Predicting outcomes in College Football games

Team members: Nate Clause and John Vanderhoff
GitHub: `https://github.com/JohnVanderhoff/college_football_prediction_project/`

Overview College football is more popular than ever, with games in 2023 broadcast on an ESPN platform averaging 1.7 million viewers per game. Further, the introduction of NIL deals and the expansion of legal sports betting across the US has expanded the amount of money in the sport. In particular, the average college football program has a revenue of $31.9 million per year, and an estimated $8 billion per year is bet on the sport (at least legally).

Motivated by the above, we want to develop a predictive model(s) for the outcomes of future college football games. We develop models to predict many of the core statistics pertaining to the outcome of a given game, all the way from a given team's rushing yards or penalties up to the final score and the winner.

Stakeholders: sports betting industry, TV networks, the individual teams, college football fans.

KPIs: accuracy of predicting the winner, mean average value of score prediction error.

Approach: There were over 50 core stats that went into the model, ranging through offense, defense, and special teams, all sourced from `https://collegefootballdata.com/`. We initially developed three regression models, "model 1" which takes in as input the core stats for each team in a game over their past 12 games and tried to predict all of the core stats for each team in the coming game. The best performing model here utilizes **ElasticNet Regression**. Next is "model 2" which takes in the given core stats of a game, with direct scoring stats removed, and predicts the final score. The best performing model here utilizes **Lasso Regression**. Lastly is "model 3", which takes in the core stats for each team in a game over their past 12 games and tries to predict just the final score. The best performing model here utilizes **Lasso Regression**.

After testing these models, we noted model 1 and model 3 not achieving desirable performance. As a result, we developed direct analogues of models 1 and 3 using a **long-short-term-memory (LSTM) deep learning neural network**. The LSTM network is a type of recurrent neural network which has demonstrated particular value in studying time-series data, which we are looking at as we input recent historical data of team's core stats into these models.

Results: We survey the results of our best models, for more details on other models see our GitHub. Our Lasso Regression model for model 2 to predict the score in a game given the realized stats was on average < 0.02 points away from the actual score which occurred in the game, and accurately predicted the winner 100% of the time. This shows that given the realized stats, we have an incredibly accurate way to predict the game outcome, so teams can use this to determine what they "need to achieve" stats-wise in order to win a game.

Our LSTM deep learning model for model 3 to predict the score based on historical stats was on average ≈ 3.45 points away from the actual score which occurred in the game, and accurately predicted the winner $\approx 90.5\%$ of the time. This is notable given that the favorites as determined by the Las Vegas sportsbooks win only $\approx 77\%$ of the time.

Future directions: Future work focuses on making improvements to models 1 and 3. There are more core stats and features we can add that we did not either for difficulties pertaining to time or finding a source, such as average field position and weather. The most notable change that would improve our model is a notion of overall opponent quality. When models 1 and 3 were notably off usually occurred when smaller schools who consistently put up high stats against weak opposition played against large schools who consistently put up mediocre stats against elite opposition. If we are able to inject a consistent measure(s) of strength of opposition into these models, it would likely substantially improve performance.