NFL Injury Data based on Turf Grass Types

**Overview:**

This page describes the datasets and variables provided to examine the effects that playing on synthetic turf versus natural turf can have on player movements and the factors that may contribute to lower extremity injuries. The data provided for analysis are 250 complete player in-game histories from two subsequent NFL regular seasons.

**Data Ideas:**

*--How early in the game do they get injured based on the surface grass type.*

*--Number of plays it takes for an injury based on position.*

*--Number of days missed based on injury*

*--What factors that may contribute to the injuries*

*--Is the Predictions of Injuries Accurate*

**Data Files:**

* **Injury Record**: The injury record file in .csv format contains information on 105 lower-limb injuries that occurred during regular season games over the two seasons. Injuries can be linked to specific records in a player history using the PlayerKey, GameID, and PlayKey fields.
* **Play List**: – The play list file contains the details for the 267,005 player-plays that make up the dataset. Each play is indexed by PlayerKey, GameID, and PlayKey fields. Details about the game and play include the player’s assigned roster position, stadium type, field type, weather, play type, position for the play, and position group.

**[Hazard Data Models](https://towardsdatascience.com/survival-analysis-intuition-implementation-in-python-504fde4fcf8e):**

Model uses to determines the likelihood of a player subject to injury based on conditions (Turf Types, Open Stadium Type and Weather Conditions)

Models are used for the following:

* [***Kaplan-Meier Estimate (Curve)***](https://www.sciencedirect.com/topics/medicine-and-dentistry/kaplan-meier-method)
* [***Cox Proportional Hazard Model***](https://www.sciencedirect.com/topics/medicine-and-dentistry/proportional-hazards-model)
* [***Logistic Regression***](https://medium.com/@dave.cote.msc/experimenting-confusion-matrix-for-regression-a-powerfull-model-analysis-tool-7c288d99d437)
* [***Random Forest Classifier***](https://towardsdatascience.com/explaining-feature-importance-by-example-of-a-random-forest-d9166011959e)

**Injury Data File Fields:**

A table with text on it

Description automatically generated

Noted that there is not a PlayKey available for every injury. This indicates that the game in which the injury occurred is known, but the specific play in which the injury occurred was not noted at the time of injury.

**Play List Data Fields:**

A table of text with text

Description automatically generated with medium confidence

\*\*Important Note: The GameID field is a unique identifier of player games but does not strictly reflect the order in which the games were played. The PlayerDay is an integer sequence that provides an accurate timeline for player game participation. To generate an accurate timeline of an individual player’s game participation, the PlayerDay variable should be used.

The interval between days in the PlayerDay field for an individual player accurately reflects the interval in days between that player’s participation in games. Every player has a PlayerDay = 1 (note that this date is not the same for all players). Some players may have negative values for PlayerDay, which simply indicates participation in a game that occurred before their individually assigned PlayerDay = 1.

**Data Results:**

Injuries shown that was abstained based on the lower parts of the body, where the Knee and Ankle are showing the most impact injuries however for the knee it’s displaying where injuries are impacted significantly with synthetic turf.

A graph of different colored bars

Description automatically generated

**Injury Data Body Part based on Player’s Position**

Two of the data charts below showing the comparison of injuries sustained by the three significant injury body parts: Ankle, Foot and Knee

A graph with different colored bars

Description automatically generated with medium confidence

VS

A graph with multiple colored bars

Description automatically generated

Knee is showing having more significant injury impact based on play type however, injury based on player position appears significate impacted by ankle and knee injuries.

Injury comparison between natural and synthetic grass turf types based on play type and player position.

A graph with orange and blue bars

Description automatically generatedVSA graph with blue and orange bars

Description automatically generated

Result displaying synthetic grass type have significant injuries by play type with punt not return, kickoff return, and kick-off not returned has most sustained injuries. With comparison with player position shows more injuries sustained from natural grass but less with synthetic grass

**Kaplan Meier Curve**

Implemented Kaplan Meier Curve model to predict the injuries uses to analyze injury occurred over period of number of days played.

Compared two curves for both natural and synthetic grass types.

A graph of different colors

Description automatically generated

Results shown that for both curves showing an increase likelihood of injuries occurring shown in the steep hill. Based on the pattern of this curve shows synthetic curve shows a small margin which indicated more injury risk for the synthetic than natural grass

**Confusion Matrix**

After examining the Kaplan Meier Curve, taking one step further and plot out the confusion matrix chart determine how accurate is the prediction and validate the data’s correctness.

Note to remember with the matrix plot:

It is a table that is used in classification problems to assess where errors in the model were made.

The rows represent the actual classes the outcomes should have been. While the columns represent the predictions we have made. Using this table, it is easy to see which predictions are wrong.

The Confusion Matrix created has four different quadrants:

False Negative (Top-Left Quadrant) | False Positive (Top-Right Quadrant)

True Negative (Bottom-Left Quadrant) | True Positive (Bottom-Right Quadrant)

A chart with blue squares and white text

Description automatically generated with medium confidence

The data shown with 26545 as false negative and 1222 False Positive which means that the prediction of this data is incorrect and inaccurate.

**Random Forest Classifier**

Lastly this model is used to evaluate the importance of each feature (or factors that contribute to injuries based on previous predictive model)

Chart below shows the results of the model:

A graph of different types of numbers

Description automatically generated

Results show that three features temperature, Player Day and Position are shown to have three significant importance which factors the injury. Interesting part shown as Field Type showing no significant importance impacting the injury prediction.