Testing the Efficiency of Sports Betting Markets: An Examination of National Football League and English Premier League Betting

---

A Thesis
Presented to
The Established Interdisciplinary Committee for Mathematics-Economics
Reed College

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Arts

Johannes Harkins

May 2014

Approved for the Committee
(Mathematics-Economics)

Jeff Parker	Irena Swanson

Acknowledgements

Thank you to my advisors, Jeff Parker and Irena Swanson, for giving me invaluable feedback so promptly and so often. I feel like I was able to make this project as good as it could be in part because of the quality of the guidance I received.

Thanks to my parents, who have been beyond patient, understanding, and supportive of me for 22 years, through times when I haven't been the same to them. I appreciate more and more every day the things you've done and continue to do for me.

Thank you Sarah, for being by my side for the past two years, and keeping me convinced that I could meet all the challenges that I've taken on in that time. Thanks for your belief and love in the times when I've needed it most. I couldn't have done any of this without you.

Thanks to my friends, especially Michael and Alex, for convincing me that I belonged here when I wasn't sure that I did. Thanks to the rugby team and the soccer team for giving me a chance to run around a few times a week. It has been vital to my sanity. Thanks Peter, Alma, and Kriya for being an amazing group to work with on Renn Fayre last year. It's the best thing I've ever done that I would never do again.

Finally, thanks to the entire Reed community for being a place which allowed and encouraged me to challenge myself in so many different ways. I feel extremely privileged to have been a part of it for four years.

Table of Contents

Introduction **1**

Chapter 1: The Efficient-Market Hypothesis and Sports Betting . . **3**
1.1 The Sports Betting Market 3
1.2 The Efficient-Market Hypothesis 4
1.3 Methodology for a Model to Predict the Outcomes of Bets 6
1.4 The Model: The NFL Betting Market 8
1.5 The Model: European football 20

Chapter 2: Application of Betting Strategy **29**
2.1 Review of Strategies 29
2.1.1 Single Bet, Two Outcomes 29
2.1.2 Many Bets, Two Outcomes 31
2.2 Results of Strategy Application 32

Conclusion **45**

Appendix A: Python Scripts **47**
A.1 Weather Data 47
A.2 Distance data 51

Appendix B: R Scripts **55**
B.1 Betting tests 55

References **59**

List of Tables

1.1 NFL Variable List 9
1.2 Summary statistics 10
1.3 NFL Simple Regression Results 11
1.4 NFL Regression Results for Point Differential Model 13
1.5 NFL Regression Results: Divisional Effects 15
1.6 NFL Regression Results: Distance Effects 16
1.7 NFL Regression Results: Weather Effects 18
1.8 NFL Regression Results: Recent Performance Model 19
1.9 Comparison of Final NFL Betting Models used for Betting Tests . . . 20
1.10 EPL Variable List 22
1.11 Initial EPL Model 23
1.12 Match Significance & Goal Differential EPL Model 24
1.13 Expanded EPL Model 25
1.14 Expanded EPL Model including Weather and Recent Performance Variables 26
1.15 Expanded EPL Model accounting for odds 27

2.1 Betting Simulation Results 33
2.2 Comparison of Different Staking Methods (2012 NFL season) 35
2.3 Comparison of Different Staking Methods (2013 NFL season) 37
2.4 Comparison of Results for Conservative Staking Methods 39
2.5 In-Sample Results for Conservative Staking Methods (NFL) 40
2.6 In-Sample Results for Conservative Staking Methods (EPL) 40

List of Figures

1.1 Predicted Probability Plot 12

2.1 Bankroll by Betting Week for the 2012 NFL Season 34
2.2 Comparison of Betting Strategies over 2012 NFL Season 36
2.3 Comparison of Betting Strategies over 2013 NFL Season 38
2.4 Outcomes of Half Kelly conservative staking method over entire NFL sample 41
2.5 Half Kelly staking method over 2012-2013 EPL Season 42
2.6 Outcomes of Half Kelly staking method over 2013-2014 EPL Season . 42
2.7 Outcomes of Half Kelly conservative staking method over entire EPL sample 43

Abstract

I test the hypothesis that the markets for National Football League (NFL) spread betting and English Premier League (EPL) fixed-odds betting are efficient. I use a probit model to predict outcomes of spread bets and find that home teams are at a disadvantage for beating the spread, being the underdog increases a home team's likelihood of beating the spread, and that to some extent, rivalry matches decrease a home team's likelihood of beating the spread. The significance of these effects amount to a rejection of efficiency for this market. I run a hypothetical betting test of my models in Chapter 2 which seeks to determine whether these inefficiencies can be exploited. The results of these tests do not give any strong evidence that the information which is not incorporated into the betting odds can be taken advantage of. My results cannot reliably determine this, based on the variance of the hypothetical betting outcomes. In testing the efficiency of the EPL model, I use an ordered probit model to directly predict results of matches. When the probabilities of each outcome as implied by the listed betting odds are included in the ordered probit regression, only the effect of the away team's point differential remains significant in predicting outcomes. In Chapter 2 I subject the EPL models to similar betting tests as the NFL models and determine that there is no reliably profitable strategy which uses the EPL models. Thus I conclude that the EPL market has little to no inefficiency, and that it could not likely be profitable in a hypothetical betting scenario.

Introduction

My thesis explores the topic of sports betting markets. The market for (legal) online and casino gambling is larger than ever, and every bettor tries to devise a system to systematically profit from this market. What they are effectively trying to do is exploit a market inefficiency; a tendency of those who set prices (odds in this case), to improperly incorporate information. The fundamental question of whether a betting market is efficient has huge implications for the economics of betting houses and those who engage in legal gambling worldwide. My thesis attempts to explore this question in two ways. First, I will devise a model to predict outcomes of bets in NFL spread betting. In this way, I will test whether any publicly available information can yield predictive power about the outcomes of bets. This would constitute a market inefficiency. I also extend this to the fixed-odds European football betting market in the English Premier League. The second way I will test market efficiency is by devising a betting strategy, based on the information provided by the outcome predicting models I will develop in the first section.

Chapter 1

The Efficient-Market Hypothesis and Sports Betting

1.1 The Sports Betting Market

In order to examine the sports betting market, it is important to establish the rules of placing a bet in each sport's market. In the United States, the most common type of bet offered on most sports is the point spread bet. For a point spread bet, the betting house (here used interchangeably with bookmaker and oddsmaker) advertises a spread, and the gambler bets on whether the difference between the score of the favored team and the underdog team will be greater than or less than the spread. Since the payoff for successfully backing either team is equal, the spread set by the bookmaker is supposed to make the bet a fair wager; the bettor should have a 50% chance of winning the bet. For example, consider a bet on an American football game between Team A and Team B. The oddsmaker sets the point spread at 4, with Team A favored to win the game by that margin. If the gambler bets on Team A, Team A's final score must be more than 4 points greater than Team B's in order for the gambler to win the bet (lines are often set in half-point increments to avoid a "tie"). In the case that Team A's final score is indeed more than 4 points above Team B's, Team A is said to have beaten the spread. The bettor who wagers on Team B wins in the case that Team A fails to beat the spread. The point spread bet is common in American football and basketball.

The betting house takes a transaction cost on each bet which is usually set such that a bettor must wager \$11 to receive a \$10 payout (on top of having their \$11 returned). It is desirable for the sports book to have an equal amount bet on both

teams, so that they risk no money on a game, instead taking transaction costs as profit. A sports book receiving betting action weighted heavily towards one team may manipulate the line to give the other team favorable odds, enticing bettors to balance the books. In this case, the bettor receives the point spread at the time of his or her bet, which is not the case for horse racing and certain other types of betting. (Thompson 2001)

Betting on baseball usually occurs in the form of fixed-odds bets. Fixed-odds betting involves the sports book giving odds for possible outcomes of games. The odds are listed in terms of payoffs on particular bets. For example, a bet on a game between Team C and Team D could have the odds listed as $+120$ for Team C winning and -130 for Team D winning (these types of odds are called money line odds). This means that a bettor wagering \$10 on Team C would receive \$12 in the event of Team C's victory, but the bettor wagering on Team D would have to put up \$13 in order to receive \$10 on a successful bet. Again, the successful bettor also has their initial bet returned along with the payout. Most baseball bets list the starting pitchers for each team, with the bet being void and money returned should either of the pitchers listed not actually play. (Thompson 2001)

American football, basketball and baseball all also offer bets on the total number of points (or runs) scored in a game. In this case, a bettor is said to take the "over" or the "under" by betting on more or less points being scored than the listed amount. Betting on European football is also done by the fixed odds system (with the added possible result of a tie game), but odds are usually displayed in decimal or fractional form. (BWIN 2013) Fractional odds are presented as a fraction with the return a bettor stands to make, less the original bet, as the numerator and the bet required to win that amount as the denominator. Decimal odds are expressed in terms of payout. Decimal odds of 1.25 for example, mean that a successful bet yields 1.25 times your original bet, and would be equivalent to fractional odds of 1/4. Some sports books will combine money line odds and point spread betting for NFL games. Additionally, proposition bets are offered in many sports, wherein fixed-odds bets are placed on specific events, such as Player X scoring a goal. I will consider point spread betting and simple fixed odds betting in my analysis.

1.2 The Efficient-Market Hypothesis

The efficient-market hypothesis is an important theory for economics and finance, but it applies equally well to sports betting. The efficient-market hypothesis (EMH)

was outlined and developed in Eugene Fama's 1970 paper. Fama formulates the EMH in terms of weak form efficiency, semi-strong form efficiency and strong form efficiency. The key concern of all forms of the hypothesis is that the market in question incorporates information efficiently. Stated generally, the EMH holds that one cannot earn systematic returns in a market above the market-wide rate given available information for forecasting. The weak form of the EMH (in the case of an asset market) holds that prices fully reflect past price and return information for those assets. The semi-strong form version is concerned with the speed of the market's adjustment to new information, while the strong form hypothesis holds that no information, public or private, can lead to excess returns earned in the market. (Fama 1970)

In order to test the application of this hypothesis to the sports betting market, it must be formulated relative to the market in question, and it must be specified which forms are subject to testing. In this case, I will test primarily weak form efficiency, since private information is by its nature unavailable to the betting public and changes in information tend not to affect betting odds (since betting occurs mostly on events within a short time horizon) as drastically as they do stocks. For NFL betting and other point spread markets, weak form efficiency holds that no information can be used to predict the outcome of a bet. This means a model that is meant to predict bet outcomes directly should have no predictive power. In terms of regression analysis, the coefficient on any variable included in a regression determining outcome should not be statistically different from zero. In the case of fixed-odds betting, the weak form of the EMH entails that the bookmakers odds be exactly correct for each of the outcomes of a particular bet (after adjusting for the transaction cost of the bookmaker). Thus, as in spread betting, no historical information about match outcomes, team performance, or any other variable should give additional insight into the future match outcomes beyond the odds set for the bet.

The EMH holds for risk-adjusted return rates, meaning that the return rate in the market, adjusted for risk, will not exceed the market-wide risk-adjusted return rates. The issue of risk is somewhat glossed over in this thesis. In Chapter 2, I derive a betting strategy which depends on a logarithmic formulation of utility. This assumes a constant relative risk-aversion parameter of one. That is, as the principal investment grows with returns, the same percentage of the new total (principal plus returns) is reinvested in the market. This assumes certain behavior on the part of the bettor, who in reality, may set aside more money as their investment grows (increasing relative risk-aversion), or alternatively, invest larger proportions of their

assets (decreasing relative risk-aversion). Realistically, we would expect most bettors to probably exhibit some degree of increasing relative risk-aversion. This is a topic for further exploration in future work.

There are two ways in which I intend to test weak form efficiency as it applies to sports gambling. One natural way to test this application of the EMH point spread betting is to develop a model predicting the outcomes of bets, where the test of the EMH will be equivalent to testing the joint significance of all variables included in modeling the outcomes. I will refer to this as the statistical test of efficiency. The weak EMH can also be tested in terms of betting returns. If a betting system can be derived with the aid of the model which consistently derives returns, the market will be found to be inefficient. This is the the economic test of efficiency. The following sections of the first chapter will attempt to model outcomes in different gambling markets and evaluate statistical tests of efficiency, while the second chapter will devise optimal strategies for using these models, testing the markets' economic efficiency.

1.3 Methodology for a Model to Predict the Outcomes of Bets

There are a number of relevant considerations when deciding how to model sports betting. Deciding whether to model betting outcomes directly or indirectly through scores is a key factor in model choice, and even within these two methods, specifications of models varies greatly. Spread betting is widely modeled by directly attempting to predict the outcomes of bets. OLS estimation has been used in the case of the NFL to predict the point differential between two teams final scores (Golec and Tamarkin 1991), but this method weights a 20 point win differently than a 14 point win, where a desirable model would consider only whether a team beats the spread. Gray and Gray (1997) instead use a probit model to estimate the probability that the home team beats the spread in any given matchup (see also Wever and Aadland 2012). Using this model, they are able to demonstrate that the NFL market is statistically inefficient, with the oddsmakers systematically setting the line too high for favored teams and undervaluing home teams.

The estimation of outcomes in European football varies greatly in terms of method. Early research attempted to model scores directly by use of a Negative Binomial distribution (see Reep and Benjamin 1968). However, it was shown (see Maher 1982;

Dixon and Coles 1997) that a bivariate Poisson distribution could be used to model the score of a particular match between two teams, with parameters for the score distribution determined by the relative attacking and defending capabilities of the teams involved in the matchup. Though relatively effective at predicting scores, this method is computationally intensive, since there are at least two parameters to be estimated for each team involved. Additionally, though predicting scores rather than discrete outcomes (win, loss, tie) gives more precise information about matches, this information is unnecessary for producing a gambling strategy from the model, since I will be considering fixed odds bets only on the discrete outcomes. Not only is predicting scores unnecessarily precise, but the phenomena accounted for in this kind of model often have no bearing on a model predicting outcomes directly. For instance, empirical evidence from papers using the Poisson score distribution method suggests that there is some correlation between the scores of two European football teams. This is a non-trivial detail for the likes of Maher (1982) and Dixon and Coles (1997), all of whom attempt to account for this in their models. However, a model predicting outcomes alone need not attempt to account for this correlation, since it would be captured in the relative frequency of ties the model predicts. Furthermore, using the score-prediction method, parameters estimating the attacking and defending capabilities for a team remain fixed over time. This an undesirable feature of the model, since during a season, a team may experience a run of poor attacking or defending performance for any number of reasons. An ideal model would be able to incorporate information in a dynamic manner that allows for new information to increase predictive ability. Karlis and Ntzoufras (2008) derive a model for estimating outcomes via modeling the difference between the scores of two teams as the difference of two Poisson distributions. They employ a Bayesian methodology which allows them to incorporate new information into their model in the form of a posterior distribution. Though this allows for some of the dynamic aspects desired, modeling score difference still requires estimation of a large number of parameters and is, again, a more specific prediction than is necessary.

The ordered probit method employed by Goddard and Asimakopolous (2004) models discrete match outcomes directly instead of through scores, making it simpler to compare to bookmaker's fixed odds for each outcome and involving estimation of fewer parameters. This is more akin to estimation employed in the NFL betting market. It allows for dynamic incorporation of information in the form of variables that account for recent performance. It is also less computationally intensive, freeing the model from having to estimate a large number of team-specific variables. For these

reasons I will employ the ordered probit method in my own estimation of English Premier League (EPL) matches.

I intend to employ both statistical and economic tests of market efficiency for the NFL and English football as a paradigm case of European football betting. I will be using a probit model for the NFL to directly predict the two fixed-odds outcomes of bets, and an ordered probit model to predict the three outcomes in English football (home win, draw, away win). In the case of European football, the ordered probit methodology is preferred due to its computational ease, direct modeling of bet outcomes, and ability to incorporate recent information. Though the methods used for the two betting markets differ in that ordered probit has three discrete outcomes instead of the binary case of the simple probit, they are both commonly estimated using maximum likelihood and employ the cumulative distribution function of the standard normal distribution in the same way. By using similar methodology for the two cases, as well as including variables that stand in for similar effects, it may be possible not only to explore the case of efficiency in each separate betting market, but also to understand why the two markets differ in efficiency.

1.4 The Model: The NFL Betting Market

In the NFL betting market, I intend to test efficiency of the simple spread bet. I will do this by constructing a probit model to predict whether a given team beats the spread. My data are for NFL seasons 1978-2012 and include scores, final point spreads set before the matchup, information about which team was home/away, as well as information on the number estimated by the sportsbook for the total number of points scored in the game. Included in the raw data are the 1982 and 1987 seasons, that featured player strikes which shortened the seasons to nine and fifteen games respectively. Given the vastly shortened schedule in 1982 and the fact that several of the games played in 1987 featured replacement players, these seasons will be excluded from analysis.

Table 1.1 shows all of the variables included in various model specifications for the NFL analysis. The table gives each variable's name, a description, and the predicted effect on the probability of the home team winning. The variables labeled (Y/N) are dummy variables. Those with predicted effects of "Neutral" are not anticipated to have effects in either direction, or to have effects only when used in interactions with other variables. Under the hypothesis that the market is efficient, the effects of any variable on the home team beating the spread should be zero. Thus these effects are

simply the effect on match outcomes. Comparing the actual point difference and the

Table 1.1: Table of Variables for NFL Analysis

Variable Name	Variable Description	Predicted Effect
Result	Betting Outcome for Home Team	Dependent Variable
Home	Home Team Effect	Positive
Underdog	Home Team Also Underdog (Y/N)	Positive
HPD	Home Point Differential	Positive
APD	Away Point Differential	Negative
Distance	Travel Distance (Miles)	Positive
Large Distance	Distance Greater than Average	Positive
Divisional	Divisional Game (Y/N)	Positive
Rain	Rain Occurred on Gameday (Y/N)	Neutral
Snow	Snow Occurred on Gameday (Y/N)	Neutral
Precipitation	Daily Rainfall (mm)	Neutral
Snowfall	Daily Snowfall (mm)	Neutral
RPAS	Recent Home Performance (Against Spread)	Positive
RPO	Recent Home Performance (Outright Wins)	Positive

point spread, it is simple to construct a binary variable for whether a particular team beat the spread. Gray and Gray (1997) assign a so-called reference team for each matchup randomly, and perfom a probit regression on whether the reference team beat the spread to provide a simple test of statistical market efficiency. Though this method works well, a random reference team need not be assigned in order to test all effects. I will use the home team as the reference team, which has similar results. Since no information should give predictive power about whether the home team beats the spread (given that the bet is offered at even odds), any variable included in estimation should have a coefficient that is not statistically significant.

In the NFL, there are pieces of prevailing wisdom about effects considered significant to the performance of a team. The idea of home-team advantage is a common trope in sports, with the team playing in their own stadium seen to have a significant advantage due to the crowd and fatigue and time-zone effects of travel suffered by the opposing team. It is possible this effect may introduce some bias into point spreads, in the case that the oddsmakers value home teams too little or too much. Another potential source of bias is valuation of favorites. These topics have been explored by Gray and Gray (1997), who showed with their data that NFL oddsmakers consistently undervalue home teams and underdogs. In addition to these factors, my model will

test different variables for recent performance of teams against the spread, as well as several other explanatory variables attempting to account for factors not related to team quality or performance. I will test the significance of a sort of rivalry effect, by accounting for a team playing another team within its division, as well as weather effects, and the effects of a team's success over the previous three games.

For the probit analysis, the functional form of the model is written as:

$$\Pr(\mathbf{Y} = 1|\mathbf{X}) = \Phi(\mathbf{Z}) \tag{1.1}$$

$\mathbf{Y}$ represents the vector of all observations of the dependent variable, in this case the dummy variable for whether the home team beat the spread. $\mathbf{X}$ is the matrix of independent variables over all observations, Φ is the cumulative distribution function for the normal distribution, and $\mathbf{Z}$ is the product of the vector of coefficients, $\boldsymbol{\beta}$, and $\mathbf{X}$. Beginning with a simple probit model, we will consider the case where, for observation i, $z_i = \beta_0 + \beta_1 x_{1i}$, with

$$x_{1i} = \begin{cases} 1 & \text{if home team is underdog in observation } i \\ 0 & \text{otherwise} \end{cases} \tag{1.2}$$

Because the dependent variable is the result of a spread bet on the home team, the constant term β_0 can be used to determine the implicit advantage in the spread betting outcome enjoyed by the home team. I will be testing the hypothesis, then, that the underdog and home team effects have predictive power for spread bets. Examining simple summary statistics for a few variables gives an insight about what to expect from this analysis. Table 1.2 compares the spread for favorites and home teams versus their actual point differentials.

Table 1.2: Summary statistics

Variable	**Mean**	**Std. Dev.**
Home Team Spread	2.493	5.891
Actual Point Diff. of Home Team	2.753	14.605
Favorite Spread	5.397	3.433
Actual Point Diff. of Fav.	5.247	13.905
N	8174	

The means of each variable show that, at a glance, the line set by betting houses undervalues home teams and underdogs, just as Gray and Gray suggest. The results of a simple regression will verify or refute this.

Table 1.3: Simple probit model for NFL bets

VARIABLES	(1) Result
Underdog	0.0385***
	(0.0121)
Home	-0.0142***
	(0.0068)
Observations	7,918

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 1.3 shows the results of the simple model, where only the effects of being the underdog and home team are taken into effect. The coefficient for the underdog variable is given as its marginal effect, which is equivalent to $\frac{\partial \Phi(\mathbf{Z})}{\partial \mathbf{X_1}}$. This means that being the underdog is estimated to increase the home team's probability of beating the spread by 0.0385. The model returns a marginal for the constant equal to $\Phi(\beta_0) - 0.50$, the base probability of a home favorite beating the spread, minus 0.50. However, we want to consider purely the difference in probability between the base probability of all home teams (not just favorites) beating the spread and 0.50. In order to obtain this value, the underdog effect, multiplied by the mean of the underdog variable, is added to the base value of $\Phi(\beta_0) - 0.50$, to obtain $\Phi(\beta_0 + \beta_1 \overline{\mathbf{X_1}}) - 0.50$, which is shown in the table. This can be interpreted as the marginal effect of being the home team. This means that, counter to the findings of Gray and Gray, home teams are estimated to have a probability of beating spread which is 0.0146 lower on average than that of the away team. The predicted probabilities for each team in different scenarios is illustrated in Figure 1.1.

This simple regression and all subsequent regressions are performed on all the NFL seasons which are spanned by the data, except the 2012 season, which is reserved for testing betting strategies. Further regressions also display marginal effects for all variables and the base probability value for the home team effect, calculated as above.

For the purpose of constructing a successful betting strategy, and in order to

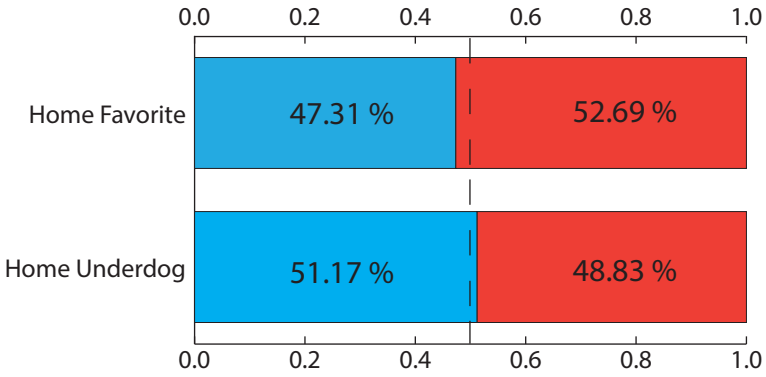

Figure 1.1: Estimated probability of home team success against the spread (in blue) versus probability of away team success (in red)

understand precisely which information has an impact in predicting $\Pr(\mathbf{Y} = 1|\mathbf{X})$, I consider a few additional models. In addition to home team and underdog effects, I explore several other factors.

Though a successful betting house takes into account the relative abilities of two competing teams when setting the spread for their matchup, it is possible that the methods used for measuring this ability can be flawed. Considering a team's Win-Loss record is a simple and conventional way of measuring ability, but there are additional methods that may improve upon this simple metric. One such measurement is point differential. Point differential can be alternatively considered as the difference or ratio between the cumulative number of points a team has scored and the number of points the team has allowed. The idea that the point differential is closely related to a team's ability is not a new one, and has been shown to apply in multiple sports, including baseball, American football and hockey. (see Barnwell 2013 and Ruggiero et al. 1997)

In order to take into account the ability of the home team, as well as its opponent, I track cumulative point differentials of both during each season. Due to fluctuation in rosters and team management staff between seasons, I consider only point differentials from a team's current season. I also normalize point differential by dividing by the number of games played. In this way, I correct for the fact that point differentials tend grow as the season goes on. After this correction however, the point differential per game variable has a larger variance for lower games played. Thus, I further correct for this non-constant variance by calculating the standard deviation for point

differential conditional on the number of matches played, then dividing all point differentials by this standard deviation at each number of games played. In this manner, I obtain a measure of a team's point differential relative to the distribution of all point differentials at the same point of each season. The expression used for calculating the normalized variable in observation i is shown below:

$$HPD\_Norm_i = \frac{HPD_i}{HGP_i} \times \frac{1}{SD(HPD|HGP_i = j)} \tag{1.3}$$

Here $SD(HPD|HGP_i)$ is notation for the sample standard deviation as calculated for all values of HPD where $HGP = j$ for the same observation (for all seasons). That is, I calculate the standard deviation separately for each of the distributions of HPD over all seasons with a given value for HGP.

I include this variable for both the home and away teams in order to measure the strength of both the home team (for whom probability of success is predicted by the probit model), and its opponent.

Table 1.4: Probit model including normalized point differential variables

	(1)
VARIABLES	Result
Underdog	0.0389***
	(0.0150)
HPD	0.00930
	(0.00660)
APD	0.00924
	(0.00654)
Home	-0.01457***
	(0.0072)
Observations	7,918

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 1.4 shows the results of this regression. Again, we see that the effect of playing as an underdog and the implicit effect of being the home team are significant for a home team's chances of beating the spread. The degree to which they affect the probability of this success is mostly unchanged, which makes sense, given that

being a home team or underdog is not expected to be collinear with performance. However, the performance variables themselves do not appear to be significant. This means that, though bookmakers appear to be underestimating the home team and underdog advantages, they accurately measure the abilities of the two teams.

An additional factor that is considered by many when placing bets on NFL games is a rivalry effect. This posits that games are more even that they otherwise might be if the matchup between the two teams is a rivalry. Teams are said to be more motivated by playing rivalry match-ups. To stand in for rivalry games, I created a dummy variable for a game being a divisional match. For the entire span of my dataset, the NFL has used a divisional structure that results in more of a team's games being played against other teams within its division than against other teams. By playing certain teams more often, rivalries often develop. In addition, teams within the same division tend to be closer together, and compete directly for playoff spots, further engendering rivalries between co-divisional teams. This measure of rivalry games is imperfect, however. Divisions have changed (though usually not drastically) over time, and teams develop rivalries outside of their divisions based on factors such as coincident periods of success and geographical proximity. My method though, avoids an arbitrary decision about what constitutes a rivalry, while capturing divisional ones, and is thus preferable.

I created a model which tested hypothesis that rivalry effects benefit the underdog by leveling the playing field, and the hypothesis that rivalry effects benefit the home team by virtue of an increased home team advantage. The results in Table 1.5 show that in neither case do rivalry matches have a statistically significant effect on the home team's probability of beating the spread. However, in a second model displayed in the table, the added underdog divisional matchup effect is dropped and the overall divisional effect improves to statistically significant at the 12.2% level, which is much closer to significance than that of the added underdog effect. There is little evidence, based on these models, to suggest that underdogs stand a greater chance of beating the spread when playing rivalry match-ups.

Another variable that is said to affect a team's probability of winning is the effects of traveling for road games. Though some of this effect may be captured by the home-team advantage considered in the simple model, quantifying the actual distance the away team travels for a matchup can more precisely pinpoint the impact of such effects as time-zone changes. I use several methods to account for these effects. I created a distance variable for the straight-line distance between the home stadiums of two teams for a given matchup. I also generated a large-distance dummy variable

Table 1.5: Probit model including point differential and rivalry effects

VARIABLES	(1) Result	(2) Result
Underdog	0.0350*	0.0397**
	(0.0195)	(0.0159)
HPD	0.00960	0.00959
	(0.00669)	(0.00669)
APD	0.00903	0.00907
	(0.00662)	(0.00662)
Divisional	-0.0214	-0.0181
	(0.0142)	(0.0117)
Divisional × Underdog	0.0104	
	(0.0251)	
Home	-0.01453*	-0.01452**
	(0.00973)	(0.00909)
Observations	7,411	7,411

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

which is equal to one if the distance travelled by the away team is greater than the mean distance traveled for all match-ups and zero otherwise. This variable accounts for the possibility that distance only has an effect for large distances travelled. Table 1.6 shows the results of regressions including the regular distance variable, as well as the large distance dummy. In both cases, the effect of travel is not significantly different than zero. Given this result, it appears that the spread fully captures any effect travel has on game outcomes.

Table 1.6: Probit model including point differential, rivalry and distance effects

	(1)	(2)
VARIABLES	Result	Result
Underdog	0.0396**	0.0397**
	(0.0159)	(0.0159)
HPD	0.00959	0.00957
	(0.00669)	(0.00669)
APD	0.00907	0.00902
	(0.00662)	(0.00662)
Divisional	-0.0182	-0.0178
	(0.0121)	(0.0118)
Distance	-3.97e-07	
	(9.22e-06)	
Large Distance		0.00249
		(0.0118)
Home	-0.01452	-0.01452*
	(0.0140)	(.01080)
Observations	7,411	7,411

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Weather is a factor that is mentioned in a number of ways in literature on American football betting. Poor weather has been mentioned as a factor which levels the playing field between two teams, since it can make it challenging for both teams to play and to score, introducing greater uncertainty in the result. Weather can also be seen to affect aspects of a team's offensive ability differently for different teams. For instance, passing plays are supposed to be more limited by weather than running plays, so a team which is more reliant on its passing ability to produce offense may be more affected by weather than a team with strong rushing ability. Since I do not

intend to break down a team's ability in terms of passing and rushing quality, I will consider the inclusion of the weather variable to be testing the hypothesis that bad weather adversely affects favored teams (the "leveling the playing field" theory). My data consist of dummy variables for whether it rained or snowed on the day of a game, and how much of each in millimeters. I also take into account domed stadiums, which negate weather effects, by introducing an indicator variable which, when multiplied with weather variables, negates their effects. The variables included in each regression are all multiplied by this dome variable.

Table 1.7 shows the results of the three models. Across all three models, the effects of weather do not differ significantly from zero. It appears then, that the effects of weather properly accounted for in the spread. If there is an effect of weather leveling the playing field, it is properly incorporated into the spread.

The last effect I intend to test for NFL spread betting is the effect of recent performance on a team's likelihood of beating the spread. Teams who have won several games in a row are often seen to be "on a hot streak" and thus considered more likely to win their next game. To account for this I created a recent performance variable which counts how many of the previous three games have been won by the home team. This variable could also capture the injury status of a team. A team missing one or more key players due to injury may suffer a string of bad results, but perform better when these players are healthy. In order to include this variable in a model, the first three observations must be dropped from the beginning of each season. This may improve previous models as well, since the first few games of each season could be harder to predict based on limited information for that season. I also created a variable for the recent performance of a team against the spread. In this case, I am testing how lines adjust to a particular team's tendency to beat or fail to beat the spread.

The results are somewhat contrary to intuition.The effect of recent performance in terms of outright victory is insignificant and the effect is near zero. However, recent wins against the spread have a negative effect on a team's probability of beating the spread in their next match, and the significance is very near 10%. Instead of interpreting this as a negative effect of recent against the spread wins on a team's performance against the spread, this effect could be thought of as a possible over adjustment on the part of oddsmakers to account for the performance of these teams in recent bets. That is, the spread is often too high in favor of the home team when it has performed well against the spread in recent games. Another curious result is the positive (and nearly significant) coefficient associated with away team

Table 1.7: Probit model including point differential, rivalry and distance effects

VARIABLES	(1) Result	(2) Result	(3) Result
Underdog	0.0394**	0.0457***	0.0397**
	(0.0160)	(0.0176)	(0.0159)
HPD	0.00894	0.00883	0.00882
	(0.00674)	(0.00674)	(0.00670)
APD	0.00850	0.00851	0.00926
	(0.00667)	(0.00667)	(0.00663)
Divisional	-0.0176	-0.0176	-0.0183
	(0.0118)	(0.0118)	(0.0117)
Rain	0.000521	0.00823	
	(0.0136)	(0.0166)	
Snow	0.0105	0.0119	
	(0.0284)	(0.0337)	
Rain × Home Underdog		-0.0240	
		(0.0292)	
Snow × Home Underdog		-0.00687	
		(0.0624)	
Precipitation			4.11e-05
			(0.00107)
Snowfall			0.00171
			(0.0177)
	-0.0153**	-0.0153**	-0.0143**
	(0.0098)	(0.0100)	(0.0092)
Observations	7,290	7,290	7,399

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 1.8: Probit model including point differential, rivalry and recent performance effects

VARIABLES	(1) Result	(2) Result
Underdog	0.0269	0.0274
	(0.0186)	(0.0186)
HPD	0.00408	-0.00101
	(0.00802)	(0.00921)
APD	0.0123	0.0122
	(0.00762)	(0.00762)
Divisional	-0.0222*	-0.0222*
	(0.0131)	(0.0131)
RPAS	-0.0105	
	(0.00787)	
RPO		0.00436
		(0.00905)
Home	-0.01043	-0.0104
	(0.0158)	(0.0172)
Observations	5,901	5,901

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

adjusted point differential. This can be thought of similar to the effect for recent against the spread performance. Additionally, the significance of the constant term disappears, suggesting some collinearity between recent performance and the home team advantage. This result is harder to interpret.

Table 1.9 contains a comparison of several models over the data minus the first three games of each season. Though the results do not not change much in terms of effect, statistical significance is affected by removing these games from the data. In order to compare the models though, we must use only these data, since comparing models estimated over different observations would have little value. Based on these results, I will test efficiency of the betting market in Chapter 2 by simulating betting over a full season using these models.

Table 1.9: Comparison of probit model specifications

	(1)	(2)	(3)
VARIABLES	Result	Result	Result
Underdog	0.0260	0.0267	0.0269
	(0.0186)	(0.0186)	(0.0186)
HPD	0.00120	0.00138	0.00408
	(0.00775)	(0.00776)	(0.00802)
APD	0.0123	0.0123	0.0123
	(0.00762)	(0.00762)	(0.00762)
Divisional		-0.0222*	-0.0222*
		(0.0131)	(0.0131)
RPAS			-0.0105
			(0.00787)
Home	-0.01918**	-0.0094	0.0063
	(0.0087)	(0.01043)	(0.01576)
Observations	5,901	5,901	5,901

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

1.5 The Model: European football

In order to explore a variety of betting markets, and test the efficiency of fixed-odds betting, I will also examine English football as an example of case of the European football betting market. As mentioned previously, there are a number of methods

that have been employed for estimating results in European football. I will construct an ordered probit model to model directly the probabilities of the 3 distinct match outcomes. In this way, I hope to offer ready comparisons to the NFL model by extending the model used in that market to 3 cases. The ordered probit model employed for the EPL data relies on an ordering of outcomes as follows:

$$Y_i = \begin{cases} 0 & \text{for an away team win in game } i \\ 1 & \text{for a draw} \\ 2 & \text{for a home team win} \end{cases} \tag{1.4}$$

Thus the bet can be considered to be from the perspective of the home team like in the NFL probit model. The probability of each outcome is calculated by first estimating the unobserved variable y_i^* where:

$$\mathbf{y}^* = \boldsymbol{\beta}\mathbf{X} \tag{1.5}$$

for the matrix of coefficients, $\boldsymbol{\beta}$ and the matrix of variables $\mathbf{X}$ over all observations. Then using two estimated cutoff values, μ_1 and μ_2, Y_i is calculated using:

$$Y_i = \begin{cases} 0 & \text{if } \mathrm{y_i^*} < \mu_1 \\ 1 & \text{if } \mu_1 < \mathrm{y_i^*} < \mu_2 \\ 2 & \text{if } \mu_2 < \mathrm{y_i^*} \end{cases} \tag{1.6}$$

So, the probability of each outcome can be calculated by the following three equations:

$$\mathrm{P(Y_i = 0)} = \mathrm{P(y_i^* < \mu_1)} \tag{1.7}$$

$$\mathrm{P(Y_i = 1)} = \mathrm{P(\mu_1 < y_i^* < \mu_2)} \tag{1.8}$$

$$\mathrm{P(Y_i = 2)} = \mathrm{P(\mu_2 < y_i^*)} \tag{1.9}$$

Unlike NFL betting, European football bets are not offered at even odds, so significance of variable coefficients will not be enough to test the EMH. Instead, estimated probabilities must be generated for each match outcome, and then evaluated against the probabilities implicit in the bookmakers' odds.

Table 1.10 lists the variables to be considered in the analysis of Premier League betting. Since the result variable is ordered from away win to draw to home win in ascending order, a variable which is said to have a positive predicted effect on the

dependent variable is predicted to increase the probability of a favorable outcome for the home team.

Table 1.10: Table of Variables for EPL Analysis

Variable Name	Variable Description	Predicted Effect
Result	Betting Outcome	N/A - Dependent Variable
HGD	Home Team Goal Differential	Positive
AGD	Away Team Goal Differential	Negative
Hsig	Match Significance for Away Team	Positive
Asig	Match Significance for Away Team	Positive
Hprom	Promoted Home Team	Negative
Aprom	Promoted Away Team	Positive
Rain	Rain Occurred on Gameday (Y/N)	Neutral
Snow	Snow Occurred on Gameday (Y/N)	Neutral

Similarly to American football, I generated a goal differential variable for each team over each season, as well as a variable for goal differential divided by number of games, in order to avoid the phenomenon of teams with goal differentials growing in absolute value as the season went on. Using the same process as in the NFL model, I normalized using standard deviations grouped by number of games played.

The match importance variable has an added wrinkle in this model, since European football involves relegation to lower leagues of teams finishing in the bottom three positions. Thus, as opposed to in American football, late-season matches can have vital importance to teams performing poorly as well as those with championship hopes. I calculated this variable by projecting end of season points totals for each team and calling their match significant if they fell within a certain range (see following page for exact calculation of this variable) of the average number of points required to win a title, qualify for European club competition, or avoid relegation. This effect applied only in the second half of the season.

The results of a simple regression are shown in Table 1.11. The table displays the marginal effects of the variables on each outcome, as well as the base probabilities of each outcome, calculated at the mean of all of the regressors. Just as in the NFL model, the most recent season is omitted from the data used for estimation in order to employ a betting test. Here we see that as expected, a higher goal differential increases the probability of a team winning. In this model, it is also estimated that increases in the home team's goal differential decreases the likelihood of draw, while

increases in the same measure for the away team increases the probability of a draw. The magnitude of the effects for each outcome are consistent with the idea that the home team has an advantage, since the marginal effects indicate that marginal increases in a home team's ability greatly decrease the probability of a home loss, and even a draw.

Table 1.11: Ordered probit model accounting for team ability

	(1)	(2)	(3)
VARIABLES	P(Result = 0)	P(Result = 1)	P(Result = 2)
Base Prob.	0.2561	0.2769	0.4671
HGD	-0.0872***	-0.0176***	0.105***
	(0.00543)	(0.00146)	(0.00631)
AGD	0.0893***	0.0180***	-0.107***
	(0.00522)	(0.00151)	(0.00613)
Observations	4,560	4,560	4,560

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 1.12 shows an extended model which includes the variables for the home team and away team's match significance. Significance is determined by a number of criteria. First, I allow matches to be significant for a team only if they have played at least half of their games already in a given season. I then calculated the team's points per game rate prior to each match, and used this, along with the remaining number of matches, to project a final point total at the end of the season. This point total was compared to thresholds which represent the average number of points needed to finish 1st (league title), 4th (threshold for European Champions League competition) and 17th (avoiding relegation). (see Grayson 2014) If a team's projected point total fell within 6 points of these threshold values, the match was deemed significant with respect to this team. The regression shows that the effects of goal differential are largely unchanged in this model. The home team's significance variable has the expected effect of adding to the probability of their victory, but the effect of a significant match for the home team fails to be significantly different from zero. This could be interpreted to mean that the home team is given an added boost in significant matches by playing in front of an especially large and enthusiastic crowd.

Table 1.12: Ordered probit model accounting for team ability and match significance

	(1)	(2)	(3)
VARIABLES	P(Result = 0)	P(Result = 1)	P(Result = 2)
Base Prob.	0.2560	0.2770	0.4671
HGD	-0.0865***	-0.0174***	0.104***
	(0.00544)	(0.00146)	(0.00632)
AGD	0.0886***	0.0179***	-0.106***
	(0.00523)	(0.00151)	(0.00615)
Hsig	-0.0274**	-0.00606*	0.0334**
	(0.0133)	(0.00323)	(0.0165)
Asig	0.0174	0.00329	-0.0207
	(0.0139)	(0.00248)	(0.0164)
Observations	4,560	4,560	4,560

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

An additional, expanded regression model is shown in Table 1.13 which includes variables to account for whether a team has been promoted to the premier league for that season. The marginal effects on the non-draw outcomes are relatively symmetric, with opposite magnitudes. It seems that overall, being a promoted team has a detrimental effect on a team's probability of winning.

In the final regression I included variables for recent team performance, and weather effects, specifically the dummy variables for rain and snow. The results of this regression are listed in Table 1.14. This regression indicates that none of recent performance or either of the weather effects have a significant effect on match outcome. Permutations of these three variables were included in separate regressions in case of collinearity, but all effects were still insignificant.

In order to test market efficiency, it is necessary to examine whether the factors included in the above models are fully accounted for by betting houses. I test this by including the probabilities of the home team winning and the away team winning, calculated using the odds given on these events. For this test I used the three models from Table 1.11, Table 1.12 and Table 1.13. The result of re-estimating all three models with these probabilities included is that the away goal differential variable becomes the only effect which is significant. Table 1.15 shows the results of including

Table 1.13: Ordered probit model accounting for team ability, match significance and promoted team influence

	(1)	(2)	(3)
VARIABLES	P(Result = 0)	P(Result = 1)	P(Result = 2)
Base Prob.	0.2458	0.2752	.4790
HGD	-0.0818***	-0.0165***	0.0982***
	(0.00568)	(0.00146)	(0.00663)
AGD	0.0842***	0.0170***	-0.101***
	(0.00545)	(0.00151)	(0.00644)
Hsig	-0.0271**	-0.00599*	0.0331**
	(0.0133)	(0.00322)	(0.0165)
Asig	0.0173	0.00329	-0.0206
	(0.0139)	(0.00248)	(0.0163)
Hprom	0.0428***	0.00712***	-0.0499***
	(0.0163)	(0.00223)	(0.0184)
Aprom	-0.0443***	-0.0108**	0.0551***
	(0.0149)	(0.00433)	(0.0192)
Observations	4,560	4,560	4,560

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 1.14: Expanded model including weather and recent performance variables

VARIABLES	(1) P(Result =0)	(2) P(Result =1)	(3) P(Result =2)
Base Prob.	0.2562	0.2793	..4645
HGD	-0.0869***	-0.0174***	0.104***
	(0.00714)	(0.00174)	(0.00837)
AGD	0.0907***	0.0181***	-0.109***
	(0.00564)	(0.00163)	(0.00663)
Hsig	-0.0256*	-0.00512*	0.0307*
	(0.0138)	(0.00277)	(0.0165)
Asig	0.0130	0.00261	-0.0156
	(0.0137)	(0.00275)	(0.0165)
Hprom	0.0369**	0.00738**	-0.0443**
	(0.0164)	(0.00332)	(0.0196)
Aprom	-0.0243	-0.00485	0.0291
	(0.0167)	(0.00334)	(0.0200)
RecP	0.000247	4.93e-05	-0.000296
	(0.00434)	(0.000867)	(0.00520)
Rain	0.00371	0.000743	-0.00446
	(0.0113)	(0.00226)	(0.0136)
Snow	-0.0236	-0.00473	0.0284
	(0.0262)	(0.00524)	(0.0314)
Observations	4,058	4,058	4,058

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

only away goal differential with the implied odds in a regression. This result indicates

Table 1.15: Odds included Model

VARIABLES	(1) Pr(Result = 0)	(2) Pr(Result = 1)	(3) Pr(Result = 2)
Base Prob	0.3640	0.2994	.3365
AGD	-0.0120*	-0.00250*	0.0145*
	(0.00703)	(0.00146)	(0.00849)
Implied Prob. (Away Win)	0.436**	0.0904**	-0.526**
	(0.198)	(0.0432)	(0.240)
Implied Prob. (Home Win)	-0.570***	-0.118***	0.688***
	(0.174)	(0.0341)	(0.207)
Observations	4,477	4,477	4,477

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

other variables are all properly captured by the odds. However, in this case, the effect of away goal differential is in the opposite direction. A higher goal differential in this case corresponds to a decrease in the probability that the away team wins, while increasing the probability that the home team wins. This can be interpreted as an indication that the odds over-adjust for quality away teams. However, this is a small inefficiency that may not be enough to make a significant profit if used in betting over the course of a season. In order to conclude that the EPL betting market was inefficient, I would need further research showing that a hypothetical betting scenario employing these models could systematically profit.

Chapter 2

Application of Betting Strategy

2.1 Review of Strategies

In addition to the statistical tests of efficiency for the NFL market provided by the modeling of bet outcomes in the first chapter, I will establish tests of economic efficiency in this betting market by devising betting strategies and assessing their success when applied to the predictive models I have developed. When determining the optimal amount to bet on a given game's outcome it is meaningful to the results if other bets are being considered alongside the bet in question. The probability of success in one game can, in some cases, affect the optimal amount to wager in another. The Kelly criterion is a betting strategy that has been proven optimal in terms of maximizing the growth rate of money. Kelly (1956) outlines this model in his paper, and Whitrow (2007) outlines a statistical test of Kelly's staking method. I will describe a formulation of Kelly staking to lay out the method I will ultimately implement in my testing. Maximizing the growth rate leads at times to betting large fractions of the bettor's bankroll, which may not be a realistic strategy for casual bettors or those who hold some negative utility with taking risk. However, this strategy can be modified to bet a fixed proportion of the Kelly-calculated stake, reducing the risk. I consider a few cases of the Kelly strategy that apply to situations with varying possible bets and outcomes.

2.1.1 Single Bet, Two Outcomes

In his review of Kelly staking, Whitrow outlines the case where the bettor is faced with a bet on a single event with two outcomes. The probability of success is denoted by, p, and odds for the bet are given by r. The value of r is the same as the would-be

decimal odds for the event (in his original paper, Kelly considers the case where $r = 2$, i.e., even odds). Denoting V_0 as the initial bankroll, and V_n as the gambler's money after n rounds of betting, Whitrow assumes logarithmic utility and defines the utility after the nth round as :

$$\begin{aligned} Q_n &= log\left(\frac{V_n}{V_0}\right) && (2.1) \\ &= log(V_n) - log(V_0) && (2.2) \end{aligned}$$

Whitrow suggests that in order to maximize $E(Q_n)$, the bettor should maximize $E(\delta Q_t)$, the change in the growth of the bankroll over a single bet, for each iteration, t, with respect to x, the fraction of the bankroll wagered. With $E(\delta Q_t) = E[log(V_t) - log(V_{t-1})]$, we obtain:

$$E(\delta Q_t) = p\ log(1 - x + rx) + (1 - p)\ log(1 - x) \tag{2.3}$$

Taking the derivative of (2.3) with respect to x, and setting it equal to zero in order to maximize, leads to an expression for the optimal stake based on p and r:

$$x = \frac{pr - 1}{r - 1} \tag{2.4}$$

In this case, p is estimated by our model, and r is given by the oddsmakers. I will refer to this as the Kelly stake for two-outcome betting. The bettor only gambles in this scenario when $pr > 1$. In NFL spread betting, for example, the transaction cost means that r is fixed at $\frac{21}{11}$ and so the bettor only wagers when $p > \frac{11}{21}$, with the fraction wagered depending on the difference between the estimated probability of an outcome and the probability implied by the odds. This staking method can be easily applied to the NFL model devised in Chapter 1 for any single matchup.

Considering a few examples of applying (2.4) to the NFL probit model developed earlier will illustrate some key points. As described above, the bettor only wagers when $p > \frac{11}{21}$. Suppose that the probit model predicts that the probability of the home team beating the spread is 0.55. In this case equation (2.4) states that 5.5% is the optimal portion of the bankroll to wager. In cases where $p < \frac{11}{21}$, (2.4) dictates that the bettor should wager a negative amount. We could simply require that the bettor wager 0 in these cases, but that would rule out the possibility of betting on the away team, since the model always predicts (which is to say that p always represents) the probability of only the home team beating the spread. Since the probability of the away team beating the spread is $1 - p$, and (2.4) suggests we bet when our estimated

probability of success is greater than $\frac{11}{21}$, the bettor should interpret a negative value for x as the fraction to wager on the away team only when $1-p > \frac{11}{21}$ Thus, if $p < \frac{10}{21}$ then a negative x should be interpreted as the fraction of the bettor's bankroll to bet on the away team, and we will require $x = 0$ for $\frac{10}{21} < p < \frac{11}{21}$. This may require the bettor to refrain from betting on some number of matches, depending on how frequently there is a discrepancy in the predicted and given odds of the size required to wager.

2.1.2 Many Bets, Two Outcomes

Consider now the case where the bettor is faced with the task of maximizing utility with respect to the fraction wagered on 2 independent, simultaneous NFL games. In this scenario, $E(\delta Q_t)$ becomes:

$$\begin{aligned} E(\delta Q_t) &= p_1 p_2 log(1 - x_1 + r x_1 - x_2 + r x_2) \\ &\quad + p_2(1-p_1) log(1 - x_1 - x_2 + r x_2) \\ &\quad + p_1(1-p_2) log(1 - x_2 - x_1 + r x_1) \\ &\quad + (1-p_1)(1-p_2)(1 - x_1 - x_2) \end{aligned} \tag{2.5}$$

Where p_i is the estimated probability of success for wagering on game i, x_i is the fraction wagered on game i, and r is the return as defined before (fixed for all NFL games). Let $q_i = 1 - p_i$ and let us reorganize terms to get:

$$\begin{aligned} E(\delta Q_t) &= p_1 p_2 log(1 + (r-1)x_1 + (r-1)x_2) \\ &\quad + p_2 q_1 log(1 - x_1 + (r-1)x_2) \\ &\quad + p_1 q_2 log(1 - x_2(r-1)x_1) + q_1 q_2(1 - x_1 - x_2) \end{aligned} \tag{2.6}$$

To maximize with respect to a given betting fraction, we take the partial derivative of $E(\delta Q_t)$ with respect to x_1. Taking the derivative of even the first term of the equation above shows that $\frac{\partial E(\delta Q_t)}{x_i}$ will have an x_j term where $i \neq j$ (these terms do not cancel). Since there are x_2 terms in the x_1 derivative and vice versa, the fraction bet on a particular event is impacted by simultaneous independent events. Thus the stake found to be optimal when x_1 was considered on its own is no longer optimal when presented with an alternative, simultaneous wager. This means we will have to devise a system for optimizing betting on many independent games at once if the probit model developed above is to be applied over multiple games in a week of an

NFL season. The equations resulting from each partial derivative of (2.6) reduce to a system of equations which can be solved to determine x_1 and x_2. However, the system quickly gets complicated as the number of games increases. Whitrow outlines a form of optimization for allocating stakes over any number of events which uses gradient ascent to maximize the log utility with respect to each stake. He applies it to two data sets and compares it to simply betting the Kelly stake as if the event was isolated, and determines that it produces optimal results. Implementing a version of this non-linear optimization is grounds for further research. For my purposes, I treat each game as if it were being considered alone, and use the Kelly stake outlined in equation (2.4)

2.2 Results of Strategy Application

In Chapter 1, I devised three models which were used to estimate the probability of the home team beating the spread in any given NFL matchup. In order to test the efficiency of the spread betting market using these models, we must establish criteria for efficiency. First, we can test efficiency in a strict sense: if any model correctly predicts enough matches to be profitable, then we can reject the hypothesis that the betting market is efficient. Given the eleven-for-ten governing NFL transaction costs, a model must correctly predict 52.38% ($\frac{11}{10+11}$) of matches in which it wagers in order to make a profit. Any model which yields this proportion of correct predictions would result in the conclusion that the market is inefficient. Efficiency can also be tested by considering gambling from a slightly different perspective. Simply breaking-even (the threshold for inefficiency described above) may be an unrealistic standard for describing the market as inefficient. Many versions of market efficiency state that gambling returns must yield unusual returns relative to alternative investments in order for the market to truly be considered inefficient (see Williams 1999). For instance, if gambling consistently yielded a 1% return, the market would be inefficient in the stricter sense that it was making a profit, but opening a savings account with the gambling bankroll would yield a higher return, with (presumably) lower risk, meaning a rational actor would never choose gambling as an investment strategy. Viewed in this way, the predictive models from Chapter 1 are held to a higher standard in that their rate of return on investment must be much higher. I will test my models under both these criteria, by simulating betting over the 2012 NFL season and looking at rates of return.

For my test of efficiency, I will use the three final NFL models from Chapter 1,

contained in Table 1.9 (from here, referred to as Models 1, 2, and 3). In estimating these models, the 2012 NFL season's data were excluded in order to maintain a betting simulation which was not influenced by attempting to predict outcomes that were already included in estimation. I first generated predicted probabilities of home team success in each matchup based on each model. Denoting the estimated probability of home team success as $\hat{p}_i$ for a given matchup i. I then determined x_i, the percentage of the bankroll to be staked on matchup i, using the Kelly formula derived in equation (2.4). The stake was reassigned to zero in the case that $\frac{10}{21} < \hat{p}_i < \frac{11}{21}$. The bankroll was initially set to \$10,000. For the purposes of comparison, I also generated a random model, that bet \$110 on a random team in each matchup over the course of a season, starting with the same \$10,000.

In testing the models, I considered games occurring on the same day to be happening simultaneously, and updated the bankroll after each day on which betting occurred. I then calculated the simulated bets placed by each model based on the proportional stake and the current bankroll. This bet was compared to the actual result, and returns were added to the bankroll.

The final bankrolls and return rates of each model are shown in Table 2.1

Table 2.1: Betting Simulation Results

Model	Final Bankroll	Return Rate
Random Model	\$ 9900	-1%
Model 1	\$7295.64	-27.04%
Model 2	\$5979.50	-40.21%
Model 3	\$14299.53	43.00%

Though the results are only for a single season, it is clear that the Kelly staking method has some extremely varied outcomes between models. The graph in Figure 2.1 shows the bankroll of each model for each betting week. As the chart illustrates, staking using the full Kelly proportion is highly volatile. Both Model 2 and Model 3 were above \$40,000 in total bankroll before losing substantial amounts in the final weeks of the season. Given that there is only one season of data to test for efficiency, this test is subject to the quirks of this particular season's outcomes, and it remains to be seen how each model would perform if tested over a number of seasons. I did, however, test different betting strategies based on partial Kelly stakes, in order to limit the size of bets being made on single games in an effort to reduce the huge gains and

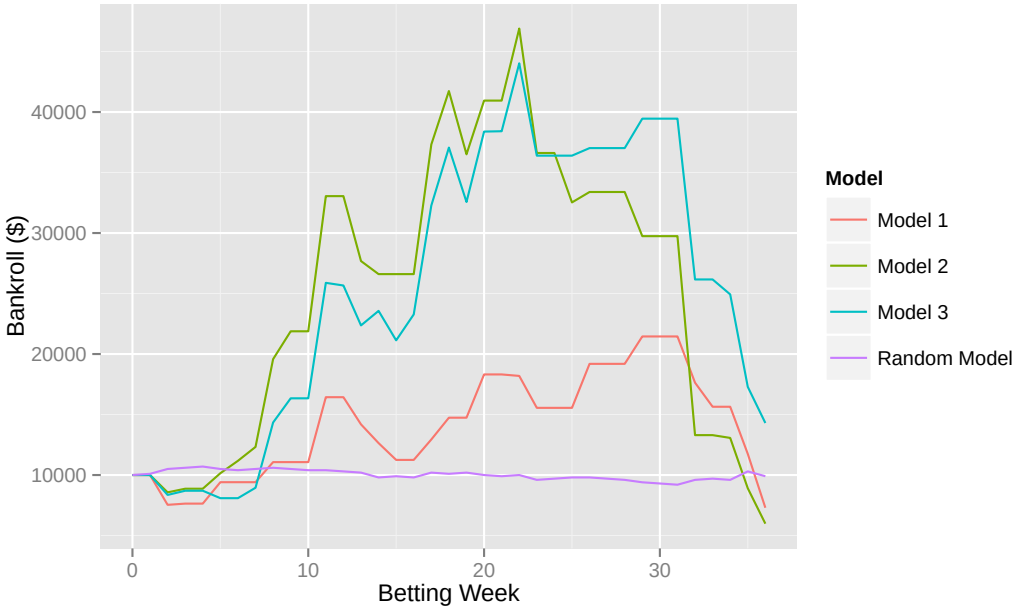

Figure 2.1: Bankroll by Betting Week for the 2012 NFL Season

losses of the original models. I implemented a test of all three models based on staking three-quarters of the proportion of the bankroll suggested by the Kelly strategy, and another test based on staking half of the optimal proportion as calculated using the Kelly criterion. The results are shown is Table 2.2. These strategies vary a great deal less in terms of week-to-week bankroll. A graphic comparison is displayed in Figure 2.2.The half Kelly staking performs best in the sense that it nets a good return with Model 3, and limits losses by a great deal for the other two models. Based on one season, Model 3 would seem to confirm a tentative rejection of market efficiency, with some hesitance due to the extreme variation in results.

In order to have a more robust test of these models, I also gathered data for the 2013 NFL season and applied the same betting model. The results are shown in Table 2.3 and Figure 2.3. The results do little to confirm any of the results from the 2012 NFL season. Again, the gains and losses under full Kelly staking are erratic, and though using partial versions of Kelly staking reduce variability of outcomes, the gains from the best model shrink again. In this season it is Model 1 which performs the best, which casts further doubt on the reliability of the results from the previous season. Viewed together, these two seasons lead to the conclusion that the betting models derived from the inefficiencies found in the first chapter are unreliable at best.

I also implemented a betting system which I call the conservative system, since it re-scales bets so that only $10000 of the bankroll is being wagered at most. Thus,

Table 2.2: Betting Simulation Results (2012)

Strategy	Model	Final Bankroll	Return Rate
Full Kelly Staking	Random Model	$ 9900	-1%
	Model 1	$7295.64	-27.04%
	Model 2	$5979.50	-40.21%
	Model 3	$14299.53	43.00%
Three-Quarters Kelly Staking	Random Model	$ 9900	-1%
	Model 1	$8582.90	-14.17%
	Model 2	$8033.06	-19.67%
	Model 3	$14489.39	44.89%
Half Kelly Staking	Random Model	$ 9900	-1%
	Model 1	$9531.63	-4.68%
	Model 2	$9550.68	-4.49%
	Model 3	$13732.86	37.33%

when the bankroll grows above $10000, any profits are set aside, and only the original $10000 is used for betting until the bankroll goes below the original amount. I tested this betting system on both seasons. Table 2.4 contains the results of these simulations. Comparing these results to the results of the ordinary and partial Kelly staking for both seasons shows that as expected, this conservative method seems to constrain outcomes a bit. Models 1 and 2 are definitely boosted using this method in the 2012 season. The effect of the conservative criterion of re-scaling the bets is reduced for partial staking methods, since the partial stakes suggest amounts already similar to the rescaled bets of the conservative system.

In order to evaluate these models further, I also tested the conservative models, at all three levels of Kelly staking, on the NFL seasons which were initially used to estimate the models (1978-2011 seasons). The models will naturally perform slightly better over these data, since they were fit to the data. With this caveat in mind, testing the models on these seasons can give a much better estimation of performance, since it increases the sample size for the results of the three models from two seasons to 36. I performed this in-sample betting test for each of 34 seasons used for estimating the models, at each level of staking (half, three-quarters, and full) for the conservative model. The results are shown in Table 2.5. The Half Kelly conservative model is preferred on the basis of minimizing the standard error of the final bankrolls, while still keeping the median outcome for each model above $10,000. The results show that the median outcome is highest for Model 3, followed by Model 2, then by Model

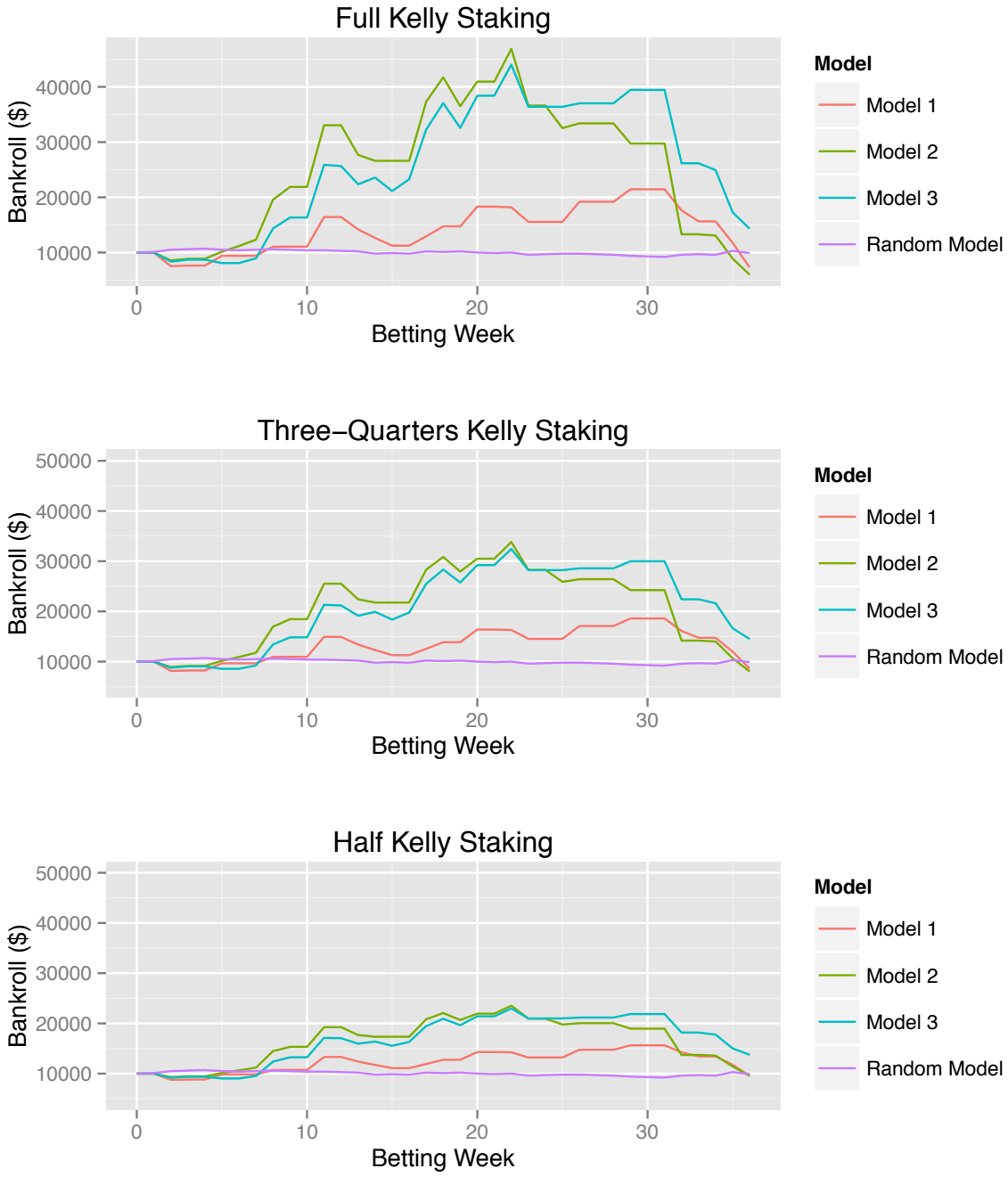

Figure 2.2: Comparison of Betting Strategies over 2012 NFL Season

1. This order is preserved for the Three-Quarters and Full Kelly staking versions of the in-sample test. Figure 2.4 shows the spread of results over the entire sample, with the starting bankroll marked in black and the median of each model in blue. Despite having the lowest standard error of final bankrolls, the outcomes are still

Table 2.3: Betting Simulation Results (2013)

Strategy	Model	Final Bankroll	Return Rate
Full Kelly Staking	Random Model	$ 10440	4.40%
	Model 1	$20290.37	102.90%
	Model 2	$4045.43	-59.55%
	Model 3	$6492.98	-35.07%
Three-Quarters Kelly Staking	Random Model	$ 10440	4.40%
	Model 1	$18308.49	83.31%
	Model 2	$ 6100.67	-38.99%
	Model 3	$7980.81	-20.19%
Half Kelly Staking	Random Model	$ 10440	4.40%
	Model 1	$15717.03	57.17%
	Model 2	$7918.97	20.81%
	Model 3	$9144.94	-8.55%

hugely varied.

I also evaluated the same betting systems on the full EPL season spanning 2012-2013, and the (as of this writing) mostly complete 2013-2014 season. The full Kelly staking method and three-quarters method both resulted in the entire bankroll being lost in both seasons. The results of the half Kelly method for the 2012-13 season are shown in Figure 2.5, while the 2013-14 season's results are shown in Figure 2.6. The poor performance of these models seems to indicate that the models built in Chapter 1 are inadequate for use in betting. I attempted to use the conservative model, as employed in the NFL case, but it had little effect other than prolonging the eventual bottoming out of the bankroll.

In order to gain further information on the performance of these models, I implemented the in-sample test in a fashion similar to that of the NFL, again using the conservative method, which re-scales bets to total $10,000 if they would otherwise total greater than this amount. The results are shown in Table 2.6 and Figure 2.7. Model 2 is has the highest median outcome across all staking proportions, while Half Kelly staking is the best option for staking proportions. The results do little to offer any more support for any inefficiency than the first two betting tests. Though the median outcomes for the Half Kelly conservative staking are much higher than those of previously tested samples and betting methods, the variability of outcomes lends little credibility to this result. Market inefficiencies for the EPL are thus unable to be exploited using the methodology of the models from Chapter 1.

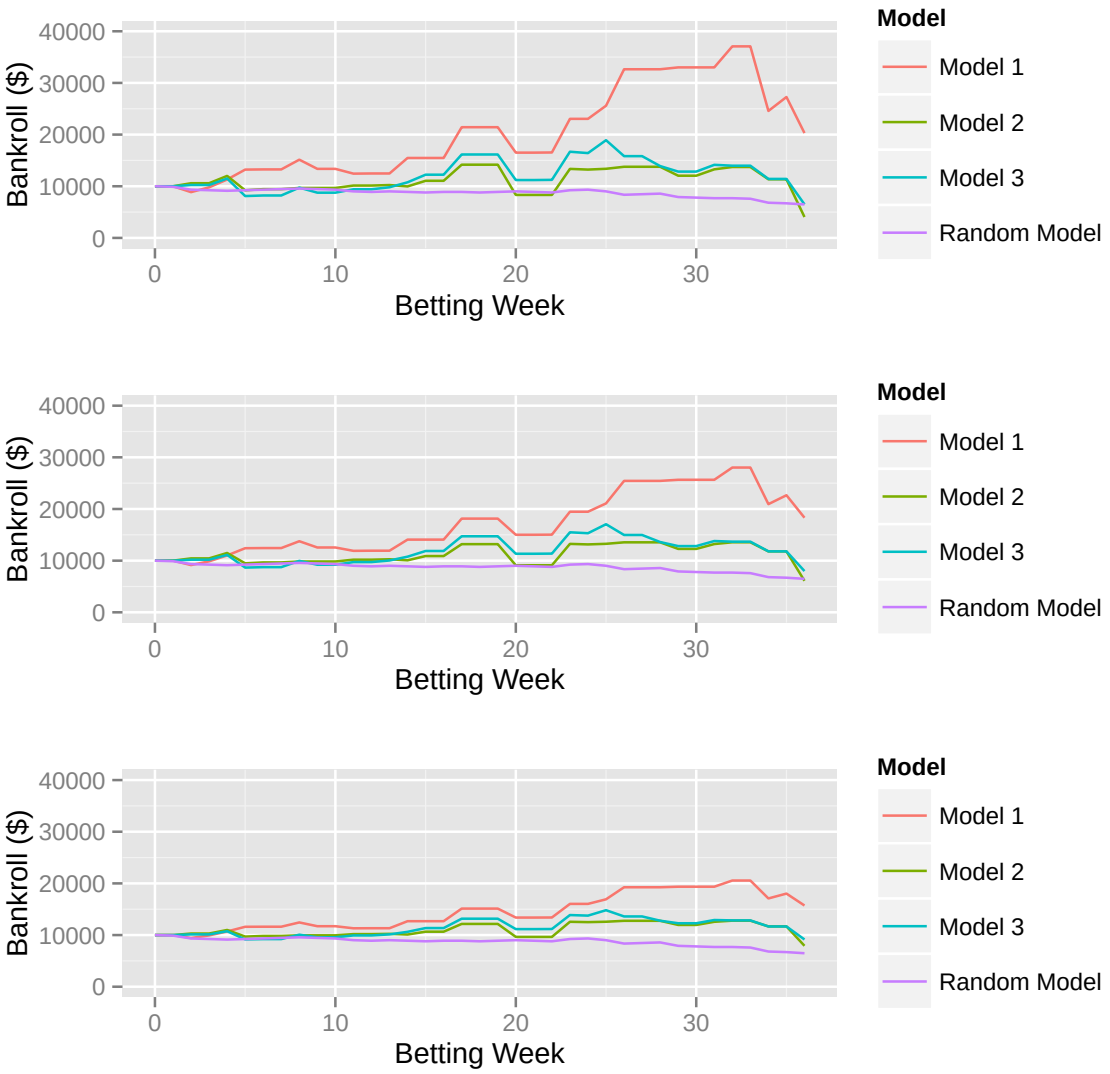

Figure 2.3: Comparison of Betting Strategies over 2013 NFL Season

Table 2.4: Conservative Betting Simulation Results

Strategy	Model	Final Bankroll	Return Rate
2012 Season			
Full Kelly Staking	Random Model	$ 10440	4.40%
	Model 1	$7247.26	-27.53%
	Model 2	$12963.99	29.64%
	Model 3	$24736.38	147.36%
Three-Quarters Kelly Staking	Random Model	$ 10440	4.40%
	Model 1	$8587.69	-14.12%
	Model 2	$ 9828.76	-1.71%
	Model 3	$16695.42	66.95%
Half Kelly Staking	Random Model	$ 10440	4.40%
	Model 1	$9534.49	-4.66%
	Model 2	$9547.44	-4.52%
	Model 3	$13940.28	39.40%
2013 Season			
Full Kelly Staking	Random Model	$ 10440	4.40%
	Model 1	$27556.84	175.57%
	Model 2	$ 6138.69	-38.61%
	Model 3	$9461.99	-5.39%
Three-Quarters Kelly Staking	Random Model	$ 10440	4.40%
	Model 1	$20612.54	106.12%
	Model 2	$ 6664.59	-33.35%
	Model 3	$9139.67	-8.60%
Half Kelly Staking	Random Model	$ 10440	4.40%
	Model 1	$15717.03	57.17%
	Model 2	$7918.97	20.81%
	Model 3	$9144.94	-8.55%

Table 2.5: In-Sample Results for Conservative Staking Methods (NFL)

Strategy	Model	Median Final Bankroll	Median Return Rate
Full Kelly Staking	Model 1	$11428.28	14.28%
	Model 2	$14403.01	44.03%
	Model 3	$15054.76	50.55%
Three-Quarters Kelly Staking	Model 1	$ 11301.04	13.01%
	Model 2	$ 13077.98	30.78%
	Model 3	$13703.21	37.03%
Half Kelly Staking	Model 1	$11545.07	15.45%
	Model 2	$12129.34	21.29%
	Model 3	$12911.98	29.11%

Table 2.6: In-Sample Results for Conservative Staking Methods (EPL)

Strategy	Model	Median Final Bankroll	Median Return Rate
Full Kelly Staking	Model 1	$1324.98	-86.75%
	Model 2	$2789.63	-72.10%
	Model 3	$2192.34	-78.08%
Three-Quarters Kelly Staking	Model 1	$ 4533.55	13.01%
	Model 2	$ 6983.16	-30.17%
	Model 3	$ 4614.87	-53.85%
Half Kelly Staking	Model 1	$10345.61	-3.46%
	Model 2	$13729.81	37.73%
	Model 3	$8868.271	-11.32%

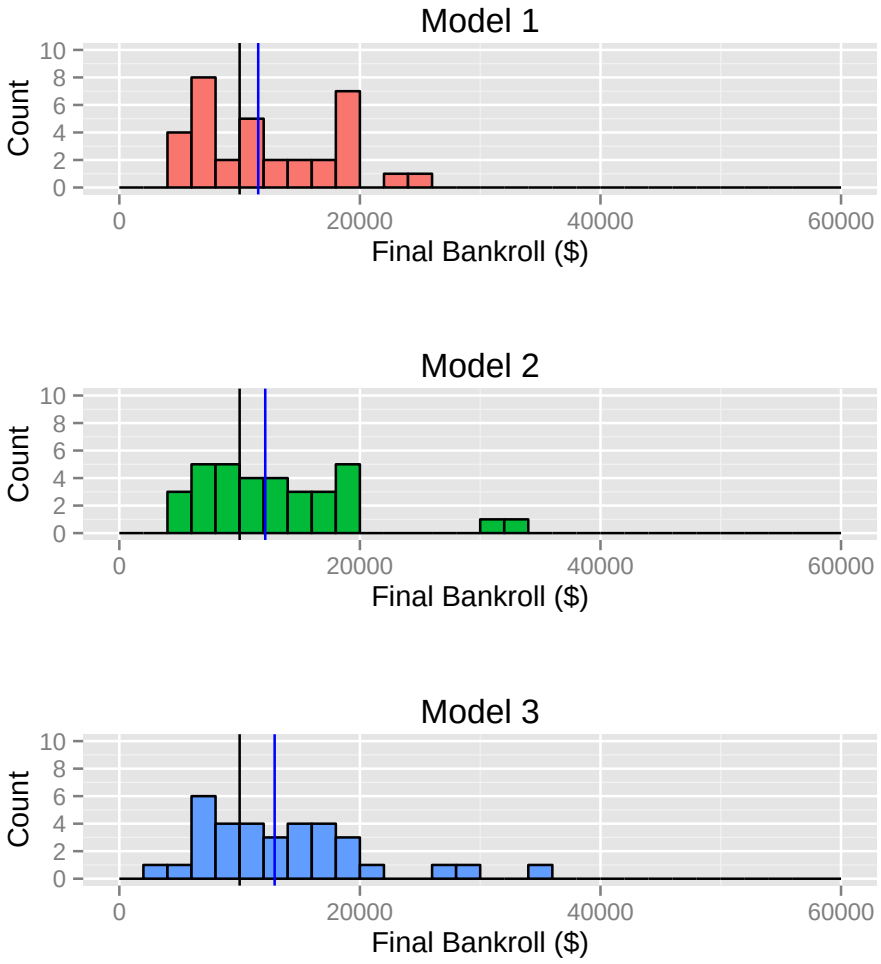

Figure 2.4: Outcomes of Half Kelly conservative staking method over entire NFL sample

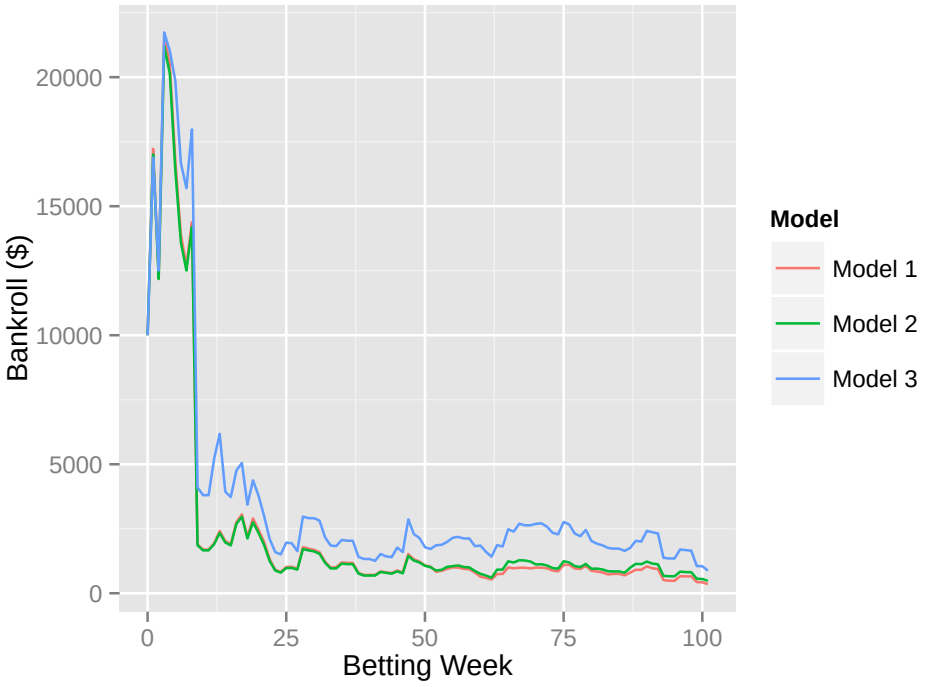

Figure 2.5: Half Kelly staking method over 2012-2013 EPL Season

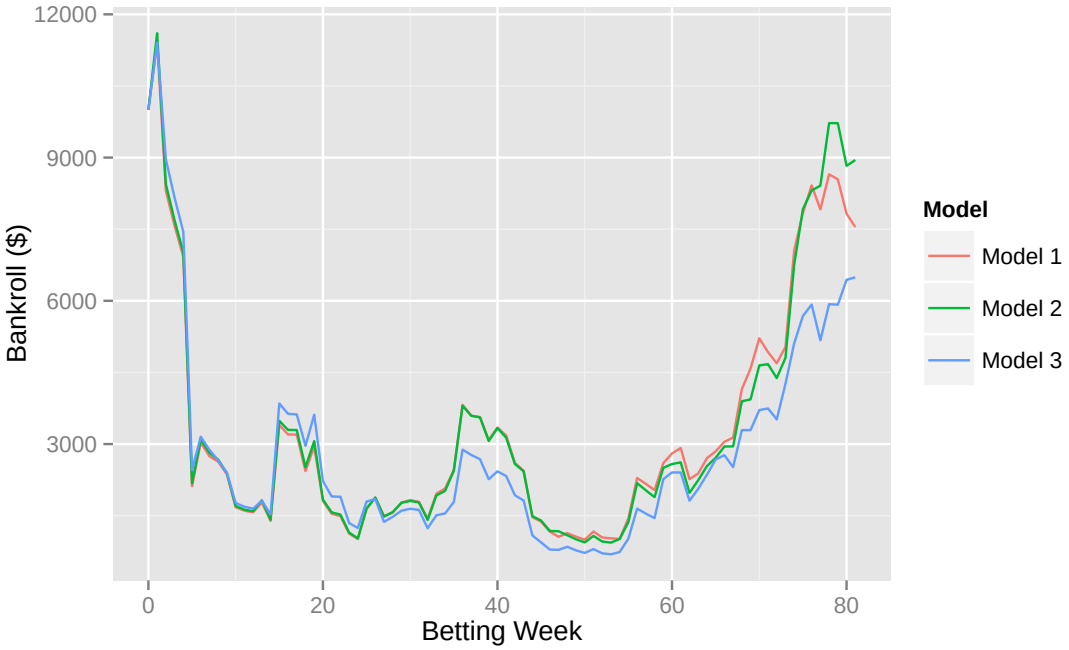

Figure 2.6: Outcomes of Half Kelly staking method over 2013-2014 EPL Season

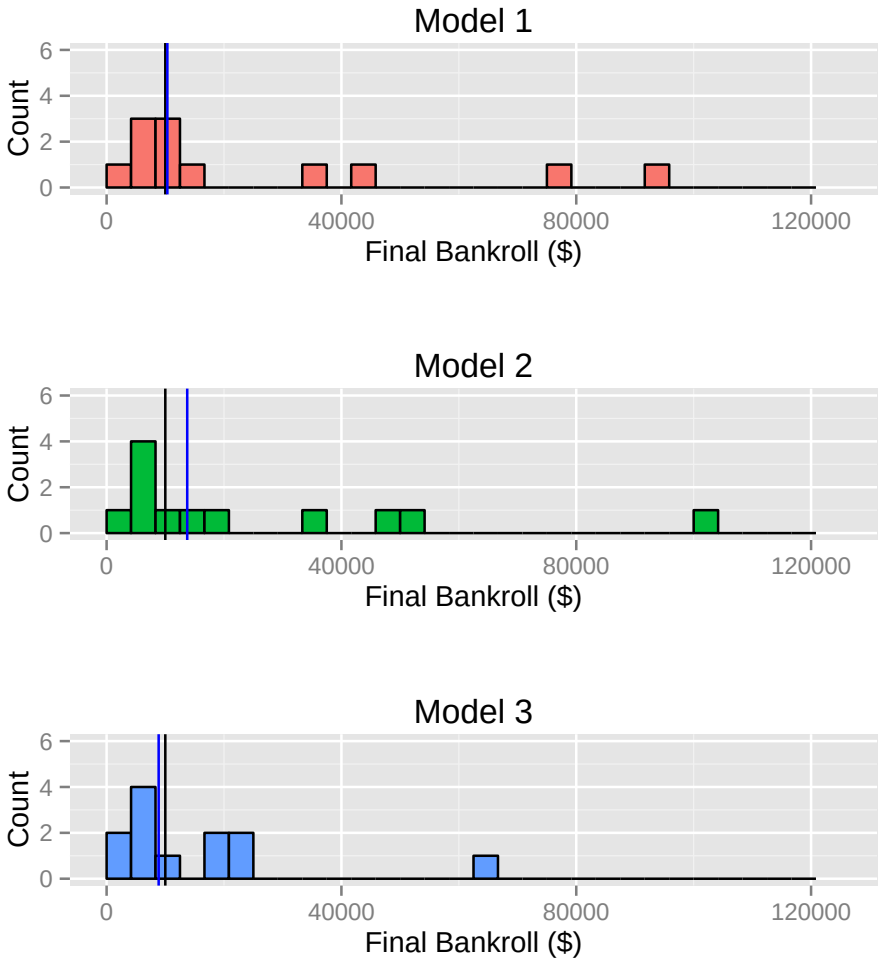

Figure 2.7: Outcomes of Half Kelly conservative staking method over entire EPL sample

Conclusion

My results for the NFL data support a rejection of market efficiency in the strictest sense. That is, the regression results illustrate the fact that information is not perfectly incorporated into setting betting spreads. Whether the inefficiency can be reliably exploited is another issue, and one that my work offers weak support for. The results of testing my predictive betting models on two NFL seasons resulted in a large amount of variation in outcomes. I further tested the models by evaluating their performance over the data from which the models were estimated. The median return was positive in all cases. Despite this, the variation in outcomes, in addition to some caution merited by the potential bias of the in-sample test leads me to conclude that further evidence is required to confirm a consistent, profitable inefficiency in the NFL betting market. In the case of the European football betting market, there is no evidence to suggest a market inefficiency, as the models which I used to test betting outcomes yielded only negative returns in testing on most of two seasons of data on which the models were not estimated. The results of a in-sample test similar to the NFL were widely varied, and generally resulted in a loss. Thus this test added no evidence that the EPL betting market contained any inefficiency. Given that I have not exhausted all possible models for estimating bet outcomes, there are potential effects that are logical to explore in further research which attempts to model outcomes. For instance, specific teams may always have an effect on the betting line due to better behavior surrounding these teams. Another example of an effect to explore in future research is the efficiency of odds in scenarios where the odds are longer, and thus the payoff is large. This applies to fixed-odds betting only. If bettors derive utility from the excitement of these long-shot bets, then they will tend to bet on them more often than they should, thus influencing the odds. Testing the models devised in this thesis on future games is a potential strategy for future research, but this suffers from having to wait for games to occur. Another potential solution is modeling the distributions of variables used to characterize NFL or EPL match ups, and using these estimated distributions to generate typical match ups and outcomes which could be used to test

the models further. A third method that could be used is to subset a larger portion of the data for betting tests, estimating the models on smaller samples. This could be extended to a rolling regression, wherein the model is evaluated under varying sizes of the sample used to estimate the model and thus different sizes of samples reserved for betting tests. The betting models for both betting markets could also benefit from fully exploring a method in the style of Whitrow, as described in Chapter 2, of optimizing fully over multiple simultaneous bets.

Appendix A

Python Scripts

In order to gather weather and distance data, I composed two separate Python files which called on web API resources. Lines preceded by the hash (#) symbol represent comments, which explain the operation of subsequent lines.

A.1 Weather Data

The python script below uses the Weather Underground API to access historical weather data for NFL games. The program defines uses Python's dictionary data structure to key team names with their stadium locations. In some cases, stadiums were in suburbs of metropolitan areas for which the online API did not have consistent historical data. In these cases, the nearby metropolitan area was used as the location associated with that team. Using a CSV (comma separated values) file containing pairs of NFL dates and home teams, the Python program looks up location by the name of the home team and uses the retrieved location to format a web URL used to access historical data. The program then writes the collected weather data to a separate CSV file. My code makes use of the json, urllib2, and cdv modules, downloaded via the Python Package Index (PyPI). For EPL data collection, the only difference was defining a Python dictionary which contained the names and corresponding locations of Premier League teams.

```
# load required modules
import json
import urllib2
import csv
```

```
# define dictionary for looking up team locations
dict = {"Washington Redskins":"Washington, DC",
"New York Jets":"East Rutherford, NJ",
"New York Giants":"East Rutherford, NJ",
"Green Bay Packers":"Green Bay, WI",
"Dallas Cowboys":"Dallas, TX",
"Kansas City Chiefs":"Kansas City, MO",
"Denver Broncos":"Denver, CO",
"Miami Dolphins":"Miami Gardens, FL",
"Carolina Panthers":"Charlotte, NC",
"New Orleans Saints":"New Orleans, LA",
"Cleveland Browns":"Cleveland, OH",
"Buffalo Bills":"Orchard Park, NY",
"Atlanta Falcons":"Atlanta, GA",
"Houston Texans":"Houston, TX",
"Baltimore Ravens":"Baltimore, MD",
"San Diego Chargers":"San Diego, CA",
"Tennessee Titans":"Nashville, TN",
"New England Patriots":"Foxboro, MA",
"Philadelphia Eagles":"Philadelphia, PA",
"San Francisco 49ers":"Santa Clara, CA",
"Jacksonville Jaguars":"Jacksonville, FL",
"Seattle Seahawks":"Seattle, WA",
"St Louis Rams":"St. Louis, MO",
"Tampa Bay Buccaneers":"Tampa, FL",
"Cincinnati Bengals":"Cincinnati, OH",
"Pittsburgh Steelers":"Pittsburgh, PA",
"Detroit Lions":"Detroit, MI",
"Arizona Cardinals":"Glendale, AZ",
"Indianapolis Colts":"Indianapolis, IN",
"Chicago Bears":"Chicago, IL",
"Oakland Raiders":"Oakland, CA",
"Minnesota Vikings":"Minneapolis, MN",
"Baltimore Colts": "Baltimore, MD",
"Tennessee Oilers": "Nashville, TN",
"Houston Oilers": "Houston, TX",
```

```
"Los Angeles Rams": "Los Angeles, CA",
"Los Angeles Raiders": "Los Angeles, CA",
"St Louis Cardinals": "St. Louis, MO",
"Phoenix Cardinals": "Tempe, AZ"}

# define dformat function which converts
# dates from format used in CSV data
# to format used for URL in weather
# API
def dformat(date):
    ind_one = date.index("/")
    mon = date[:ind_one]
    date = date[(ind_one+1):]
    if len(mon) != 2:
        mon = "0" + mon
    ind_two = date.index("/")
    day = date[:ind_two]
    date = date[(ind_two+1):]
    if len(day) != 2:
        day = "0" + day
    if int(date) < 50:
        date = "20" + date
    else:
        date = "19" + date
    return date + mon + day

# define lformat function which converts
# locations from format received from dictionary
# to format used for URL in weather
# API
def lformat(team):
    place = dict[team]
    ind_one = place.index(",")
    state = place[ind_one+2:]
    city = place[:ind_one]
    if city.find(" ") != -1:
```

```
        city = city.replace(" ", "_")
    return state + "/" + city
 # defines apical function used to generate properly
 # formatted URLs to retrieve data
 #
 # The section of the url1 variable
 # containing the string "APIKEY"
 # should be replaced with the unique
 # key associated with Weather Underground
 # API users
def apicall(team, date):
    url1 = "http://api.wunderground.com/api/APIKEY/history_"
    url2 = "/q/" + lformat(team) + ".json"
    day = dformat(date)
    return url1 + day + url2

# main operation

weather = []
# reads input from csv file
with open('nfl_in.csv', 'rU') as f:
    reader = csv.reader(f)
    for row in reader:
        datea = row[0]
        teama = row[1]
        # print home team to unix terminal for error tracking
        print teama
        urlcall = apicall(teama,datea)
        # json and urllib2 modules must be used
        # to parse data returned in JSON format
        j = urllib2.urlopen(urlcall)
        js= json.load(j)
        ds = js["history"]["dailysummary"]
        # Error handling for cases where the API
        # is missing data
        if not ds:
```

```
            rain = " "
            snow = " "
            mmprecip = " "
            mmsnow= " "
        else:
        rain = js["history"]["dailysummary"][0]["rain"]
            snow = js["history"]["dailysummary"][0]["snow"]
            mmprecip = js["history"]["dailysummary"][0]["precipm"]
            mmsnow = js["history"]["dailysummary"][0]["snowfallm"]
        weather = (row + [str(rain),str(snow),str(mmprecip), str(mmsnow)])
        # write data out to csv file, attaching to the end of the file
        with open('nfl_out.csv', 'ab') as csvfile:
            spamwriter = csv.writer(csvfile, delimiter=',', quotechar='|',
            quoting=csv.QUOTE_MINIMAL)
            spamwriter.writerow(weather)
```

A.2 Distance data

This python script follows a similar method to the weather data collection, taking away and home teams as inputs from a CSV file, calculating distance between locations, and writing out to a new CSV file. The GeoPy package is used to calculate distances by accessing the Google Maps API. This package is also available on PyPI.

```
# load required modules
import csv
import time
from geopy.distance import vincenty
from geopy.geocoders import GoogleV3

# create dictionary for location lookup
dict = {"Washington Redskins":"Landover, MD",
"New York Jets":"East Rutherford, NJ",
"New York Giants":"East Rutherford, NJ",
"Green Bay Packers":"Green Bay, WI",
"Dallas Cowboys":"Arlington, TX",
"Kansas City Chiefs":"Kansas City, MO",
```

```
"Denver Broncos":"Denver, CO",
"Miami Dolphins":"Miami Gardens, FL",
"Carolina Panthers":"Charlotte, NC",
"New Orleans Saints":"New Orleans, LA",
"Cleveland Browns":"Cleveland, OH",
"Buffalo Bills":"Orchard Park, NY",
"Atlanta Falcons":"Atlanta, GA",
"Houston Texans":"Houston, TX",
"Baltimore Ravens":"Baltimore, MD",
"San Diego Chargers":"San Diego, CA",
"Tennessee Titans":"Nashville, TN",
"New England Patriots":"Foxborough, MA",
"Philadelphia Eagles":"Philadelphia, PA",
"San Francisco 49ers":"Santa Clara, CA",
"Jacksonville Jaguars":"Jacksonville, FL",
"Seattle Seahawks":"Seattle, WA",
"St Louis Rams":"St. Louis, MO",
"Tampa Bay Buccaneers":"Tampa, FL",
"Cincinnati Bengals":"Cincinnati, OH",
"Pittsburgh Steelers":"Pittsburgh, PA",
"Detroit Lions":"Detroit, MI",
"Arizona Cardinals":"Glendale, AZ",
"Indianapolis Colts":"Indianapolis, IN",
"Chicago Bears":"Chicago, IL",
"Oakland Raiders":"Oakland, CA",
"Minnesota Vikings":"Minneapolis, MN",
"Baltimore Colts": "Baltimore,MD",
"Tennessee Oilers": "Nashville, TN",
"Houston Oilers": "Houston, TX",
"Los Angeles Rams": "Los Angeles, CA",
"Los Angeles Raiders": "Los Angeles,CA",
"St Louis Cardinals": "St. Louis, MO",
"Phoenix Cardinals": "Tempe, AZ"}

# define travel function which takes two teams as
# arguments, and looks up their locations, calculates
```

```
# latitude and longitude for each, and returns the distance
# between the two teams
geolocator = GoogleV3()
def travel(teamA, teamB):
    locA, (latA, longA) = geolocator.geocode(dict[teamA])
    locB, (latB, longB) = geolocator.geocode(dict[teamB])
    coordA = (latA, longA)
    coordB = (latB, longB)
    return vincenty(coordA, coordB).miles

#main operation
vtravel = []
# reads input from csv file
with open('nfl_in.csv', 'rU') as f:
    reader = csv.reader(f)
    for row in reader:
        #calculate distance between input teams
        teama = row[0]
        teamb = row[1]
        d = round(travel(teama, teamb),0)
        vtravel = (row + [str(d)])
         # write data out to csv file, attaching to the end of the file
        with open('nfl_out.csv', 'ab') as csvfile:
            spamwriter = csv.writer(csvfile, delimiter=','
            , quotechar='|', quoting=csv.QUOTE_MINIMAL)
            spamwriter.writerow(vtravel)
        # print home team to unix terminal for error tracking
        print teama
        # pause before next lookup to avoid overloading API
        time.sleep(0.5)
```

Appendix B

R Scripts

B.1 Betting tests

The following R script was used to perform the betting test of the 3 NFL probit models described in the second chapter on the 2012 season. Similar scripts are used for the EPL betting test, with modifications to account for the three possible outcomes and calculating returns under fixed odds betting. Lines preceded by the hash (#) symbol represent comments, which explain the operation of subsequent lines.

```
# in order to run this script, the data must be loaded as a data
# frame called nfl, with a "day" column formatted in the R date
# format, the game outcome, betting spread, and kelly stake
# calculated for each game based on each model's predicted
# probabilities

# Subset data frame to last season for betting test
nfl$y <- sapply(nfl$day, function(x) as.numeric(format(x, "%Y")))
yr1 <- subset(nfl, y == 2012)
yr1$m <- sapply(yr1$day, function(x) as.numeric(format(x, "%m")))
yr1 <- subset(yr1,  m > 1)
# store all different "betting weeks" in which games occur
wks <- unique(yr1$day)
# store number of betting weeks to set loop length
m <- length(wks)

# Initialize betting bankrolls for each
```

```
# model and bankroll vectors which
# track week-to-week bankroll levels
bankroll <- 10000
bankroll1 <- bankroll
bankroll2 <- bankroll
bankroll3 <- bankroll
bankroll.rand <- bankroll
bankroll1.v <- bankroll
bankroll2.v <- bankroll
bankroll3.v <- bankroll
bankroll.rand.v <- bankroll

# Loop through all betting weeks in a season,
# caculating bets for each game, and
for (k in 1:m) {
  gmwk <- subset(yr1, day == wks[k])
  l <- nrow(gmwk)
  # initialize variables for betting
  # return for each model for this
  # betting week
  rtrn1 <- 0
  rtrn2 <- 0
  rtrn3 <- 0
  rtrn.rand <- 0
  # loop over all games in the betting week,
  # calculating kelly bet in dollars and
  # corresponding return, based on outcome
  # bets are scaled by 0.75 or 0.50 for
  # half and three-quarters staking methods
  # for the conservative betting models,
  # bets are rescaled before calculating return
  # by summing amount bet over a week,
  # then multiplying by 10000 divided by that amount
  # if it is larger than 10000
  for (j in 1:l) {
    gmwk$bet.amt1[j] <- bankroll1 * gmwk$kstake1[j]
```

```
    gmwk$bet.amt2[j] <- bankroll2 * gmwk$kstake2[j]
    gmwk$bet.amt3[j] <- bankroll3 * gmwk$kstake3[j]
    if (gmwk$beat_home[j] == 0) {
      rtrn1 <- rtrn1 - gmwk$bet.amt1[j]
      rtrn2 <- rtrn2 - gmwk$bet.amt2[j]
      rtrn3 <- rtrn3 - gmwk$bet.amt3[j]
      rtrn.rand <- rtrn.rand - gmwk$bet.amt.rand[j]
    } else {
      rtrn1 <- rtrn1 + gmwk$bet.amt1[j]
      rtrn2 <- rtrn2 + gmwk$bet.amt2[j]
      rtrn3 <- rtrn3 + gmwk$bet.amt3[j]
      rtrn.rand <- rtrn.rand + gmwk$bet.amt.rand[j]
    }
  }
  # update bankroll quantities and vectors
  bankroll1 <- bankroll1 + rtrn1
  bankroll1.v <- append(bankroll1.v,bankroll1)
  bankroll2 <- bankroll2 + rtrn2
  bankroll2.v <- append(bankroll2.v,bankroll2)
  bankroll3 <- bankroll3 + rtrn3
  bankroll3.v <- append(bankroll3.v,bankroll3)
  bankroll.rand <- bankroll.rand + rtrn.rand
  bankroll.rand.v <- append(bankroll.rand.v,bankroll.rand)
}
```

References

Barnwell, Bill. "The NFL's Numbers Game.", 2013. http://grantland.com/features/bill-barnwell-breaks-2012-numbers-get-feel-2013/.

BWIN. "Help-General Information.", 2013. https://help.bwin.com/general-information/legal-matters/general-terms-and-conditions/sports-betting.

Dixon, Mark J., and Stuart G. Coles. "Modelling Association Football Scores and Inefficiencies in the Football Betting Market." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 46, 2: (1997) 265–280.

Fama, Eugene. "Efficient Capital Markets: A Review of Theory and Empirical Work." *The Journal of Finance* 25, 2.

Goddard, John, and Ioannis Asimakopoulos. "Modelling Football Match Results and the Efficiency of Fixed Odds Betting." *Journal of Forecasting* 23, 1: (2004) 51–66.

Golec, Joseph, and Maurry Tamarkin. "The Degree of Inefficieny in the Football Betting Market." *Journal of Financial Economics* 30: (1991) 311–323.

Grayson, James. "A quick and dirty estimate of the points that'll be scored by the teams finishing 1st, 4th, and 17th in the Premiership this season.", 2014. http://jameswgrayson.wordpress.com/2014/01/15/a-quick-and-dirty-estimate-of-the-points-thatll-be-scored-by-the-teams-finishing-1st-4th-and-17th-in-the-premiership-this-season/.

Karlis, Dimitris, and Ioannis Ntzoufras. "Bayesian Modelling of Football Outcomes; Using The Skellam's Distribution for the Goal Difference." *IMA Journal of Management Mathematics* 20, 2.

Kelly, J. L., Jr. "A New Interpretation of Information Rate." *Bell System Technical Journal* 35, 4: (1956) 917–926.

Maher, M. J. "Modelling Association Football Scores." *Statistica Neerlandica* 36, 3: (1982) 109–118.

Reep, C., and B. Benjamin. "Skill and Chance in Association Football." *Journal of the Royal Statistical Society. Series A (General)* 131, 4.

Ruggiero, John, Lawrence Hadley, Gerry Ruggiero, and Scott Knowles. "A Note on the Pythagorean Theorem of Baseball Production." *Managerial and Decision Economics* 18, 4: (1997) 335–342.

Thompson, William N. *Gambling in America: An Encyclopedia of History, Issues, and Society*. ABC-CLIO, 2001.

Wever, Sean, and David Aadland. "Herd Behavior and Underdogs in the NFL." *Applied Economics Letters* 19, 1: (2012) 93–97.

Whitrow, Chris. "Algorithms for Optimal Allocation of Bets on Many Simultaneous Events." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 56, 5: (2007) 607–623.

Williams, Leighton Vaughn. "Information Inefficiency in Betting Markets: A Survey." *Bulletin of Economic Research* 51, 1: (1999) 1–39.