Advancing Injury Prediction in Football via Machine Learning

Siyan WANG Bushra TASNEEM Amon STHAPIT

Cornell Tech, New York, NY

sw2296@cornell.edu, bt353@cornell.edu, ars435@cornell.edu

Abstract – Enhancing injury prevention in American football is crucial for athlete safety and success. While making progress in this area can be a formidable task, the integration of machine learning technologies and data science shows great potential in providing valuable insights. Our primary goal was to harness the power of machine learning to investigate the factors related to lower extremity injuries in NFL athletes, taking into account player attributes, game dynamics, and environmental factors, based on historical injury data. In this research, we intend to employ a diverse set of machine learning algorithms, including logistic regression, support vector machine, random forests, and gradient-boosted decision trees to construct a predictive model. To assess the model's performance, we will utilize key metrics such as accuracy, precision, recall, F1-score, and the ROC AUC score.

1 Motivation

American football, known for its intense physical contact, is often cited as one of the world's most hazardous sports. Protective gear like helmets and pads provide some safeguarding, yet injuries remain prevalent. Since the 1970s, there's been an uptick in severe injuries and truncated careers among NFL athletes, coinciding with players becoming larger and faster. Historically, research has concentrated on head injuries, recognizing them as particularly perilous. Yet, it's crucial to note that injuries to the midsection and lower body also have significant consequences, potentially ending careers or causing chronic issues later in life. Furthermore, there has been a tendency to overlook the relationship between external factors such as field conditions and weather and the movements of athletes, which, if addressed, could offer athletes valuable insights into adapting their movements to minimize injury risks.

Our research aims to address these two overlooked areas. We plan to delve into the relationship between external elements, such as weather conditions and the state of the playing field, and internal dynamics, including the athletes' movements and positioning. Our objective is to apply different supervised learning algorithms on historical injury record data to discern patterns that may lead to lower-limb injuries. By understanding these correlations, we aim to diminish the incidence of these injuries, thereby safeguarding athlete well-being.

2 Background

2.1 Dataset

Our study utilizes data from the "NFL 1st and Future - Analytics" competition hosted on Kaggle [1]. This dataset encompasses 250 comprehensive in-game histories of players across two consecutive NFL regular seasons. It is compiled into three distinct datasets that contain varied information:

- **Injury Record**: The injury record file 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.
- **Player Track Data**: player level data that describes the location, orientation, speed, and direction of each player during a play recorded at 10 Hz (i.e. 10 observations recorded per second).

2.2 Models Used in the Previous Work

Injury prediction in football is critical, yet, regrettably, the majority of existing research on injury prediction relies on empirical and descriptive statistical analyses, including studies from top performers in Kaggle competitions. The application of machine learning techniques in this domain is notably rare.

3 Method

In this investigation, we are planning to implement logistic regression, support vector machines, random forests, and gradient-boosted decision trees.

3.1 Logistic Regression

Logistic regression is the most widely used supervised learning binary classification algorithm, which would produce a linear

boundary. Logistic regression uses a model f_θ in the form of

$$f_\theta(x) = \delta(\theta^T x) = \frac{1}{1 + exp(-\theta^T x)},$$

where $\delta(z) = \frac{1}{1+exp(-z)}$ is the sigmoid or logistic function.

Logistic regression is easy to implement and interpret, and it can serve as a baseline for performance comparison with more complex models.

3.2 Support Vector Machines

Support Vector Machine (SVM) is a supervised machine learning algorithm that can solve classification problems as well. It operates by finding a hyperplane that best divides a dataset into classes with the widest margin between support vectors. For non-linearly separable data, SVM employs kernel functions to project the data into higher-dimensional spaces where a hyperplane can be drawn.

The mathematical formulation of a linear (SVM) for a binary classification task is centered around finding the optimal hyperplane that separates the data points of different classes with the maximum margin. The hyperplane can be described by the equation:

$$\theta^\intercal x + \theta_0 = 0$$

Here, θ represents the weight vector, x represents the feature vectors, and θ_0 is the bias term.

The objective of the SVM is to minimize the norm of the weight vector $||\theta||$ while ensuring that the data points x_i are classified correctly, which means to minimize

$$\frac{1}{2}||\theta||^2,$$

subject to for every i, $y_i(x_i^\intercal \theta + \theta_0) \geq 1$, where y_i represents the label of each data point, which is either +1 or -1 depending on the class. The constraint $y_i(\theta^\intercal x_i + \theta_0) \geq 1$ ensures that all data points are on the correct side of the margin, or on the margin itself for the support vectors.

For cases where the data is not linearly separable, a soft-margin SVM introduces slack variables ξ_i to allow some data points to be within the margin or even on the wrong side of the hyperplane. The optimization problem then includes a trade-off parameter C that controls the penalty for these violations, which becomes to minimize:

$$\frac{1}{2}||\theta||^2 + C\sum_{i=1}^{n}\xi_i,$$

subject to for every i, $y_i(\theta^\intercal x_i + \theta_0) \geq 1 - \xi_i, \xi_i \geq 0$. Here, a higher value of C puts more emphasis on minimizing the classification error, while a lower value of C puts more emphasis on maximizing the margin.

Support Vector Machines (SVMs) excel in high-dimensional spaces and are renowned for their robust generalization capabilities. This strength lies in their ability to minimize overfitting, a feat accomplished through fine-tuning the regularization parameter. This tuning process carefully balances bias and variance, enhancing the model's predictive accuracy. However, SVMs have limitations, particularly in handling imbalanced datasets. In such scenarios, they may exhibit a bias toward the majority class. Thus, we were implementing tree-based models, which are well-regarded for their efficacy in managing imbalanced datasets, in this project as well.

3.3 Random Forests

Random forests are a versatile supervised learning algorithm suitable for both regression and classification tasks. This method operates by building numerous decision trees during the training phase. In classification tasks, such as the one addressed in this project, random forests make predictions based on the mode of the classes output by individual trees. This ensemble approach enhances the model's generalizability, effectively mitigating the overfitting issue commonly associated with single decision trees.

Random forests are robust to overfitting, effective for handling complex feature interactions with little data preparation (no rescaling, handle continuous and discrete features) and accurate predictions. All of those features made random forests a solid choice in injury datasets.

3.4 Gradient-Boosted Decision Trees

Gradient Boosted Decision Trees (GBDT) are an ensemble learning technique that builds a model in a stage-wise fashion by combining the predictions from multiple decision tree models. Each tree is trained to correct the errors of the previous one, effectively improving the model's accuracy with each iteration. GBDT uses the gradient descent algorithm to minimize the loss when adding new models. GBDT is particularly known for its effectiveness in handling a variety of data types, robustness to outliers, and ability to model complex non-linear relationships.

Gradient-boosted decision trees are powerful for predictive tasks and often perform very well in structured data problems. In particular, it is known for its speed and performance, and its ability to handle missing data and capture complex patterns in the data effectively.

4 Setup

4.1 Data Exploration

From some rudimentary descriptive analysis, we could see that we have almost equal number of observations for synthetic and natural surface. The majority (74%) of the stadiums were outdoor stadiums. The temperature varied widely. During the game, it ranged from 10° Fahrenheit to 90° Fahrenheit with mostly ranging from 35-90° Fahrenheit.

When we explore the injury count by the injured body part and the surface type on which the injury happened, the most frequent type of injuries were knee injuries and we observe equal number of knee injuries on both synthetic and natural

playing surfaces. In contrast to this, for the second most frequent type of injury is an ankle injury, more ankle injuries are observed on a synthetic surface compared to a natural playing surface. Additionally, when we visualize the number of injuries under different weather conditions on different turfs, the proportion of injuries on natural turf in rainy weather is relatively high. This gives us some insight that weather and type of turf may influence the chances of getting injured.

Also there are only 105 injury records in total, which means that the dataset has a very large imbalance of class labels.

4.2 Data Pre-Processing

For data pre-processing of the NFL Injury dataset, we performed manual lemmatization of categorical features like the weather during a game and the type of stadium that was used for the game. This helped us filter the data from redundant and descriptive texts into a small selection of classes that were representative of the different types of weather and stadium types present during those games.

Then we used one-hot encoding to represent all categorical features as binary vectors so that they could be used by the machine learning models for training.

To handle missing values, we impute missing values with the appropriate values (either 0 for false or the average of valid values from the same column) for completeness of data so that we do not get biased results when we run our model.

We also refined the target variable through introducing a new column, "Injury", as a flag indicating the occurrence of an injury. This was derived from other injury columns serving as a critical dependent variable.

We consciously decided not to implement dimensionality reduction techniques in this project due to two primary considerations. First, the dataset exhibits significant imbalance; our focus extends beyond mere dimensionality reduction to ensuring comprehensive representation and learning of the minority class by the model. Second, in addition to logistic regression, which serves as our baseline model, we plan to deploy other models that are known to perform optimally in high-dimensional settings.

5 Preliminary Experiments and Metrics

5.1 Cross-Validation and Resampling

Because our dataset has a large imbalance in the class labels for player injuries, we used stratified 5-fold cross validation to separate our dataset into training sets and test sets, so that the proportion of injury labels in the original data is maintained across each of the training and test sets.

However, even after the stratified splits our data still has a large imbalance of class labels. There are over 250,000+ records with no injuries and just 105 injuries recorded, which may teach the classification models to predict no injuries for everything. To address that issue, we used the Imbalanced-learn library's RandomOverSampler method to oversample the training set so that the minority class label (of actual injuries) is represented more. This will help the classification models to better train and learn from the data.

5.2 Metrics

In this project, we monitored 5 metrics to measure the performance of our model: accuracy, precision, recall, F1-score, and ROC AUC score. We define True Positive (TP) as correctly classifying injured as injured, True Negative (TN) as correctly classifying uninjured as uninjured, False Positive (FP) as misclassifying uninjured as injured, and False Negative (FN) as misclassifying injured as uninjured. The metrics are defined below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - score = 2 \cdot \frac{precision \cdot recall}{precision + recall}.$$

The ROC curve is a plot of the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The AUC measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1). ROC AUC score provides an aggregate measure of performance across all possible classification thresholds. The higher the score, the better the model is at predicting true positives while minimizing false positives, across all possible thresholds.

Among 5 metrics, we prioritized the F1-score and the ROC AUC score due to the imbalanced nature of our dataset, especially the ROC AUC score. The F1-score, which harmonizes precision and recall, becomes a more critical measure under these circumstances. ROC AUC score evaluates the model's ability to distinguish between classes without being influenced by the distribution of those classes. It provides a balanced metric for assessing performance irrespective of class imbalance.

Additionally, recall is of particular importance in our analysis, particularly from the standpoint of athlete safety. While accuracy and precision are also monitored, they serve more as supplementary metrics in our model evaluation process.

5.3 Baseline: Logistic Regression

Our baseline model is logistic regression as it is one of the simplest yet most widely used model for binary classification.

Over 5 different cross validations, our simple logistic regression model achieved, on average, an accuracy score of 0.71, precision score of 0.0005, recall score of 0.56, F1-score of 0.0009, and ROC AUC score of 0.6.

The logistic regression model predicted a larger number of injuries than there were in the actual test set. However, 0.5 is

the baseline of ROC AUC score for random guessing; thus, the model still performs better than random guessing. This model serves as our performance benchmark.

6 Further Experiments and Metrics

6.1 Support Vector Machines

In our next experiment, we employed a support vector machine (SVM) model with a radial basis function (RBF) kernel. The choice of the RBF kernel was because: firstly, it offers flexibility in the decision boundary, enabling the mapping of input data into a higher-dimensional space for a non-linear boundary. Secondly, its versatility and lower overfitting risk make it an ideal first choice in varied scenarios, especially when the feature-label relationship is ambiguous.

Unfortunately, the SVM did not perform better than expected. The SVM showed an accuracy of 0.91, but its precision was notably low at 0.0002, coupled with a recall of 0.08 and an F1-score of 0.0004. Interestingly, the ROC AUC score was 0.49, indicating a performance close to random guessing. This suggests that the SVM struggled to correctly classify the minority class in our imbalanced dataset.

The result was disappointing, but not entirely out of our expectations. As introduced above, the SVM might be sensitive to imbalance and hence biased towards the majority class, leading to poor classification performance for the minority class. This outcome underscores the critical need for advanced strategies in handling imbalanced datasets to enhance the model's ability to accurately identify instances of the less prevalent class.

6.2 Random Forests

Random forests, which combines multiple decision trees, is effective in handling imbalanced datasets since the model is able to capture the complexity of different classes by constructing multiple trees. In order to optimize the model, prevent overfitting, and improve generalization, we implemented a pruning strategy of controlling Maximum Tree Depth.

We started from setting the max_depth = 5. Over 5 different cross validations, the model achieved, on average, an accuracy score of 0.71, precision score of 0.0006, recall score of 0.57, F1-score of 0.0012, and ROC AUC score of 0.67. Although the general accuracy was not improved, two prioritized metrics, F1-score and ROC AUC, have been improved, which demonstrated its ability to catch minority groups. A minor improvement in recall is also favorable for our objectives, as our primary goal is to minimize the occurrence of injuries. This improvement in recall indicates a better capability of the model to correctly identify actual injury cases, which is crucial in preventive strategies.

When we increased the max_depth parameter to 10 in our model, we observed a significant improvement in accuracy, reaching a high of 0.94. However, this enhancement in accuracy came with certain trade-offs. Specifically, the model showed a precision of only 0.0006 and a recall of 0.11, which indicates a considerable challenge in correctly identifying the minority class. The F1-score, a measure of a test's accuracy, was also low at 0.0012. Furthermore, the ROC AUC score, which represents the ability of the model to distinguish between classes, decreased to 0.63. This suggests that while the model became more accurate overall, its performance in identifying the minority class was compromised.

Further adjustments were made by increasing the max_depth to 15. This change led to an even higher accuracy rate, reaching 0.99. However, the model's ability to identify the minority class was completely lost, as indicated by the precision, recall, and F1-score all dropping to 0.0. Additionally, the ROC AUC score experienced a slight decline, falling to 0.61. These results are indicative of severe overfitting to the majority class present in the training data. In essence, the model achieved high accuracy at the expense of its generalization capability and its ability to correctly identify instances of the minority class.

In juxtaposition with the baseline logistic regression model, the random forests showed a notable enhancement in both F1-score and ROC AUC metrics at various tree depths. This improvement signifies a more balanced performance in terms of precision and recall, particularly crucial for the F1-score which harmonizes these two metrics. Moreover, the enhancement in the ROC AUC score reflects a better overall capability of the random forests model in distinguishing between the classes, especially relevant in the context of our imbalanced dataset. When the maximum depth increases, the random forests demonstrated an ability to achieve higher accuracy at greater tree depths. However, this increase in accuracy was consistently offset by a decrease in the model's capability to correctly classify the minority class, as indicated by the declining recall and ROC AUC scores. These results underscore the critical balance between model complexity and its ability to generalize, especially in the context of an imbalanced dataset.

6.3 Gradient-Boosted Decision Trees

Lastly, we employed the gradient-boosted decision trees (GBDT) model, leveraging its strengths in handling complex and non-linear data relationships. Theoretically, this model is capable of capturing intricate class dynamics but with a distinctive approach of sequentially building and combining trees to minimize errors.

Initially, we configured the maximum tree depth of the GBDT model to 3. Across 5 iterations of cross-validation, this configuration resulted in an average accuracy of 0.83, accompanied by a precision of 0.0004, a recall of 0.22, an F1-score of 0.0007, and a ROC AUC score of 0.61. Compared to the logistic regression baseline, the gradient-boosted decision trees model (max_depth = 3) has a higher accuracy and ROC AUC value while a lower precision, recall, and F1-score. This suggests that when max_depth = 3, while GBDT is overall more accurate and slightly better at distinguishing between classes (as indicated by the ROC AUC) in this case, it is less effec-

tive in correctly identifying the minority class, a critical aspect in the context of our imbalanced dataset. These findings highlight the nuanced trade-offs inherent in model selection and the importance of considering a range of performance metrics, particularly in imbalanced scenarios where high accuracy does not necessarily equate to overall model efficacy.

Adjusting the model's max_depth to 7 led to a notable increase in accuracy (0.98), but this didn't extend to other metrics: precision was very low at 0.0003, recall at just 0.0125, and the F1-score at 0.0006, indicating poor performance in identifying the minority class. The ROC AUC score remained at 0.60, the same as the logistic regression. These results suggest that the model's effectiveness in capturing and correctly classifying the minority class was limited.

Increasing max_depth further to 10 achieved near-perfect accuracy (0.9996), but with significant drawbacks. Precision, recall, and F1-score dropped to zero, indicating a failure in detecting the minority class, likely due to overfitting. The slight improvement in the ROC AUC score to 0.6120 still showed difficulties in class distinction, implying that the higher accuracy did not equate to better overall model performance, especially in handling the minority class.

Compared to the the baseline logistic regression model, the GBDT model, particularly at lower depths, showed some improvements in accuracy and ROC AUC. However, as the maximum depth increased, we observed a pattern and challenge similar to that in random forests: higher accuracy but at the expense of the model's sensitivity towards the minority class.

7 Conclusion and Future Work

In this project, we pioneered the use of machine learning techniques to predict potential injuries in NFL athletes, an approach that was largely overlooked in prior research. However, the prior empirical and descriptive analysis provided us with foundational insights into our datasets, guiding our approach to feature engineering.

Our exploration revealed significant variations in the performance of different models.The logistic regression served as a fundamental baseline. Its results highlighted its limitations in complex, imbalanced scenarios. This prompted us to explore alternative machine learning methods. However, not all models surpassed this baseline. The performance of SVM was notably poor, falling short of even the logistic regression baseline.

Particularly noteworthy was the performance of random forests, which displayed a notable potential in handling the class imbalances inherent in our dataset. The random forests model, especially at lower tree depths, showed improvements in correctly identifying actual injury cases. However, a crucial observation was the diminishing returns in model performance with increasing tree depth, particularly in terms of precision, recall, and F1 scores. These outcomes underscore the importance of carefully calibrating tree complexity to maintain a balance between accuracy and the ability to detect the minority class.

On the other hand, the gradient-boosted decision trees, theoretically an excellent choice for imbalanced datasets due to its sequential, error-correcting approach, did not markedly enhance performance in our specific context. This deviation from expectation highlights a critical aspect of machine learning in practice: theoretical advantages do not always translate into superior performance in real-world scenarios. Factors like hyperparameter settings, the nature of the data, and the complexity of the model can significantly influence the outcomes. Therefore, while gradient-boosted decision trees are powerful tools, their application requires careful tuning and consideration of the dataset's unique characteristics.

Table 1: Model Performance Summary (max_depth = 5 for Random Forests and max_depth = 3 for GBDT)

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
Logistic Regression	0.71	0.0005	0.56	0.0009	0.6
SVM	0.91	0.0002	0.08	0.0004	0.49
Random Forests	0.71	0.0006	0.57	0.0012	0.67
GBDT	0.83	0.0004	0.22	0.0007	0.61

For future work, we are planning to concentrate on:

a. **Model Optimization**

- Incorporating additional data sources: increasing the number of data points, particularly those pertaining to injuries, and adding additional features, for example, player histories, training routines, and detailed game conditions.
- Implementing more sophisticated oversampling or undersampling techniques: Techniques like SMOTE or ADASYN might prove beneficial in providing more balanced training data.
- Further model training: for example, a deeper investigation into random forests, with a focus on advanced techniques like feature selection, more sophisticated tree pruning strategies, and exploring various ensemble methods, might yield further improvements. More extensive hyperparameter tuning using methods like grid search or Bayesian optimization would be beneficial as well. This is particularly relevant for complex models like gradient-boosted decision trees.

b. **Application**
With a robustly performing model, we can then delve deeper into features strongly correlated with athlete injuries. This analysis will enable us to offer targeted injury-prevention recommendations to both athletes and the NFL, aiming to further mitigate injury risks.

8 Reference

[1] NFL 1st and Future - Analytics, https://www.kaggle.com/competitions/nfl-playing-surface-analytics/overview