CVPR
#

ML Project Proposal

INJURY RISK MINIMIZATION

1. Abstract

Sports Injury has been the biggest threat amongst the players, teams and sporting management since the advent of the sporting culture. Our objective here is to use this data to examine factors that may contribute to lower extremity injuries. In this paper, we present machine learning based solutions to predict the injury risk factor and the number of days required to recover in case an injury occurs.Predicting injury beforehand would be a huge help to the players, ultimately revolutionising the sports industry and knowing the resting period in advance , would help teams strategize in a better manner for future tournaments. We also aim to tell the players in advance about the body part more likely to be injured so that the players can take prior measures in order to prevent those injuries. The github repository can be found here here .

2. Introduction

Billions of dollars are invested in sports industry in order to enhance the sports performances and reduce the risk of Injuries. A lot of factors influence the risk factor in an injury. The given dataset helps us to analyse the playing conditions like field type, weather, and temperature along with the various plays of every single player along with their movements, position, turf, speed etc. Our Classifier takes into account all these factors to predict if an injury would occur or not. Our Regressor model then estimates the number of days required to recover from the injury.Further we also predict the body part most likely to be injured using our multi class classification models.

3. Literature Survey

3.1. Thesis on Predictive modelling of football injuries:

The goal of this thesis is to investigate the potential of predictive modelling for football injuries. This work was conducted in close collaboration with Tottenham Hotspurs and the PGA European tour. In this review,mainly three investigations were conducted[1]:

- **Predicting the recovery time of football injuries using the UEFA injury recordings:** For this investigation, three datasets of UEFA injury recordings were analysed by different machine learning algorithms to build a predictive model.
- **Predicting injuries in professional football using exposure records:** The relationship between exposure (in training hours and match hours) in professional football athletes and injury incidence was studied. The primary task was to predict the number of days a player can train before he gets injured.
- **Predicting intrinsic injury incidence using in-training GPS measurements:** A significant percentage of football injuries can be attributed to over training and fatigue which can be detected using GPS.This research aims to predict when an injury is most likely to take place for different players of THFC team using the GPS data which was gathered during their training sessions.

3.2. Modelling the Risk of Team Sport Injuries: A Narrative Review of Different Statistical Approaches-:

Injuries are a common occurrence in team sports and can have significant financial, physical and psychological consequences for athletes and their sporting organisations. There are a number of methods that can be used to identify injury risk factors but choosing the right method is trivial as wrong statistical approaches can lead to incorrect inferences and decisions.[2]

This narrative review aims to -:

- Outline commonly implemented methods for determining injury risk
- Researchers should carefully consider the different types of variables that were examined in relation to injury risk and how the analyses pertaining to these different variables were interpreted.
- Describe advances in statistical modelling and the current evidence relating to predicting injuries in sport.

CVPR
#

3.3. Kaggle Notebook on Feature Analysis of the NFL Dataset :

Involves the visualization of different Data Frames together for the better understanding of features.

4. Dataset Details:

The dataset is divided into two parts: Injury Recors and Playlist. 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. 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.

- **Analysis of PlayList :**
  1)There are 250 players in the dataset.
  2)There are 5712 games in the dataset.
  3)There are 267005 plays in the dataset.

- **Analysis of Injury Record:**
  1)There are 105 injuries records in total
  2)100 unique players have been injured and hence cases of multiple injuries to the same player is present.
  3)28 Playkey values are missing.

5. Dataset Preprocessing:

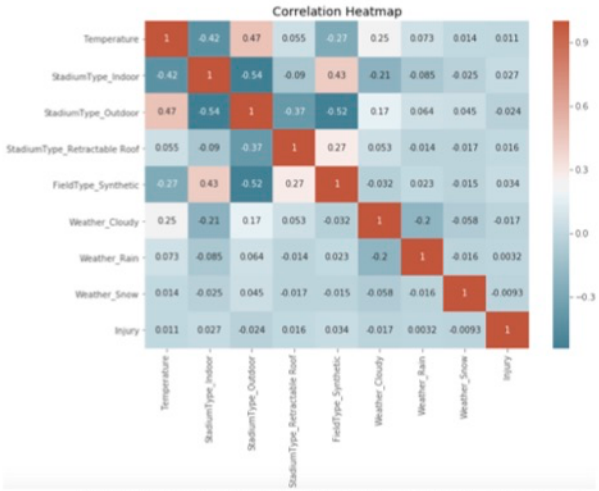
- We replaced the missing values of temperature with the average temperature value.
- Our objective was to detect whether the player has suffered an injury or not however there was no field in the original dataset for the following , so we made use of the attributes DMM1, DMM7,DMM28,DMM42 , and if any one of them had a non-zero value , henceforth an injury has occured and therefore it is an instance of injury.
- Some play days were negative due to some erroneous data so their absolute values were taken as play days can never be negative.
- Non-numeric data of certain features were converted to binary data using one hot encoding for effective model training.
- We joined Injury data and play data to increase the number of attributes in our final dataset and then took care of all the null values.
- We performed oversampling in order to increase the instances of injury to perform our task 2 and task 3.

6. Problem Statement:

- **Task1-:**Predict whether the player suffered from injury or not based on attributes like Stadium Type,Field Types,Weather,etc.
- **Task2-:**Predict the number of days of rest needed by the player in an occurrence of injury.
- **Task3-:**In an occurrence of injury,predict the body part most likely to be injured based on other attributes.

7. Methodology:

- **Task 1: Classification Problem**
  - Aggregation of Data:
    - With a total of 267005 playsKeys and only 105 injuries, the dataset is highly biased. We solved this problem by aggregating the records with the GameID (number = 5712) instead of playKey.
    - Since the class distribution is biased towards no injury,so we resampled the data before feeding it to the classification model.
  - We analysed the features and made a correlation map to identify the dependent features.Correlation was found between-:
    - Weather and Temperature
    - Stadium type and Temperature



  - Final features selected for injury prediction are-:
    - RoasterPosition
    - PlayerDay
    - Player Game
    - Stadium Type

CVPR
#

* Player Position
* Field Type
* Temperature
* Weather
* Play Type

– Models used for performing the classification task to predict whether an injury has occurred or not based on other features of the player,game and stadium-:

* **Gaussian Naive Bayes**
* **Logistic Regression**
* **Decision Tree**
* **XGBoost**[6]

– K Fold cross validation was finally used to evaluate the model.

- **Task 2: Regression Problem**

– We modified the column "Injury" such that it contains the sum of values contained in DMM1,DMM7,DMM28,DMM42.Now the column "Injury" has values ranging from 0 to 4.A 0 value indicates an occurrence of no injury while a non zero value indicates an occurrence of injury.

– We further created a new column which indicates the days of rest a player has to take in occurrence of an injury. For each occurrence of 1,2,3 and 4 in injury column a random value generated using Gaussian distribution in the intervals of 1-7,7-14,14-28,28-42 was correspondingly added in the new column.For a instance of non injury number of days of rest is 0.

– Then we ran the following regression models on the data to predict the number of days of rest needed in an occurrence of injury-:

* **Linear Regression**
* **Support Vector Machine**
* **Random Forest Regressor**

- **Task 3: Multi Class Classification Problem**

– We combined the injuryRecord.csv with PlayList.csv along the 'GameID' column using inner join.We took care of the null values which appeared while joining the two datasets.Then we dropped certain unnecessary columns and oversampled the data in order to increase the number of rows in the data.

– We took the body parts Knee,Foot,Ankle,Heel and Toes in account and predicted the body part most likely to be injured based on other attributes.We used the following algorithm to perform the Multi class classification problem-:

* **Logistic Regression**
* **Decision Tree**
* **Multi Layer Perceptron**

8. Results

- **Task1: Classification Problem**

Classification Task

Classifiers	Accuracy	Precision	Recall	AUC
Naive Bayes	0.41532976827	0.3581213307	1.0	0.60
Logistic	0.67914438502	0.51311953352	0.32058287	0.74
DecisionTree	0.780848787	0.708830548	0.953281027	0.84
XGBClassifier	0.85264408	0.80194174	0.7388193	0.92

Table1: Results for Binary Classification problem

- **Task2: Regression Problem**

Regression Task

Classifiers	MSE
Linear Regression	134.80333333333334
SVR	34.99333333333333
RandomForestRegressor	8.609047619047619

Table2: Results for Regression problem

- **Task3: Multi Class Classification Problem**

Model	Macro Precision	Macro Recall	Macro F1 Score	Accuracy
Logistic Regression	0.8206	0.6630	07098	0.6952
Decision Tree	0.8687	0.8791	0.8643	0.8952
Multi layer Perceptron	0.9734	0.9488	0.9600	0.9428

Table3: Results for Multi-label Classification problem

9. Analysis

- **Task1:**

– **Gaussian Naive Bayes**.It performed poorly as was expected from our analysis before, as the features were highly correlated.

CVPR
#

- **Logistic Regression:** In our dataset after binary conversion of non-numeric attributes the number of attributes are 60 and the data becomes highly sparse and multi dimensional as a result of which logistic regression is failing to converge. Also Logistic regression fails to optimally fit the data as the data is not linearly separable.
- **Decision Tree:** Accuracy and precision found using the decision tree classifier was better as compared to LR and NB as it was a decision based problem and the Decision trees provided a clear indication of which fields were important for classification.
- **XGBoost:**[6] The best accuracy and precision for the data set was found using the XGBoost which is an ensemble learning technique.It is because ensemble based learning techniques produce better results than a decision tree.

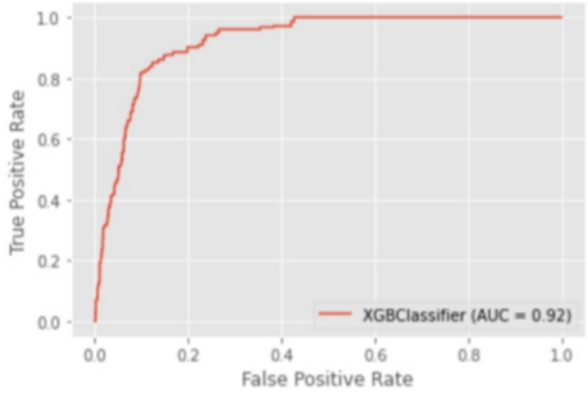


Fig1: ROC AUC plot for XGBClassifier

- **Task2:**
  - **Linear Regression** We used Linear Regression as a baseline for our regression problem. The mean squared error was highest for this model. Since the data is highly dimensional, a linear regressor does not produce satisfactory results.
  - **Support Vector Regressor** Essentially it creates a single optimal hyperplane based on the support vectors thus producing a better result compared to Linear Regression . However the problem is better addressed by ensemble learning classifiers.
  - **Random Forest Regressor** This ensemble learning algorithm helped us in achieving the best results that gave minimum error.

- **Task 3:**
  - **Logistic Regression** Similar to the Task1, the non-numeric attributes were converted to binary and the total number of features fed into the model were 57. Thus the data becomes too sparse and highly multi-dimensional for a logistic regressor to handle. The model fails to converge. The model fails to optimally fit this non-linearly separable data.
  - **Decision Tree Classifier** The Decision Tree performed better than the Logistic Regressor. Also, decision trees are more prone to overfitting the data and this was the reason that we got a higher performance on the training data as copared to the testing data.
  - **Multi Layer Perceptron** Our multi layer perceptron performed the best with the following ROC curve. This is because of the use of non-linear activation functions that allows the model to capture complex non-linear relations in the data.

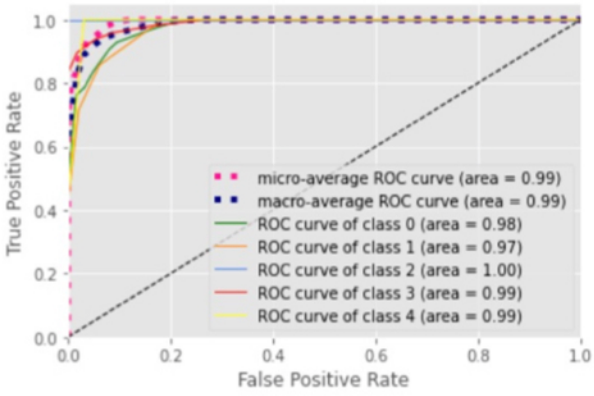


Fig2: ROC AUC plot for MLPClassifier

10. Conclusion

The problem allowed us to work with real player injury data which was a highly imbalanced data . We applied various pre-processing techniques for removing the NA values and made use of correlation matrix to drop highly correlated features and applied dimensionality reduction for faster convergence before feeding it to the models. Furthermore to tackle imbalanced dataset, various re sampling techniques were applied and certain tweaks were made to combine the datasets .Variety of baseline models like Naive Bayes,Logistic Regression and Decision Tree were tested with different parameters.Further we applied ensemble learning methods like XGBoost to reduce the variance and thus resulting in better performance.

CVPR
#

ROC curve of different models to find out that for the classification problem XGBoost classifier performed the best amongst the tested models under optimal model parameters.The model parameters for XGBoost were found using GridCV.

For the Regression problem various models like Linear Regression,SVM and Random Forest Regressor were applied to achieve minimum MSE loss. Random Forest Regressor along with optimal model parameters performed the best amongst the tested models.

In the last phase of this project we worked on creating a model which tells us which body parts are more prone to injury given the set of features related to the field, weather and player's health. This was a multi-class classification problem. Logistic Regressor and Decision Tree were used for baselining this problem. Finally, a Multi Layer Perceptron Classfier was used to capture the complex relations in the data.We observed that MLP Classifier gave the most satisfactory results for this dataset.

11. Individual Contribution

- **Data Cleaning and Organisation:** Larika, Yashdeep
- **Literature Review:** Shreeya
- **Exploratory data analysis** Larika, Shreeya, Yashdeep
- **Creating Learning Model:**
  1. Injury Prediction(Binary Classification Task): Larika, Yashdeep
  2. Estimating recovery period(Regression Task): Shreeya, Yashdeep
  3. Predict the Bodypart most prone to injury(Multi-Label Classification Task):Larika, Shreeya
- **Analysing Accuracy among all models and reasoning for the best output:**
  1. Binary Classification Problem: Larika, Shreeya
  2. Regression Problem:Shreeya, Yashdeep
  3. Multi-Label Classification problem Problem: Larika, Yashdeep

12. References

1.Kampakis, S; (2016) Predictive modelling of football injuries. Doctoral thesis , UCL [thesis]

2.Joshua D. Ruddy,1,* Stuart J. Cormack, Modeling the Risk of Team Sport Injuries: A Narrative Review of Different Statistical Approaches. [ncbi]

3.A Machine Learning Approach to Assess Injury Risk in Elite Youth Football PlayersResearchGate

4.Akobeng A. K. (2007a). Understanding diagnostic tests 1: sensitivity, specificity and predictive values. Acta Paediatr. 96 338–341. 10.1111/j.1651-2227.2006.00180.x [PubMed]

5.Altman N., Krzywinski M. (2015). Points of significance: association, correlation and causation. Nat. Methods 12 899–900. 10.1038/nmeth.3587[PubMed]

6.Tianqi Chen, Carlos Guestrin (2002) ,XGBoost: A Scalable Tree Boosting System

7.R. Bekkerman. The present and the future of the kdd cup competition: an outsider's perspective