Using One Dimensional Convolutional Neural Networks to Predict Sports Outcomes

Logan Noonan

Fall 2021

Abstract

Predicting the outcome of a sports match before it has even begun seems nearly impossible, yet millions of people every year make predictions like this, and with high stakes. This project aims to implement a 1D-Convolutional neural network (1D-CNN) to predict the outcome of NFL games. Match data from nearly every game spanning 10 seasons, 2009 through 2018, was collected to implement the deep learning model. After cleaning and pre-processing the data it was split into training, validation, and testing sets to verify it could generalize to new data. Our deep learning model was built using Python's deep learning library Tensorflow. The model was able to achieve 60% testing accuracy.

Contents

List of Figures

Introduction

The problem we want to solve is not contained within the scope of just a few seasons so we obtained play by play data over an entire decade of NFL games. The data provides an in depth breakdown of every NFL game from 2009-2018. The data was originally obtained via a web-scraping program written in R. The program gathered game data on a play-by-play level for each game, every season. This allowed us to explore statistics relative to individual games, individual teams, unique match ups, and entire seasons. The data was originally collected for the same purpose we are using it or now, to better analyze and predict the outcomes of NFL games and can be accessed at https://www.kaggle.com/maxhorowitz/nflplaybyplay2009to2016?select=NFL+Play+by+Play+2009-2018+%28v5%29.csv.

Preparing the Data

The original data set was far to in depth to effectively train a deep learning model on a personal computer. There were 255 columns and over 400,000 entries. Conveniently, many of the columns of the original data set were not thoughtfully named nor well documented hence, they were useless for our purposes. Additionally, the data needed to be reduced row-wise.

Data Cleaning

The following list describes the steps taken to clean the data.

- The data contained information on 35 teams but there are only 32 in the NFL. Some team name abbreviations had changed over time, but they're still the same team.
- There were some data entry errors such that the home team was also the away team. Clearly this is an error and these rows were removed.
- Some columns were nearly completely empty i.e., they contained actual data in only a small fraction of their entries. These columns needed to be dropped.
- There were various instances of repeated columns in the sense that they provided the exact same information, just in a different way. Still, since these columns could be derived from one another there was no need to keep them all.

Data Reduction

The following list describes the steps taken to reduce the data.

- As previously mentioned many of the columns were uninterpretable and so were removed.
- We removed the few games that ended in a tie since we only want data on strict wins.
- A single column was created to store unique game ids. This way we could summarize the data of a particular game as a single row entry.
- The rows were then reduced again slightly by summarizing the data of the last three games, for each team, as a single row. Thus, the model will make the win/loose prediction based on results from the last three games respective to each individual team.

Data Transformation

The following list describes the steps taken in transforming the data.

- Home and away team final scores are determined from play-by-play score information.
- Various column attributes are separated and summed together to to provide a single number game attribute description for the home team and the away team, as separate columns. Examples of such columns include: *Total 4th down conversions*, *Total 3rd down conversions*, *Total interceptions thrown*, *Defensive sacks*, *Penalty yards accumulated* and *QB pass completion percentage*, among several others.
- In order to indicate which team won a response column was created with value = 1 if home team wins otherwise 0.
- There are 32 teams total, they were mapped to a numeric value, in the range of 0-31.
- Information was extracted to create a feature column containing the ratio of sums of last seasons wins for the home team vs the away team. Since 2009 is the first year I will just set every value in that year to 1. The purpose of this column is to compare the previous years standings of the home and away team each game.
- All of the data was scaled using *min-max-scaling*.

Visualizing the Data

Now that the data has been fully cleaned and processed we can explore and visualize it. The correlation matrix in Figure 1. is useful in exploring how the predictor variables are related to the response variable. Figures 2 and 3, respectively, verify our assumptions that the team with larger completion percentage, and the team who gains the most yards will generally win the game. Figure 4 tells us that if the away team doesn't have any recent wins, then the home team is very likely to win regardless of their number of recent wins. Similarly, if the home team has won all their recent games and the away team has also won all their recent games, the home team is still slightly more likely to win. And this is what we would expect and define as home-field advantage. Figures 5 and 6 show somewhat unexpected trends. Third Down conversions are less important to a win when comparing home vs away, as the home teams have their greatest density of wins at a lower conversion rate than the away teams. Similarly, the away teams have their greatest density of wins at a lower conversion rate than the home teams. Similar to 3rd down conversions, Penalty yards are a slightly more important indicator regarding who will win the game. For the home team the greatest density occurs when they have less penalty yards than the opponent. This is not the case for the away team.

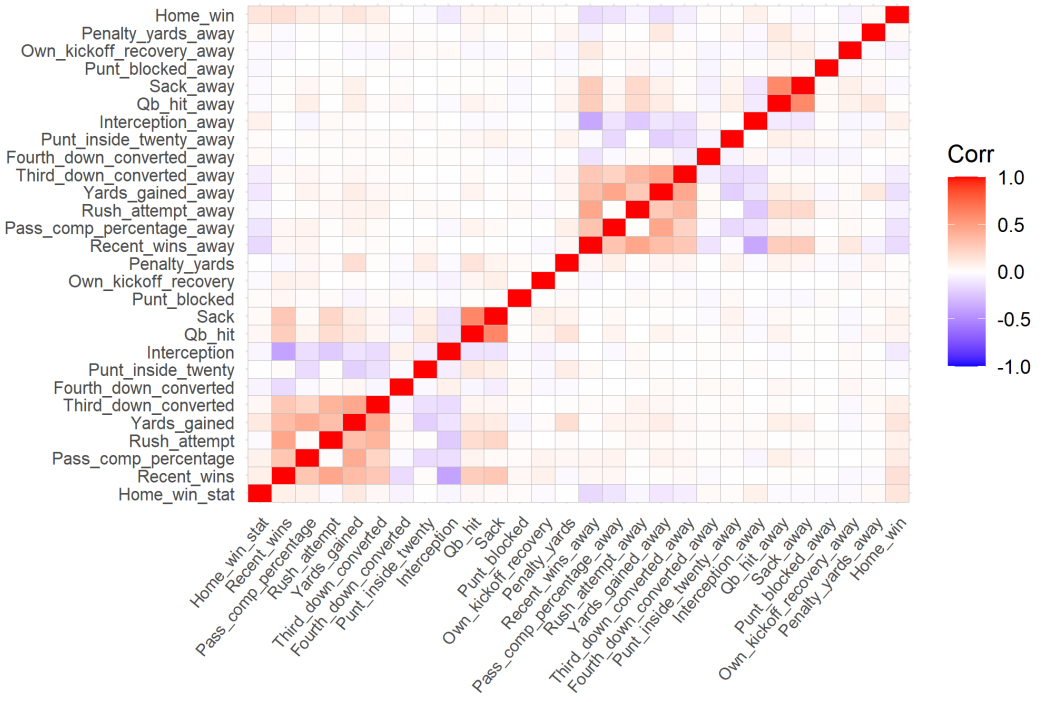

Figure 1: *NFL Data Correlation Matrix.*

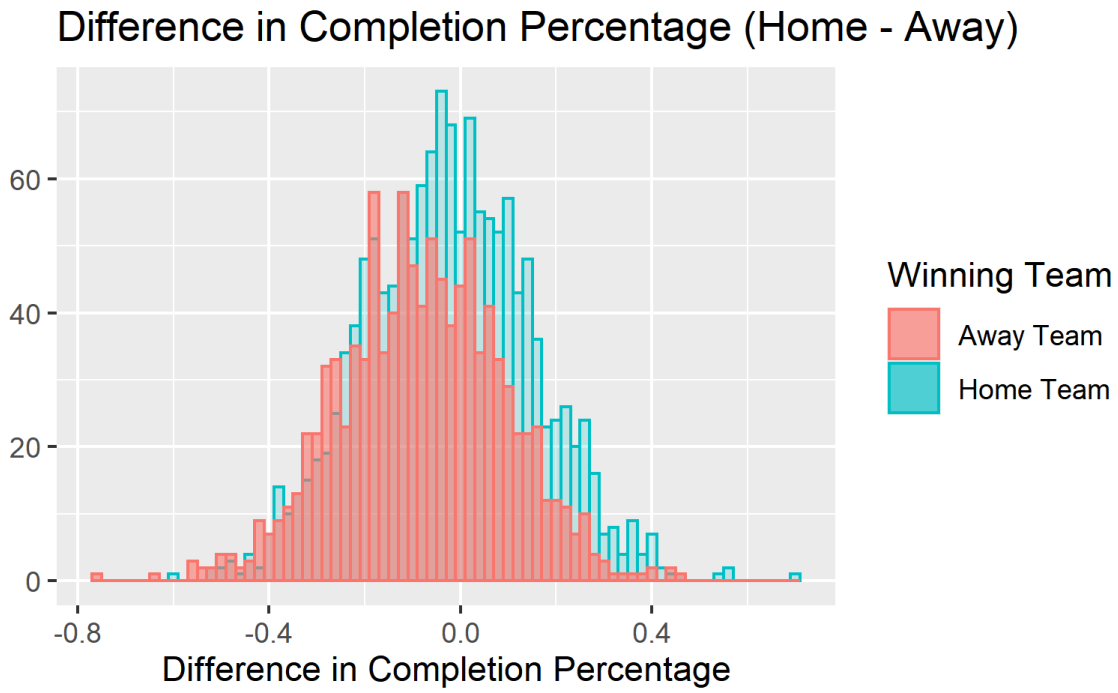

Figure 2: *The Scaled Difference in the Completion Percentage Between the Home Team and Away Team*

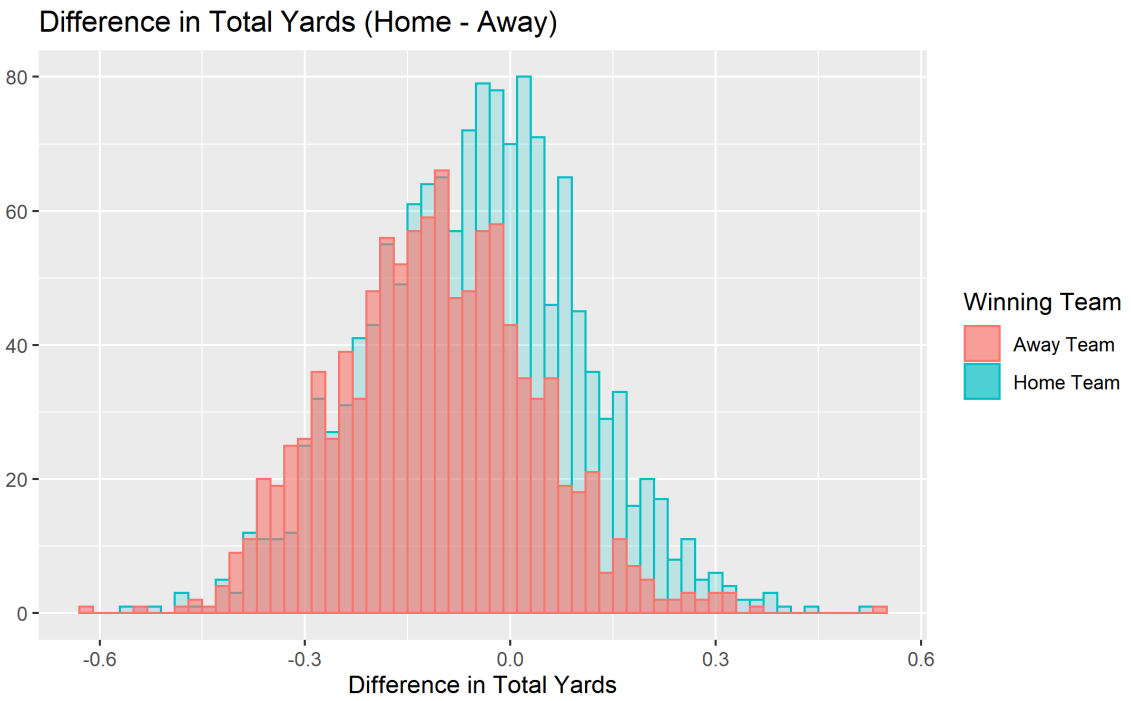

Figure 3: *The Scaled Difference in the Total Yards Between the Home Team and Away Team*

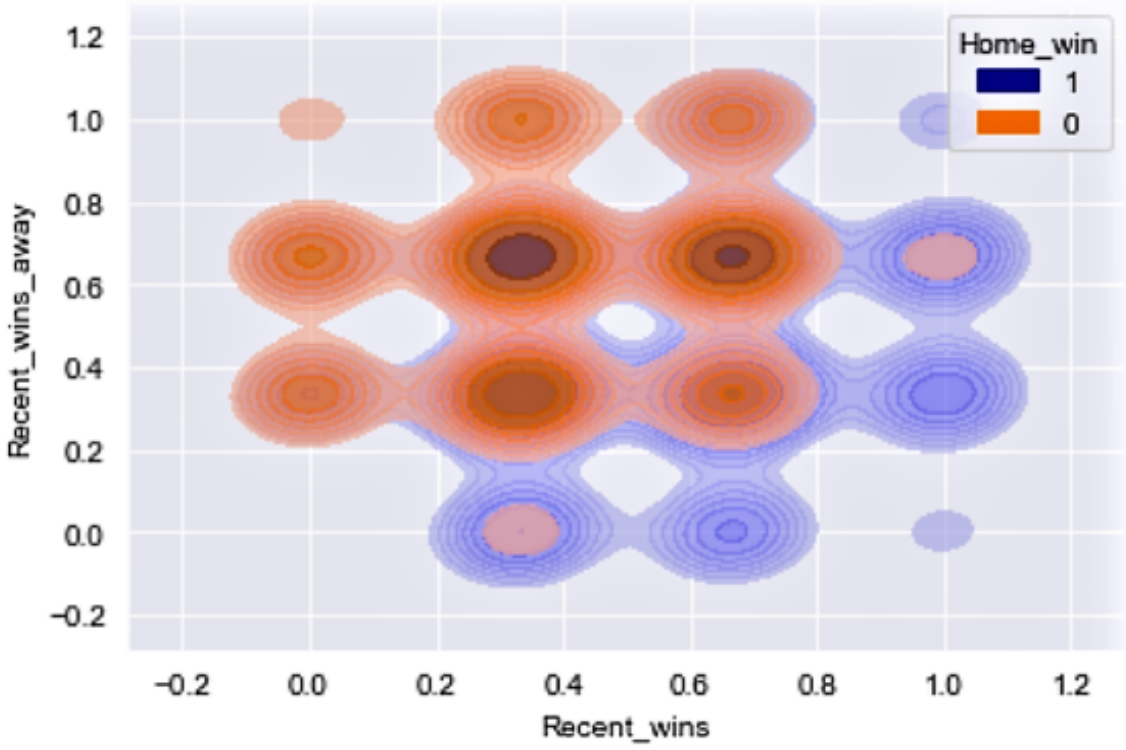

Figure 4: *A Comparison of the Relationship Between Recent Wins and Winning the Current Game* 8

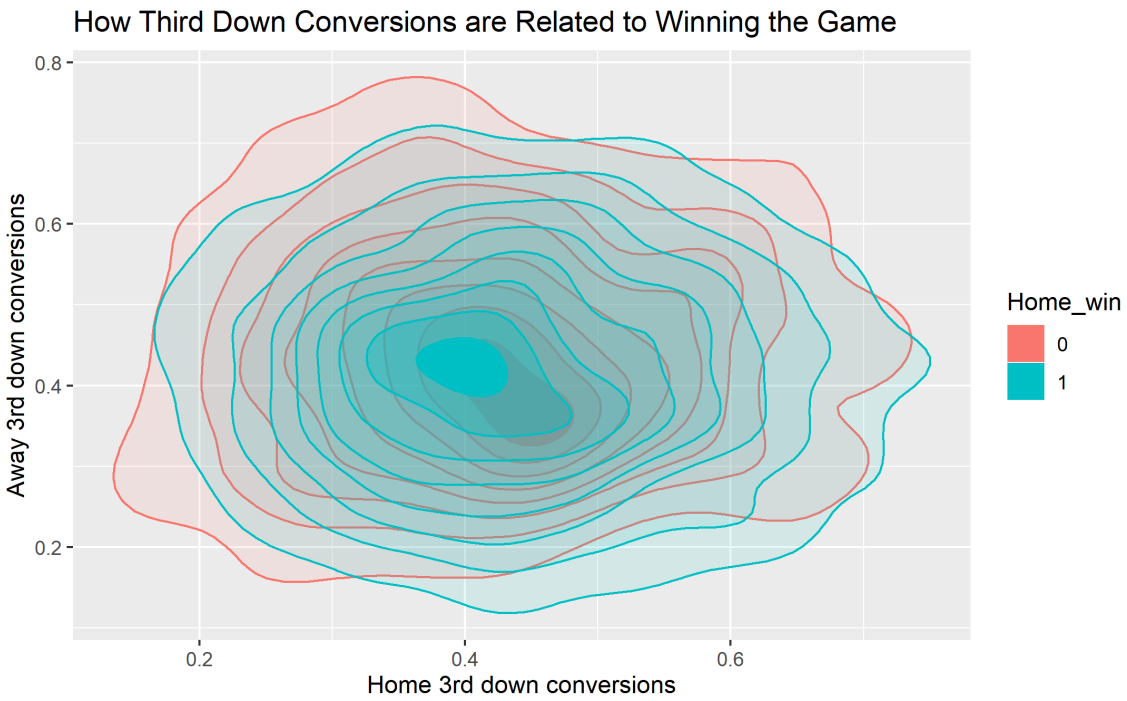

Figure 5: *Density plot of away team recent 3rd down conversions vs the home team.*

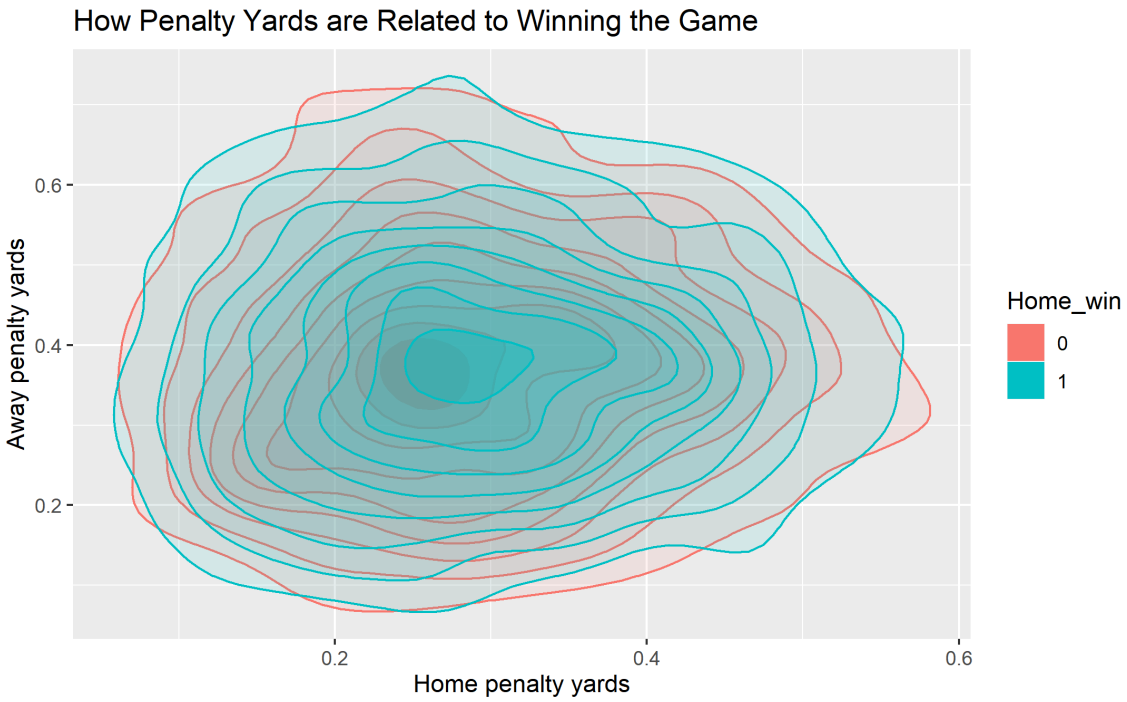

Figure 6: *Density plot of away team recent penalty yards vs the home team*

Building and Evaluating the Deep Learning Model

You will generally find three types of layers in any CNN which are the Convolutional layer, Pooling layer, and Dense, or Fully-connected layer. Combine these layers and a CNN architecture will be formed. Additionally, the Dropout layer and Activation Function are important tune-able parameters for any CNN. Figure 7 shows the model summary.

	Layer (type)	Output Shape	Param #
0	InputLayer	[(64, 27)]	0
1	Reshape	(64, 27, 1)	0
2	Conv1D	(64, 27, 8)	32
3	MaxPooling1D	(64, 27, 8)	0
4	Conv1D	(64, 27, 16)	400
5	MaxPooling1D	(64, 27, 16)	0
6	Dropout	(64, 27, 16)	0
7	Flatten	(64, 432)	0
8	Dense	(64, 64)	27712
9	Dense	(64, 1)	65

Figure 7: *Density plot of away team recent penalty yards vs the home team*

The final version of the model was trained and validated over 1000 epochs in batches of size of 64. During training the model achieved a training accuracy of 66.56%. The model did generalize well to the unseen testing data achieving an accuracy of 60.07%. Figures 8 and 9 show the testing loss and accuracy, and confusion matrix, respectively.

	loss	acc
0	0.684607	0.600694

Figure 8: *Testing Loss and Accuracy*

	Real (1)	Real (0)
Predicted (1)	212.0	104.0
Predicted (0)	126.0	134.0

Figure 9: *Confusion Matrix*

Conclusion

The outcomes of NFL games were predicted by collecting 10 years worth of data on all 32 teams and using that data to train a 1D-CNN deep learning model. Various combinations of model parameters were tested and their resulting validation accuracy's compared. The final model's testing accuracy was just over 60%. According to ("Sports Betting Math - How To Win Money at Sports Betting") any sports betting winning record above 52.4% will earn a profit therefore, this model may be useful to sports enthusiast and team staffing alike. The amount of data used in this paper was relatively small so it would be reasonable to conclude that an even better testing accuracy's could be achieved with an a larger sample of data.

References

"Sports Betting Math - How To Win Money at Sports Betting." The Sports Geek, https:www.thesportsgeek.com/sports-betting/math/. Accessed 3 Dec, 2021.