Synthetic Turf's Impact on NFL Injuries

Amado Uyehara
Lally School of Management
Rensselaer Polytechnic Institute
uyehaa@rpi.edu

1. Executive Summary

The National Football League is not only the highest-grossing sports market in the United States and world-renowned for it is thrilling displays of athleticism and strategic expertise but unfortunately marked by a recurring concern of on-field-related injuries that have cast a shadow over the gridiron spectacle. The NFL consists of a 17-week-long regular season (excluding the playoffs). The consistent occurrence of injuries linked to playing surfaces has sparked discussions on player safety and the impact of field conditions. From the pristine grass fields to the rise of synthetic turf, the NFL's battle with recurring injuries has become a subplot that demands attention. The data under examination is a 250 complete player in-game history from two subsequent NFL regular seasons and all the pertinent details relating to those injuries. Using two datasets, PlayList and InjuryRecord, specific data for each play that involved an injury were recorded under the feature columns.

Figure 1.

(Playlist)	**(Injury Record)**
PlayerKey - unique identifier for players **GameID** - unique identifier for player and game (separated by dashes) **PlayKey** - unique identifier for player, game, and play (separated by dashes) **PlayerGame** - integer unique identifies a player's games throughout the season, increases as the season progresses **FieldType** - character string identifier for field type (synthetic or natural) **Roster Position** - character string identifier for position (ex. Quarterback) **Weather** - character string identifier for the weather conditions of the game (indoor/ outdoor included) **Temperature** - float for the on-field temperature at the start of the game	**Body Part** - character string identifier for the injured body part **Surface** - character string identifier for the field type the injury occurred on **DM_M1** - one hot coding indicating how many days the player was injured (1 day) **DM_M7** - (7 days) **DM_M28** - (28 days) **DM_M42** - (42 days)

1.1 Class Labels

The class labels that were focused on are the number of injuries that occur on synthetic surface playing fields, temperatures at the start of the game, how many games a player has played during the season, and how long those players have been out.

Figure 2.

PlayerKey	GameID	PlayKey	RosterPosition	PlayerDay	PlayerGame	StadiumType	FieldType	Temperature	Weather	PlayType	PlayerGamePlay	Position	PositionGroup

Figure 3.

PlayerKey	GameID	PlayKey	BodyPart	Surface	DM_M1	DM_M7	DM_M28	DM_M42

2. Benchmarking of Other Solutions

In exploring various Kaggle solutions addressing NFL player injuries concerning playing surfaces, three distinct analyses revealed insights into injury correlations, predictive models, and factors such as surface type, playing location, and player activity, shedding light on the multi-dimensional dynamics of injuries in professional football.

2.1 First Solution: NFL Injury Analysis

Link: https://www.kaggle.com/code/derekxue/nfl-injury-analysis

The first Kaggle solution is solving a solution very similar to the problem faced in this report. They go into a deeper analysis including an additional database that keeps track of the exact location the injury happened on the field to find when, where, and why they happened. This solution however doesn't use machine learning techniques, their solution consisted more of a data analysis rather than prediction models. They used most of the same features my report is emphasizing. The PlayList databases' PlayerKey, Game ID, PlayKey, FieldType, StadiumType, PlayType, and Roster Position feature columns are being utilized the most in conjunction with the Injury Record's PlayerKey, Game ID, PlayKey, BodyPart, and Surface. The modeling approach this Kaggle solution had was to set a null hypothesis and prove or disprove the null hypothesis by calculating p-values and z-scores on standardized data. The two hypotheses that were stated were determining if synthetic or natural grass had a positive or negative effect on injuries. The Kaggle solution proved that there is a higher positive correlation of injuries on synthetic playing surfaces. This solution also provided more detail and used the exact precise location in their analysis more than other solutions. This Kaggle solution was more successful than others in visualizing the data and locations where the injuries occurred.

2.2 Second Solution: NFL Injuries Prediction in Python

Link: https://www.kaggle.com/code/jonbown/nfl-injuries-prediction-in-python

The second Kaggle solution wanted to find the correlation in NFL player injuries and see if there is a way to predict injury numbers based on common factors. Their approach followed this order: visualize the field, heatmaps of different types of injuries, heatmaps with the player position on days missed on synthetic vs real turf, trends when connecting player injuries to the attributes of the play, connect it to a surface factor, build a logistic model that predicts the probability of injury based on surface type, how early in the game do they get injured based on the surface, number of plays it takes for an injury based on position, and number of days missed based on injury. They used two different models to perform analysis, the Kaplan-Meier Model and the Cox Proportional Hazard Model. Using logistic regression, the solution found that the biggest impact on player injuries across both synthetic and natural surfaces was PlayerDay, which is the day in the season they played. The longer the player was active in the season the higher the chance they were injured. The logistic regression score the solution got for their train test split was 0.99. Proving the model was accurately predicting the likelihood of an injury happening dependent on factors such as field type, temperature, stadium type, days a player has played, the type of play a player was in, and their position. After logistic regression, the solution used a random forest classifier technique to predict injuries using the same criteria used in the logistic regression. This model produced a score of 0.99 as well, proving that it was accurately predicting the likelihood of an injury happening dependent on factors such as field type, temperature, stadium type, days a player has played, the type of play a player was in, and their position. This was the only solution I was able to find that used machine learning techniques to predict NFL Injuries, therefore it was the most successful kernel compared to other kernels in Kaggle.

2.3 Third Solution: NFL Playing Surface Data Analysis

Link: https://www.kaggle.com/code/priankravichandar/nfl-playing-surface-data-analysis

The third Kaggle solution follows the same analysis as the report and the first Kaggle solution. An analysis of three datasets related to the effects that playing on synthetic turf versus natural turf can have on player movements and the factors that may contribute to lower body injuries. The analysis that this solution conducted an exploratory data analysis to identify any trends within the data. The solution also provided a few visualizations, but one stood out on how many days a player has missed when suffering an injury on natural or synthetic surfaces. The results provided were that players have a relatively higher likelihood of suffering short and medium-duration injuries on Natural surfaces and players have a relatively higher likelihood of suffering long and indefinite-duration injuries on Synthetic surfaces. Since this solution didn't have any machine learning techniques, there weren't any scores to determine if it was more successful than other solutions. Overall, injuries on synthetic surfaces result in players missing more days than those on natural surfaces. This solution was better in understanding why the players injured themselves concerning changes in direction and movement speeds on playing surfaces. They concluded that the majority of plays seem to require some type of instantaneous change in the player's direction of motion and orientation in the first 40 seconds of the game, there are 18.72% more instances of potential injury during plays on Synthetic surfaces as

opposed to Natural surfaces, and the average time for a potential instance of injury to occur is slightly lower on natural surfaces as opposed to Synthetic.

3. Data Description and Initial Processing

Using the two initial datasets Playlist and Injury Record, I focused on features columns that were important to my prediction analysis. Those features in the PlayList dataset are PlayerKey, GameID, PlayKey, PlayerGame, FieldType, Roster Position, Weather, and Temperature. As for the InjuryRecord dataset, where my basis for analysis is focused, these are the features I used Body Part, Surface, DM_M1, DM_M7, DM_M28, and DM_M42. Upon uploading both CSV files, I began a data check to see what specific values I was dealing with. There were a lot of null values and bad data that needed to be cleaned. Null values are bad and improper handling can lead to unexpected results. The bad data that I came across were in the temperature column where I found unique values such as “-999”. Bad data is especially bad when conducting any form of data analysis because it can greatly skew the results you are trying to achieve. Null values are a form of bad data as they are typically treated as 0’s. Next, I performed some data cleaning to the features I wanted to use and anything extra outside to make the dataset cleaner and easier to work with in case I wanted to introduce more features or variables to my analysis.

The first form of data cleaning I conducted was mapping the respective positions to specific players in the PlayList dataset. In the Position and Position group, there were numerous null values, on the other hand, I noticed that the RosterPosition feature column had no null values. I used the RosterPosition group to map the respective positions to the position and position group feature columns, eliminating all null values. There were no other null values in the database, but to be sure I decided to drop the rest since they wouldn't be able to be mapped as they were all unique values or values that mattered greatly to the analysis. If there were null values for a feature such as temperature, I didn’t want to fill null values with averages as they would impact the result I wanted to achieve. Next, I removed any bad data I found which happened to be in the temperature column. Some rows had temperature values of “-999”, which is scientifically impossible, so I deemed those rows not important anymore as the data could not be normalized properly and would affect the rest of the temperature data when I did normalize it. The next form of data cleaning I performed was to map the weather conditions to respective bins as there were too many unique values. I created 8 different bins, including one for indoor stadiums, and mapped them to a bin that closely matches the weather conditions. The eight bins I used were Clear, Partly, Cloudy, Rain, Indoor, Hot, Cold, and Snow. This would make it easier to categorize certain weather conditions that may affect injuries in the event I needed to analyze them. Due to the many unique values that the Temperature column has, I wanted to normalize them and prepare the values for any prediction methods I would use. I used the MinMaxScaler function to normalize temperature values between 0 and 1. Since I was creating a new column for the normalized temperature values, I then mapped those values to the rows they were assigned to. Finally, the essential element of my prediction analysis, I wanted to see how many injuries that lasted at least a week due to other variables. I would achieve this by

using the DM_M1, DM_M7, DM_M28, and DM_M42 feature columns. Seeing how it would be difficult to use surface type in a prediction analysis as a string, I decided to one-hot encode this feature column as well and append it to the data frame. I mapped the natural surface type to 0 and the synthetic surface type to 1. Finally, I created a new feature column called “Injured >= 7 days” where I used one hot encoded data to find if a player suffered an injury that lasted a week. After all the data cleaning and preprocessing were finished, I inner joined the two datasets together on the PlayerKey as it is a unique identifier for each player.

Following my data cleaning and preprocessing, I used histogram plot visualizations to help me better understand the data I was working with. These plots also helped me better understand what hypothesis I can formulate before proceeding with my prediction analysis.

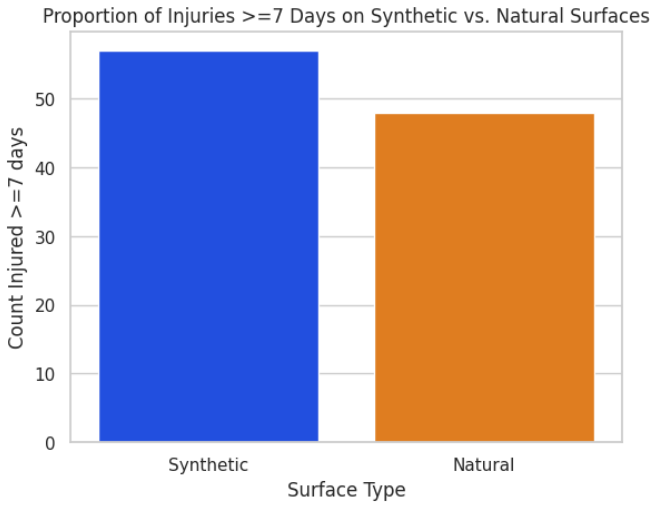

Figure 3.

3.1 Plot 1

The first plot shows the Proportion of injuries that lasted at least a week and the field surface type they were. The initial purpose was to find which surface type contributes more to injuries for players and their positions across the board. I was able to draw minimal conclusions from this plot as it was only between two specific features. However, the plot did show a higher proportion of injuries on synthetic turf, but the difference between natural surfaces was not that significant.

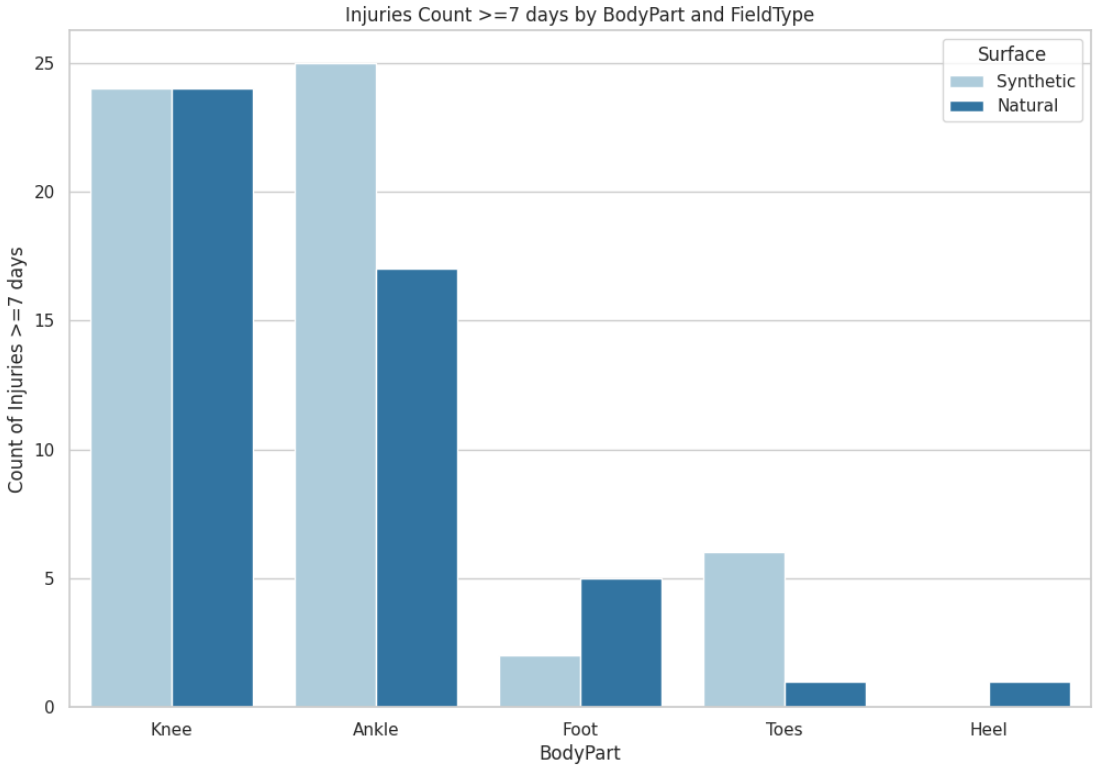

Figure 4.

3.2 Plot 2

I wanted to dive deeper into the surface type and how it affects specific types of injuries to see which contributes more to synthetic and natural surface types. This histogram plot used the BodyPart feature column to separate injuries that lasted at least a week by surface type. Most of the injuries outside of the ankle body part are pretty event, so I wanted to dive deeper into why ankle injuries are much higher on synthetic surfaces. I conducted some outside research and found that according to the National Center for Health Research, ankle injuries on synthetic turf are the most common injury at a much higher rate than on natural grass. This study also aligns closely with what the histogram plot is showing. Moving forward I can conclude with my data analysis that more injuries that last a week occur on synthetic turf than on natural grass.

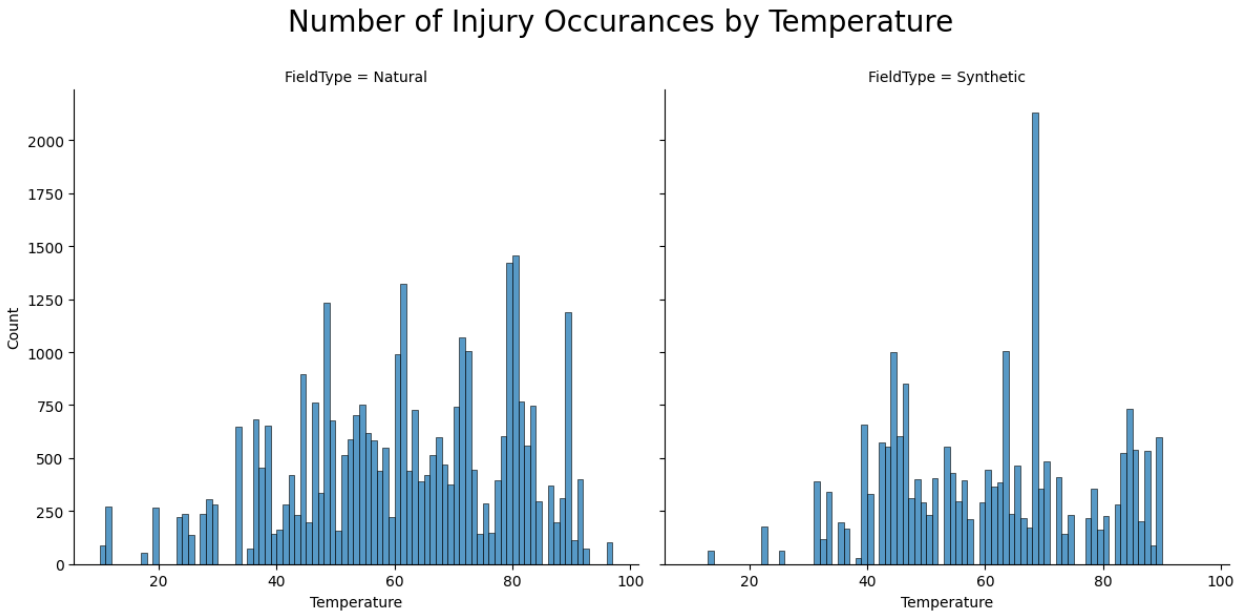

Figure 5.

3.3 Plot 3

The last plot I used before moving into my prediction analysis was the number of injury occurrences by temperature compared to synthetic and natural surfaces. I wanted to investigate deeper as to why more injuries would occur on synthetic turf. Temperature plays a huge role in injury occurrences in sports as a whole. The graph is split between synthetic turf and natural grass and shows the number of injuries per degree of temperature. The natural grass side is more evenly distributed as more games are played on natural surfaces. As for the synthetic turf side, we can see a high spike in the 60-80-degree range. The data points towards a lot more injuries in indoor stadiums that keep their internal climate within that range throughout the season. Another factor could be that more injuries happen earlier in the season when temperatures are warmer. After conducting my data analysis on three different factors, I had a better understanding of which variables I wanted to include in my prediction analysis.

4. Modeling

In my analysis of using predictive models for injury prediction on synthetic turf, the exploration of various independent variables is crucial. After conducting analysis and looking at trends displayed by the histogram plots, I decided to proceed with three features that I believe correlate and contribute the best to predicting player injuries and whether or not they last at least a week. Of the features in the datasets, I decided to use the Temperature, SurfaceType, and PlayerGame features. Each has its effect on injuries individually, but together they will help me better understand their impact. Looking at the different ways to use predictive modeling, I decided to use three different algorithms Logistic Regression, Decision Tree Classifier, and Random Forest Classifier.

4.1 Logistic Regression

The first predictive modeling technique I used was a logistic regression using the three independent features compared against whether they affect injuries lasting at least a week. The reason I chose this as my first modeling technique was mainly for how it can handle binary classification effectively. Since my “Injured >= 7 days” and “Surface_Encoded” features are both one hot encoded (or binary), I wanted to see how their prediction methods would handle the dataset. I first split the dataset into the independent features I wanted to compare against “Injured >= 7 days”. I then included a train, test, and split to divide the dataset into training and testing datasets. Applying the linear modeling logistic regression technique to the training data, I was able to get an accuracy, precision, and recall score. The scores were accuracy: 0.68, precision: 0.69, and recall: 0.97. These scores were decent to start with, as the recall was pretty high showing the model was correctly predicting that injuries lasting at least a week have a strong correlation to the independent variables. Next, I wanted to see which features had the biggest impact on predicting those injuries. The feature importances were listed such as Temperature_Normalized: 0.996, PlayerGame: 0.053, and Surface_Encoded: 0.052. Proving that temperature played the biggest role in predicting whether an injury lasted a week or longer. The other features had some effect on the prediction, but the temperature was significantly higher. This technique wasn’t able to paint a full picture of what exactly I was looking for so I decided to look deeper.

4.2 Decision Tree Classifier

The second predictive modeling technique I used was a Decision Tree Classifier. The reason why I decided to use this technique instead of others was mostly to see how a different model from logistic regression could handle the independent features. The Decision Tree Classifier algorithm is different in how it interprets the data and also better at handling non-linearity. Seeing how my accuracy scores were a little low in the logistic regression, I wondered if the data I wanted to predict was linear or non-linear. Using the Decision Tree Classifier method would help me understand if it was or wasn’t. As for the adaptability with binary encoded features, the way the Decision Tree Classifier predicts the data is much more different and thorough than logistic regression. To start I used the same train, test, and split as the logistic regression as I wanted to compare the scores. Applying the predictive model to the train and test data, I was able to achieve scores such as accuracy: 0.91, precision: 0.92, and recall: 0.94. These scores were much better as a whole than logistic regression, as the model is better at predicting whether an injury lasts at least a week or not. The recall score is lower, but the accuracy of true positives is much higher, which is what I wanted to accomplish when implementing the Decision Tree Classifier. As for the feature importances and which ones affected the predictive analysis the most these were their results Temperature_Normalized: 0.47, PlayerGame: 0.42, and Surface_Encoded: 0.12. Once again temperature is the most important feature in analyzing why injuries last at least a week or longer.

4.3 Random Forest Classifier

The third and final predictive modeling technique I used was the Random Forest Classifier. I chose this technique as I wanted to refine my accuracy, precision, and recall scores further. The Random Forest Classifier is much more intricate than the Decision Tree Classifier as it uses multiple decision trees and averages their predictive scores. Like the Decision Tree Classifier, it is much better at handling non-linear relationships in the dataset. It is much better at handling overfitting as well, which would handle outliers much better than Logistic Regression and the Decision Tree Classifier. The main focus of using the Random Forest Classifier was to find a better score for recall and keep my accuracy and precision scores very close to the previous scores in the Decision Tree Classifier. Using the same train, test, and split from the previous two predictive techniques, I was able to achieve these scores accuracy: 0.91, precision: 0.92, and recall 0.95. The accuracy and precision scores were identical to the ones found in the Decision Tree Classifier with a better recall score. This technique was much better at predicting true positive values than the Decision Tree Classifier and Logistic Regression. I wanted to look deeper as to what the confusion matrix would look like for this model and I got these values True Positives (TP): 35008, True Negatives (TN): 15491, False Positives (FP): 3052, and False Negatives (FN): 2036. There were a lot of true positives showing that the model was able to predict that the independent features do have a significant impact on whether an injury lasts at least a week. Next, I wanted to see how the model prioritized feature importances, and these were the results Temperature_Normalized: 0.59, Surface_Encoded: 0.01, and PlayerGame: 0.4. As seen in the other predictive models, the temperature has the highest importance with the number of games a player plays as second, and the surface type as third.

5. Conclusion

The NFL is a world-renowned league and is the largest sports market in the United States. The league brings in millions of dollars of revenue every year and displays a spectacle of athletic and strategic talent very different from many other sports across the globe. Recently, a call for change in the league's playing surfaces has gained popularity as many players, especially star players, have been suffering long-term to career-ending injuries. The main driving force for this reform would be to mandate natural grass fields for all stadiums as data has pointed at it playing a crucial role in injuries. The data examined is a complete report of players in game history from two subsequent NFL seasons and the injuries they suffered during games. Conducting data analysis and predictive modeling analysis, there is a significant correlation between medium to long-term injuries occurring on synthetic turf. Factors such as temperature, games played, and field surface type play huge roles in whether a player is prone to a serious injury while playing. As this data is already proven, the NFL has started taking measures to switch their synthetic turf fields to natural grass to minimize the amount of serious injuries that occur during the season.

6. References

[1] Derekxue. “NFL Injury Analysis.” Kaggle, Kaggle, 12 June 2023, www.kaggle.com/code/derekxue/nfl-injury-analysis.

[2] “Injuries Related to Artificial Turf.” National Center for Health Research, 19 Sept. 2023, www.center4research.org/injuries-related-to-artificial-turf/#:~:text=Ankle%20Injuries,-A%20study%20published&text=They%20found%20that%20the%20rate,turf%20compared%20to%20natural%20grass.

[3] Jonbown. “NFL Injuries Prediction in Python.” Kaggle, Kaggle, 31 July 2022, www.kaggle.com/code/jonbown/nfl-injuries-prediction-in-python.

[4] Priankravichandar. “NFL Playing Surface Data Analysis.” Kaggle, Kaggle, 20 Sept. 2021, www.kaggle.com/code/priankravichandar/nfl-playing-surface-data-analysis.