premierleague.com/en-gb/matchday/league-table.html.