Question

The head coach of a team is concerned with the number of injuries sustained by their players in games. This task is to explain the drivers of this injury risk in the given dataset.

About Dataset

The dataset contains 3 files with information about the players metrics, game workload and injuries with attributes like athlete ids, date, workload, metric (hip mobility, groin squeeze).

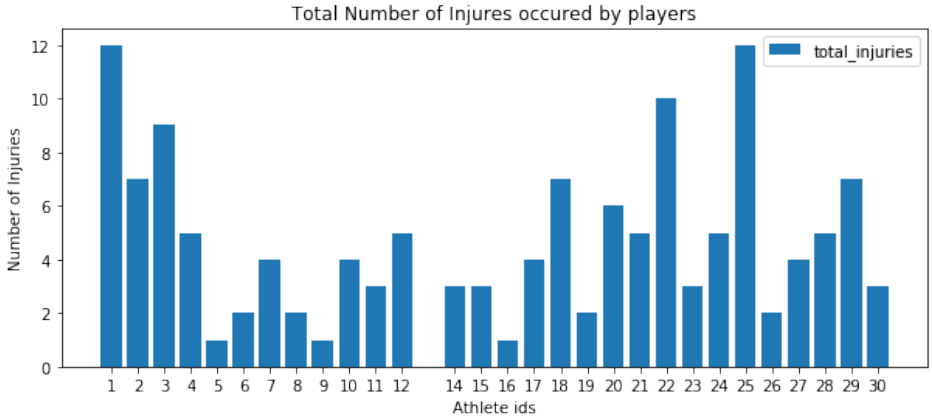
Dataset description

1. Metrics dataset: provides data of hip mobility and groin squeeze of 30 athletes recorded every day from 01/05/2016 to 30/04/2018.
2. Workload dataset: provides workload taken by each athlete on a game day.
3. Injuries dataset: provides the dates on which the athletes were injured during a game.

Exploratory Data Analysis

I have performed the basic summary statistics on the datasets and below are the few snapshots. For the dataset provided:

1. Most number of injuries were occurred to athlete with id 12. Athlete with id 13 has not been injured at all.
2. Below is the graph with number of injuries per athlete



3. Maximum and minimum workloads taken by an athlete in a game are 534 (by athlete id: 10) and 225 (by athlete id: 13) respectively.
4. Maximum and minimum number of games played are 101 (by athlete id: 3) and 65 (by athlete id: 5) respectively.
5. Calculated Fitness ratio for each player, from the total number of games played and the number of injuries sustained in those games using the formula:

   fitness_ratio = (1- (num_of_injuries / total_num_of_games))*100

6. From the above formula, it was observed that athlete 13 had the best (100 fitness ratio, since he does not have any injuries) and athlete 1 had the least fitness ratio.

Visualization and Interpretation

I have plotted data over time to identify any relation between the attributes.

1. I plotted the workload and injuries over time for each player individually, please find a sample graph of an athlete below:

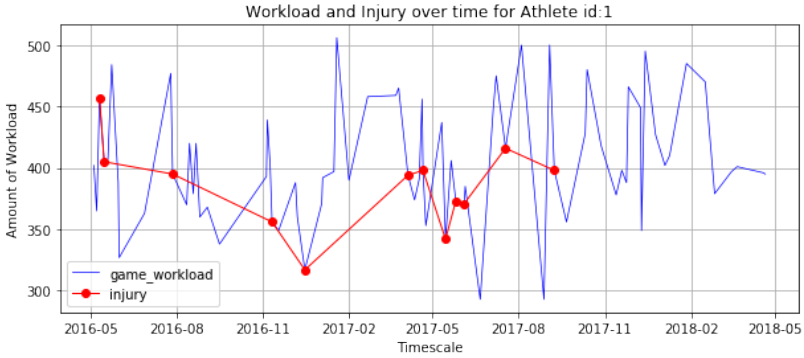


Figure 1: workload and injury over time for athlete 1

Observation 1: Visually, there is no obvious relation between the amount of workload taken up by a player and injury sustained in a game. Therefore, performing statistical analysis might provide a better understanding.

2. Similarly, I plotted number of resting days before each game and injuries sustained over time to identify any relation:

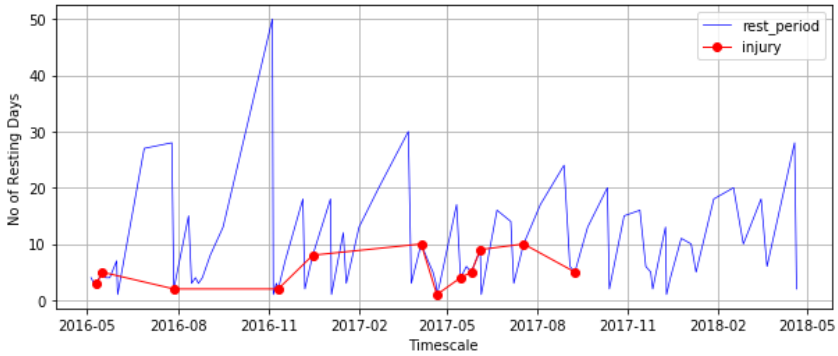


Figure 2: Resting period and injuries over time for athlete 1

Observation 2: Visually, there seems to be a slight relation between resting days and injuries. Injuries sustained are more when the resting days are few. More can be understood by performing statistical analysis.

3. I have plotted groin squeeze and injuries over time as well

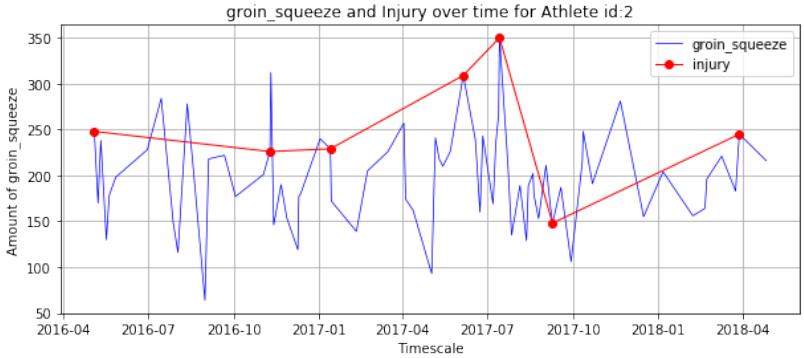


Figure 3: Groin squeeze and injuries over time for athlete 2

Observation 3: Even with the groin squeeze and injuries, there seems to be a slight relation. When groin squeeze is high, more injuries have seem to occurred when compared with low groin squeeze.

4. Number of games played by each player is plotted as below to see the distribution

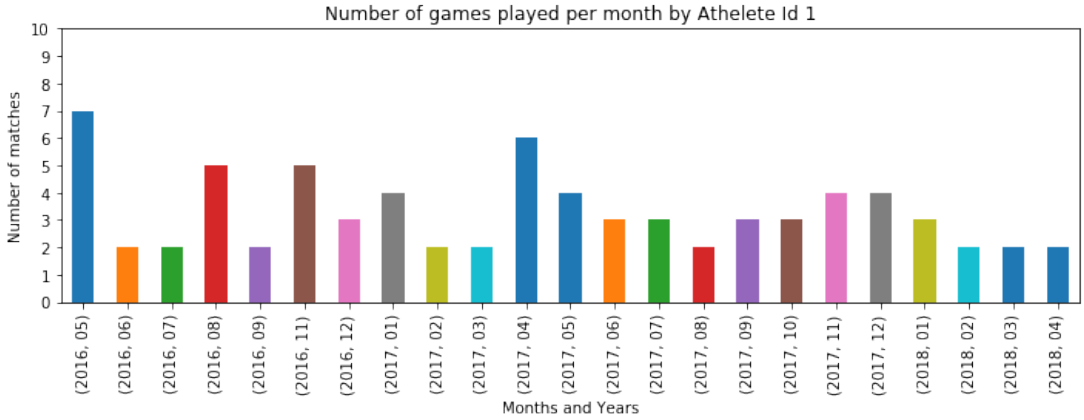


Figure 4: Number of games played by athlete 1 over months

Observation 4: No cyclicality or seasonality is visible.

Note: Graphs related to other players and other visualisations are available in the exploratory_data_analysis notebook submitted.

Feature extraction

1. Created new column named “injury” with values [“yes”, “no”] which was important to merge the workload and injury datasets.
2. Transposed the metric attribute into 2 new attributes as [hip_mobilty and groin_squeeze] with values in them.
3. Created new column attribute named “**rest period**”. Typically, rest period is the difference in days between two consecutive games played by each athlete.
4. However, “rest period” for the first game played by each athlete was calculated using a standard date (01/04/2016) as this was the first date in the dataset.
5. Therefore, below is the final data frame structure used for further statistical analysis after merging all the tables provided.

	athlete_id	date	game_workload	injury	groin_squeeze	hip_mobility	rest_period
0	1	2016-05-05	402	No	284	35	4.0
1	1	2016-05-08	365	No	250	41	3.0
2	1	2016-05-11	457	Yes	331	33	3.0
3	1	2016-05-16	405	Yes	260	38	5.0
4	1	2016-05-20	407	No	378	60	4.0

6. Correlation between the above attributes was checked and none of the columns were significantly correlated.

	athlete_id	game_workload	groin_squeeze	hip_mobility	rest_period
athlete_id	1.000000	-0.030159	-0.051175	-0.014437	-0.016292
game_workload	-0.030159	1.000000	-0.022571	0.000956	-0.023001
groin_squeeze	-0.051175	-0.022571	1.000000	-0.001996	0.039563
hip_mobility	-0.014437	0.000956	-0.001996	1.000000	0.006614
rest_period	-0.016292	-0.023001	0.039563	0.006614	1.000000

Statistical Analysis

The task here is to define the statistically significant independent variables that impact on the dependent variable. That is, changes in independent variable that cause the shift of dependent variable (in our case shift from “yes” to “no” class or vice-versa).

1. Performed Logistic Regression – Maximum Likelihood Analysis

```
                           Results: Logit
================================================================
Model:              Logit            Pseudo R-squared: 0.117
Dependent Variable: y                AIC:              996.0700
Date:               2020-02-04 21:25 BIC:              1192.6996
No. Observations:   2400             Log-Likelihood:   -464.03
Df Model:           33               LL-Null:          -525.28
Df Residuals:       2366             LLR p-value:      3.2402e-12
Converged:          0.0000           Scale:            1.0000
No. Iterations:     35.0000
----------------------------------------------------------------
          Coef.    Std.Err.      z    P>|z|     [0.025     0.975]
----------------------------------------------------------------
x1       -0.0007     0.0020 -0.3399 0.7339     -0.0045     0.0032
x2        0.0074     0.0019  4.0002 0.0001      0.0038     0.0111
x3       -0.0024     0.0091 -0.2667 0.7897     -0.0202     0.0154
x4       -0.1094     0.0195 -5.5992 0.0000     -0.1477    -0.0711
x5       -2.6794     1.0829 -2.4744 0.0133     -4.8018    -0.5570
x6       -2.6292     1.0144 -2.5920 0.0095     -4.6174    -0.6411
x7       -2.9460     1.0300 -2.8603 0.0042     -4.9647    -0.9273
x8       -3.8380     1.1322 -3.3898 0.0007     -6.0572    -1.6189
x9       -4.4242     1.3759 -3.2155 0.0013     -7.1209    -1.7275
x10      -3.6988     1.1378 -3.2509 0.0012     -5.9288    -1.4688
x11      -2.6084     1.0334 -2.5240 0.0116     -4.6339    -0.5829
x12      -4.6265     1.2673 -3.6507 0.0003     -7.1103    -2.1426
x13      -5.4413     1.4460 -3.7630 0.0002     -8.2755    -2.6072
x14      -3.3745     1.0811 -3.1214 0.0018     -5.4934    -1.2556
x15      -2.9731     1.0730 -2.7707 0.0056     -5.0762    -0.8700
x16      -2.2468     1.0092 -2.2263 0.0260     -4.2249    -0.2688
x17     -24.7826 15309.8790 -0.0016 0.9987 -30031.5940 29982.0289
x18      -4.6693     1.2055 -3.8733 0.0001     -7.0321    -2.3066
x19      -3.9683     1.1288 -3.5154 0.0004     -6.1808    -1.7558
x20      -5.2568     1.4296 -3.6770 0.0002     -8.0589    -2.4548
x21      -4.2697     1.1382 -3.7512 0.0002     -6.5006    -2.0388
x22      -2.0863     0.9927 -2.1015 0.0356     -4.0320    -0.1405
x23      -3.7463     1.1931 -3.1400 0.0017     -6.0847    -1.4079
x24      -2.3546     0.9850 -2.3903 0.0168     -4.2852    -0.4239
x25      -3.0681     1.0633 -2.8854 0.0039     -5.1522    -0.9840
x26      -2.4515     1.0290 -2.3825 0.0172     -4.4683    -0.4348
x27      -3.4986     1.1291 -3.0985 0.0019     -5.7116    -1.2856
```

Observation: Variables x2 and x4 are significant compared to variable x1 and x3.
In this case (**x1 = workload, x3 = hip mobility**) and (**x2 = groin squeeze and x4 = resting days**). Rest all the variables are dummy variables created for each athlete id. They are significant as well other than **x17** (that is the variable of athlete id 13, who was never injured).

- Clearly, “workload” and “hip mobility” are not statistically significant. (i.e., p value greater than 0.05). Therefore, we can suggest that hip mobility and workload taken by each player is not the major reason for an athlete to get injured in a game.
- **“Groin squeeze” value and “resting days”** are statistically significant variables with p-value less than 0.05. These variables contribute the most towards the changes in dependent variable (in this case injury). Therefore, changes in the values of groin squeeze and resting days will have impact on the decision *‘whether an athlete is being injured or not?*’ in other words they are the drivers for the injury.
- Increase in the values of “groin squeeze” (indicated by positive coefficient) and reduce in “resting days” (indicated by negative coefficient) is impacting the athlete for being injured.
- Also, x17 = Athlete id 13 is also not statistically significant. Obvious reason is that the athlete was never injured and therefore is not contributing much in decision making process for dependent variable.

Similarly, I have performed OLS and found the similar results that only groin squeeze and resting period are the only significant attributes.

Note: Results of OLS are present in the injury_data_statistical_analysis notebook.

Modelling

Data provided has two classes (Yes, No). “Yes” meaning athlete is injured and “No” being the athlete is not injured. Out of 2400 instances of game workload there were only 137 instances of injury sustained. This clearly brings in a bias making “Yes” class a minority.

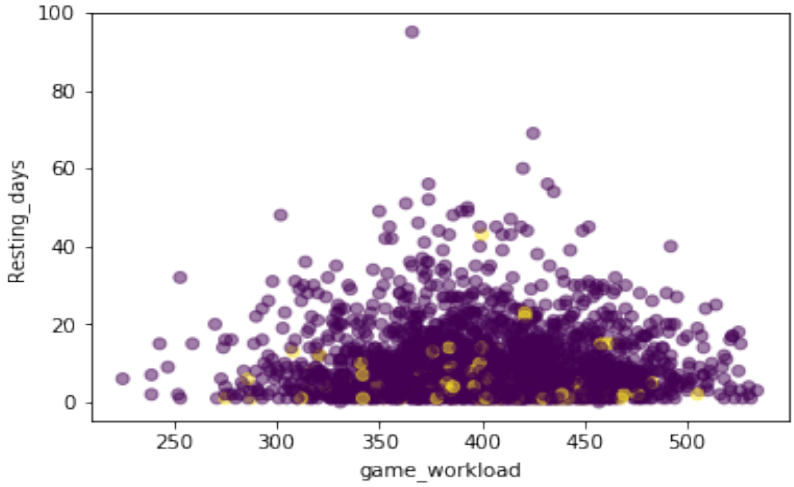


Figure 5:Classes in the dataset (yellow- Injury and Purple - Not injured)

To eliminate the bias, I have used oversampling technique named Synthetic Minority over sampling technique (SMOTE). This technique will create some rows similar to that of the minority class (without influencing the statistics of the class). Please find the below image of the classes after sampling.

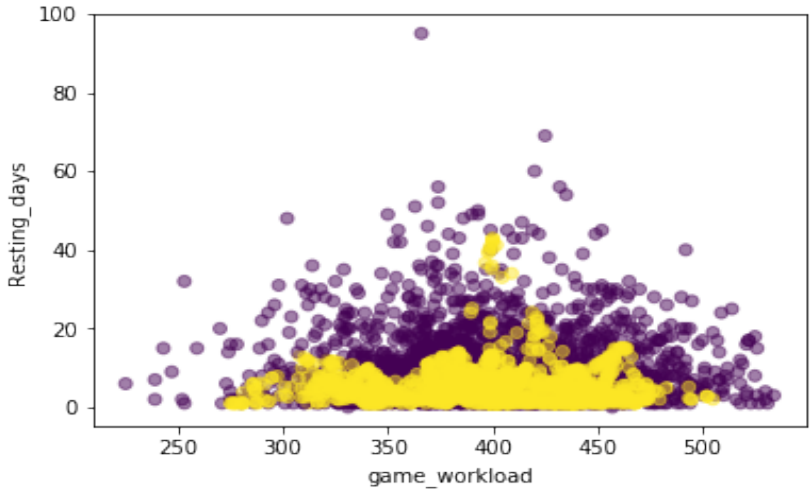


Figure 6:Classes in the dataset after Sampling(yellow- Injury and Purple - Not injured)

Classifier: This is a classification problem. From the given attributes, we need to define if a player is injured or not injured. Therefore, I used traditional supervised models to train on this data. Please find the results from various models:

Model Name	Precision	Recall	F1 Score
Logistic	0.73	0.72	0.72
KNN	0.87	0.85	0.85
SVC	0.92	0.92	0.92
Decision Tree	0.94	0.94	**0.94**

Observation: Decision Tree classified the data accurately compared to other models. More analysis can be performed to identify the deeper insights.

Conclusion

The major task in this analysis was to identify the reasons behind athlete being injured based on the provided data. As per my analysis, it is observed that **Groin squeeze** and **resting days (gap between two consecutive games)** are the two driving factors that impact the occurrence of injury to an athlete. With high metric value of groin squeeze and less number of resting days, athletes are more likely to be injured in a game. Therefore, head coach should focus on providing more rest (gap between games) if athletes have high groin squeeze.

Attachments:

1. Injury_data_statistical_analysis.ipynb
2. Exploratory_data_analysis.ipynb

Note: Additional processing and analysis are present in the above notebooks. So please refer to the notebooks if you need further information on any item mentioned in this document. Else, you can reach out to me for clarifications or questions.

Future Analysis

Another feature could be extracted from the datasets provided: **resting period between two injuries,** and the impact of this feature on likelihood of injury could be studied. However, due to time constraints I have not pursued this analysis.

Signature:
Swathikiran Srungavarapu

Date:
04/02/2020