ST 495 Final Project – Bootstrapping a model to predict sports betting

INTRODUCTION

Casinos are one of the most profitable businesses in the world because they give people the illusion of making money but set up their games so that they, the casino, always come out on top statistically. Caesar’s Palace, for instance, boasted a net gain of 1.7 billion dollars in the year 2021 alone, and that was while the market was still recovering from the coronavirus pandemic. Gamblers often believe that they can make money and will rely on previous outcomes to predict the future in independent statistical phenomenon’s, in a mindset called the Gambler’s fallacy. In popular gambling games like Blackjack, Craps, and Roulette, even with a perfect strategy, the house still will boast a statistical edge. However, sports betting is something to be explored because of the factors such as homefield advantage, historically good teams, and that there is not a guaranteed loss of money if one is well educated on the sport.

The biggest appeal of sports betting is that it is statistically impossible to perfectly predict a game, since there are certain human elements that are harder to account for. March Madness, for example, has never had a perfect bracket, with most brackets being busted in the first round due to upsets that most people would never predict. While no prediction model can be perfect, there are factors that can push your odds above the line to make money. To do this, we will be using NFL data from the last 10 years to see if there are factors, or certain bets, that can push your odds over the line and give you the edge over the house.

This study will explore whether the logistic model introduced in the original study can have the variables modeled by bootstrapping the data. Bootstrapping won’t give us as precise an answer as a logistic model, but it doesn’t rely on us estimating the standard deviation when we don’t really know what it is. So, in a way, it is a better way to model our data, especially given the relatively limited sample size.

VARIABLES AND MODEL

The model for this study, as previously stated, is a bootstrapped model. A bootstrap model allows us to make conclusions about the data without having to make assumptions about the distribution of the data or the standard deviate of the data. For the first half of the study, we assumed the data followed a logistic regression model, which it may very well have, but with bootstrapping we do not have to assume this. This data was run with 300 different samples, with the seed matching the current iteration of the for loop.

To maintain similarity, the exact same variables from the first study and the second study will be analyzed as having an affect on predicting an away cover. These variables, in order are opening away odds, opening home line, games that take place at New England, and games where the away team is the Cleveland Browns. Although the original data boasted an ROC curve only slightly above the mean, these variables have the most observations and are the most useable, and still predicted correctly more often than incorrectly, even if by a small margin.

The first variable would be the Beta 0, or the y intercept on a non-New England home game or Cleveland away game. As we can see, the bootstrapped data demonstrates a normal distribution, and the true beta 0 value based on the gradient descent algorithm estimate is still within this bootstrapped interval. This shows that the beta 0 value is accurately estimated given this data, at least to a relatively high level. The estimate for the gradient descent algorithm was .535. The average of the bootstrapped data was .633. While this is different, the estimate does hold mostly true. Also, this is just the intercept, not a specific predictor that was chosen from the data set, albeit affected by these predictors.

Chart, histogram

Description automatically generated

The first variable was the away odds estimate, or the odds of the away team. The bootstrapped data seemed much more centered around 0, which would imply no significant correlation between away odds and away victory. This estimate for this reason as well as the Beta MLE estimate being outside the interval shows that this might not be a significant variable, especially in a bootstrapped model. The beta estimate was -.1304. whereas the bootstrap estimate was -.033. The beta estimate being almost 4 times as large as the bootstrap estimate also leads me to believe that this variable is not represented well. Since the away odds mean in 2.863933, this means that this variable on average will lower the odds of the away team covering by 9.5%, compared to 37.4%, which is a significant difference.

Chart, histogram

Description automatically generated

The second variable was the home line estimate, which basically means the amount of points the home team is given against the away team. Although this is also close to 0, the majority of the points fall underneath 0, which tells me even with bootstrapped data, opening home line is a significant variable in this predicted. It seems as though the more points the home team is given, the away team has a lower chance of covering. The bootstrap estimate for this data was -.00677, the -.026 maximum likelihood estimate is close enough to show that this variable has similarity between the estimates.

One observation about this data is that 0 is still in the interval, which means in terms of a hypothesis test, we would fail to reject that the beta is different from 0. Therefore, we should proceed with caution in using this variable as a significant factor in impacting away covers.

Chart, histogram

Description automatically generated

Bootstrapping the New England Patriots home games can be a risky, and likely inaccurate estimate, which is a reason we see such a large range. With each bootstrap sample holding 300 observations and there being 32 NFL teams means that each bootstrap sample holds on average under 10 New England home games. This could lead to a large extrapolation and inaccurate representation since the range could be so wide with such a small sample size. This is shown in the bootstrap estimate below, with the data ranging from under -.6 to over .2. In general, this data still is generally negative, and the beta estimate is a part of the bootstrap histogram, which is encouraging to our sample data.

The beta estimate for bootstrap is centered around -.12, whereas the maximum likelihood is -.44. This is a significant difference, especially when we consider that this is on a scale of predicting a 0 or a 1. The Maximum Likelihood Estimator states that playing against the Patriots at home decreases the odds of the away team covering by 44%, compared to 12% with the bootstrap estimate.

Chart, histogram

Description automatically generated

The same risks we run bootstrapping the New England Patriots data is prevalent in the away Cleveland Browns games. We also see the large distribution of data from above .4 to below -.6. This large data can be set to each observation having 10 or so observations of away Cleveland Browns games. For this reason, I would be more apt to trust the Maximum Likelihood Estimate above the bootstrapped data.

A similar trend is present between these two in comparing the estimates as well. The bootstrap mean is -.057 compared to the Maximum Likelihood estimator of -.23. This also shows a significant difference, and would suggest that with bootstrapped data, although Cleveland being away still lowers your odds of covering on the road, it is not as significant as the likelihood estimator suggests.

Chart, histogram

Description automatically generated

One key observation about these variables is that in all of the intervals of bootstrapped data, 0 is contained within it, which in a hypothesis test would lead to the conclusion of an insignificant variable. This would once again lead me to be cautious in being overly trusting of these results. However, as stated in the previous study, we cannot expect perfect estimates, as this is sports data and we are competing against the casino, with them trying to beat us more often than they lose to us. Their whole idea is to have the predictor below the 50% mark, so our best attempt is to raise it above the 50% mark. Aft

RESULTS

Our new estimates, with using the bootstrap mean’s mean as the center for each yields us this result:

Y(pred away cover ) = .6333 - .03 ( Away odds estimate ) - .006 ( home line odds) -.12 ( at NE) -.056 ( away CLE )

This model shows numbers that reflect significantly smaller beta values than the logistic regression model. Plugging in our averages for each estimate yields 53.8%-win rate for non-Cleveland road games/ new England home games. This more reflects expected results than the other study. The previous study, as well as this one, tried to take into account that Cleveland is outlier bad and New England is historically good, so this removed these outliers of team.

Overall, I think that this model yields better, and more realistic results in predicting outcomes. The Maximum Likelihood Estimators seem a little overblown and almost overestimate the data. To test this, I plot this data on the False Positive Rate/ True Positive Rate line. Results are plotted below.

Chart, line chart

Description automatically generated

With both seeds set to 1, this is how the bootstrapped estimator compares to the Maximum Likelihood estimators on the ROC Curve. As we can see, even though the estimates are significantly different in many cases from one another, it is difficult to determine which estimate is more useful in predicting outcomes. Since the Bootstrapped estimates are slightly lower values than the Maximum Likelihood, I would use these in a model for future data.

Chart, line chart

Description automatically generated

One take away from both models is that games that take place at New England should almost always be bet on for New England, with the average New England game having a 40% predicted away cover. This means betting the spread on New England will give you the win 60% of the time, based on this model, on an average game. This makes sense, as the Patriots are a historically good team and have a historical home field advantage.

Another observation from this data is that for every beta value except for Home line, the predictor confidence intervals were significantly tighter, shown by the Beta interval being overlayed on the bootstrapped histogram. It would be interesting to see why the home line interval is so much tighter. I believe this could be because Away odds and Home line are relatively invertible, so they have an impact on each other which could explain this.

CONCLUSION

All in all, although the variables differ from the Maximum Likelihood estimators, often significantly, we can see in the ROC curves that they yield similar results and this is without making any assumptions about our data in terms of standard deviation or guessing a distribution. In terms of using these estimators, I would still be cautious in trusting the results since a sports betting algorithm is never going to be perfect, as you are competing against the “bookie” or someone else who is purposefully making these harder to predict. Counting cards is one of the most frowned upon activities by casinos, and this only raises the players odds of victory to 52%. Measuring against this metric shows how our model can predict data at a relatively high rate.

Betting is addictive and illegal in many states, so this study is in no way an endorsement on sports betting. If anything, this study shows that using basic stats, it is hard to predict a game based on past games. Also, the house by default takes a -110 edge, basically for every dollar you bet, you get 90.9 cents back on a perfect 50/50 scenario. If anything, this study should discourage sports betting and further prove that gambling is unpredictable and should be done extremely responsibly.

In reality, flipping a coin can lead to 10 straight heads or 10 straight tails. Using this same logic, with sports betting, a 53% correctness rate could lead to similar results, which reiterates why it is important to be responsible and not bet too much money, as even with the most advanced model, dry streaks and hot streaks can occur.

LINK TO DATASET:

https://www.aussportsbetting.com/data/historical-nfl-results-and-odds-data/