**Capstone Development Milestone Three**

Lee McFarling

Boston University, Department of Computing and Data Science

OMDS DX699 O2 - Module B - AI for Leaders (Spring 2025)

Joshua Von Korff

March 9, 2025

**Capstone Development Milestone Three**

On February 9, 2025, viewership for Super Bowl LIX surpassed the record for viewership in the United States, with more than 127.7 viewers tuning in (Associated Press, 2025). While this might seem impressive, the record that the game broke was set by the Super Bowl LIII the previous year (Associated Press, 2025). With fan engagement and television audiences consistently reaching new heights—alongside league revenue climbing into billions of dollars annually (Mickle, 2024), it would seem to some that the NFL can do no wrong. However, there is a darker side to the game as well, and one that both the National Football League and the game of football at large have grappled with for years: Injuries.

Injuries present a significant challenge at all levels of the game of football. Injured players must continuously battle to regain their health; the team must test their depth by relying on backup players in key positions, and entire franchises risk losing both points and revenue when their best players are sidelined. And this impact extends even beyond the professional leagues—with injuries also affecting college players and players of all ages. That’s why, when I had the opportunity to apply data analysis and modeling to this issue, I was eager to take it.

This project seeks to build upon the work of a fellow student who conducted an initial analysis last fall. Their research focused on identifying factors that contribute to concussion rates among NFL players and exploring potential rule changes or other interventions that could help mitigate these impacts (Boston University, 2025). To that end, three key datasets were selected:

**The NFL First and Future dataset** – This dataset compiles multiple years of NFL play data, specifically aimed at analyzing the impact of playing surfaces on both injuries and player performance. (Howard, J, Langdon, Cormier, & Huddleston, 2019)

**The NFL Big Data Bowl** - this dataset includes 2021 Next Generation Statistics (NGS) tracking data, incorporating player tracking, game, play, and player information, as well as Pro Football Focus (PFF) scouting data for passing plays from Weeks 1–8 of the NFL season. The dataset was compiled with a couple of possible 'problems' to tackle in mind – and the data was presented to be open-ended and have users analyze the data to provide clear, actionable insights. (Howard et al., 2022)

Finally, the **NFL Punt Analytics Competition** was compiled from the 2015 – 2017 seasons to provide actionable insights and rule changes to prevent concussions during punt plays and kickoffs. (At the time, the league had learned that while these plays accounted for only 6% of total plays, they comprised nearly 12% of concussions in the league. (J et al., 2018)

The student initially presented these datasets with the aim of compiling them and using their combined insights to investigate trends impacting high-impact plays with the aim of preventing concussions – this exploratory data analysis began by integrating their insights to identify key factors that could potentially influence injury risks.

The first step in this process was to acquire the relevant datasets and perform an initial cleaning and pre-processing. Each dataset consisted of 3 to 12 CSV files, each with varying structures, primary and foreign keys, etc. The result of this preprocessing is as follows:

The **NFL First and Future Dataset** – was comprised of three different datasets, an Injury dataset, a player tracking dataset, and a playlist dataset –as an overview, the dataset contained 267,005 plays, with 105 of those plays leading to injuries. Additionally, 76,336,748 rows of player tracking data were obtained which captured positional coordinates (x, y), orientation (angle), speed, and other metrics in 1/100th-of-a-second increments. However, identifying the exact moment of an injury was a challenge, as injuries were not logged as specific ‘events’ in the player tracking dataset, and plays often continued after an injury. To address this, the mean average of the position, speed, direction, and orientation of a play was recorded to incorporate these movement trends into the analysis.

A screen shot of a computer code

AI-generated content may be incorrect.

**Figure 1:** Python excerpt from cell [55] from NFL\_First\_And\_Future.ipynb file. Test Administered: IF there was injury information in the PlayerTrackData[‘event’] column, then the cell would return > 0. Because of this, injury information had to be obtained from the ‘play’ level dataset, which led to the aggregation of NGS data.

Additionally, further preprocessing was required to handle inconsistencies. Some temperature values were recorded as -999°F and were imputed with the median temperature. The “Weather” and “Stadium Type” columns in the *PlayList* data frame also lacked standardization, requiring manual grouping (e.g., "Cloudy, 55 degrees" was simplified to "Cloudy", for example). Additionally, in the Injury data frame, the "Injury Length" column was one-hot encoded, which was useful for modeling but less so for direct analysis. To improve usability, a new column was created to store the actual injury duration for affected plays.

Once preprocessing was complete, the cleaned datasets were merged into two aggregated data frames: *All\_Plays* (containing all plays) and *Merged\_Injury* (a subset of only injury-related plays). Both data frames maintained identical columns, ensuring consistency between injury and non-injury plays for further analysis.

For the **Big Data Bowl** dataset, 12 original CSVs were aggregated and merged with the other datasets, leaving two final data frames: one containing all plays (injury and non-injury) and another consisting only of injury-related plays. While this dataset included some Next Gen Stats (NGS) data, integrating it with the rest of the dataset proved challenging. The NFL's play-level data provides detailed descriptions of individual plays, including player involvement (e.g., "Player A passed to Player B"), key events (e.g., "Player C was injured"), and outcomes (e.g., "touchdown, interception, or incomplete pass"). However, a single play -- such as a screen pass -- may be executed multiple times across different weeks. Since injuries are recorded only in play descriptions rather than as specific events in the NGS data, merging this would lead to averaging the movement data of players every time they ran a screenplay, for instance, because *one* of these screenplays led to an injury -- therefore, to avoid any misleading conclusions, week 1-8 NGS data was excluded from the analysis. (Further demonstration in discussion section).

For the **NFL Punt Analytics** dataset, 13 different CSV files were included, but many were excluded from this analysis for various limitations. One key challenge was the *Next Gen Stats* (NGS) data, which presented the same issue as in the Big Data Bowl dataset – injuries were recorded at the play level, but each play could be executed multiple times over a season, and it was impossible to tell which play instance resulted in the injury indicated. Furthermore, aggregation techniques were tested with this data in an earlier exploratory data analysis of the NFL First and Future dataset and no significant correlations between these columns and the target variable *(Inj\_Occurance)* were found, reinforcing the decision to exclude them.

Other types of data cleaning were necessary here as well -- two columns contained similar information related to weather conditions on game day, and both of datasets had the same standardization issues as the First and Future dataset. To ameliorate this, null values were backfilled using the available data from the other column. Then a dictionary as used to standardize the weather in the columns (e.g. ‘Cloudy, 65 degrees’ was changed to ‘Cloudy’). Finally, there was a *game*\_*start* column that ended up being significant later on – to improve the usability in this column, values were grouped into bins (14:00 ad 14:59 were categorized together, for example.). Finally, while the *punt*\_*vid*\_*review* and *punt*\_*vid*\_*footage*\_*injury* datasets contained valuable details, including video footage of injury plays, they were not included in the final analysis. Video data while powerful was not relevant to the statistical approaches applied in this analysis, and these datasets covered only about one-third of all recorded injuries, making their inclusion impractical due to the limited sample size.

Two sets of visualizations were generated for each dataset during univariate analysis. First, histograms were created for all numeric variables—one set representing all plays and another focusing only on injury-related plays. Additionally, box plots were produced following the same approach to compare distributions across the full dataset and the injury subset.

For the NFL first and Future dataset, histograms showed a wider spread of data in the *all* *plays* histograms relative to the *injury* *only* subset, which is expected given the larger sample size. For the temperature histograms, the *all plays* subset was normally distributed around a peak of 60-70 degrees Fahrenheit, while the injuries subset tends to have more data points at the extremes, which suggested a higher injury occurrence in very hot or very cold temperatures. (Figure 2).

Additionally, in the X and Y coordinates (taken from the next gen statistics) both X and Y coordinates centered around the center of the field, while in the injuries subset there was more of a spread, suggesting that there are certain parts of the field where players tend to get injured more. While direction seems somewhat evenly distributed in both cases, speed seems heavily skewed toward the lower end indicating that most plays on average involve a lower speed (Figure 2). However, in injury plays, there are some spikes towards higher speeds, indicating that wide receivers, running backs, and other fast-moving players might be prone to injuries high risk and high speed plays. That said, the large peak at the lower end of the spectrum could indicate that lineman are also at risk (though the 'types' of injuries in these situations might differ). Additionally, time showed a similar distribution pattern to speed and distance, indicating that these variables are likely correlated.

A graph of injury and injury

AI-generated content may be incorrect.

**Figure 2:** Histogram univariate analysis of numeric data for the NFL First and Future dataset compared between ‘all plays’ and ‘injury plays only’ subsets. Plots are meant to demonstrate the subtle deviation between the All Plays dataset, and the Injury Plays Only subset. Notable deviations are noted in ‘x’, ‘y’, ‘distance’, and ‘speed’. (NFL\_First\_And\_Future.ipynb)

In the All Plays dataset, outliers were observed for several variables including temperature (low end), x and y coordinates (low end and high end), speed (high end), distance (high end), and time on play (high end). While the presence of these outliers may seem concerning, they align with common phenomena in football. For example, while most games occur at a moderate temperature, some might occur in snowstorms. Similarly, while most plays happen in the middle of the field, but some happen on the extremities of the field (in terms of x and y), and so on. Filtering the dataset to include only injury-related plays reduced the prevalence of outliers, but all remaining patterns were consistent with the nature of the game.

For the *Big Data Bowl* dataset, while all plays seemed evenly distributed between the four quarters, there was a slight trend upwards in the injuries dataset, with injuries increasing as the game progressed. There was an additional uptick in the number of injuries in 3rd and 4th downs, possibly indicating more risky plays as a team’s chances to retain possession decreased. Although injuries did not significantly correlate with scoring metrics or yardage results, there was an increase in injuries during plays that involved penalty yards, suggesting a possible connection to personal fouls or other infractions. Additionally, injury rates slightly increased as the number of defenders in the box rose, indicating that blitz-heavy situations might carry a higher injury risk. As for outliers, both the *all plays* and the *injury only* datasets, there were outliers in the *yards to go*, *presnap score* metrics*, penalty yard* metrics*, and penalty play results.* While the numberof outliers decreased in the injury-only dataset, there were no significant changes in overall trends between the two datasets.

Finally, the NFL punt analytics dataset only contained four numeric features in the dataset – *season* *year*, *week*, *temperature*, and *quarter* – Of these, *season* *year* looked evenly distributed over the two-year period and didn’t seem to change between the *all* *plays* data and the *injury*-*only* subset. Injury rates did tend to fluctuate during the season, with three spikes in injury rates at the beginning of the season, the middle, and the end. Additionally, there seemed to be a spike in injury occurrence around the 30 – 40 degrees Fahrenheit range for *temperature*. Finally, a spike in 4th quarter injury rates was observed, which could indicate that punt plays become riskier as games progress—either due to increased strategic aggression in late-game situations or player fatigue.

Regarding outliers, they were primarily observed in the fields, temperature (snow storms), quarter (overtime), and injury occurrence (which is itself a rare event). However, filtering for injury plays largely eliminated these outliers, suggesting that anomalies in regular play did not have a significant impact on injury rates.

For the NFL First and Future dataset, there were strong correlations between *PlayerDay* and *PlayerGame*, which indicate that these columns should be carefully managed before building any models. Additionally, there were correlations between *speed*, *play time*, and *distance* which make sense logically as longer plays with significant yardage gains tend to involve higher player speeds (e.g., breakaway runs vs. short gains or tackles at the line of scrimmage).

Because the occurrence of injuries was considered very rare in the dataset (0.03% in this dataset) the correlations between features and this target variable were extremely low (0.01 to -0.01 on average). To ameliorate this class imbalance, we performed downsampling, which uncovered a potential negative correlation between *PlayerDay* (the number of days a player has played in a season) and injury occurrence. Meanwhile, *PlayerGame* (an identification field) also showed some correlation with injuries – which should be taken with suspicion.

To account for categorical variables, we conducted a chi-squared analysis between injury occurrence and all the categorical features. *Type of Play* had the most significant impact on injury occurrence (with a p-value of less than 0.04%) with pass rush and punt plays accruing the most injuries. The chi squared test also determined that *Position Group, Position, Field Type*, and *Weather* all had p-values around 25% or lower, suggesting a moderate relationship with injury occurrence.

Although the Big Data Bowl was clean and well-structured overall, it suffered from a severe class imbalance between *injury* and *non-injury* plays, and it had to be under sampled like the First and Future dataset. Correlation analysis revealed several highly correlated and probably redundant variables. *Pre snap home score* and *pre snap visitor score* are both highly correlated with *playId* and *quarter* (which makes sense as all these things tend to go up with time). Additionally, *playId* and *quarter* were highly correlated to each other which is likely due to sequential assignment of *playId* rather than any meaningful relationship. Because of these redundancies, some of these columns should probably be dropped during the feature selection.

After under sampling, new correlations were ‘uncovered’ most notably involving NFL *Id* for players committing a second foul on a play. However, as this is an identification column, its correlation with multiple numeric variables suggests it is likely just noise rather than a meaningful predictor, and the fact that this has happened twice in bivariate analysis suggests that perhaps under sampling to ameliorate the class imbalance was not a good strategy either.

Categorical variables proved more insightful, with nine features showing a potential link to injury occurrence (determined by a p-value ≤ 25%). The most notable risk factors included:

* Offensive formation styles
* Drop back types
* Defensive pass coverage schemes
* Team identity (both offense and defense), with some teams appearing more injury-prone or more likely to cause injuries

A screen shot of a computer program

AI-generated content may be incorrect.

**Figure 3:** Output of Chi-Squared analysis -- (Results converted to percentages for interpretability. Taken from NFL\_BigDataBowl\_Cleaning\_and\_Analysis.ipynb

Finally, for the punt data analytics, the only significant correlation between variables was found to be between the week of the season and the temperature logged in the game. Even after under-sampling, there were still no correlations between *numeric* variables and injury occurrence. However, like the other datasets, categorical variables provided more insights. *game site*, the *home team*, the *start* *time*, and the *game* *weather* all posed potential factors to potential injury occurrence. Again, any feature with a p-value less than 25% -- indicating that the relationship was unlikely due to chance -- was plotted in the data analysis Jupyter notebook *(NFL\_Punt\_DataAnalytics.ipynb)*.

When going over the data in both univariate and bivariate analysis, one unexpected finding was that numerical factors – such as x and y coordinates, temperature, and player direction – did not have a significant impact on injury prevalence. Even after addressing class imbalances through under-sampling non-injury plays, these variables remained statistically insignificant. (And in this case, under-sampling often amplified noise in ID columns rather than uncovering meaningful patterns in the data.)

A screenshot of a graph

AI-generated content may be incorrect.A screenshot of a heatmap

AI-generated content may be incorrect.

**Figure 4**: Correlation heatmap of **NFL Punt Return** numeric data. The far right column regards injury prevalence. The left image depicts the correlation heatmap before remediation. The right figure represents the same dataset correlations after non-injury data was under sampled. (NFL\_Punt\_DataAnalytics.ipynb)

Additionally, working with NGS data across the three datasets proved surprisingly difficult. As discussed earlier, NGS data did not indicate when injuries occurred – instead focusing on events like passes, completions etc. Instead, injury data was often assigned at the play level, which complicated analysis because plays could be run dozens of times. This pattern became evident when merging the play data with NGS data – and consequently, NGS data was often aggregated so heavily that it became statistically insignificant and was subsequently excluded from other datasets. A screenshot of a computer

AI-generated content may be incorrect.

**Figure 5:** From NFL Big Data Bowl Cleaning and Analysis ipynb file. This excerpt shows that a playId, 788, was played twice in game 20211090900. Readers are encouraged to navigate to the notebook to see the full breakdown and analysis of why this data was excluded.

Another unexpected issue was the punt dataset’s incomplete injury information. Despite over 90 injuries being recorded, only a third had corresponding video and review data. While this data was extremely detailed and could have been very useful, its limited availability compared to the rest of the dataset led to its exclusion from review as well.

Finally, a major challenge that was faced in all three datasets was the inconsistency in formatting. Despite all three datasets being compiled from the NFL – and presumably drawn from the same source, these datasets exhibited some pretty stark differences:

* Primary and foreign key structures (*playId*, vs. *PlayKey*) with id fields ranging from sequential integers to hyphen separated values (play 200 vs. 39753-4-32 for example).
* Column naming conventions varying both between and within datasets, shifting between snake case (*play\_id*), PascalCase (*PlayId*), and camelCase (*playID*).
* The number of plays recorded differing wildly between datasets, with one dataset recording 432,000 plays over an 8-week period, whereby another would report 267,000 plays over multiple years and a third reporting 75,000 plays in one year’s time.
* Injury rates did not align between the datasets despite multiple verification steps and varied from 1.2% in one dataset to 0.03% percent in another.

These discrepancies suggest that the datasets were either pulled from different sources or extracted selectively by the NFL to highlight certain pre-determined patterns in the data, rather than providing a comprehensive sample of NFL data over a certain period. Despite multiple verification steps, these datasets did not align, making it nearly impossible to integrate them into a unified dataset without introducing some sort of severe bias. So, although the authors aimed to combine these three datasets into a unified set and use it in conjunction with NGS data to track concussions, the data were largely unusable for these purposes due to formatting and other inconsistencies. Even filtering for concussions – or even further to lineman’s concussions for example – would be infeasible as broader injury rates were already very low (ranging from 2 % to 0.03%). Meaningful conclusions in this regard would require a much larger dataset, such as a comprehensive 10-year database of every NFL play.

However, these datasets still provide plenty of value by offering different perspectives into factors that influence injury trends more broadly. Instead of incorporating them into a ‘super dataset’ their value lies in their use as a multifactored analysis whereby injury trends could be investigated from multiple different angles. For instance, unsupervised methods like clustering could be used to identify overarching trends in one dataset, and then supervised methods could be used to further investigate those patterns in other datasets. For example, if trends were identified between play type and injury prevalence in the First and Future dataset, we could use this information to run a supervised learning analysis on the identified play types in the Big Data Bowl dataset and look for differences in offensive and defensive formations, pass coverage types, offensive drop back types, etc. In this instance, because the Big Data Bowl dataset had a greater degree of granularity in player positioning, and play style than the First and Future dataset, it could be used to drill down into details that the First and Future dataset lacked. In this sense, patterns identified in one dataset could be used to narrow in on or augment details found in another. Additionally, while the lack of correlation to the target variable was evident in the analysis, the plethora of different categorical variables lends itself very well to the use of decision trees (and types of models built off them like random forests, etc.) moving forward. Overall, I am excited to see where the continued analysis of this dataset leads!

Works Cited

Associated Press. (2025, February 11). Super Bowl LIX averages record audience of 127.7

million viewers. NFL.com. <https://www.nfl.com/news/super-bowl-lix-averages-record-audience-of-127-7-million-viewers>

Boston University (2025). NFL Player Health Analytics. [https://docs.google.com/document/](https://docs.google.com/document/d/1M9srLpE-eHKoCYtX_zAMS3Eqtp-rmKg3KylncdWVWto/edit?tab=t.0)

[d/1M9srLpE-eHKoCYtX\_zAMS3Eqtp-rmKg3KylncdWVWto/edit?tab=t.0](https://docs.google.com/document/d/1M9srLpE-eHKoCYtX_zAMS3Eqtp-rmKg3KylncdWVWto/edit?tab=t.0)

ChatGPT. (2025, March 9). Personal conversation with AI assistant regarding citations and

sentence fluency improvements. <https://chatgpt.com/share/67ce4356-dbe4-800f-8474-8a137997225f>

Howard, A., J, A., Langdon, C., Cormier, J., & Huddleston, S. (2019). NFL 1st and Future –

Analytics [Dataset]. Kaggle. <https://kaggle.com/competitions/nfl-playing-surface-analytics>

Howard, A., Blake, A., Patton, A., Lopez, M., Bliss, T., & Cukierski, W. (2022). NFL Big Data

Bowl 2023 [Dataset]. Kaggle. <https://kaggle.com/competitions/nfl-big-data-bowl-2023>

J, A., Caitlin, C., Crawford, C., Sherwood, C., Cormier, J., & O'Connell, M. (2018). NFL Punt

Analytics Competition [Dataset]. Kaggle. <https://kaggle.com/competitions/NFL-Punt-Analytics-Competition>

Mickle, T. (2024, March 7). NFL's national revenue hits $13 billion in 2023, driven by TV deals.

Sportico. <https://www.sportico.com/leagues/football/2024/nfl-national-revenue-2023-tv-13-billion-1234786762/>