Predictive Capabilities of Sports Market Line Movement

Abstract:

The decision making behind the design of propositions and payoffs for the various types of investments is one that can decide winners and losers far before any individuals decide to participate. With vast multi-variable models designed to understand patterns unpredictable to the human eye, the designs of these investment markets are meticulous, complex, but efficient. The reason for this is simple, to guarantee profit for those who are doing the proposition setting. Much study has gone into variable-odds games such as the stock market and horse racing where an individual will set an initial guiding price however the actions of the public decide the final payoffs that each of the strategies will return. These games provide benefit to the individual as they only have to outplay the general public, while the proposition setters produce profit by keeping a portion of the losing investors risk that is to be given to the profiting investors. Alternatively closed-odds betting allows for the public to place risk on propositions set by a single oddsmaker. Here the oddsmaker will also withhold a percentage of all bets at the time of placement, reducing the payoff for the investor, however this does not guarantee profit for the oddsmaker. Instead, to guarantee a profit the oddsmaker will need to keep the difference in risk on one side of a proposal to the other side within the percentage withheld from each bet.

While it may seem intuitive to set odds that will result in an even distribution between both sides of a proposition, the task of understanding the public perception of a proposition is often just as difficult as predicting the outcome of a proposition. To control for times when the oddsmaker's perception of the public is incorrect, the oddsmaker can shift the payoffs to the proposition or the proposition as a whole, in order to entice the public to divide more evenly. Alternatively, there are times when we see the oddsmakers fail to adjust their propositions despite massive imbalances in the risk distribution, or even shift the proposition in favor of enticing more of the public to the risk dominant side. This produces the question, why do oddsmakers act this way, and is this act against risk-aversion a sign of potential profit for the investor?

Introduction:

Sportsbetting kingpins such as FanDuel, William Hill, and DraftKings have dominated the market, accumulating yearly profits in the hundreds of millions. While much of this dominance in the industry can be attributed to aggressive advertising and a demand for brand relevance, an even larger role may be played by what is not in the public eye. Much like their advertising budgets, the major sportsbooks spend millions of dollars yearly researching, building, and maintaining models to predict the outcomes and propose lines for a wide array of events and matches. Due to changes in perception and obtained information, payoffs and propositions can change in moments. In an attempt to observe the predictive capabilities of these changes we will make use of common machine learning models such as linear regressions and random forests to better understand what these changes represent. The results from these models will help determine not only if elements of the propositions presented to an average consumer are diagnostic but also if the results are accurate enough to provide profit for an individual investor.

Observations across long periods of time and various sports will require various types of analysis in an attempt to accommodate for various hidden variables that can affect our model's perception of changes in the propositions.

Dataset:

sportsbookreview.com provides information required on the propositions, results, and changes in odds that will be used to build our models. For each proposition, the initial and closing odds are provided as well as the percentage of investment placed on each side. The results of the proposition are provided with precise detailing on the intermediate results. The following dataset will provide me with years worth of data across various forms of sport the propositions can be based upon. From the datasets provided 5 primary variables were chosen to represent a single team of a contest in any major sport. These included, the opening point spread, the closing point spread, the corresponding 'home' or 'visitor' value, the numerical change in point spread, and the result of the proposition. To provide context, the point spread of a game is a numerical value used by a sportsbook in order to create propositions that appear equally likely to occur. Given a contest between teams A and B, if the sportsbook predicts team A to win by 10 points, they may submit a point spread for team A of '-10'. Here investors can place bets on team A, where they collect on their investment if team A wins by more than 10 points. Conversely an investor can bet on team B using the same point spread where they will collect on their investment if team B wins or loses by less than 10 points. If the game ends where team A wins by exactly 10, bets on both sides will be refunded.

Due to the nature of some sports, adjustments to the variables obtained had to be made. For sports where scoring is less prevalent such as baseball (MLB) and hockey (NHL), the point spread or the amount of points a team must win by remains constant throughout much larger changes in information obtained and public reception. Due to this, for these sports we will represent opening and closing lines in money line (ML) values. Here to collect on an investment an individual simply has to bet on the winner of a contest however the payoff will commonly not be 1:1. Instead teams that are favored will pay out at a price less than even money, and teams that are considered underdogs will pay out at a price greater than even money. We see that this form of proposition is often considered more volatile and accurate due to the level of precision that is able to be achieved. Unlike point spread which can change minimally by half point values, money lines can change by one percent of the amount wagered.

The use of change in line in tandem with opening and closing lines was done in an attempt for the model to understand the nature of the sport as well as the behaviors of the models that produced the opening and closing lines in the first place. While the idea of line movement is supposed to represent a change in understanding in the eyes of the sports model, the size and significance of these shifts is dependent on the sport itself. Extending beyond sports where scoring is less prevalent leading to shifts in lines representing large changes in perception, the nature of the scoring can also provide insight. For example, in american football where scoring primarily comes in the form of three and seven points for field goals and touchdowns, shifts across those numbers display a much larger change in perception then similarly sized shifts

elsewhere. For example, a one point shift from -6.5 to -7.5 displays a much larger change in beliefs than a shift from -1.5 to -2.5. This idea can be seen in Figure 1 of the appendix where it shows not only that a greater percentage of games end in margins of victory of 3 and 7, but also their summations such as 6, 10, and 14. Ultimately, we want the models to learn that while the size of the shift is important, the locations of the shift can be just as big of an indicator of changing beliefs.

Analysis:

One of the first intuitions in the analysis of the predictive capabilities of sportsbook lines was the use of decision trees and random forests. Specifically the goal was that the output from a decision tree could provide a greater understanding of the relationship of the variables that we were included in our model. The readability of decision trees allowed for the confirmation of which of our explanatory variables were most correlated with the result of contests and determine if any of our variables demonstrate non-linear properties. The choice of decision trees was also guided by their invulnerability to multicollinearity. Due to relationships such as teams playing at home being more likely to be considered a favorite, and the distribution of closing lines being skewed towards the numerical value of the opening line, we see that our data is likely to have high co-variance. Considering this relationship and our relatively small number of explanatory features the desire to avoid overfitting resulting in the use of random forests. Here the collection of decision trees, random selections of observations and explanatory variables are used to build multiple decision trees. Together this ensemble of trees average their results in order to label data that is to be classified.

With the original procedure the four explanatory variables of location, opening line, closing line, and change in line, were used to simply predict what side of the closing line would result in payout. An example of such a tree is shown in Figure 2 of the appendix that was produced using data from the 2020-2021 NFL season. From this tree and others observed some key patterns begin to form. Specifically, we see that the two most ‘influential’ explanatory variables are the closing line and the shift in the line from opening to closing. The location of the contest is also present in the trees, however the opening line fails to be found. From this we expanded on the decision tree to now produce random forests. Using the procedure discurred, random forests were created using training data for five of the major sports organizations respectively. The models were then tested across testing data from the previous five seasons. The results are shown in Figure 3 of the appendix. The chart shows much of what was expected. Specifically models based on organizations such as the NHL and MLB perform slightly worse due to only using money line figures. These figures represent the games more linearly, failing to convey information on the significance of the shift that can be represented in some point spread movements. Additionally, we see that while some sport organizations produce models that perform choose the correct side of the point spread on average more than fifty percent of the time they fail to be able to produce profit for the individual investor. This is because in addition to the point spread, major sportsbooks will commonly only payout ten-elevenths of every dollar of profit that is won. This cut of all winnings taken by the sportsbook is known as ‘juice’. The result

of ‘juice’ is that in order for an individual investor to remain profitable they need to hold a winning percentage of fifty-five percent or better. Due to this condition the created models fail to succeed in producing long-term profits resulting in a mandatory change in the way data is handled and our model is produced.

Moving forward the dependent variable used in our model will no longer be a binary representative of what side of a point spread to bet on. Instead we will be predicting our own point spread that will be used to later determine investment. This shift allows us to weigh decisions made by the model proportionally, attempting to maximize on patterns that the produced model is most confident on. Additionally, the selection process for the data used to train and test the model has become stricter. Specifically, we now no longer consider contests where the point spread has shifted by more than a predetermined cutoff amount. This cutoff is determined as being one-third of the average margin of victory of a game in the respective organization. The cutoff is implemented to remove contests where major changes to the game or its participants have occurred. Whether due to injury, personnel change, or change in location, these scenarios represent an entirely different game than the one that was occurring when the opening line was set. In the scenario where a major participant in a contest is removed, we want to observe the line movement from the moment after the line adjusted for the removed player to the closing line, not from the opening line to the closing line.

With these changes we shift the model to a linear regression. While previously, normality of residuals may have posed an issue with modeling using a linear regression the current format should now produce accurate results. Additionally, potential issues such multicollinearity pose a threat only to determining ‘individual predictor variable’s impact on the response variable’, not predictive power as a whole (Frost). Using the matrix inversion method to produce model weights, our findings are shown in Figure 4. Using these weights, a separate points spread has been created that can be compared to the one created by the sportsbooks. Using the kelly criterion we can determine the optimal bet sizes in order to avoid ruin. The kelly criterion bet size formula is a ratio consisting of the bet payoff, the projected probability of a win, and the current bankroll or budget. In order to determine our probability of winning the bet we will use a ratio of the difference between the cutoff point spread used when determining considered data and the difference in closing point spread to our predicted point spread. Using this model and these determined bet sizes the resulting profits for the previous ten NBA and NFL seasons are displayed in Figure 5.

Conclusion:

The ultimate conclusion reached from observing the results from the linear regression model is one of restricted optimism. While the results for an individual investor may not guarantee long-term profit they do seem to show an aversion to ruin, something that the sportsbook models are designed to lead to. Most notably the changes that provided the most benefit included the restriction of contests where the line had shifted by less than a third of the average margin of victory, as well as using both opening and closing lines, as well as the difference between the lines as explanatory variables in our models. The next steps of this project

include combining these results with that of models of other aspects of sports gambling. Specific team or player models could provide greater insight as to specific matchups and play styles which cause points spreads to move. More than anything however, accurate data on public betting tendencies could provide immense data clarification on the reasoning and patterns of line movement. Insight into when movements are with or against public betting patterns reveals hidden variables of the sportsbook model's tendencies allowing for a dar more accurate prediction of 'true' lines. Together the combination of these three models of point spread movement, team and player production, and public betting records could provide a full-scale view on sports gambling and the possibility of individual investor achievement.

Appendix

Figure 1:

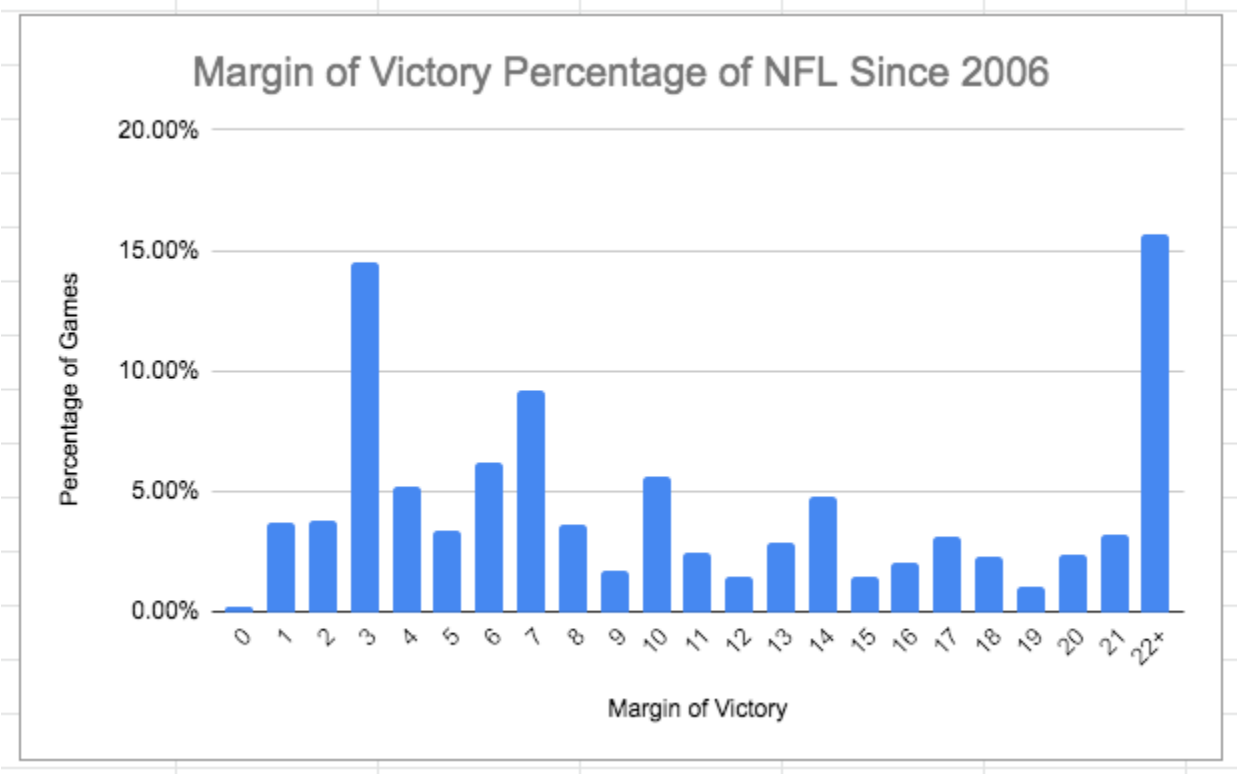


Figure 2:

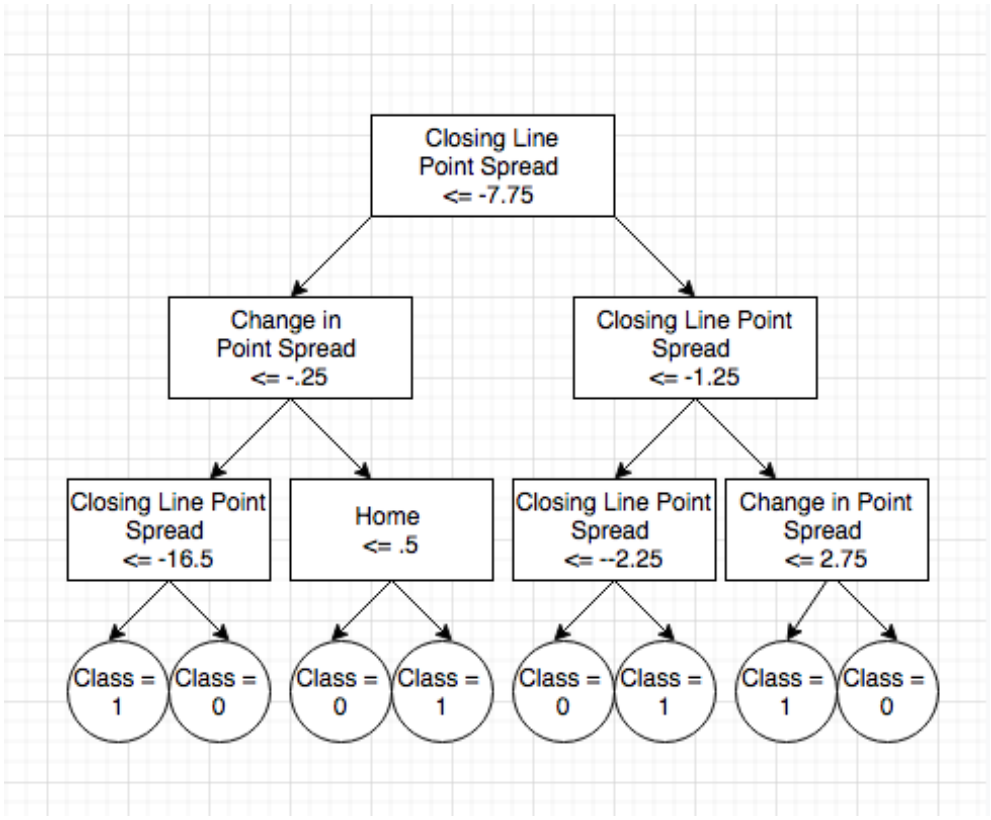


Figure 3:

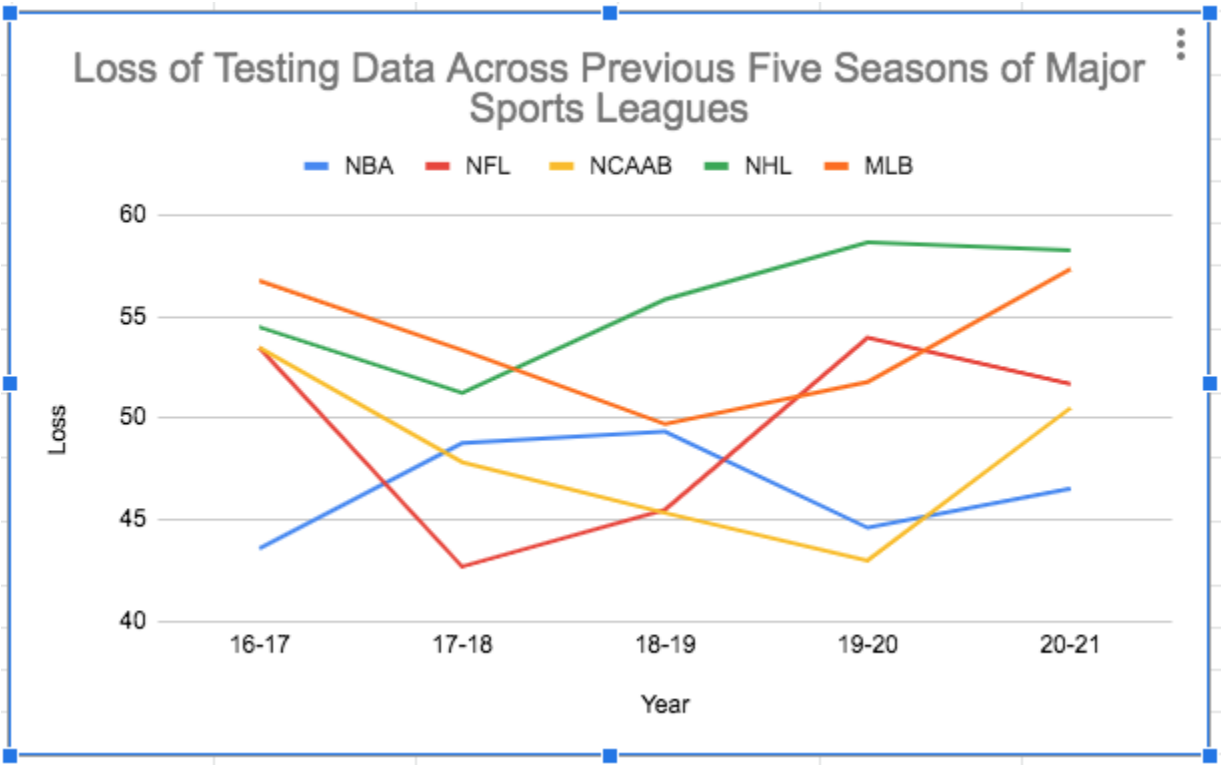


Figure 4:

Weights correspond to explanatory variables, Location, Opening, Closing, Change, and Bias

(NBA)

```
---- LINEAR REGRESSION w/ Matrix Inversion ---
[-0.61446568 -0.46882218 -0.46649978 -0.0023224   0.35256183]
```

(NFL)

```
---- LINEAR REGRESSION w/ Matrix Inversion ---
[-1.55223459 -0.53733057 -0.44005012 -0.09728045  2.0320272 ]
```

Figure 5:

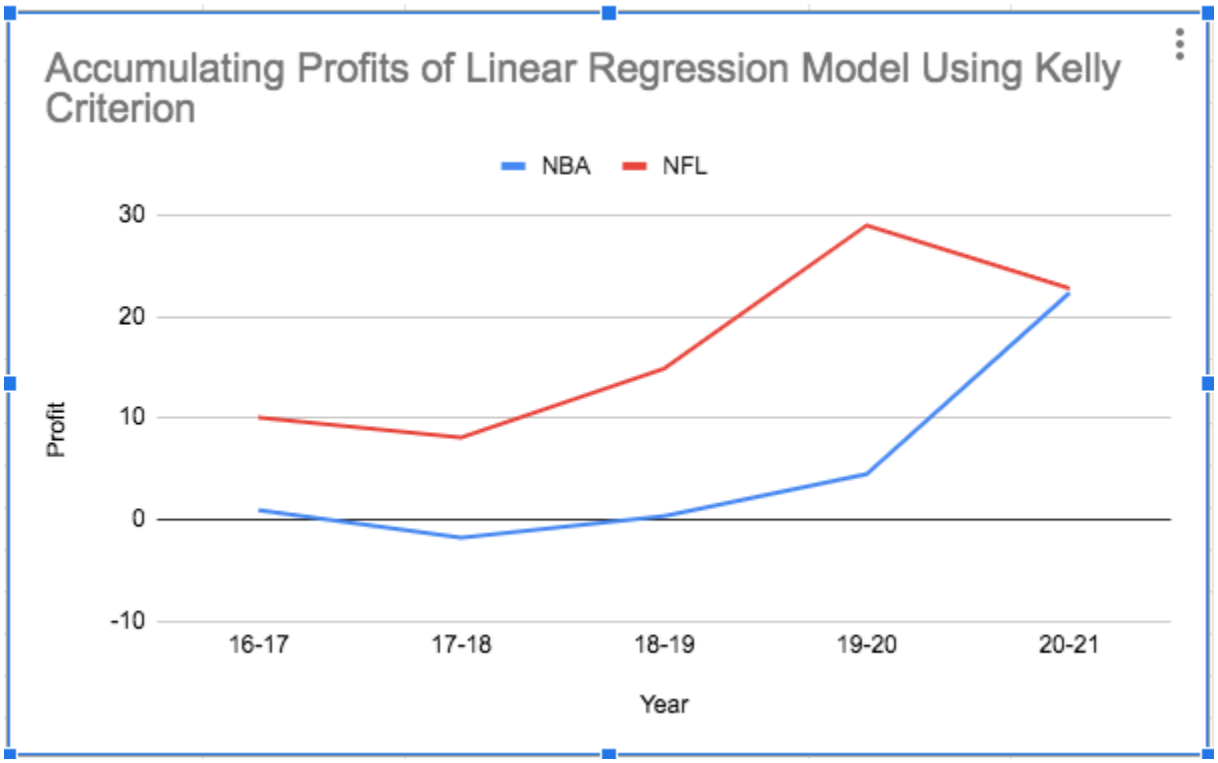
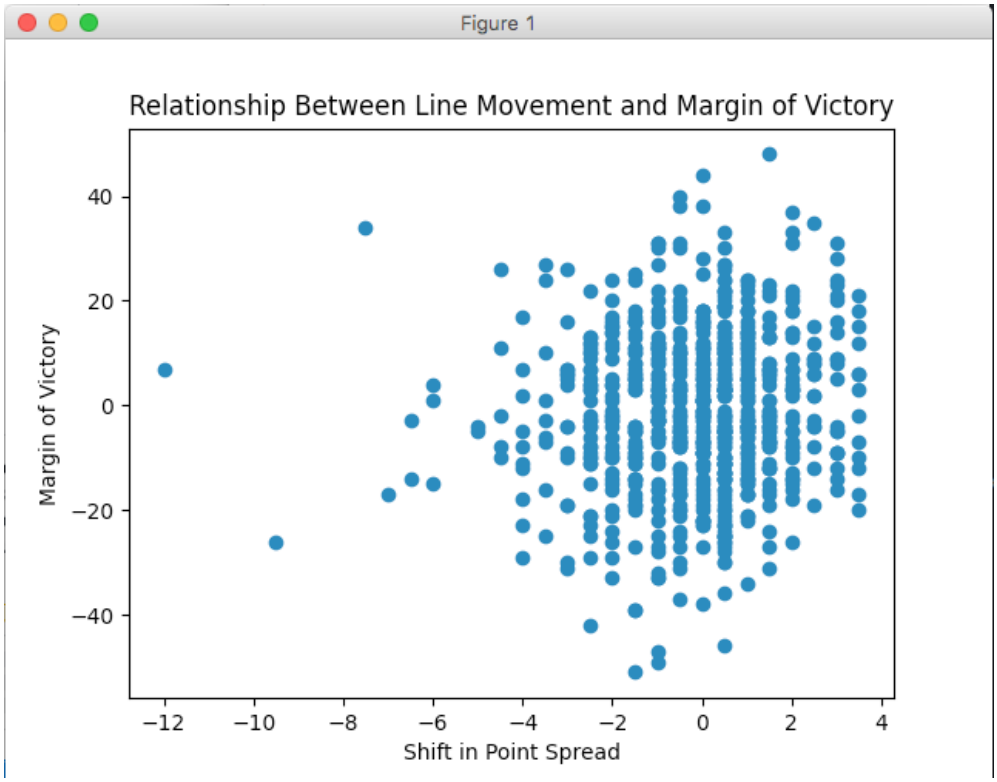


Figure 6:



Citations

- https://sportsbookreviewsonline.com/scoresoddsarchives/scoresoddsarchives.htm
- Multicollinearity in Regression Analysis: Problems, Detection, and Solutions, Frost, Jim
- Econometrics, Hansen E., Bruce pg. 124-126 (https://www.ssc.wisc.edu/~bhansen/econometrics/Econometrics.pdf)
- Fortune's Formula: The Untold Story of the Scientific Betting System That Beat the Casinos and Wall Street, Pounstone, William