Project EDA : Sports Bet like a Hedge Fund Manager

Nico Manzonelli, Jeffrey Mayolo, Joseph Zuccarelli, Simon Lam

Due Date: 11/23/2021

Introduction

In the past, sports gamblers had to place bets at the casino or with a local bookie who operates illegally. Now, with recent relaxations in gambling laws across the United States and the emergence of online sports books, sports betting is more accessible than ever. Gamblers can easily place bets online (in states where sports gambling is legal) before or during sporting events. The materialization of live sports betting opens up a completely new branch of opportunities to make money through gambling. Unlike ever before sports gamblers have the opportunity to observe a change in odds on their opening bets and hedge when the opportunity is right to maximize profits or minimize losses.

The Math that Backs Hedging Bets

Given that live betting is so accessible, sports gamblers have the opportunity to "hedge" almost every bet they make. One can "hedge" by placing an additional competing bet to maximize profits or minimize losses based on how the odds have changed since they placed the initial bet.

After a bettor hedges their bet there are two outcomes: 1) the initial bets wins or 2) the competing bet (i.e. "the hedge") wins. Profits for either outcome are as follows:

$$P_1 = b_i * r_i - b_h$$

$$P_2 = b_h * r_h - b_i,$$

where P_1 and P_2 are profits for outcome one and two, respectively. Variables b_i and b_h represent the dollar amount of money wagered in the initial and hedge bet, respectively. Finally, r_i and r_h represent the return on dollar (based on the odds) from the initial and hedged bet.

In order to ensure that one wins or loses the same amount of money no matter which team wins, they should hedge with a bet where $P_1 = P_2$. Therefore, by solving for the amount required to hedge (as a function of initial odds, initial bet and half-time odds), we get the following:

$$b_h = \frac{b_i r_i + b_i}{r_h + 1}.$$

Whether the bet will guarantee profits or minimize losses is dependent on the direction that the line moved. The conditional $r_i > r_h$ guarantees profits, while the conditional $r_i < r_h$ minimizes losses. However, r_i and r_h are not fixed values. Given the multitude of bookmakers that one can choose from, r_i and r_h can take on up to 10 different values. This characteristic of sports betting increases one's opportunity to bet and hedge with the optimal bookmakers to potentially guarantee profits.

Study Objectives

- Determine if there is a significant difference in how various bookmakers set their pregame and halftime odds for different teams like home team, away team, underdog and favorite.
- Analyze the association of change in odds with various game-related factors such as difference in score, number of turnovers, number of penalties, etc.
- Build a best predictive model for what we expect the change in odds to be based off the predictors at hand and use this model to determine if a sportsbook over or under corrected.
- Build a best predictive model to determine the probability of one's initial bet winning to recommend the use of hedging.

Exploratory Data Analysis

The population under observation in this study is NFL football games during the 2021 season. The data set that we plan to use in our analysis is composed of 67 complete records with the following variables: a unique GameID for each game, the home team, the away team, the home team's pregame odds, the away team's pregame odds, the home team's pregame score, the away team's pregame score, the home team's odds at halftime, the away team's odds at halftime, the home team's halftime score, the away team's halftime score, the difference in score at half (home team - away team), various football-related statistics for both the home and away team at halftime (yards, first downs, fumbles, sacks, interceptions, penalties), ESPN's predicted win percentage for the home team at half, a binary indicator of whether or not the home team wins, and the change in odds for the home team from pregame to halftime. The response variable in our study is the change in the home team's odds, which represents the difference between the home team's odds at the start of the game and the home team's odds at halftime. Note that as football season progresses, we plan to collect more data, including NCAAF data.

First, let's explore the variables included in our data set. The table included below displays the first six rows of our data set.

```
# Import necessary libraries
library(dplyr)
library(lubridate)
library(ggplot2)
```

```
# Read in data set
avgOdds_df <- read.csv("average_odds.csv")

# Data cleaning
avgOdds_df <- avgOdds_df %>%
  select(-X)

# Mutate variable 'home odds change'
avgOdds_df$home_odds_change <- avgOdds_df$home_odds_start - avgOdds_df$home_odds_half

# Mutate variable 'score differential half'
avgOdds_df$score_differential_half <- avgOdds_df$home_score_half - avgOdds_df$away_score_half

# Head of data set
head(avgOdds_df)
```

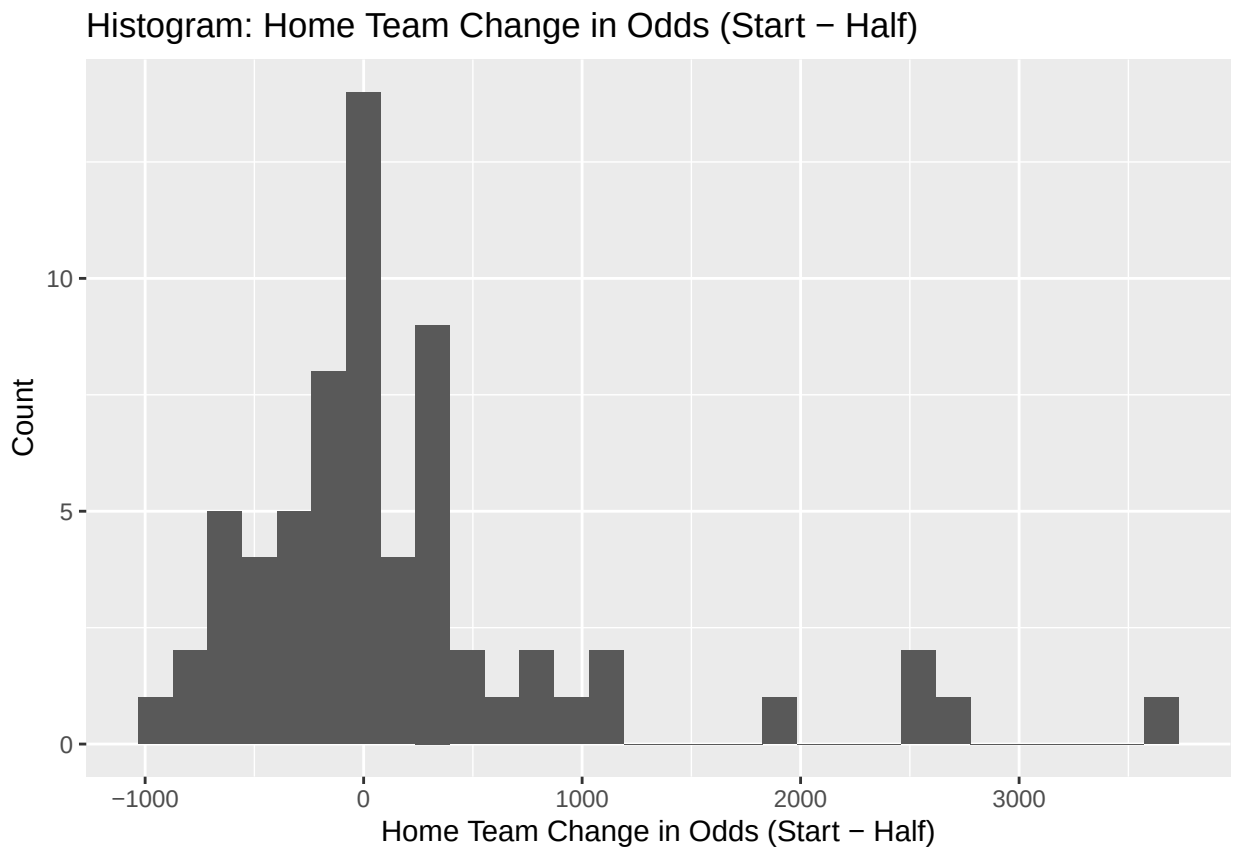
```
##           home_team            away_team home_odds_start away_odds_start
## 1 Arizona Cardinals    Green Bay Packers       -221.0000        199.0000
```

```
## 2 Arizona Cardinals       Houston Texans      -1964.5000       1300.0000
## 3   Atlanta Falcons    Carolina Panthers       -132.2308        113.0769
## 4   Atlanta Falcons New England Patriots        258.3333       -314.2222
## 5  Baltimore Ravens   Cincinnati Bengals       -274.8462        227.6923
## 6  Baltimore Ravens Los Angeles Chargers       -159.8462        127.0000
##   home_score_start away_score_start home_odds_half away_odds_half
## 1                0                0      -111.2727       46.54545
## 2                0                0     -3065.5000     1530.50000
## 3                0                0      -198.5385      154.53846
## 4                0                0       973.3333    -2749.11111
## 5                0                0      -166.3077      132.00000
## 6                0                0     -1299.7692      696.30769
##   home_score_half away_score_half     GameId home_yards away_yards
## 1               7              10 2021102800        102        157
## 2              17               5 2021102408        207         87
## 3              10               9 2021103100        121        190
## 4               0              13 2021111800         71        181
## 5              10              13 2021102400        183        199
## 6              17               6 2021101701        210        128
##   home_first_downs away_first_downs home_fumbles away_fumbles home_sacks
## 1                5                13            1            0          0
## 2               19                 6            2            1          2
## 3               11                10            0            1          0
## 4                6                11            0            0          2
## 5               14                11            0            0          0
## 6               20                11            2            0          1
##   away_sacks home_ints away_ints home_pen away_pen espn_win_pred home_win
## 1          2         0         0        3        4         0.362        0
## 2          4         0         0        4        4         0.947        1
## 3          1         0         1        2        4         0.560        0
## 4          3         0         0        6        2         0.073        0
## 5          3         0         0        2       NA         0.453        0
## 6          3         1         1        1        1         0.860        1
##   home_odds_change score_differential_half
## 1       -109.72727                      -3
## 2       1101.00000                      12
## 3         66.30769                       1
## 4       -715.00000                     -13
## 5       -108.53846                      -3
## 6       1139.92308                      11
```

Next, let's refine our exploration to the response variable–the change in the home team's odds at halftime. The histogram included below displays the distribution of the change in odds from pregame to halftime for the home team. Notice that there appears to be a few outlying games in which the home team's change in odds at halftime was greater than 1500. These outlying games are causing the distribution to be heavily right-skewed, as the mean home team change in odds at halftime is 197.939 and the median home team change in odds at halftime is -6.526. However, note that if we ignore these few outlying games the distribution appears to be unimodal, fairly symmetric, and centered around zero.

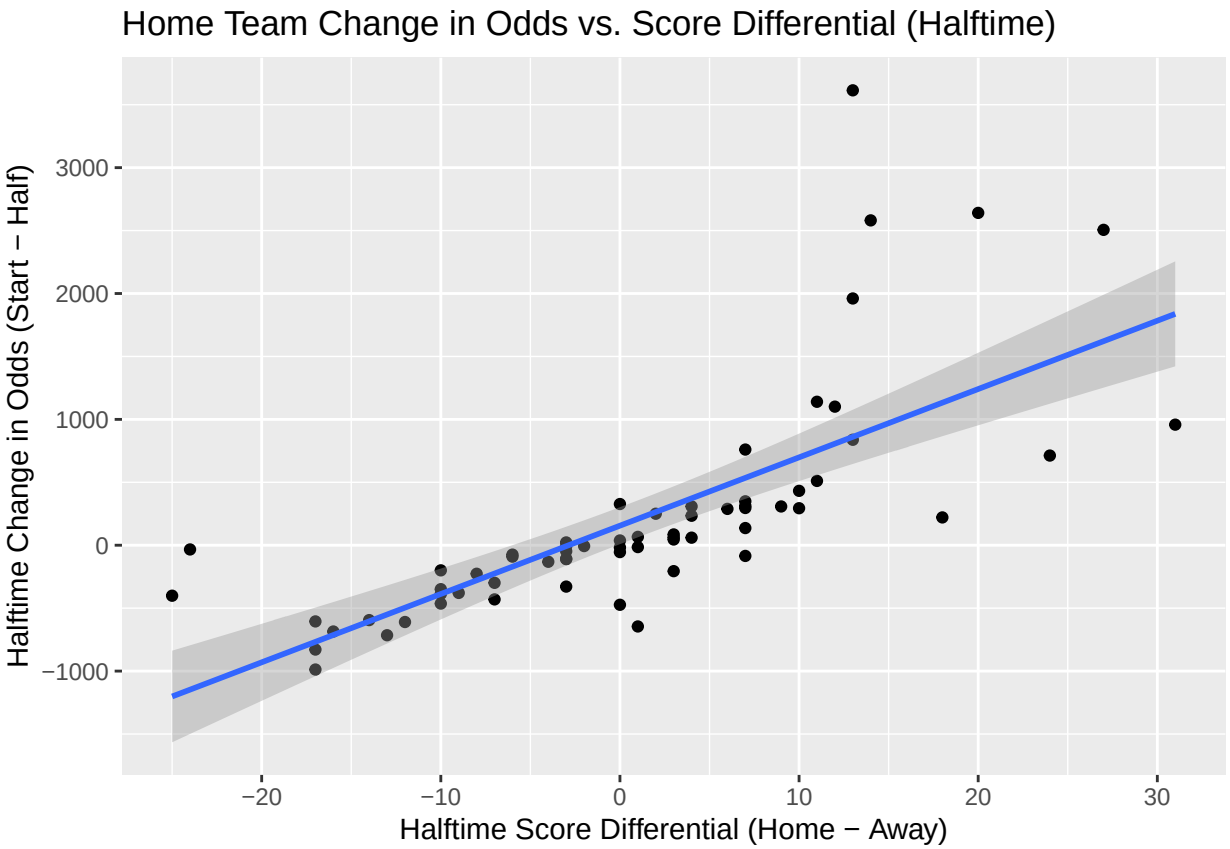
```
# Histogram: Change in Home Team Odds (Initial - Halftime)
avgOdds_df %>%
  ggplot(aes(home_odds_change)) +
  geom_histogram(bins = 30) +
  labs(x = "Home Team Change in Odds (Start - Half)", y = "Count",
```

```
       title = "Histogram: Home Team Change in Odds (Start - Half)")
```



Finally, let's explore the relationship between the response variable, the home team's change in odds at halftime, and one of our predictors, the score differential at halftime (home score - away score). Notice that there appears to be a positive linear association between the home team's change in odds at halftime and the score differential at halftime.

```
# Plot: Home Team Change in Odds (Start-Half) vs. Halftime Score Differential
avgOdds_df %>%
  ggplot(aes(x=score_differential_half, y=home_odds_change)) +
    geom_point() +
    geom_smooth(method = 'lm') +
    xlab("Halftime Score Differential (Home - Away)") +
    ylab("Halftime Change in Odds (Start - Half)") +
    ggtitle("Home Team Change in Odds vs. Score Differential (Halftime)")
```



Preliminary Model

Now that we have completed some initial exploratory data analysis, let's build a baseline model to quantify the relationship between the variables in our data set. Our baseline model is a simple linear regression model that explores the relationship between the home team's change in odds at halftime and the score differential at halftime (home score - away score). This model is structured as follows:

$$\hat{HomeOddsChange} = \beta_0 + \beta_1 HalfScoreDifferential.$$

Fitting this model using our entire data set leads to the following results:

```
# Base Model (home_odds_change ~ score_differential_half)
baseModel <- lm(home_odds_change ~ score_differential_half, data = avgOdds_df)
summary(baseModel)
```

```
##
## Call:
## lm(formula = home_odds_change ~ score_differential_half, data = avgOdds_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -911.45 -239.40 -101.03   38.26 2753.22
##
```

```
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                155.347     73.525   2.113   0.0386 *
## score_differential_half     54.283      6.467   8.394  7.3e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 591.4 on 63 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.528,  Adjusted R-squared:  0.5205
## F-statistic: 70.47 on 1 and 63 DF,  p-value: 7.301e-12
```

Based on the output above, $\hat{\beta}_0$ is estimated to be 155.347: the average home team's change in odds at halftime when the score differential (home score - away score) at halftime is zero is approximately 155.347. $\hat{\beta}_1$ is estimated to be 54.283: a one point increase in the score differential at halftime is associated with approximately a 54 point increase in home team's change in odds at halftime on average. For $H_0 : \beta_0 = 0$ vs. $H_a : \beta_0 \neq 0$, (where β_0 is the coefficient for the intercept), we are able to reject the null hypothesis at the $\alpha = 0.05$ level ($t = 2.113$, $p = 0.0386$). For $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$, (where β_1 is the coefficient for the score differential at halftime), we are able to reject the null hypothesis at the $\alpha = 0.05$ level ($t = 8.394$, $p = 7.3 \times 10^{-12}$). Therefore, we conclude that the score differential at halftime is a significant predictor of the change in the home team's odds at halftime in this model. Note that the R-squared value included above is 0.528, indicating that our base model with one predictor, the score differential at halftime, explains approximately 52.80% of the variability in the response variable, the home team's change in odds at halftime.

Future Work

At this point, our model predicts the change in odds of the home team based off solely the score differential at halftime (home team - away team). In reality, prospective sports gamblers care much more about underdogs and favorites than home team and away team. In future work we would like to further wrangle our data to predict based off underdog and favorite and add home team as an additional categorical variable. We also plan to include various other predictors in our model from our current data set. While the score differential is one of many predictors that we would like to explore, we suspect that the other potential predictors included in the data set will also increase the predictive power of our model. Therefore, we plan to build a multi-linear and polynomial regression model with additional predictors from the data set.

As indicated by the preliminary model included above, there are heavily weighted outliers as the difference in score gets larger. We plan to adjust for this by adding a weighted regression model that fits the majority of the data better and is less affected by the outliers. We also plan to explore the use of random forest models, as these may lead to a better predictive model than linear regression at the expense of interpretability.

In addition to the in-game predictors included in our data set, we expect a difference in the response variable based off the bookmaker who sets the odds. To evaluate this further, we plan to create a mixed effects model where we pool the data based on bookmaker. This will allow for us to see the different effects from the different bookmakers and possibly predict which bookmaker to use for initial bets and for hedged bets. The data set used for this EDA includes averaged odds across all bookmakers, but the existence of significant differences between bookmakers may lead us to evaluating on the most advantageous odds across all bookmakers (Note: In sports betting this concept is often called "line shopping").

In the end, these models will all be compared using RMSE to determine the best fit prediction model for the change in odds at halftime. This model will tell us if the odds over or under corrected for the variables in the first half and thereby indicate if a certain team in undervalued or overvalued.

Finally, we plan to create a logistic regression model that uses the first half data to predict the probability that a team will end up winning the game. This model will be useful in informing a sports gambler's strategy:

should they hedge or should they let their initial bet ride out. Through simulation we will compare profits when following the model's betting strategy, when following a random betting strategy, and when following a constant betting strategy.