NFL Concussions Data Analysis

Patrick Connell, Skylar Furey, and Zack Licursi

**1. Introduction**

**1.1 Overview**

We are looking at data on concussions in the NFL and their many impacts. We are first looking at how a player getting injured with a concussion in a game impacts the outcome of the game, win or lose. We know that a player’s position and average playtime impacts the influence a player’s injury has on the result of the game. We will also look into which positions are more likely to get hurt, and how well different teams do in coaching their players how to reduce concussions. We can also see how these change over the years and how much of the variance is based on randomness. We can also test how getting injured affects the players' playtime before and after the injury. Further, we can see if players who miss more games have more affected playtime.

**1.2 Variables**

Player- The player that got injured during the game. This is a supervised categorical variable and will be used in classification analysis.

Team- The team that the player who got injured played on at the time of injury. This is a supervised categorical variable and will be used in classification analysis.

Position- The position that the player who got injured played during the time of injury. This is a supervised categorical variable and will be used in classification analysis.

Winning\_Team- Whether or not the injured player’s team won the game. This is a supervised binary variable and will be used in classification analysis.

Season- Which season the game was that the player got injured. This is a supervised categorical variable and will be used in classification analysis.

Games\_Missed- How many games the player missed after the game the player was injured in. This is a supervised numerical variable and will be used in regression analysis.

Playtime\_Before- The number of downs the player played in games on average before suffering the injury. This is a supervised numerical variable and will be used in regression analysis.

Playtime\_After - The number of downs the player played in their return game after suffering the injury. This is a supervised numerical variable and will be used in regression analysis.

Playtime\_Lost- The number of downs less the player plays in their return game than they did on average in games before their injury. This is a supervised numerical variable and will be used in regression analysis.

**1.3 Data Preview**

*Table 1.1 Concussions Data*

A screenshot of a cell phone

Description automatically generated

**2. Explanatory Data Analysis**

**2.1 One Categorical Variable- Graphical Summary**

A picture containing screenshot

Description automatically generated

*Figure 2.1 Pie Chart of Seasons*

A screenshot of a cell phone

Description automatically generated

*Figure 2.2 Downs Played Before and After Injuries*

A close up of a device

Description automatically generated

*Figure 2.3 Games Missed and Playtime Lost*

A picture containing drawing

Description automatically generated

*Figure 2.4 Position Distribution*

Figure 2.1 depicts my biggest concern with this data set, the lack of equal distribution of concussions between seasons. The 204/2015 season had many fewer reported concussions than the other years. This could have been for a variety of reasons; a couple being the players didn’t report their injuries, or the players were safer about how they played.

Figure 2.2 above shows that players tended to play close to 50 downs in games before their injuries on average. Some players averaged up to 80 downs per game before their injuries while others didn’t see the field. After injuries players only averaged about 40 downs in their game returning after the injury. Some players never returned that season while others played up to 100 downs in their return. Therefore players played about 10 games less on average in their return games than they did in games leading up to the injury. This is most likely due to players and teams wanting to make sure the player has fully recovered and does not suﬀer another concussion.

From the left panel of Figure 2.3, one can see that majority of the players did not have to sit out a game after suﬀering a concussion. The distribution is highly skewed to the right as expected, with only seven players missing double-digit games.

The right panel of Figure 2.3 shows that majority of the players weren’t signiﬁcantly impacted by their injury and were able to play within ten downs of what they averaged in games before the injury. The majority of players gained play time in their return but the plot is skewed left since a few players lost more playing time than anyone gained.

From Figure 2.4, one can see that most positions make players equally susceptible to concussions. Cornerbacks are the most susceptible to nearly 100 cornerbacks sustaining a concussion over the four years. The fullbacks count is very low due to the position not being used as much in the NFL anymore due to the changes in the playstyle of professional football. Defensive Back is also lacking because these players are versatile and can play both Safety and Cornerback whereas most players are position speciﬁc.

A screenshot of a cell phone

Description automatically generated

*Figure 2.5 Correlation*

From Figure 2.5 we can see due to the low p-value and high t-statistic that the number of downs that a player averages before their concussion impacts how many downs they play afterward. This makes sense because players who tend to get more playing time will continue to play more. At the same time, we know that players who rarely touch the ﬁeld won’t increase their playing time by getting injured.

**2.1 One Categorical Variable- Numerical Summary**

*Table 2.1 Concussion Distribution by Team*

A screenshot of a cell phone

Description automatically generated

*Table 2.2 Five Number Summary*

A screenshot of text

Description automatically generated

*Table 2.3 Correlation*

A screenshot of text

Description automatically generated

From Table 2.1, one can see that most teams had around 10-15 players get hurt over the four years. The Cleveland Browns suﬀered the most with 33 concussions, averaging about 8 a year, which is about a concussion every other game. There staﬀ needs to learn from the Falcons whose players only suﬀered 9 in the whole four year period. Only one more than the Browns averaged per season.

Table 2.2 shows that players played 4.715 fewer downs in their return game versus what they averaged in games before the injury. This is most likely because at least 75 percent of players missed one or no games so their concussions weren’t very signiﬁcant.

From Table 2.3 we can see there is little correlation between Games Missed, Playtime Before, Playtime After, and Playtime Lost. The biggest correlation is between Playtime After and Playtime Lost. This makes sense because players that play a lot in their return game won’t be able to lose that many downs as they could have only played a certain amount of downs in the games before the injury. We also see some positive correlation in Playtime Before and Playtime Lost, this is because the more downs one average before the injury then one has more downs to lose due to the injury.

*Table 2.4 Position vs. Season Contingency*

A screenshot of a cell phone

Description automatically generated

*Table 2.5 Position vs. Games Missed Contingency*

A screenshot of a cell phone

Description automatically generated

*Table 2.6 Team vs. Season Contingency*

A screenshot of a cell phone

Description automatically generated

*Table 2.7 Team vs. Games Missed Contingency*

A close up of a piece of paper

Description automatically generated

Table 2.4 shows that some positions were more likely to suﬀer positions in more recent years versus later years. Cornerbacks for example had much more reported concussions in the most recent season of 2015/2016 versus the three seasons prior. The 38 injuries in 2025/2016 were more than the previous 2 years combined. One can also see once again the full-back position becoming less relevant as no full backs suﬀered concussions in the last 2 years of the study.

Table 2.5 shows although cornerbacks suﬀered the most injuries (90), over half of those injuries did not cause the player to miss a game. Quarterbacks' concussions seemed to be the worst as about a third of the injuries resulting in two or more games missed. This is most likely because when quarterbacks suﬀer concussions they are being caused by players 100 pounds bigger than them. Running backs typically suﬀer their concussions in similar situations yet time missed typically isn’t as severe except for an outlier where a running back missed 13 games.

Table 2.6 reinforces that the Cleveland Browns suﬀered the most concussions during the studied period. Yet of the 33 concussions they suﬀered, nearly half of them were suﬀered in the 2015/2016 season. The most concussions suﬀered by a team in a year was barely another team in the same year of 2015/2016 by the San Diego Chargers with 16 concussions. Many teams did not have a concussion in the 2014/2015 season but only the Buﬀalo Bills were fully healthy in any other year.

Table 2.7 shows that the Kansas City Chiefs had a player miss 14 games in a season. Two teams, the Jaguars and Colts, had players miss 13 games in a season. The Steelers players either played through injuries or only suﬀered minor concussions; out of the 18 concussions the team suﬀered, 15 players didn’t miss a game while the other 3 only missed a single game. Many of the Brown’s concussions were luckily not very serious as about 2/3 of the injuries didn’t sideline a player for more than a week and only 2 injuries lasted longer than 5 games.

**3. Variable Selection**

**3.1 Best Subset Selection**

*Table 3.1 Best Subset Variable SelectionA picture containing text, table

Description automatically generated*

**Chart, diagram, box and whisker chart

Description automatically generated**

*Figure 3.1 Best Subset Variable Selection*

*A picture containing timeline

Description automatically generated*

*Figure 3.2 Best Subset Variable Selection using Adjusted R-Squared*

*A picture containing timeline

Description automatically generated*

*Figure 3.3 Best Subset Variable Selection using Cp*

*A picture containing diagram

Description automatically generated*

*Figure 3.4 Best Subset Variable Selection using bic*

In Table 3.1 one can see that Playtime After is best predicted by the variable Playtime Before, then Games Missed, and finally by Winning Team. From Figure 3.1, one can see that for all model selectors two variables are said to be used for the best model. One can see for Figures 3.2 through 3.4 that the two variables that would create the best model are Playtime Before and Games Missed. In none of the figures, the Winning Team is highlighted and therefore we can conclude adding that predictor to the model weakens the model.

**3.2 Forward Subset Selection**

*Table 3.2 Forward Subset Variable Selection*

**A picture containing table

Description automatically generated**

From Table 3.2 one can see that when using the forward model selection we get the same conclusion as we saw in Table 3.1. This table also shows that Playtime Before is the best predictor variable and Winning Team is the worst.

**3.3 Backward Subset Selection**

*Table 3.3 Backward Subset Variable Selection***A picture containing table

Description automatically generated**

From Table 3.3 one can see that when using the backward model selection we get the same conclusion as we saw in Tables 3.1 and 3.2. This table also shows that Playtime Before is the best predictor variable and Winning Team is the worst.

**3.4 Variable Selection Coefficients**

*Table 4.4 Coefficients*



In Table 3.4 one can see the impact that each predictor has on Playtime After. As one would expect players who played more before the injury played more after leading to a positive coefficient for playtime before. One could also presume that the coefficient for games missed would be negative since the more games someone sits out the more likely they to have a limited role in their return. Finally, the winning team is also positive this is not as important since as we saw before it is minimal if any correlation between the variables playtime after and winning team.

**4. Regression Models**

**4.1 Simple Linear Regression Models**

*Table 4.1 Playtime After from Playtime Before*

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

*Figure 4.1 Playtime After from Playtime Before*

We can see from the top left panel of Figure 4.1 that the line of best fit that on average most points had a low residual. One can notice also that players with the lowest playtime after the injury fit the model the least. This is because some players who played a bunch had season-ending injuries and therefore their playtime in a return game that season was zero because they didn't return. From the top right panel of Figure 4.1, one can observe that the data fits the line very well with only some extremities at the top due to the season-ending injuries explained before. The bottom left panel of Figure 4.1 shows that even after standardizing and square rooting the residuals the shape is similar to the same outliers. Finally, from the bottom right panel of Figure 4.1, one can discover that most of the high leverage points have small standardized residuals, and the only points that truly affect the line of best fit are the season-ending outliers once again.

A close up of a map

Description automatically generated

*Figure 4.2 Playtime After vs Playtime Before*

In Figure 4.2, the line of best fit shows a clear positive correlation between playtime before and playtime after. One can see along the bottom the players that had season-ending injuries as well as more serious injuries that caused them to be limited in their respective return games. Along the left side of the graph, one sees a couple of players who got hurt in their first game and therefore didn't have any playtime that season before the injury. The variance of those players being from a backup only playing about 25 downs in their return game versus a couple of guys who played close to a hundred and were most likely key players for their teams that year. One can see that slope is close to one meaning that on average players played close to the same amount in their return games as they did in games before their injuries.

**4.2 Interaction Terms**

*Table 4.3 Playtime After from Playtime Before, Games Missed*

A screenshot of a cell phone

Description automatically generated

Table 4.3 presents another variable that would seem logical to influence a player's playtime after an injury is how many games they missed. Typically when a player misses multiple games they have a limited workload in their return game. These statistics show otherwise though. There is minimal if any correlation between playtime after and games missed alone. When combined with playtime before the games missed variable only hurts the t-statistic and p-value compared to playtime before on its own.

A screenshot of a cell phone

Description automatically generated

*Figure 4.3 Playtime After from Playtime Before, Games Missed*

Figure 4.3 shows when the variable games missed are added one can see a few changes in the plots but not many. The top left panel keeps its same shape and has the same outliers of points 176 and 500. The top right and bottom left panels are also very similar with once again the same outlying season-ending injuries as the troublesome points. The biggest change is in the bottom right panel. The overall shape is very similar but now there is a new highest leverage point. That player missed the most games of any player in the dataset with 13. Therefore with the addition of that variable, his season stuck out more. Although that point has high leverage its residual is close to zero and therefore doesn't affect the line of best fit too much.

*Table 4.4 Playtime After from Playtime Before, Games Missed ANOVA*

A screenshot of a cell phone

Description automatically generated

Table 4.4 shows that although the t-statistic was small for the relationship between playtime after and games missed the f-statistic was much higher leading to a tiny p-value. Yet when combined with playtime before that relationship to playtime after severely weakens. Therefore individually one can predict playtime after playtime before and games missed but when combined they don't do well. This is because season-ending injuries are outliers in both and when combined they can highly skew the overall dataset. With the addition of the extra variable, we see even more of the mean square being based on the predictor variables and less on the residuals.

**4.3 Quadratic Linear Regression Model**

*Table 4.5 Playtime After from Games Missed, Games Missed Squared*

A screenshot of a cell phone

Description automatically generated

In Table 4.3 we saw that games missed with games before the games missed variable has a low t-statistic when related to playtime after. Yet in Table 4.5 we see that when games missed are combined with the squared version of itself, the non-squared portion has more effect than the squared portion.

A close up of a map

Description automatically generated

*Figure 4.4 Playtime After from Games Missed, Games Missed Squared*

The top panel of Figure 4.4 represents the relationship between playtime after and game missed, whereas the bottom includes games missed squared. In both relationships, we see in the first graph a severe cone shape meaning that as games missed increased the absolute value of the residual increases. The Normal Q-Q plots are very similar in shape and have the same extremities. Once the squared term is added the Scale-Location graph gets much less linear as expected. When everything is squared and one square root all of the residuals are more likely to show a trend. Finally, adding the squared term increased the leverages of all the points but the high leverage points had much smaller residuals and therefore kept the line of best fit more horizontal.

*Table 4.6 Playtime After from Games Missed, Games Missed Squared ANOVA*

A screenshot of a cell phone

Description automatically generated

In Table 4.6 comparing with and without the squared term, we see that adding the term has a relatively small f-statistic and therefore larger p-value. From the codes, we can infer that this would only be significant in a test using 99% confidence or less.

**4.4 5-Fold Cross-Validation**

**Chart, box and whisker chart

Description automatically generated**

*Figure 4.5 Boxplot for MSE*

**Chart, histogram

Description automatically generated**

*Figure 4.6 Barplot for Runtime*

*Table 4.7 Average MSE and Running Time for Regression Methods*

Table

Description automatically generated

Figure 4.5, shows that the different methods overall had an average of around 500-600 for the MSE. Figure 4.6, shows a big range in running time between the methods with half taking less than half a second and the other half taking almost or more than two seconds. B is the worst method for predicting a regression line for the data despite being the slowest. Whereas Decision Tree, Quadratic, and Random Forest all did much better. KNN had the most range between the different folds whereas quadratic was able to predict a regression line that was good for all folds. Overall, Table 4.7 shows that the Quadratic method seems to be the best model as it takes the second lowest time and has the lowest average MSE.

**5. Classification**

**5.1 Classification Methods using full model and 50% data**

*Table 5.1 Side by Side Comparison of Classification Methods*

Table

Description automatically generated

Table 5.1 compares multiple different classification methods attempting to predict whether or not the team won the game that each player was injured in. The predictions are based on how many games the player missed and how many downs the player averaged in games before the injury that season. We chose to use how many games the player missed because it helps determine whether or not the player was able to return to the game or not as well as how serious the concussion was. We chose to use how many downs the player averaged in games before the injury because it shows how much impact losing that player would have on their team. Key players would have more playtime before the injury than backups.

In the running time columns of Table 5.1, one can see that due to our dataset only having 516 points all of the classification methods only took about 13-14 milliseconds on average. In the accuracy columns, one can view that the KNN technique did not work nearly as well when the data was split into a test and train set as compared to when the test and train sets were both the entire dataset. Overall none of the accuracies were very high, this is because there are many more factors that go into whether a team wins a game besides a single player’s injury. All of the accuracies based on a train and test set sit around 50% meaning that someone could guess whether or not a game was won about as well as our model could.

From the sensitivity and specificity columns, one can view that LDA was especially good at identifying correct losses but was bad at identifying correct wins for both models. The Logistic Regression on the other hand was better at identifying correct wins for the full model and correct losses for the train and test set model. The Logistic Regression technique had the worst identifying rate of either losses or wins of all the methods by predicting less than twenty percent of correct losses using the full model. All of the KNN models were able to predict about sixty percent of correct wins and seventy percent of correct losses using the full model. Using the test and train set, the KNN technique averaged correctly identifying 45% of wins and 55% of losses.

**5.2 Classification Methods using 5-fold cross-validation**

*Table 5.2 Side by Side Comparison of Classification Methods using Cross-Validation*

**Table

Description automatically generated**

Table 5.2 shows that attempting to predict whether or not the team won the game that each player was injured in based on how many games the player missed and how many downs the player averaged in games before the injury that season was once again unsuccessful. The Logistic Regression method was the best and still only predicted the correct outcome of 56 percent of the time. This is only 6 percent better than flipping a coin since there are only two outcomes, win or loss.

Chart, box and whisker chart

Description automatically generated

*Figure 5.1 Boxplot for Accuracy*

Figure 5.1 shows the ranges and means for each method side by side graphically. From this figure, we can see that QDA had the worst accuracy for one of the cross-validations compared to all other methods singular cross-validations. We can also see that the Decision tree was the best in terms of all cross-validations were close to fifty percent or above. The Random Forest method had three of five cross-validations that produced similar accuracies and then one way below the mean and another way above it. The boosting and KNN methods were both very consistent but the KNN had a better max, min, and mean accuracy.

Chart, bar chart

Description automatically generated

*Figure 5.2 Barplot for Accuracy*

Figure 5.2 shows that although in Figure 5.1 there seemed to be great discrepancies between the different accuracies, they all hover around fifty percent. There is not much difference between any of the methods.

Chart, bar chart

Description automatically generated

*Figure 5.3 Barplot for Standard Error*

Figure 5.3 further shows my point that KNN and boost were the most consistent methods. It also highlights that the logistic regression method despite having the highest average accuracy also had the most discrepancy between the cross-validations.

Chart, histogram

Description automatically generated

*Figure 5.4 Barplot for Runtime*

In figure 5.4 we better see the great divide in runtimes. We see that the tree methods took much longer than the non-tree methods. An exception being the decision tree whose runtime was similar or better than all the non-tree methods.

Chart, bar chart, histogram

Description automatically generated

*Figure 5.5 Barplot for Sensitivity*

Chart, bar chart

Description automatically generated

*Figure 5.6 Barplot for Specificity*

In figures 5.5 and 5.6 we, see that all of the methods except QDA were able to predict losses better than they were able to predict wins. This makes sense because injuries don’t typically contribute to a team winning a game but can be detrimental to a team causing them to lose. We see that the decision tree was able to use its splits to very efficiently identity which conditions of a player's injury led to their team losing a game. The QDA on the other hand was alright at predicting both wins and losses but not exceptional at either.

**6. Conclusion**

Our analysis has shown that concussions are very prevalent in the game of football yet rarely do players miss time for the injury. We found that the amount of games the player misses is very unpredictable a relies more on how bad the injury is and not necessarily the player's contribution to their team.

We did find that the amount of playtime that a concussed player plays in their return game is most dependent on how many games they miss and how much they played before their injury. The most important predictor is their playtime before the injury. This makes sense because players typically get similar playtime before and after injuries.

We found that a quadratic regression method best predicted whether or not a team won the game the player was injured in based on how much the player played before the injury and how games the player missed. Although quadratic was the best it wasn’t the best by much with most methods having an accuracy of around fifty percent which is as good as guessing. This is since a lot of factors go into the outcome of a game other than a single player's injury. We saw that our model was much better at predicting losses than wins because injuries contribute more to losses than wins.

Overall, our analysis shows that predicting the number of games a player misses or the outcome of a game requires many more factors than our dataset possessed.

**7. References**

ali-ce. “Head Injured Players.” *NFL*, November 21, 2014. Github. https://github.com/ali-ce/datasets/blob/master/NFL/Head%20Injured%20Players.csv

Breslow, Jason M. “Concussion Watch.” *Frontline*, PBS. http://apps.frontline.org/concussion-watch/#positions\_2015.