REPORT 2: Injury Prevention Analysis

Alex Castro, Isabella Malek, & Tamara Nenninger

MIS 401 Business Intelligence and Analytics

Professor Lance Cameron

San Diego State University

Table of Contents

Introduction

The analysis particularly delves into several key aspects, such as the examination of player performance metrics, injury history, and team practices/ training over a defined period. The examination of these variables is crucial in uncovering the fundamental reasons behind recurring injuries that have plagued the team in recent years. This report will elucidate the background and purpose of this analysis and the methods and instrumentation employed in the evaluation. It will also present the results of the analysis and, lastly, offer long-term recommendations to enhance player health and reduce the likelihood of future injuries.

Researchers, Alex Castro, Isabella Malek, and Tamara Nenninger led the comprehensive analysis detailed in this report. All three roles encompassed a meticulous examination of the data from the provided data set, ensuring the data's accuracy and efficiency. They further conducted a thorough analysis using RapidMiner Studio version 10.2, employing a specific machine-learning classification model with a focus on evaluating the relationship between player performance metrics, injury history, and various external factors. Their efforts culminated in the formulation of four actionable recommendations designed to augment player well-being and reduce the likelihood of injuries in the team.

Background & Purpose

The purpose of this Business Intelligence Report revolves around the overall outlook of causes of injuries for the San Diego Padres baseball team. Following a noticeable increase in injuries this season, Dr. Catherine M. Robertson, Head Team Physician, has requested that researchers Castro, Malek, and Nenninger create a comprehensive analysis of potential root causes of this uptick. Dr. Robertson intends to mitigate future injuries to improve overall individual well-being on the team.

Data Source and Descriptive Analysis

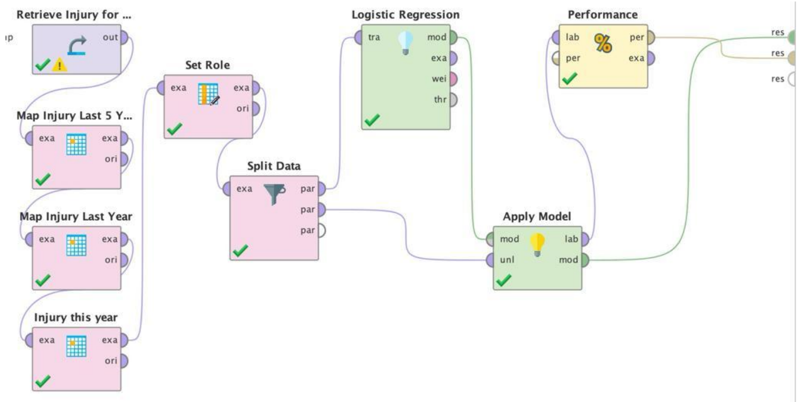
To perform the analysis, all three researchers leveraged the Injury(1).xlsx dataset. This comprehensive dataset incorporates San Diego Padres players' metrics, such as: injury this year, injury last year, last 5 years, training per week, average weekly training, year with team, last performance evaluation, and last endurance evaluation.

While observing and analyzing the dataset, they came across an error, a null value. Which chose to disregard and continue further with the dataset. For the most part, the variables are all integers, except for the average weekly training, which is real.

Methods & Instrumentation

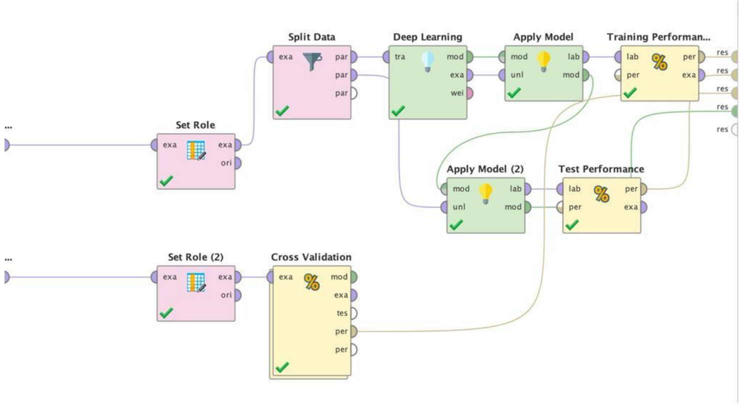
As previously mentioned, the research team used RapidMiner Studio (10.2) to conduct an in-depth analysis. They utilized specific models for the first run, employing Logistic Regression, Performance, Set Role, Split Data, Apply Model, and three Map operators (Map Injury Last 5 Years, Map Injury Last Year, Injury This Year) to arrive at their analysis. The initial step involved incorporating the Injury dataset into their process. Subsequently, they integrated three maps, each representing the duration of the occurrence of the injury, such as Map Injury Last 5 Years, Map Injury Last Year, and Map Injury This Year.

Model 1.1



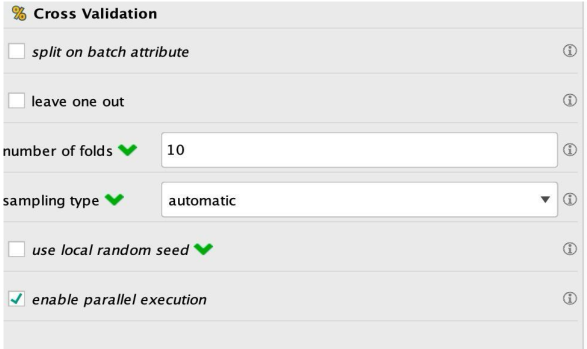
Next, the team linked the model with Set Role, utilizing parameters for the attribute name, “Injury This Year”, and the target role to label. Additionally, they incorporated the Split Data operator, with implemented parameters of 0.70 and 0.30. Within the Split Data operator, connections were established to both the Logistic Regression and the Apply Model. There were no changes in parameters for Logistic Regression, which remained consistent with those for Apply Model.

Model 1.2



For the second run, the research team utilized both the Deep Learning and Cross-Validation operators. Within the Cross-Validation, they opted for 10 folds (Model 1.3) and enabled parallel execution, as the remaining parameters were kept at their default settings.

Model 1.3



Cross Validation

□ split on batch attribute

□ leave one out

number of folds: 10

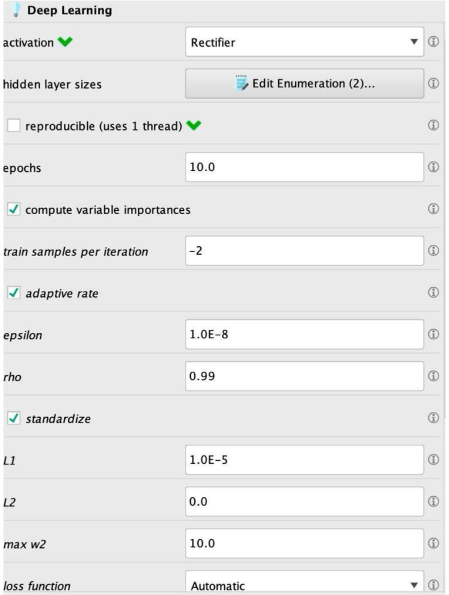
sampling type: automatic

□ use local random seed

☑ enable parallel execution

For the Deep Learning model, the team implemented parameters with two hidden layer sizes of 50. Additionally, they set the epochs to 10.0 and the train samples per iteration at -2. The Split Data operator was used with a split ratio of .7 and .3 to partition the data between training and testing for performance outputs.

Model 1.4



Deep Learning

activation: Rectifier

hidden layer sizes: Edit Enumeration (2)...

□ reproducible (uses 1 thread)

epochs: 10.0

☑ compute variable importances

train samples per iteration: -2

☑ adaptive rate

epsilon: 1.0E-8

rho: 0.99

☑ standardize

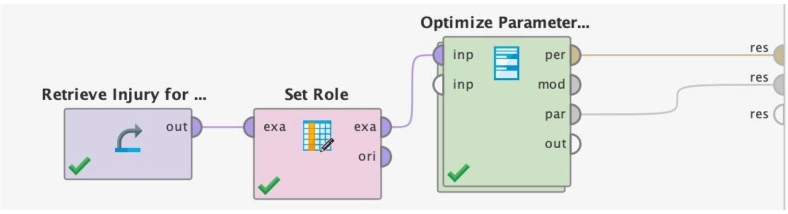
L1: 1.0E-5

L2: 0.0

max w2: 10.0

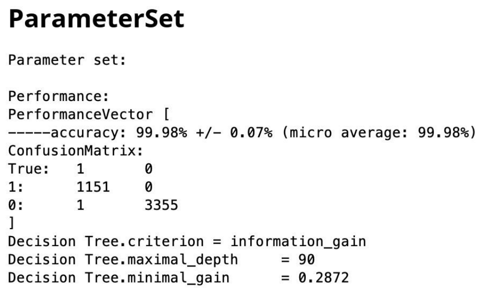
loss function: Automatic

Model 1.5



The modeling process commenced with the application of the dataset, followed by connecting it to Set Role, where the label was designated as 'injury this year.' The Optimize Parameter operator was utilized to identify the parameters that would produce the optimal output for our dataset in the context of the decision tree.

Model 1.6



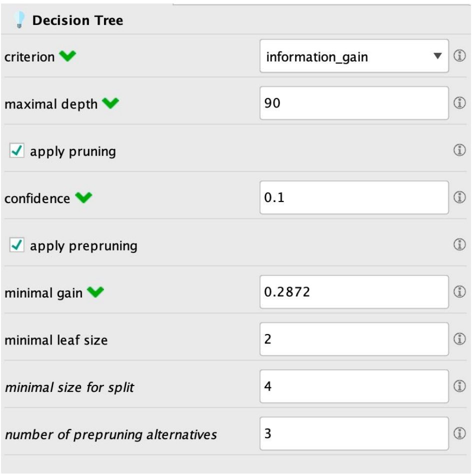
ParameterSet

```
Parameter set:

Performance:
PerformanceVector [
-----accuracy: 99.98% +/- 0.07% (micro average: 99.98%)
ConfusionMatrix:
True:   1       0
1:      1151    0
0:      1       3355
]
Decision Tree.criterion = information_gain
Decision Tree.maximal_depth     = 90
Decision Tree.minimal_gain      = 0.2872
```

The output of the optimization model run prompted the team to adjust the parameters in the decision tree. They integrated the optimal input parameters, specifically setting the criterion to information_gain, maximal_depth to 90, and minimal_gain to 0.2872 (Model 1.7).

Model 1.7



Decision Tree

criterion: information_gain

maximal depth: 90

☑ apply pruning

confidence: 0.1

☑ apply prepruning

minimal gain: 0.2872

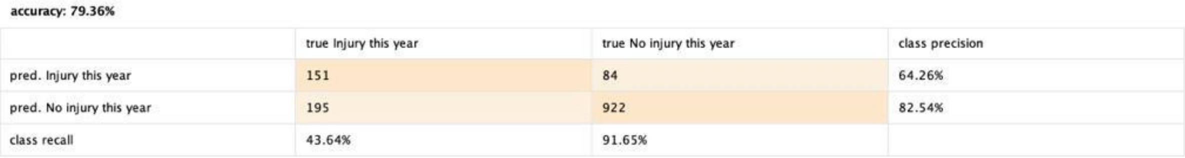
minimal leaf size: 2

minimal size for split: 4

number of prepruning alternatives: 3

Results

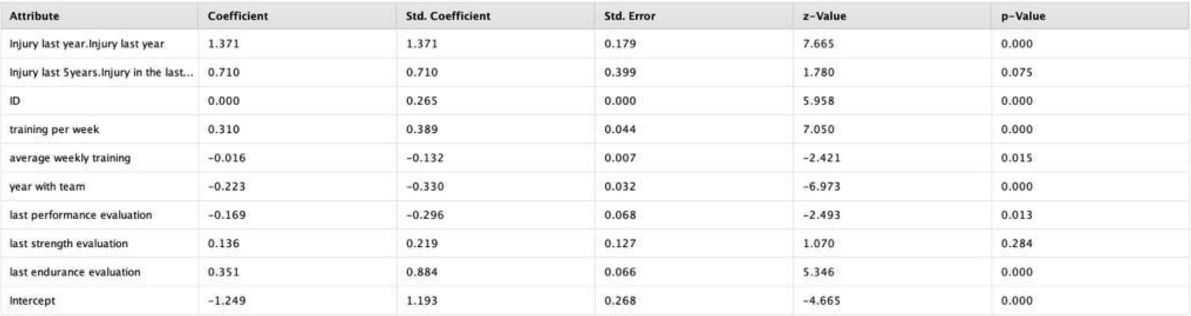
Model 1.8 (Logistic Regression)



accuracy: 79.36%

	true Injury this year	true No injury this year	class precision
pred. Injury this year	151	84	64.26%
pred. No injury this year	195	922	82.54%
class recall	43.64%	91.65%	

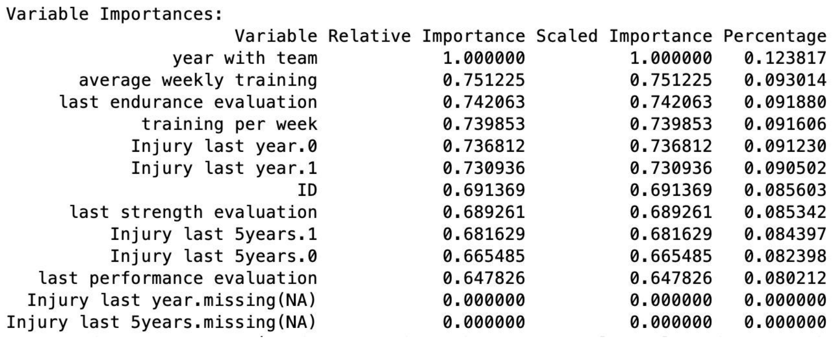
Model 1.9 (Logistic Regression)



Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
Injury last year.Injury last year	1.371	1.371	0.179	7.665	0.000
Injury last 5years.Injury in the last...	0.710	0.710	0.399	1.780	0.075
ID	0.000	0.265	0.000	5.958	0.000
training per week	0.310	0.389	0.044	7.050	0.000
average weekly training	-0.016	-0.132	0.007	-2.421	0.015
year with team	-0.223	-0.330	0.032	-6.973	0.000
last performance evaluation	-0.169	-0.296	0.068	-2.493	0.013
last strength evaluation	0.136	0.219	0.127	1.070	0.284
last endurance evaluation	0.351	0.884	0.066	5.346	0.000
Intercept	-1.249	1.193	0.268	-4.665	0.000

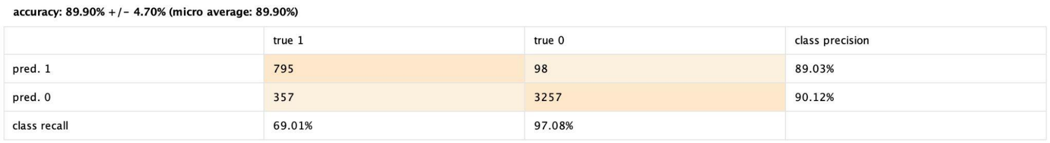
Based on the Logistic Regression model, the research team identified the top attributes as injury last year, training per week, and last endurance evaluation. These variables were determined to have a significant impact on the rate of injury for the current year within the model. However, the overall accuracy level, as indicated by the accuracy matrix, stands at 79.36%, which is not considered very high for achieving optimal output with the dataset.

Model 1.10 (Cross Validation, Importance)



```
Variable Importances:
                        Variable Relative Importance Scaled Importance Percentage
                  year with team            1.000000          1.000000   0.123817
         average weekly training            0.751225          0.751225   0.093014
       last endurance evaluation            0.742063          0.742063   0.091880
               training per week            0.739853          0.739853   0.091606
               Injury last year.0           0.736812          0.736812   0.091230
               Injury last year.1           0.730936          0.730936   0.090502
                              ID            0.691369          0.691369   0.085603
        last strength evaluation            0.689261          0.689261   0.085342
             Injury last 5years.1           0.681629          0.681629   0.084397
             Injury last 5years.0           0.665485          0.665485   0.082398
      last performance evaluation           0.647826          0.647826   0.080212
   Injury last year.missing(NA)             0.000000          0.000000   0.000000
 Injury last 5years.missing(NA)             0.000000          0.000000   0.000000
```

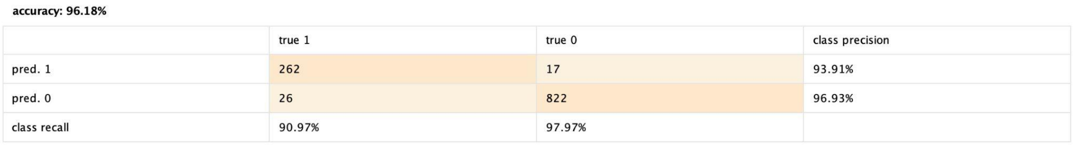
Model 1.11 (Cross-Validation, Performance)



accuracy: 89.90% +/- 4.70% (micro average: 89.90%)

	true 1	true 0	class precision
pred. 1	795	98	89.03%
pred. 0	357	3257	90.12%
class recall	69.01%	97.08%	

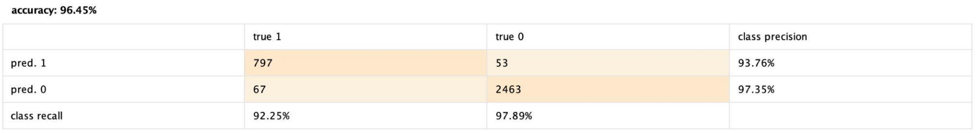
Model 1.12 (Cross-Validation, Testing Performance)



accuracy: 96.18%

	true 1	true 0	class precision
pred. 1	262	17	93.91%
pred. 0	26	822	96.93%
class recall	90.97%	97.97%	

Model 1.13 (Cross-Validation, Training Performance)



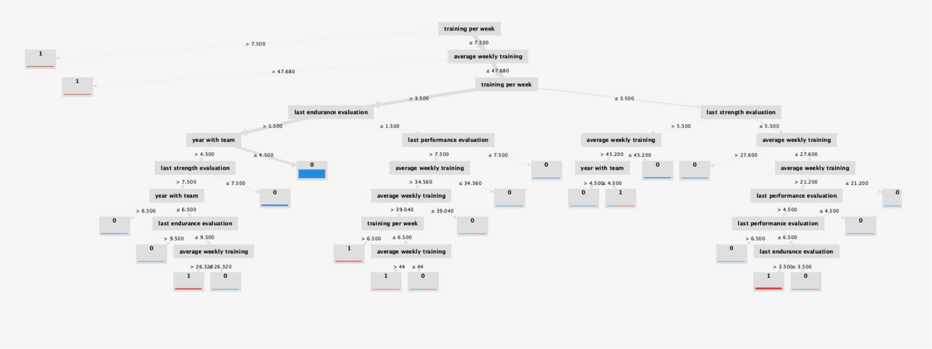
accuracy: 96.45%

	true 1	true 0	class precision
pred. 1	797	53	93.76%
pred. 0	67	2463	97.35%
class recall	92.25%	97.89%	

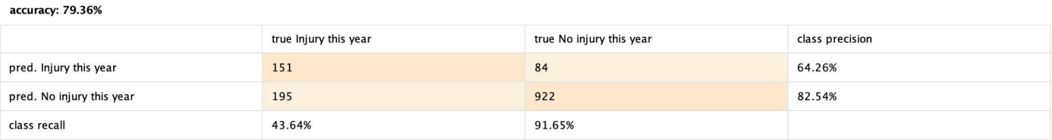
Based on the Cross-Validation model that was executed (Model 1.10), the research team identified the top three important variables as average weekly training, last endurance evaluation, and training per week. The performance matrices demonstrate the predictability of both true negative and true positive rates concerning the prediction of injury (1) or no injury (0). The

Cross-Validation performance output yielded an overall accuracy level of 89.90%, with a 4.70% margin of error, indicating a relatively high performance for our dataset.

Model 1.14 (Decision Tree)



Model 1.15 (Decision Tree Performance)



accuracy: 79.36%

	true Injury this year	true No injury this year	class precision
pred. Injury this year	151	84	64.26%
pred. No injury this year	195	922	82.54%
class recall	43.64%	91.65%	

This decision tree and its associated performance output were derived from the parameters established in our optimal performance model. Consequently, it achieved accuracy performance levels consistent with our logistic regression model, reaching an accuracy of 79.36%. Primarily, the athletic program successfully predicted that 922 athletes would not sustain injuries and would remain injury-free throughout the season. While this is relatively commendable, it is important to note that in terms of accuracy and comparison to other models, the accuracy performance is comparatively low.

Recommendations

Based on the results of the deep learning model, which had the highest top-line performance across testing, training, and Cross-Validation, several variables may impact the likelihood of injury for the San Diego Padres baseball team. In terms of relative importance, the “Average Weekly Training” and “Last Endurance Training” variables appear to have the highest impact on the model making accurate predictions. This could indicate that the amount of training endured by each athlete has a significant effect on the rate of injury for each athlete.

Further regression analysis could be performed on these two variables to see what their relationship is with the independent variable, potentially providing insight into what volume of weekly training, or how varying endurance levels, could potentially lead to an increase or reduction in team injuries.

1. The Padres athletic trainers need to monitor athletes' overall training endured every week to ensure athletes are not overextending themselves which seems to be leaving them prone to injury.
2. Endurance performance evaluations are also a unit of measure that athletic trainers should be looking into more closely as the model outputs show that it has a statistically significant impact on the rate of injury.
3. Prioritizing athletes based on those who had an injury from the previous season (last year) is also another factor that has led the team to injury as those who had been injured previously are more likely to be injured this year. This can be a recurrence of the previous injury or another injury that correlates back to the initial point of injury. Advising each athlete about how their previous injury can affect their future performance even if it’s believed the injury has completely subsided.

4. As within any workplace, becoming a part of a new team can take some adjusting. The number of years within the team affects the overall comfort and communication between players and the athletic training staff. For new players, an additional recommendation is one-on-one meetings to understand the full history of a player's prior injuries and any onset injuries they are currently dealing with.

Limitations

In terms of limitations, the research team identified several. The dataset, Injury(1).xlsx, is overall broad and general. The available data did not provide enough information to draw precise conclusions. Information on factors such as the severity of the injury, its location, recovery time, and the place where the injury occurred was not present and could be incredibly informative in predicting injuries and identifying trends. However, due to the lack of more specific and detailed data, the team was unable to arrive at an exact conclusion. With an expanded dataset containing more specific details, the team could formulate a closer conclusion and offer precise recommendations to prevent future injuries. Additionally, the research team encountered a null value, which could have been a very or least important variable, potentially influencing the outcome.