NFL Sports Betting

Sports betting has increased dramatically in the last few years. In May 2018, the Supreme Court ruled that the 1992 federal ban on sports betting in most states violated their rights. Since then, 22 states have legalized sports betting and more are looking into legalizing it. Sportsbooks are amazingly adept at providing numbers where 50% of bettors choose each side of a bet. Then they profit by taking a small amount from each bet. The object of this project is to use data to predict a winner more than 50% of the time which will give a slight profit over randomly choosing a bet.

Data

Historical season data is found on Kaggle. Current season data is found on TeamRankings. Current season data needs to be added as the season progresses to provide the information needed for the next weeks games.

Kaggle

TeamRankings

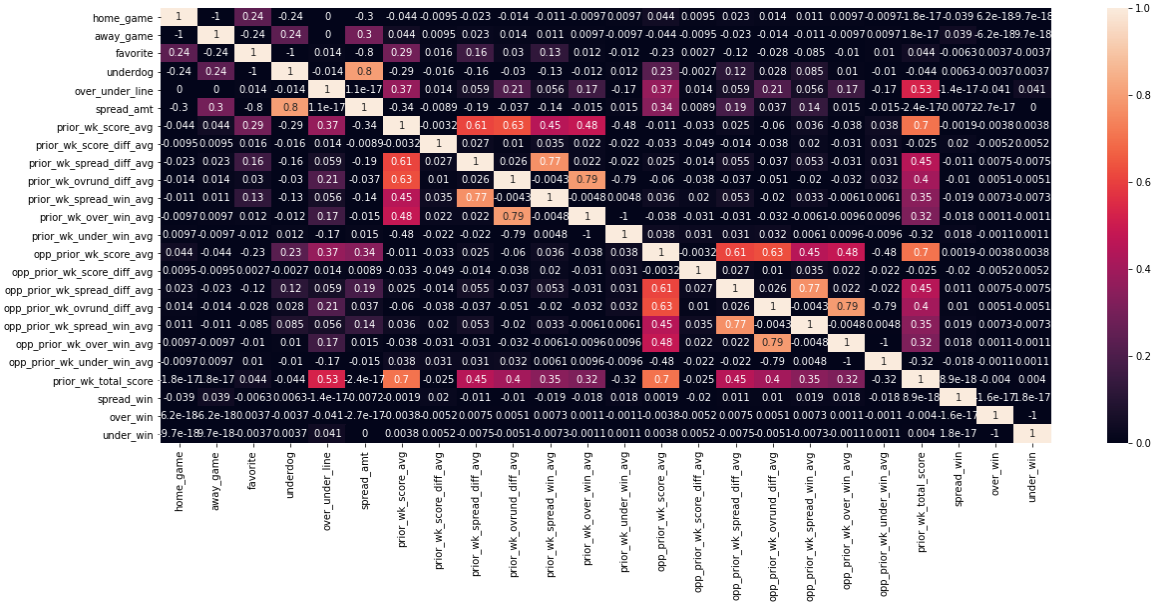
Data Wrangling

1. Many columns from the Kaggle dataset are not needed and can be dropped.
2. Older data does not seem to pertain to the current team so data prior to 2015 is dropped.
3. Current season data needs to be rebuilt each week incorporating the prior week's data.

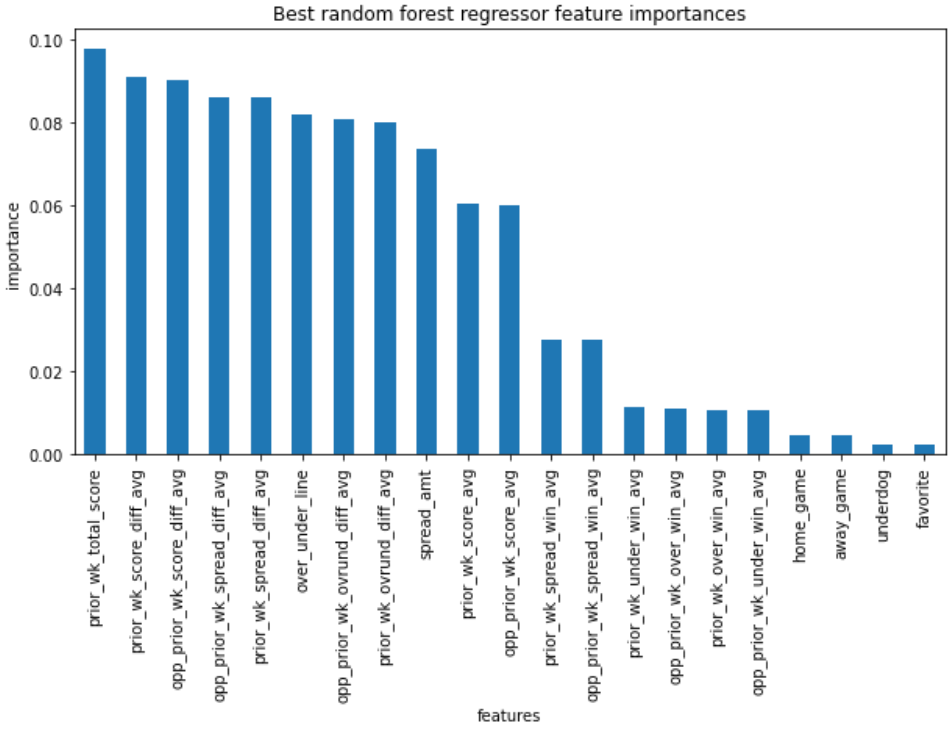
Exploratory Data Analysis

1. over_win: Team that wins an over bet.
2. under_win: Team that wins an under bet.
3. over_under_diff: Amount the actual scores are over or under the over/under line.
4. spread_diff: Amount the team's score differs from the spread.
5. Prior week stats: Recent data may make more of a difference in the prediction than older data. Prior week stats take a rolling average of the last 3 games.

6. A heatmap of the features didn't provide much help in determining useful features.



7. Random Forest Regressor feature importance shows prior week rolling stats may be useful.



Modelling

For training, I used years 2015-2018 data, and for testing, I used 2019-2020 data. Linear Regression and Random Forest Regressor both provided similar less than satisfactory results.

	Accuracy
Linear Regression	49.5%
Random Forest	49.3%

Summary

Unfortunately, both models returned results less than 50%. This would not help a person beat the odds and should not be used. Additional features and modelling should be done to provide better results.