Analysis of Professional Football Gambling Line Performance

Team: Thomas Colicchio (tac69), Dylan Paul (dhp14), William Pierce (wcp17), Jack Steel (jas333)

Part 1: Introduction and Research Questions

As the sports gambling industry continues to grow, there is lots of money to be made or lost on every NFL Sunday. For this reason, the oddsmakers have a difficult task on their hands as their decisions can be the difference between hitting the jackpot or losing it all for many sports gamblers. Therefore, we are interested in the following research question:

How accurate are oddsmakers at setting lines for NFL games and what significant general trends, if any, can we derive?

Due to the broad nature of our research question, we have identified specific trends that we will measure to analyze the predictive accuracy of oddsmakers. This will include analysis on the performance of oddsmakers in general, their performance over time, their performance in different conditions (weather or field type), and for home favorites and away favorites. Understanding the factors that may affect the predictive accuracy of gambling lines is **relevant** to developing a deeper understanding of the inner workings of the sports gambling industry. We find that analyzing these trends in the accuracy of line-making will be worthwhile to potentially uncover room for exploiting inaccurate lines based on the environment, team, or matchup of a specific game.

Part 2: Data Sources

We found a dataset on Kaggle called "NFL scores and betting data" that includes 3 .csv files we are able to download.[1] In the primary .csv file called spreadspoke_scores, each observation represents an NFL game spanning from 1966 to the present. For each game, the dataset includes many variables such as the home and away team, final scores, spread, the over/under, etc. From 1978-2013, the betting lines are sourced from http://www.repole.com/sun4cast/data.html, and from 2013, they come from sportsline.com and aussportsbetting.com. From these variables, we can easily calculate the difference between the true spread and total score and the predicted spread and total score. These two variables represent the oddsmaker's accuracy and can be used to answer our set of research questions that pertain to the accuracy of betting lines, how they have changed over time, and what factors impact their accuracy. The second .csv called nfl_teams contains information about each franchise. We first merge these datasets using a left merge on the spreadspoke_scores data set using the team_home column and pairing it with the team_name column in the teams dataset, and then repeating this for team_away column and pairing it with the team_name column in the teams data set. The resulting columns are named home_id and away_id with the 3 letter codes for each team. Once merged, we created columns for the difference in predicted and actual spread and O/U which required the 3 letter team codes. Finally, we merged the third .csv file on the stadium_name column to get data about the field type for each game. Using this merged and cleaned dataset, we have access to many columns that include relevant information to answer our research question. This includes but is not constrained to home team, away team, actual and predicted lines, weather data, field type information, and dates of games played.

[1] Spreadspoke. (2023). NFL scores and betting data. Kaggle. Retrieved from https://www.kaggle.com/datasets/tobycrabtree/nfl-scores-and-betting-data

Part 3: Modules Being Used

Module 5 (Statistical Inference): This module is the most essential module in our testing and statistical analysis process as it features a variety of tests and methods (*norm.cdf(),* t-tests, bootstrapping) for breaking down and drawing conclusions from data. We used stats.norm.cdf() to calculate the p-values for the hypothesis tests in the **O/U Analysis** and **Spread Analysis** sections of our results. This function was appropriate because we are running a hypothesis test on data that is approximately normally distributed and has more than 30 data points. We also used the stats.ttest_ind() function to perform t-tests in the **Playing Conditions Analysis** and **Home/Away Analysis** sections of our results. This was the best function for these sections because we tested whether the O/U or spread data was from the same underlying distribution depending on the games playing conditions or the home/away teams.

Module 6 Combining Data: This module is most prominently used in the beginning stages of our report, when cleaning and combining the datasets. In this section, we use the pd.merge() function to join our three .csv files to create a single file that contains all of the relevant information. Using the common columns such as team names and stadium names, this Pandas function is the most logical way to combine these datasets. We also use the pd.groupby() function along with the .mean() aggregate function in the **Performance of Oddsmakers over Time** section. This is an easy way to group the games by season to analyze the mean performance of spread and O/U lines.

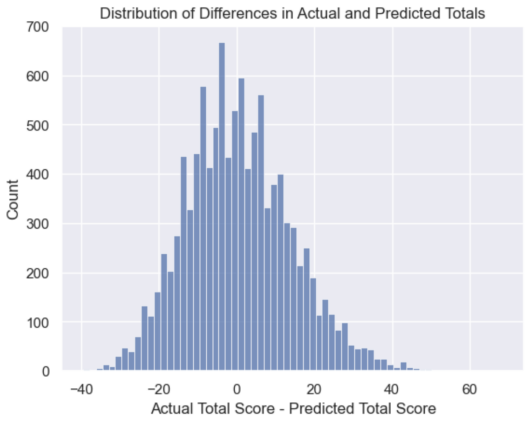
Module 8 Visualization: We used the tools developed in Module 8 throughout the entirety of our Results and Methods section of our project. There are multiple functions from the Seaborn library such as sns.histplot() and sns.barplot() that are used to create visualizations depicting the distributions of performative measures of gambling lines. These are nice accompaniments to traditional hypothesis tests because they provide a more qualitative understanding of the data which includes a sense of normality and the measure of center for a given distribution.

Part 4: Results and Methods

For each of the trends analyzed below we used what we called the ‘games’ dataframe which is the final result of the merging described at the end of Part 2.

O/U Analysis: The first graph displayed in the “Data Analysis and Hypothesis Testing of O/U Line Performance" is a histogram with “actualTotal - predictedTotal" on the x-axis and count on the y-axis. This plot displays a seemingly normal distribution with the peak of the distribution surrounding 0 with a mean of 0.6. This indicates that on average, the difference between the actual total score and the predicted total score is close to 0. We used the seaborn histplot function alongside the set function to create the plot and label the axes.

To get quantitative results of this relationship, we performed a hypothesis test with the null hypothesis being the mean difference between actual total score and predicted total score is less than or equal to 0 and the alternative hypothesis being the mean difference between actual total score and predicted total score is greater than 0. This yielded a p-value of around 2.06×10^{-6} using the stats.norm.cdf() function. Since this p-value is less than the alpha value of 0.05, we can reject the null hypothesis and we have statistically significant evidence to believe that the mean difference between actual and predicted total score is greater than 0. This makes sense in that there are outlier games that well exceed the predicted total score, however, the degree by which an actual total be less than the predicted total is bounded by the fact that the theoretical minimum amount of points that can be scored in a game is 0.



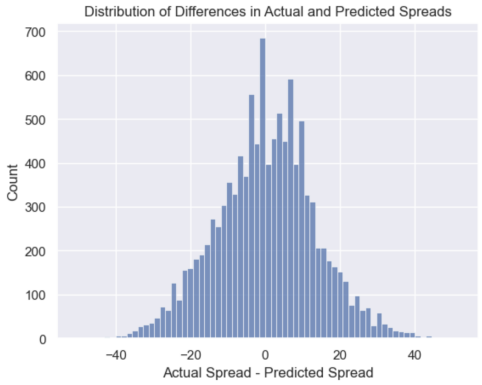
$H_0 : \mu_T \leq 0$

$H_a : \mu_T > 0$

```
tempMean = np.mean(games["actualTotal - predictedTotal"])
tempSD = np.std(games["actualTotal - predictedTotal"])
n = len(games)

zScore = np.sqrt(n) * (tempMean - 0) / tempSD
pVal = 1 - stats.norm.cdf(zScore)
print("P-Value: " + str(pVal))
```

```
P-Value: 2.069435050655599e-06
```

Spread Analysis: This histogram displays an approximately normal distribution with a mean of around 0.156, indicating that on average the difference between the actual spread and the predicted spread is close to zero. However, we do not have convincing evidence to believe the true difference in actual spread and predicted spread is zero without hypothesis testing. To get quantitative results for this relationship, we performed a hypothesis test with the null hypothesis that the difference in the actual spread and the predicted spread is equal to 0 and the alternative hypothesis being that the mean difference in actual spread and the predicted spread is not equal to 0. This yielded a p-value of 0.222 which is greater than the alpha value of 0.05, so we cannot reject the null hypothesis. Unlike the O/U difference, the difference in real and predicted spreads is not biased in one direction. This failure to reject the null underscores the simple point that oddsmakers are very good at their jobs.



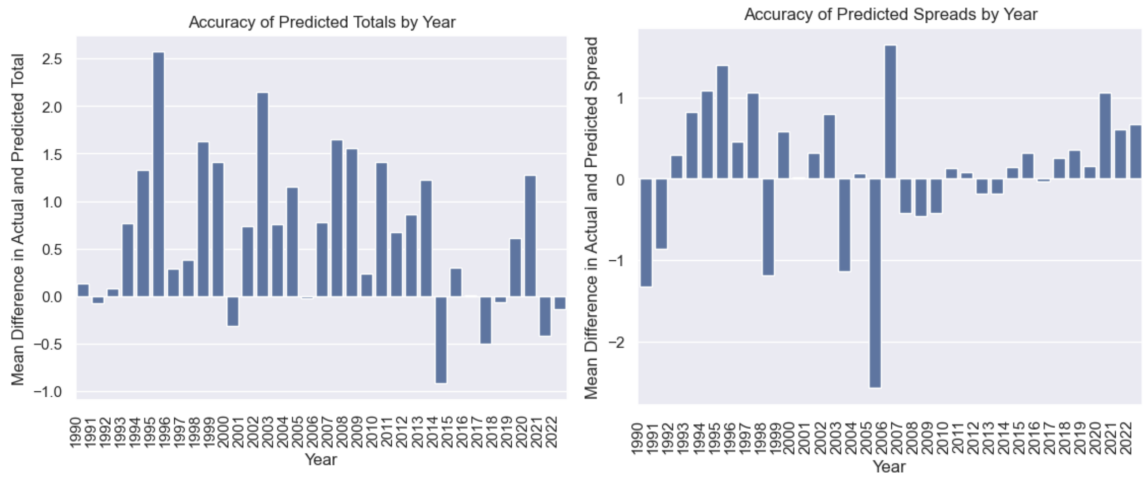
$H_0 : \mu_S = 0$

$H_a : \mu_S \neq 0$

```
tempMean = np.mean(games["actualSpread - predictedSpread"])
tempSD = np.std(games["actualSpread - predictedSpread"])
n = len(games)

zScore = np.sqrt(n) * (tempMean - 0) / tempSD
pVal = (1 - stats.norm.cdf(zScore)) * 2
print("P-Value: " + str(pVal))
```

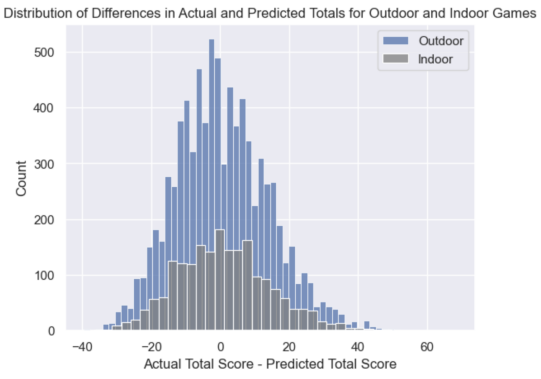
```
P-Value: 0.22207118210241372
```

Performance of Oddsmakers over Time: By grouping the games by year and calculating the mean difference in actual and predicted total score/spread, we created the following bar plots. As you can see by the visualizations, most years the mean difference in actual and predicted total was relatively positive with a few exceptions, while the mean difference in actual and predicted spread was relatively split between positive and negative values. The fact that most of the differences in actual and predicted total scores are positive provides visual corroboration of our previous finding that the true mean of this difference is greater than 0.



While the previous results confirm that the predictive accuracy of gambling lines is very strong, there does appear to be some slight inaccuracies. Therefore, we will now focus on specific factors that may affect the performance of the spread or O/U lines.

Playing Conditions Analysis: We separated the games into those played in outdoor stadiums versus those played in indoor stadiums to see whether this affected the difference between the actual and predicted O/U. The rationale here is that offenses perform better all else equal in indoor stadiums. The distribution of differences in actual and predicted total score are plotted for both sets of games with outdoor games in blue and indoor games in gray. As shown, many more outdoor games have been played, however the dispersions and central tendencies of these distributions are similar. We conduct a one-sided t-test to test the null hypothesis that the difference in actual and predicted total scores in outdoor games is less than or equal to that for indoor games.



$H_0 : \mu_{TO} \geq \mu_{TI}$

$H_a : \mu_{TO} < \mu_{TI}$

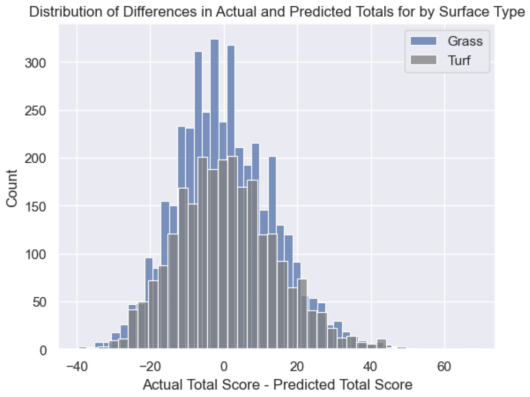
```
t_stat, p_val = stats.ttest_ind(outdoor["actualTotal - predictedTotal"],
            indoor["actualTotal - predictedTotal"],alternative='less')

print("t-statistic:", t_stat)
print("p-value:", p_val)
```

```
t-statistic: -1.6084281847657138
p-value: 0.05388607343764481
```

We fail to reject this null hypothesis at the significance level of 0.05, although barely. Under these conditions, these results corroborate the idea that there are more high scoring outlier games that far exceed their over at indoor stadiums.

Next, we run an analogous test on games played on turf and grass fields. Here, the differences in actual and predicted totals for turf games are plotted in blue and those for grass game in gray. We see that the distributions appear very similar except for the fact that more games have been played on turf. Here, we conduct the one-sided t-test to test the null hypothesis that the true difference for grass games is

greater than or equal that of turf games. The intuition here is that all else equal offenses will perform better on turf leading to more outlier games that well exceed their predicted totals.



$H_0 : \mu_{TG} \geq \mu_{TT}$

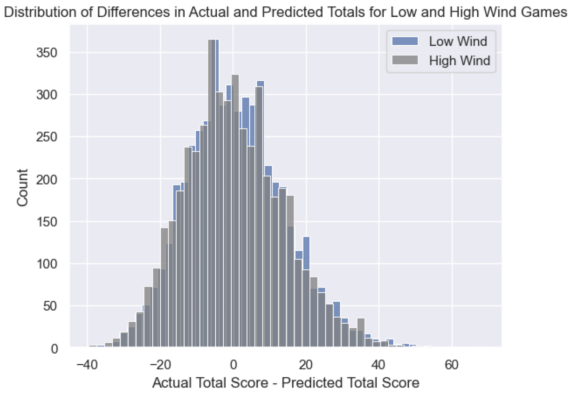
$H_a : \mu_{TG} < \mu_{TT}$

```
t_stat, p_val = stats.ttest_ind(grass["actualTotal - predictedTotal"],
            turf["actualTotal - predictedTotal"],alternative='less')

print("t-statistic:", t_stat)
print("p-value:", p_val)
```

```
t-statistic: -1.4533265506619222
p-value: 0.0730902332475748
```

We fail to reject the null hypothesis at the 0.05 significance level given the p-value of 0.07. In these ways, we do not find overwhelmingly strong evidence that there is a difference in oddsmakers accuracy of O/U lines depending on whether the game is indoors or outdoors or played on turf versus grass. It is also worth noting that this data has a hidden time component as well. As time has passed, more and more stadiums have transitioned to turf fields from grass fields.

Subsequently, we conduct an analogous analysis on wind type. Here, we tag each game as either low wind or high wind game depending upon whether the wind measured during the game is less than or equal to the median wind speed in the dataset. Below, the low wind games are plotted in blue and the high wind games are plotted in gray. Although the distributions appear similar, the mean difference in actual and predicted O/U for low wind games is more than four times larger than that for high wind games. This implies that there are more outlier high scoring games in low wind conditions where all else equal offenses are more successful. Thus, we test the null hypothesis that the mean difference for low wind games is less than or equal to that for high wind games.



$H_0 : \mu_{TL} \leq \mu_{TH}$

$H_a : \mu_{TL} > \mu_{TH}$

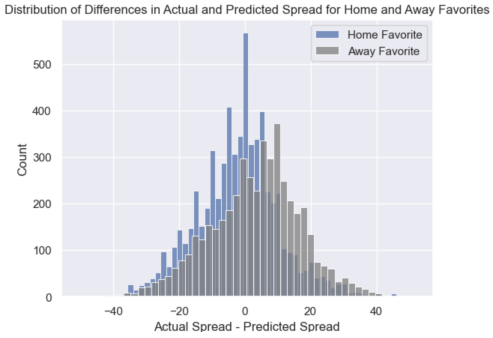
```
t_stat, p_val = stats.ttest_ind(wind_low["actualTotal - predictedTotal"],
    wind_high["actualTotal - predictedTotal"],alternative="greater")

print("t-statistic:", t_stat)
print("p-value:", p_val)
```

```
t-statistic: 2.9027281184909532
p-value: 0.0018536686451014301
```

We reject this null hypothesis at the significance level of 0.05 given the p-value of approximately 0.001. We can say that there is less than a 0.2% chance we would observe this data if the null hypothesis is true. In this way, if one's strategy is to bet “overs” we suspect all else equal that it is more advantageous to bet overs in low wind games compared to high wind games.

Home and Away Favorite Analysis: Homefield advantage has a big impact on a game’s spread. Sometimes the ‘worse’ team may be slightly favored because they are home. To analyze the extent to which this homefield adjustment is justified we split the data into two dataframes: one which includes

games where the home team is favored and the other which includes games that the away team is favored. We plot the distributions of the differences in actual and predicted spreads from these two dataframes below where the games where the home team is favored are plotted in blue and those for the away team in gray. As shown, the home team is favored more often than not. However, the mean for the home distribution is approximately -2 while the mean for the away distribution is approximately 3. Thus, in this dataset home favorites on average under perform the spread by 2 points while away favorites on average over perform the spread by 3 points. One can observe that the home favorite distribution has a fatter left-hand tail while the away favorite distribution has a fatter right-hand side tail. While these raw observed means suggest that oddsmakers overweight home field advantage, we conduct a two-sided t-test to determine whether or not there is a true difference between the true means for both samples.



$H_0 : \mu_{SH} = \mu_{SA}$

$H_a : \mu_{SH} != \mu_{SA}$

```
t_stat, p_val = stats.ttest_ind(home_fav["actualSpread - predictedSpread"],
                                away_fav["actualSpread - predictedSpread"])

print("t-statistic:", t_stat)
print("p-value:", p_val)
```

```
t-statistic: -21.80868175977445
p-value: 2.959922599494843e-103
```

Here, we can reject the null hypothesis that the true means are the same with very strong evidence and state that there would be well less than a 1% chance we observe data this extreme if the null hypothesis was true. Digging deeper we want to test whether the home true mean is negative and the away true mean is positive, so we conduct two one-sided hypothesis tests to test these alternative hypotheses.

$H_0 : \mu_{SH} >= 0$

$H_a : \mu_{SH} < 0$

```
tempMean = np.mean(home_fav["actualSpread - predictedSpread"])
tempSD = np.std(home_fav["actualSpread - predictedSpread"])
n = len(home_fav)

zScore = np.sqrt(n) * (tempMean - 0) / tempSD
pVal = stats.norm.cdf(zScore)
print("P-Value: {:.5f}".format(pVal/2))
```

```
P-Value: 0.00000
```

$H_0 : \mu_{SA} <= 0$

$H_a : \mu_{SA} > 0$

```
tempMean = np.mean(away_fav["actualSpread - predictedSpread"])
tempSD = np.std(away_fav["actualSpread - predictedSpread"])
n = len(away_fav)

zScore = np.sqrt(n) * (tempMean - 0) / tempSD
pVal = 1 - stats.norm.cdf(zScore)
print("P-Value: {:.5f}".format(pVal/2))
```

```
P-Value: 0.00000
```

We can reject both null hypotheses. Accordingly, we have strong evidence that oddsmakers on average overestimate the performance of home favorites and underestimate the performance of away favorites. These findings are very notable and form a strong foundation from which a more nuanced data-driven gambling strategy could develop from.

Part 5: Limitations and Future Work

While we have discovered some relevant trends in the performance of gambling lines through certain factors, we must recognize that our work assesses these individual factors (wind, stadium type, field type, home/away) separately to reveal extremely pointed results that cannot be aggregated to develop a “one size fits all” gambling strategy. We must also note that the purpose of oddsmakers is not exactly to predict the exact outcome, but rather to generate lines such that the public evenly distributes their money on both sides of the line. Therefore, the work of oddsmakers is as much about public perception as it is accurately predicting the spread or total score. While these two things are often highly correlated, they aren’t always equivalent.

Further exploration of our research topic could greatly benefit from more specific data associated with each game played. Whether this be injury reports, specific player data, or matchup performance history, we could extract much more meaningful results with these metrics. These are all hidden variables with our current dataset and would be useful factors to consider with respect to the accuracy of lines. The limitations of our project regarding aggregation of our individual results could be addressed by further filtering our data into groups that contain certain values for relevant factors (wind, stadium type, field type). After doing this, it would be easy to reveal which combination of these factors yields the best/worst gambling line performance.

Part 6: Conclusion

By analyzing trends in the differences between actual and predicted spread/total scores over time and across various factors, we were able to draw statistically significant conclusions regarding our overarching research question. In general, almost all of our histograms were approximately normal with a mean just above zero, with the exception of indoor stadiums vs. actual - predicted total score and low wind speed vs. actual - predicted total score, that had means around one. Our first major conclusion was that the mean difference in actual and predicted total score is greater than zero, meaning that on average, oddsmakers tend to underestimate the total score of NFL games. First we tested the difference in actual and predicted total score against stadium type and found that we did not have statistically significant evidence that the difference for indoor stadiums is greater than for outdoor stadiums. We then tested for turf vs. grass field types and also found a lack of statistically significant evidence to prove that the difference in actual and predicted total score is higher for turf fields in comparison to grass fields. For the next test we found convincing evidence that the difference in actual and total score is greater for games with low wind in comparison to games with high winds. The first test regarding home and away favored games provided statistically significant evidence that the mean difference in actual spread and predicted spread is not equal for home and away favored games, meaning that the accuracy of oddsmakers in predicting the difference of scores for two teams in a particular game is affected by whether the game is home-favored or away-favored. The last two tests provide statistically convincing evidence that the mean difference in actual and predicted spread is less than zero for home-favored games and the mean difference in actual and predicted spread is greater than zero for away-favored games.

Ultimately, through significance tests, our group was able to provide convincing evidence that the factors of wind speed and whether the game is home-favored or away-favored have a statistically significant effect on the accuracy of oddsmakers in creating predictions for total score and spread. Sports gamblers can exploit these miscalculations made by oddsmakers by betting the over for spread in away-favored games and betting the under for spread in home-favored games. Although we did not find convincing evidence that all factors had an impact on the accuracy of oddsmakers in making lines, there is some indication there may be some relationship with factors such as stadium and field type.

Link to Google CoLab

https://colab.research.google.com/drive/1kXfDLJYVcFQ2xFONI5tmub3hgT8tx3Gb?usp=sharing