LULEÅ
UNIVERSITY
OF TECHNOLOGY

Project: Sports Injury Prediction

Predictive Analytics D7056E

Group 3:

Athanasios Koutoulas

Bruno Reis Da Silva

Isabella Södergren

Venkata Sreenadh Movva

1. Introduction

This document refers to the Sports Injury Prediction project with focus on US NFL for the Predictive Analytics course. The document summarizes the steps followed to complete this project from the perspective of the CRISP-DM process and all its phases. The practical part of the project like data exploration and preprocessing, modeling, training and evaluation, were performed in Anaconda platform using Jupyter Notebook and Python 3.10.

2. CRISP-DM Process

In this section we describe the six phases of the CRISP-DM process that were followed in this project

2.1 Business Understanding

A sports organization revenue heavily relies on some factors like ticket and subscription, jersey sales, profits from player transfers etc. Those essential factors can be negatively impacted by player injuries, leading an organization to extra costs and declining revenue. The goal of this study is to tackle the problem of player injury in sports, by exploring and suggesting different predictive solutions. In order to understand the problem of sport injuries more in depth, we list some of the issues a sport organization could face, focusing on the financial aspect:

- The club will have to invest extra money for covering his/her operation and treatment costs.
- The club has to still pay a player's salary while he cannot really offer to the team goals while he/she is injured.
- Possibly the club will have to dedicate extra resources (doctors, special trainers) for ensuring that the player follows an individual training program throughout his recovery time to ensure a smooth integration back to the team selection. Potentially buy new equipment as well to ensure player's smooth and quick recovery. Extra resources or equipment implies extra costs for the club.
- When a player suffers from injuries during his service at a certain club, it is highly likely that the club won't be able to make a profit out of a potential future transfer. Very possibly, the club will lose some money from a potential future transfer (comparing the price of the player to buy initially and to sell after several injuries).
- One of the factors a club "buys" a player is the player's popularity in the market. Transfers of popular players in many cases increase ticket selling, yearly subscriptions, team jerseys selling etc. While a player is out of service for a long period of time, due to injuries, his popularity in the market drops. That can affect ticket selling or jersey selling and in general can have a negative impact on the club's revenue.
- Injured players are unavailable for selection before games. Missing important players for long periods due to injuries might impact game results negatively. Negative results have a negative financial impact on the organization, since ticket selling, jersey selling, subscriptions might be reduced in such case.
- Occurring player injuries in high frequency at a certain sport club, might be a serious concern for other players and can negatively affect the possibility of future player transfers.

2.2. Data Understanding

For our project we chose a dataset with injuries collected from the US National Football League (NFL) over two seasons. We worked with two initial datasets (submitted as part of the practical work, PlayList.csv and InjuryRecord.csv), which were eventually merged into one final dataset, in the data preparation process. The "PlayList" dataset contained data related to game and player specifics such as weather, temperature, stadium and field type, player participation time in the game, position, players play style in the game, number of continuous matches played and total participation etc, while the "InjuryRecord" provides information related to injuries such as the game id when the injury occurred, the player id, the body part injured, the duration of the injury after occurrence (recovery time) etc. In first place we became familiar with NFL terminology in order to understand certain features and values in our datasets and then we created the ABT tables for both continuous and categorical features for both datasets. Through the data quality report, we identified the issues in our datasets, issues such as missing values, negative values for features that could not be negative, features with cardinality equal to 1, categorical feature with unusually high cardinality and we addressed those issues in the data preparation phase. Moreover, we explored the datasets through visualizations in order to understand them in depth, we searched for outlier values and relations between features.

2.3. Data preparation

For the data preparation part, we fixed the data quality issues addressed in the data quality reports during the data understanding process for both datasets we used in this project. More specifically for the continuous features in both dataset we replaced negative values with their absolute values, for features that is not possible to contain negative values considering those as a mistake during data insertion, we replaced missing values such as temperatures (-999) with the median value. Moreover, we removed continuous features with cardinality equal to 1 and we searched for outlier values, but the datasets didn't contain any after normalization. For the categorical features in the two datasets we replaced missing values with the mode, we normalized features with unusually high cardinality to contain more meaningful values, for example for the weather categorical feature in the PlayList dataset we transformed certain values into one meaningful value (example: Rain, Rainy, Light rain replaced by Rain) and we did the same for the stadium type feature (example Indoor, Indoors, Closed replaced by Indoors) and thus we derived more meaningful cardinality for those features. Finally, in order to achieve better performance with our models we derived additional features representing categorical values with binary values, in other words, we performed one-hot encoding to the categorical features, for example from the Weather feature we derived features like Weather_Rainy, weather_Snowy, etc. and we removed duplicated entries in all datasets and we merged the two datasets into one final dataset before providing it to our models.

2.4. Modeling

We explore and suggest three different predictive approaches that can used to tackle and minimize the sports injury problem. The first approach is to build models capable of predicting possible player injury in the next game based on next game's specifics and uncorrelated factors like stadium and field type, weather conditions, temperature, players position and play style etc. For this binary classification task, we trained several predictive models including K-nearest Neighbors, Naïve Bayes, Logistic Regression, Decision Tree Classifier, Multi-layer Perceptron, Random Forest, XGBoost Classifier and some deep learning models like LSTM and GRU Neural Networks. Moreover, for this task we used a dataset of 59 descriptive features used to predict the derived injury target feature (0: no injury, 1: injury). The target feature was derived out of four features in the initial dataset indicating different time ranges a player was out of service due to injury. Moreover, we performed oversampling technique in the dataset in order to achieve a more balanced dataset in terms of injuries and no injuries. The ratio after oversampling was still 1:4 for the no-injury datapoints, and the final dataset consisted of 7010 data points. Finally, we split the dataset to 80:20 for training and testing for the algorithmic models and 70:30 for the deep learning models. The practical part for this approach is submitted as task_1.ipynb. The second approach is to build models able to predict the body parts likely to be injured based on the same uncorrelated factors and next game's specifics as described above. Similarly to the first approach, we trained several predictive models such as Logistic regression, Decision Tree Classifier, Multi-layer Perceptron, Random Forest, XGBoost Classifier, LSTM and GRU Neural Networks. For this multi-class classification problem, we used 55 descriptive features to predict the multi-label target feature consisting of injured body parts (ankle, foot, heel, knee, toes etc.). For this task we used only the injured player cases of the initial dataset and we perform oversampling in order to increase the number of datapoints (1050), while still maintaining the initial ratio of injured body parts as they occur in the initial dataset. Here before providing the dataset to the Neural Networks, we performed one-hot encoding for the categorical target feature as well. Similarly to the first approach, here we also split the dataset to 80:20 for training and testing for the algorithmic models and 70:30 for the deep learning models. The practical part for this approach is submitted as task_2.ipynb. The third approach we suggest for this problem is to build predictive models able to predict a sufficient number of resting days after injury, in order to prevent it from reoccurring. For this approach we build several models like Linear Regression, Random Forest Regressor, Support Vector Regressor, LSTM and GRU. For this task we used 51 descriptive features to predict the continues target feature indicating number of resting days. Based on the four features of the initial dataset indicating different time ranges a player was out of service due to injury, we generated random numbers of resting days within those time ranges, comprising the target feature of this task. Furthermore, similarly to the other tasks, oversampling was performed here as well in respect to the time ranges occurring in the initial dataset. The dataset consisting of 1050 datapoints was provided to the models and the dataset was split to 80:20 for training and testing for the algorithmic models and 70:30 for the deep learning models. The practical part for this approach is submitted as task_3.ipynb.

2.5. Evaluation

We evaluate our models based on different metrics such as Accuracy, F1 score, ROC curve, Confusion matrix and we consider their training and prediction times as well in order to choose the best model for each of the three approaches described in the previous section. Based on these criteria, for the first approach, injury prediction, we chose as best model the Random Forest, for the second approach, predict body part to be injured, we chose as best model the Multi-layer Perceptron and for the third approach, prediction of sufficient resting days after injury, we chose the Random Forest Regressor as the best model for this task. One general observation we make is that although deep learning models presented sufficient results, their training and prediction time is significantly higher compared to the other models. Finally, the relevant comparison graphs related to model performance for each approach are provided in the presentation slides and at the end of the submitted Jupytep Notebook files for each task.

2.6. Deployment

The predictive models could be integrated as a part of a customized internal application where the end user (doctor, coach, assistant, or physiotherapist) could interact through a User Interface and choose some options regarding next game's specifics such as the expected temperature and weather conditions, field and stadium type, each player's position and planned play type for the next game etc. The result of such query would be a list with the players of the team most likely to be injured (first approach), and/or the body parts most likely to be injured based on those specifics (second approach), and thus the necessary actions could be performed. For the third approach the user (mostly team doctors) could chose the game specifics of the previous game (when the injury occurred), the injured body parts and get a result with the number of suggested resting days for the player which could be sufficient for preventing the injury from reoccurring soon. The challenge for the sport organization would be to adopt a data-driven culture and learn how to use the predictive models in their decision-making process when it comes to player injuries. Moreover, data collection in regular basis in order to improve model capabilities and improve prediction quality, is another challenge for the organization. Finally, the sport organization should be ready to accept the extra costs that come with the utilization of such predictive models, data collection and storage and necessary infrastructure.

3. Workload Distribution

The workload distribution for this project was 15% business understanding, 35% for data understanding and preparation and finally 50% for modeling and evaluation. The modeling and evaluation phases took the most since we build several modes for each of the three approaches, we prepared comparison graphs for all metrics for the evaluation, tries several combinations of hyperparameters to achieve best performance etc. Finally, the business understanding phase took the least since we are familiar with sports.

4. Strong Points

One of the strong points of our project is that we tackled the problem of sport injuries by exploring and introducing three different approaches such as predicting injury before next game, predicting body parts likely to be injured and suggesting a number of sufficient resting days after injury to prevent reoccurrence. Another strong point, is that during this project we build several models for each approach, including deep learning solutions such as neural networks. Finally for the model evaluation part, we considered several metrics to choose the best model for each approach including training and prediction times. Finally, as part of the business understanding process, we came up with many potential problems that might arise in a sports organization from player injuries.

5. Weak Points

One of our project's weak points is that the dataset we used contained only 105 injuries over 5712 games, thus oversampling was imperative for training and testing our models. It would be certainly better to work with a dataset containing more injury cases, but we couldn't find one. Moreover, to tackle the problem of sport injuries more efficiently, we would like our dataset to include data collected from wearable devices as well, but again we could not find one.