Predicting NFL Outcomes

Jack Huffman and Dave Nguyen

ABSTRACT:

This project aims to predict the outcomes of NFL games based on various team performance metrics using machine learning techniques. We gathered game data from the last 10 NFL seasons for all 32 teams, recording 136 different data points from each game. We combined this data into a singular file which recorded 19 of the most important performance statistics and organized the data by home team and away team. A neural network was developed and trained on this dataset to learn patterns associated with winning teams. Our model was evaluated on a test set of unseen data and consistently showed around 95% accuracy at predicting the winning team.

INTRODUCTION:

On May 14, 2018 the United States Supreme court issued a decision which struck down a federal ban on state authorized sports betting, since then 38 states have legalized sports betting in some fashion. This has resulted in a massive increase in the market cap for sports betting and has become an extremely lucrative industry for professional sports books. These sports books drive their predictions off complex machine learning algorithms and are backed by teams of professional data scientists. While sportsbooks keep their methods behind closed doors, you can find many open source examples of NFL predictions on the internet. Some of these use linear regression, such as Austin Streitmatter's implementation[1] or Stephen Bouzianis's implementation[2], and other implore more complex methods such as using a neural network. Both of these are tailored to predicting NFL games but there is also significant work for predicting other sports that served as a excellent resource for us, such as this paper[3] detailing an implementation of neural network to predict English Premier League outcomes.

Our goal with this was to build a machine learning tool which could produce accurate win predictions and be used to singlehandley bankrupt each and every professional sportsbook. Our goal is to have a neural network which accepts various statistical data points from an NFL game, separated by home and away team, and outputs a value between 0 and 1 which represents the probability that the home team wins.

Dataset:

DATA CAN BE FOUND HERE:
https://drive.google.com/drive/folders/1MSmBLEbfuA0UCAK5KkbVwL--MwrhzPpo?usp=drive_link

[1] https://www.samford.edu/sports-analytics/fans/2023/How-I-Built-a-Competitive-NFL-Prediction-Model-with-Only-Five-Statistics

[2] https://scholars.unh.edu/cgi/viewcontent.cgi?article=1472&context=honors

[3] https://arxiv.org/abs/2307.13807

You can also find the data as part of the zip folder

For our data, we gathered it all from stathead.com which has an extensive collection of NFL data on every season from 1920 to the present. We gathered game data from the last 10 NFL seasons for all 32 teams, recording 136 different data points from each game. The specific data points we collected for all 32 teams are as follows (each teams CSV file with this data can be found in the 2014-2023teamData folder):

Game Order,Team,Date,Passing Yards,Passings TD,Rushing Yards,Rushing TD,Turnovers,Opponent Passing Yards,Opponent Rushing Yards,Time of Possession,Day of Week,Game Number,Week,Home/Away,Opponent,Result,Passing Completions,Passing Attempts,Passing Incompletions,Passing Completion Percentage,Passing TD,Interceptions,TD Percentage per Pass,Interception Percentage per Pass,Passer Rating,Times Sacked,Sack Yards,Sack Percentage,Yards per Attempt,Net Yards per Attempt,Adjusted Yards per Attempt,Adjusted Net Yards per Attempt,Yards per Completion,Opponent Rushing Attempts,Opponent Rushing Yards per Attempt,Opponent Rushing TD,Opponent Pass Completions,Opponent Pass Attempts,Opponent Completion Percentage,Opponent Passing TD,Opponent Times Sacked,Opponent Sack Yards,Opponent Interceptions,Opponent Passer Rating,Rushing Attempts,Total Yardage,Offense Number of Plays,Yards per Offensive Play,Defense Number of Plays,Yards Allowed per Defensive Play,Turnovers Lost,Total Time,Opponent Turnovers,Penalties,Penalty Yards,Opponent Penalties,Opponent Penalty Yards,Field Goals Made,Opponent Field Goals Made,3rd Down %,Oppoenent 3rd Down %,1st Downs,1st Downs by Rush,1st Downs by Pass,1st Downs by Pen,3rd Down Attempts,3rd Down Conversions,3rd Down Conversion %,4th Down Attempts,4th Down Conversions,4th Down Conversion %,Opponent Total Yards,Opponent Turnovers.1,Penalties.1,Penalty Yards.1,Opponent Penalties.1,Opponent Penalty Yards.1,Combined Penalties,Combined Penalty Yards,Opponent TD,Opponent XPA,Opponent XPM,Opponent Field Goals Attempted,Opponent Field Goals Made.1,Opponent Safeties,Total Yards,Offensive Plays,Yards per Play,Defensive Plays,Defensive Yards per Play,Turnovers.1,Time,Opponent 1st Downs,Opponent 1st Downs by Rush,Opponent 1st Downs by Pass,Opponent 1st Downs by Penalty,Opponent 3rd Down Attempts,Opponent 3rd Down Conversions,Opponent 3rd Down Conversion %,Opponent 4th Down Attempts,Opponent 4th Down Conversions,Opponent 4th Down Conversion %,Total TD,XPA,XPM,Field Goals Attempted,Field Goals Made.1,2PA,2PM,Safeties

We then took all these CSV files and produced a combined data set (in the shared data folder as 'combined_team_data.csv') which has a record for every game in the last 10 seasons with statistics sorted by Home and Away team. The values that we stored in this file are as follows:

Date,Week,

Home Team Name,Away Team Name,
Winner,
Home Team Total Yards,Away Team Total Yards,
Home Team Passing Yards,Away Team Passing Yards,
Home Team Passing TD,Away Team Passing TD,
Home Team Rushing Yards,Away Team Rushing Yards,
Home Team Rushing TD,Away Team Rushing TD,
Home Team Turnovers Lost,Away Team Turnovers Lost,
Home Team Passing Attempts,Away Team Passing Attempts,
Home Team Passing Completion Percentage,Away Team Passing Completion Percentage,
Home Team Times Sacked,Away Team Times Sacked,
Home Team Rushing Attempts,Away Team Rushing Attempts,
Home Team Rushing Yards per Attempt,Away Team Rushing Yards per Attempt,
Home Team Penalties,Away Team Penalties,
Home Team Field Goals Attempted,Away Team Field Goals Attempted,
Home Team Field Goals Made,Away Team Field Goals Made,
Home Team 3rd Down Attempts,Away Team 3rd Down Attempts,
Home Team 3rd Down %,Away Team 3rd Down %,
Home Team 4th Down Attempts,Away Team 4th Down Attempts,
Home Team 4th Down Conversion %,Away Team 4th Down Conversion %,
Home Team 1st Downs,Away Team 1st Downs

Excluding game stats such as date, week and team names we had 19 statistics of interest recorded for both teams (38 total). For our model training and testing we did a 80-10-10 split, that is 80% training, 10% validation and 10% testing.

Methods:

For our project we used a sequential neural network with 5 dense layers with 20, 80, 160, 80 and 1 layers respectively. For the first 4 layers we used a Relu activation function and for the last layer we used a sigmoid activation function (as we want to express our result as a probability). We experimented with changing a variety of different parts of our model and systematically testing them with iterative loops. We found that after 5 or more layers we didn't see a significant enough increase in accuracy to warrant the increased training time the model took. Similarly we experimented with changing activation functions but came to a similar conclusion that Relu achieved the goals we were looking for. Changing from 4 to 5 layers we also changed the structure of our neural network in terms of how many nodes were in each dense layer. We did some experiments and ended up finding that 20-80-160-80-1 gave good resultss and was sufficient to identify complex trends hidden in our data.

As far as our data goes we chose these specific 19 statistics for a variety of reasons. Some of them seemed to intuitively represent significant aspects of a game, such passing yards and turnovers. Some other statistics provided some key insights into the efficiency of the teams that were playing, such as third and fourth down conversion percentages. An important factor that we considered when selecting all of our data points was that there is no gaps in the data. Some advanced statistics weren't recorded until relatively recently and as such not all teams have complete data, as such we declined to incorporate these into the building of our model as we wanted a too which could be generalized to a wider range of data. Additionally, we saw a high enough degree of accuracy in our testing that we felt comfortable with the set of features we landed on.

When evaluating our model, we consistently saw around 95% accuracy at predicting the outcomes of game in our test set. At one point we were concerned that given a winner is determined based off points scored that if we provided the model with all the methods of scoring points it would become biased to that data and use it only to predict outcomes. Because of this we tested our model while excluding TD's and field goals but saw no significant drop in testing accuracy. To further evaluate our model we asked ourselves how this performs against some more rudimentary prediction approaches, such as simply picking the team with more total yards. Doing this approach

Discussion and future work:

In order to accurately compute predictions for real time games, we want to do some work with how we are computing our predictions. Our initial idea was to simply use the average values of teams in game leading up to the one in question (from the same season) but there are certainly better ways to do this. One idea would be to use some type of conditional probability for what we expect particular statistics to be, one example of this could be conditioning our expectation of a teams passing yards as a function of the number of passing yards the team they are facing has let up on average. Taking this a step further, instead of conditioning this on the average number of passing yards let up by the team they are facing we could condition it on some passing defense statistic which itself is calculated not as an average but as a weighted value that is also conditioned on the strength of the offenses that defense has played in its prior games. I give this example to illustrate the rabbit-hole we can go down to create more and more accurate predictions of game statistics. Ultimately, the tool we have developed still does a good job (if given accurate statistical values) at predicting the winner of a game.