Group 2 - Project 1

# NFL Betting Data

Gage Anderson

Brandon Ibarra

Quintele Jackson

## Introduction

Our group chose to do an exploratory data analysis over NFL Betting Data, with the data being found on Kaggle. The overall data consisted of three datasets: all NFL teams that have existed since 1960, all NFL stadiums used since 1960, and all NFL scores for every game played since 1960. Information given in these datasets ranged from stadium type and stadium weather to the spread of the game and the over/under.

Most NFL betting data revolves around two concepts: spread and over/under. The over/under (O/U) is a number that the betting markets decide on that takes the predicted scores of the home and away teams and combines them. For example, an O/U of 46.5 means that the betting market thinks that Team A and Team B will combine for around 46 points. The over would be over 46.5 combined, and the under would be under 46.5 combined.

The spread is the amount of points the favored team is projected to win by. There are ways for the non-favored (underdog) team to cover the spread, but this analysis will just focus on the favored team. An example of the spread in action would be the Cowboys vs Vikings game that occurred on 11/20/2022. The spread was Cowboys by 2, written as -2. The final score was Cowboys 40 Vikings 3, with the score differential being in the Cowboys favor by 37. 37 is greater than 2, so the Cowboys covered the spread.

With datasets being merged together on a case by case basis, our research questions were as follows:

- How often do favored teams win?
- Do different types of stadiums have more average points scored?

## Data Breakdown and Cleanup

There were over 30 NFL stadiums, 32 NFL teams, and over 4500 games in the data that were analyzed from 2000 to 2019. Our sample is around a quarter of all NFL games ever played, and around 40% of all NFL games that have betting data associated with them.

For the NFL Stadiums dataset, stadiums that simply had name changes over the years were consolidated under whatever the most recent stadium name was. The close and open dates were converted from floats to integers. The stadium address and zip code were dropped, as well as the field type due to 43% of stadiums having a NULL value for this column. For the NFL Teams dataset, teams that didn’t play a game between 2000 and 2019 were dropped.

For the NFL game data, the biggest shortening of the data was only including games from 2000 to 2019. Neutral site and playoff games were also dropped. Both score columns had to be converted from floats to integers. The home team and away team were converted from their long form names (ie: Dallas Cowboys) to their team ids from the Teams table (ie: DAL).

## How often do favored teams win?

Through the merged data of a list of all NFL Stadiums used, excluding neutral sites, a list of all 32 NFL teams, a list of all NFL games, excluding neutral site games and playoffs, including the scores, the favored team, the spread, and the over/under it was determined that favored teams win 66.1% of the time and lose 33.9% of the time.In total there were 5061 games played. Out of those 5061, 3347 favored teams won while 1714 favored teams lost. 20 games were defined as Pick which meant no team was favored.

There was much more data that could have been analyzed that most likely could have changed the percentage of favored teams that won and lose,The data that i feel could have changed it was the weather because it plays a huge role in how a team plays as well as stadiums that were in a dome that could avoid harsh weather.The scores on average are different but not by much which is surprising since I thought weather would be a big factor in football. Indoor scored more on average which is expected but not by much.

In Our data we were able to analyze the data needed to help us know how often do favored teams win as well as understanding the data that was given to help us find out the percentage as well as a list from least to greatest of teams who have won the most being the favored team from the patriots being the greatest while cleveland being the least with only 20 while the patriots have 197.The datasets that we used had a huge part in helping us analyze this data.

## Logistic Regression Analysis

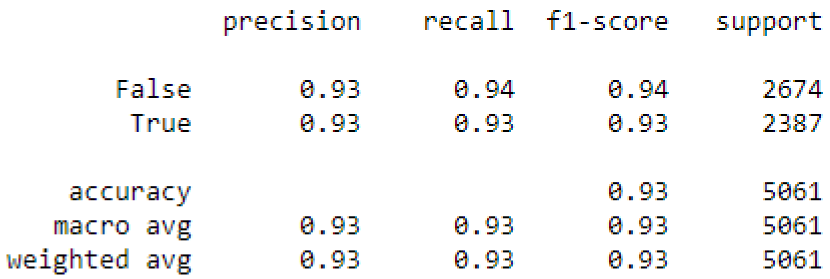
Logistic Regressions work well when the dependent variable is dichotomous, or can only be one of two values. Because we wanted to find how likely a favored team would cover the spread as a true/false boolean, we determined that using a logistic regression would better fit our data than a linear regression.

To prep the data for regression, additional columns were needed:

- regression_df["home_favorite"] = regression_df.team_favorite_id == regression_df.team_home
- regression_df["home_won"] = regression_df.score_home > regression_df.score_away
- regression_df["home_ties"] = regression_df.score_home == regression_df.score_away
- regression_df["home_loss"] = regression_df.score_home < regression_df.score_away
- regression_df["score_diff"] = abs(regression_df.score_home - regression_df.score_away)
- regression_df["favorite_won"] = (regression_df.home_favorite & regression_df.home_won) | (~regression_df.home_favorite & regression_df.home_loss)
- regression_df["spread_covered"] = regression_df.favorite_won & (regression_df.score_diff > abs(regression_df.spread_favorite))

We then set the features for our regression. The following data was included: year of game, the spread, if the home team was the favorite, and who the favored team was. To convert who the favorite team was to a boolean flag, we used pd.get_dummies on the feature. This created extra columns such as team_favorite_id_ARI, team_favorite_id_ATL, etc., to assist with our predictive model later. Our target column was called “Spread Covered.”

RandomForestClassifier was used for our regression model. Random Forest's goal is to split the data into subsets and sort them through decision trees to try and find patterns to fit the data.

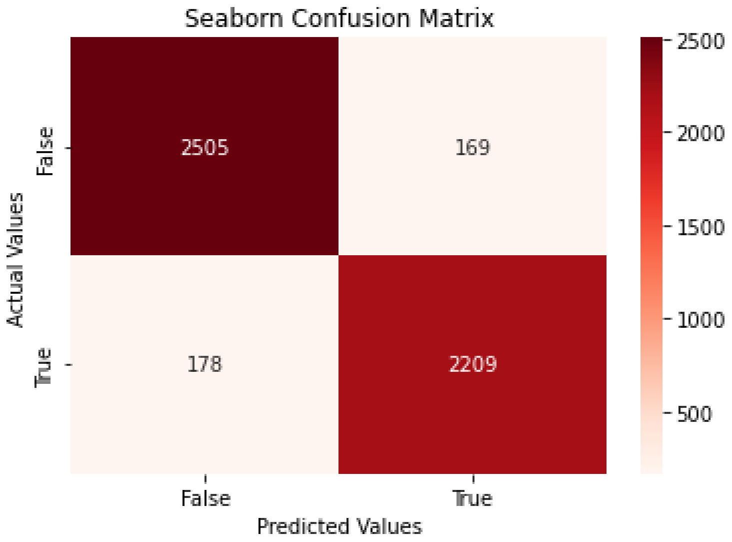


```
              precision    recall  f1-score   support

       False       0.93      0.94      0.94      2674
        True       0.93      0.93      0.93      2387

    accuracy                           0.93      5061
   macro avg       0.93      0.93      0.93      5061
weighted avg       0.93      0.93      0.93      5061
```

As our classification report shows, our model correctly predicted if the favorite team would cover the spread 93% of the time.



Going off the correct values in our confusion matrix, the favored team did not cover the spread 53% of the time, which lines up with our real life data that showed the favored team didn't cover 52% of the time. A feature in our regression that could be used in the future is the ability to set the values of the feature columns to predict whether a team could cover.

## Limitations

A massive limitation of our dataset was the lack of weather data which could have played a big impact in our analysis that one of group members would have done regarding if any stadiums have a bigger home field advantage than others, in addition to our analysis on average points scored in different stadium types. The weather could've worked with the previously dropped "Game Date" column to also detect seasonality. If we had more time, an analysis on the O/U data could have been interesting to delve into as well. For the regression, keeping track of a teams win streak in general and at home could have had an impact on how successful the model was.

## Conclusion

To answer our research questions:

- Favored teams win around 66.1% of the time. In betting terms, you are more likely to win money if you bet on the moneyline (betting on the favored team).
- It appears that different stadiums have more average points scored than others. Indoor stadiums on average score more total points, but the difference is marginal.

The predictive model was highly accurate, and may have been too accurate if the RandomForestClassifier identified "patterns" that weren't actually patterns. Running the model again on all games with betting data (20,000+ games) could determine if our model was, for lack of a better term, accurately accurate. If this model is proven accurate, it could be used by sports bettors to better enhance their chances at making successful bets.