Training Trainers

By: Nigel Nelson and Collin Quinn

Background Information

International Journal of Sports Physiology and Performance, 2021, 16, 1522-1531
https://doi.org/10.1123/ijspp.2020-0518

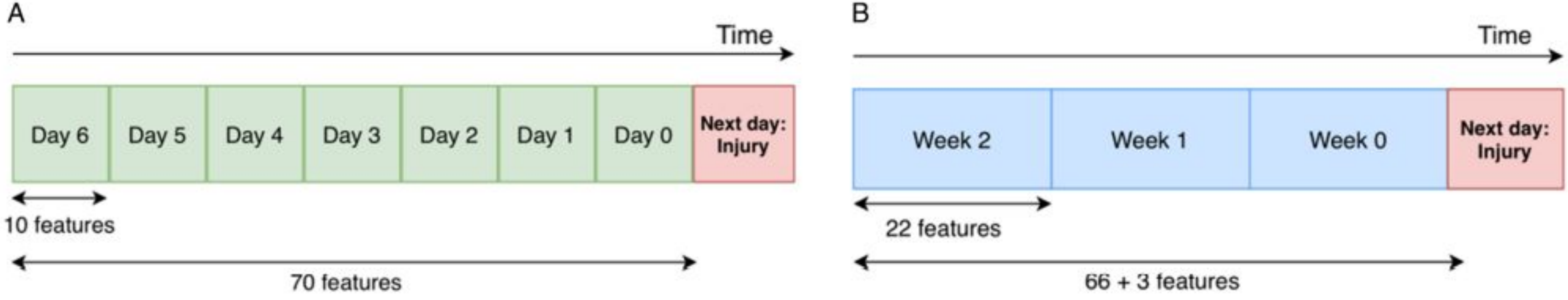
Human Kinetics
ORIGINAL INVESTIGATION

Injury Prediction in Competitive Runners With Machine Learning

S. Sofie Lövdal, Ruud J.R. Den Hartigh, and George Azzopardi

The Data Set

- Training logs for team of 74 high-level medium/long distance runners
- Collected over 7 years
- 42,766 entries, 583 injuries
- Two data sets: weekly logs and daily logs:



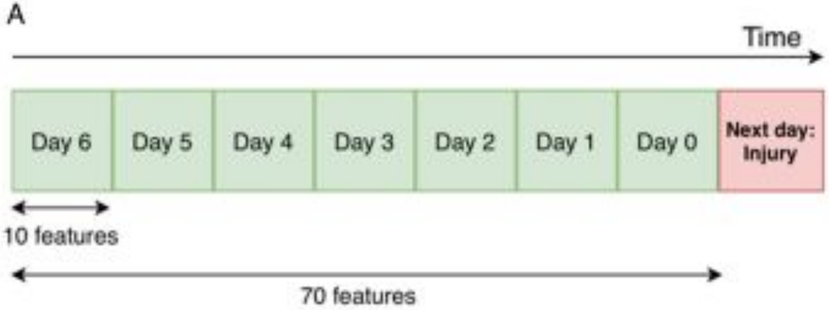
Day approach

Week approach

The Day Approach

- Greater predictive ability than week approach
- 73 features in total: 10 features for each day, ahtleteID, date, and Injury flag to indicate if injury occured

No	Day feature	Range
1	Number of sessions	[0, 2]
2	Total distance	[0.0, 25.0]
3	Sum of distance in Z3–Z4	[0.0, 15.0]
4	Sum of distance in Z5, T1, and T2	[0.0, 10.0]
5	Distance sprinting	[0.0, 1.5]
6	Number of strength sessions	[0, 1]
7	Hours alternative training	[0.0, 3.0]
8	Perceived exertion	[0.0, 1.0]
9	Perceived training success	[0.0, 1.0]
10	Perceived recovery	[0.0, 1.0]



Day approach

Research Questions

- What features are most predictive of injuries in long distance runners?

- Can a injury prediction model be made accurate enough to provide meaningful insight on training protocols?

Hypothesis

- The greater the perceived exertion an athlete reports, the higher the probability that later training sessions result in injury.

Experimental Design: Pre-processing

- No empty values in data set
- Perceived independent features normalized for each athlete
- Dropped non-useful columns (i.e. Athlete ID)
- Renamed feature names that used jargon:
    - Km Z3-4 : km low-intensity
    - Km Z5-T1-T2 : km medium-intensity
    - Km sprinting : km high-intensity
- Float64 -> Categorical (# of sessions, # of strength training sessions)

```
Data columns (total 73 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   nr. sessions                 42766 non-null  category
 1   total km                     42766 non-null  float64
 2   km low-intensity             42766 non-null  float64
 3   km medium-intensity          42766 non-null  float64
 4   km high-intensity            42766 non-null  float64
 5   strength training            42766 non-null  category
 6   hours alternative            42766 non-null  float64
 7   perceived exertion           42766 non-null  float64
 8   perceived trainingSuccess    42766 non-null  float64
 9   perceived recovery           42766 non-null  float64
 10  nr. sessions.1               42766 non-null  category
 11  total km.1                   42766 non-null  float64
 12  km low-intensity.1           42766 non-null  float64
 13  km medium-intensity.1        42766 non-null  float64
 14  km high-intensity.1          42766 non-null  float64
 15  strength training.1          42766 non-null  category
 16  hours alternative.1          42766 non-null  float64
 17  perceived exertion.1         42766 non-null  float64
 18  perceived trainingSuccess.1  42766 non-null  float64
 19  perceived recovery.1         42766 non-null  float64
...
 70  Athlete ID                   42766 non-null  int64
 71  injury                       42766 non-null  category
 72  Date                         42766 non-null  int64
dtypes: category(15), float64(56), int64(2)
```

Experimental Design- Before Model Creation

- Use Statistical Tests to evaluate features
  - Kruskal Wallis used on continuous features
  - Chi Square Contingency on categorical features
- Use Data visualization to understand features
  - Box plots used on continuous features
  - Heat maps used on categorical features

= Categorical Feature

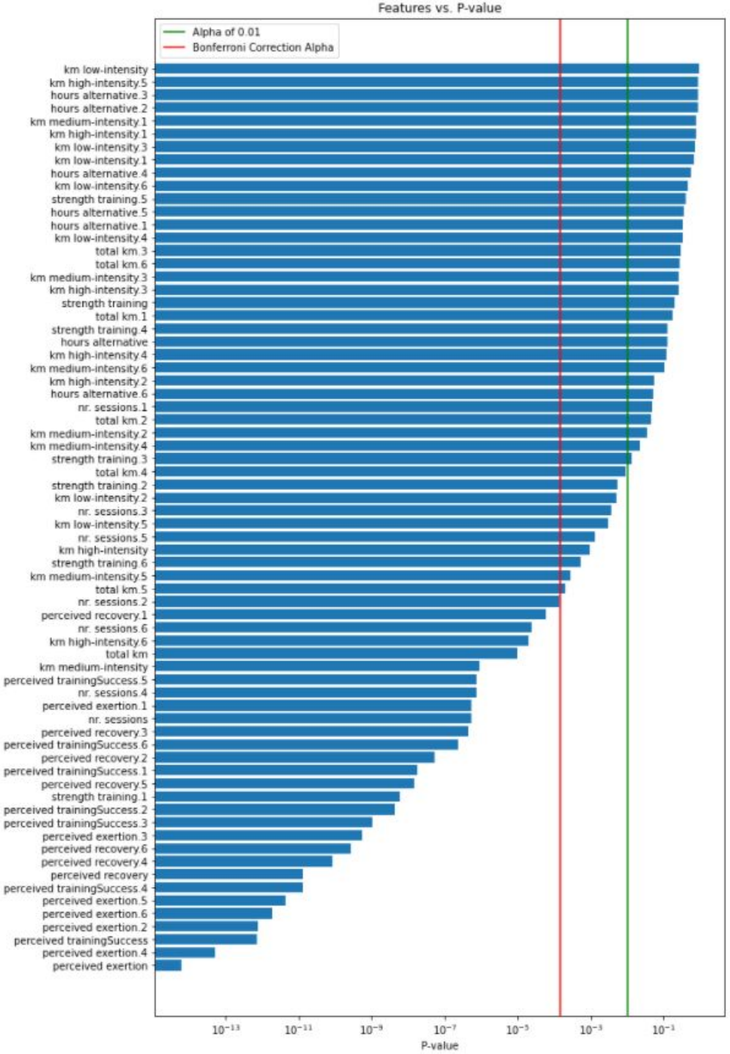
No	Day feature	Range
1	Number of sessions	[0, 2]
2	Total distance	[0.0, 25.0]
3	Sum of distance in Z3–Z4	[0.0, 15.0]
4	Sum of distance in Z5, T1, and T2	[0.0, 10.0]
5	Distance sprinting	[0.0, 1.5]
6	Number of strength sessions	[0, 1]
7	Hours alternative training	[0.0, 3.0]
8	Perceived exertion	[0.0, 1.0]
9	Perceived training success	[0.0, 1.0]
10	Perceived recovery	[0.0, 1.0]

Experimental Design: Model Creation

- Classification Problem
- Numerous types of classification models received 98% accuracy
  - Why did this happen!? (42,766 features - 583 injuries) / 42,766 features = 0.98636768
- As a result, created a balanced (injury/non-injury) training/testing set
  - Elected to choose Random Forest
  - Used Synthetic Minority Oversampling Technique (SMOTE)

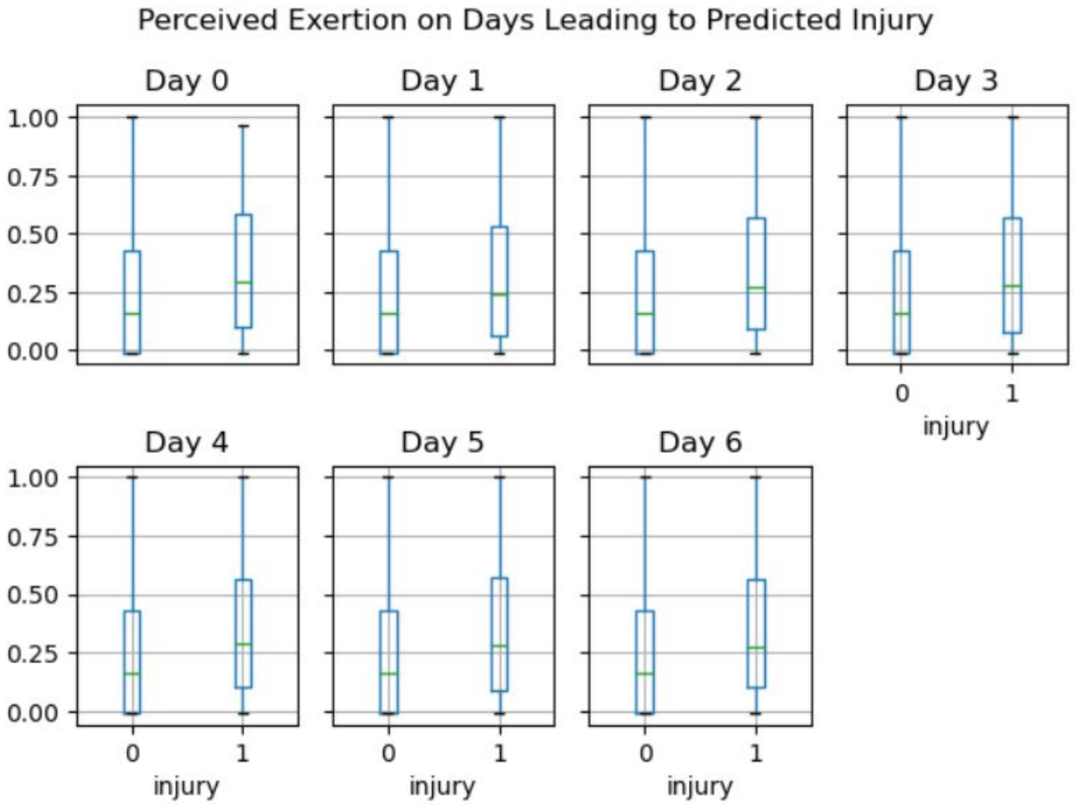
Results: Most Predictive Features

- Most Statistically significant Features:
  - 1. Perceived exertion
  - 2. Perceived training success
  - 3. Perceived recovery
- Most Significant Features from RF model:
  - 1. Perceived exertion
  - 2. Perceived training success
  - 3. Perceived recovery
- Interesting Observations:
  - Most recent day logs not always most predictive:
    - Perceived exertion, strength training, etc.



Results: Plots of Most Predictive Features

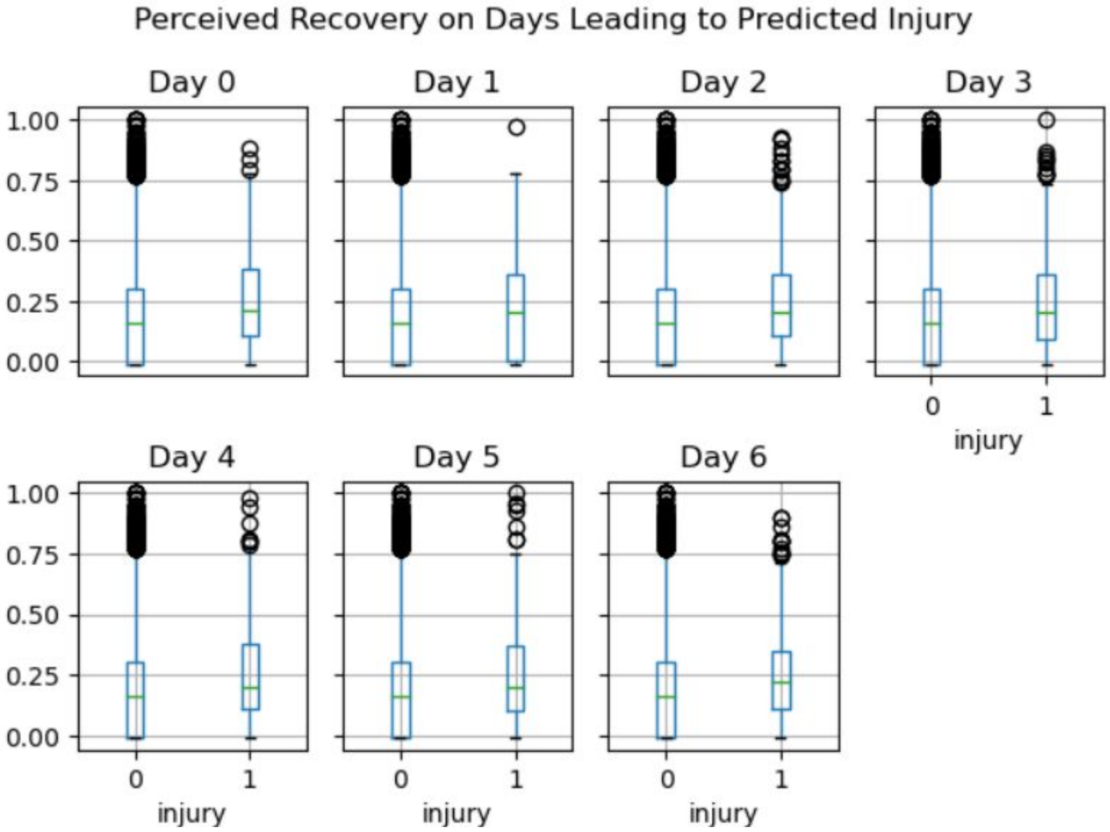
- Consistent throughout day logs
- Unsurprising results, days with higher perceived exertion had higher incidents of injuries



	perceived exertion	injury
perceived exertion	1.000000	0.039748
injury	0.039748	1.000000

Results: Plots of Most Predictive Features

- Consistent throughout day logs
- High number of outliers
- Surprising results, days with higher perceived recovery had higher incidents of injuries



Results: Plots of Most Predictive Features

- Consistent throughout day logs
- Surprising results, days with higher perceived success had higher incidents of injuries

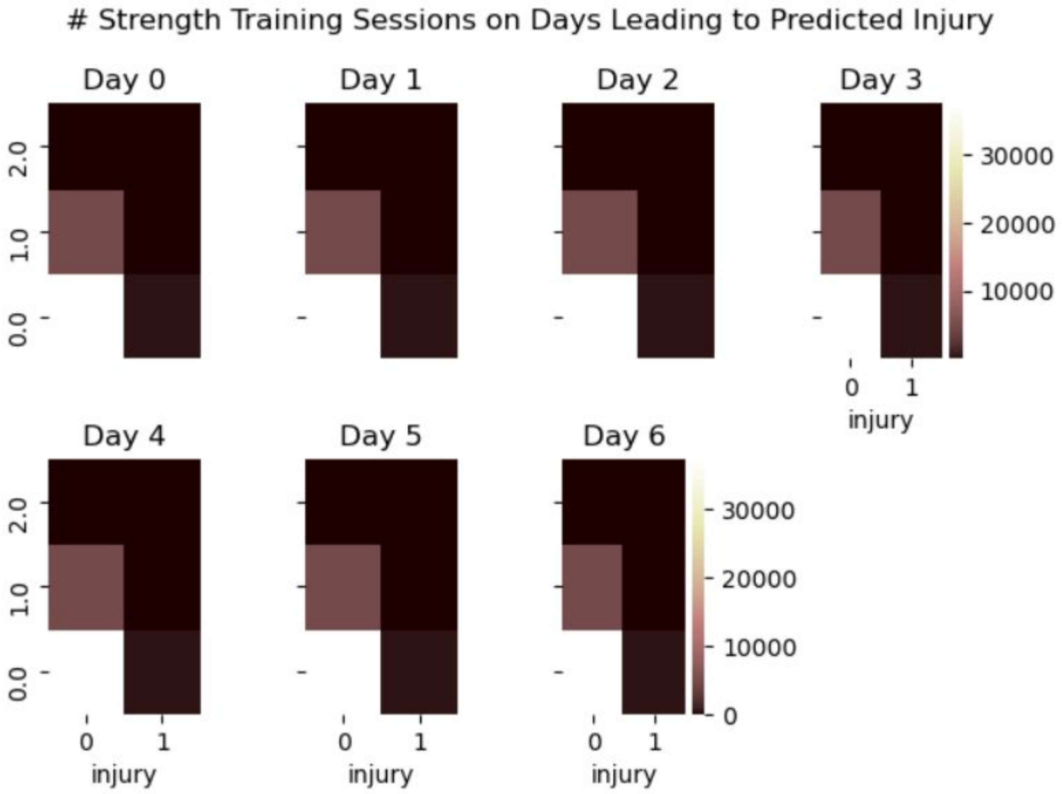
Results: Plots of Most Predictive Features

- Consistent throughout day logs
- Plots don't tell whole story:

```
day_data['strength training'].value_counts()
```

✓ 0.8s

```
0.0    37871
1.0     4819
2.0       76
```



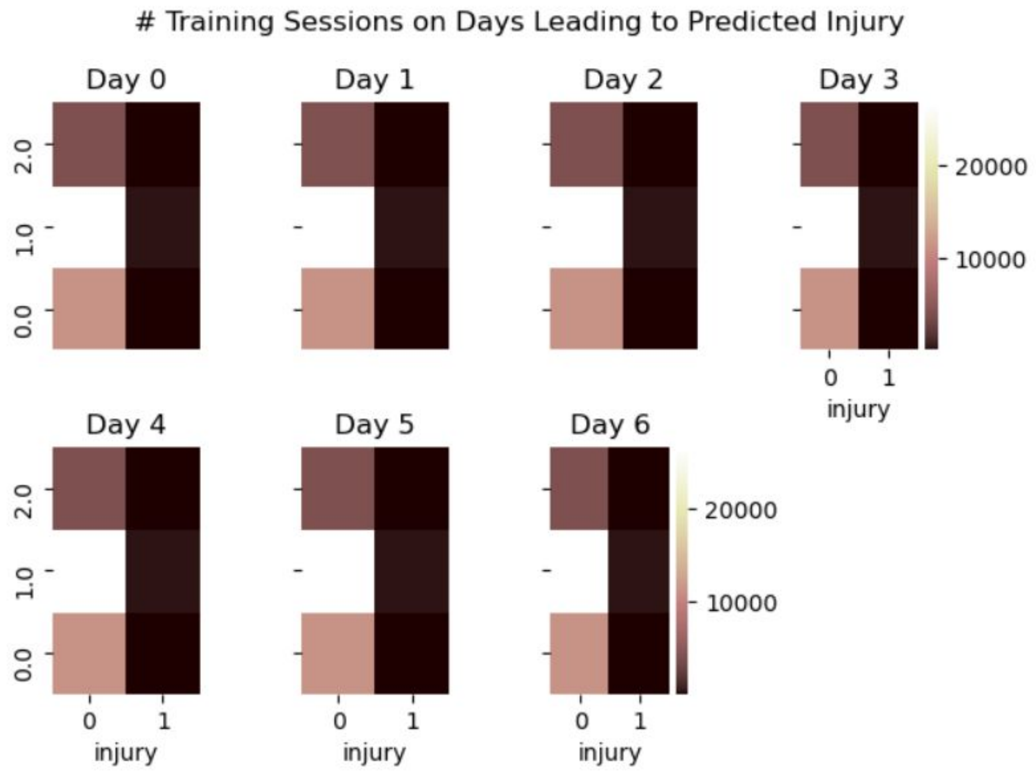
Results: Plots of Most Predictive Features

- Consistent throughout day logs
- Plots don’t tell whole story:

```
day_data['nr. sessions'].value_counts()
```

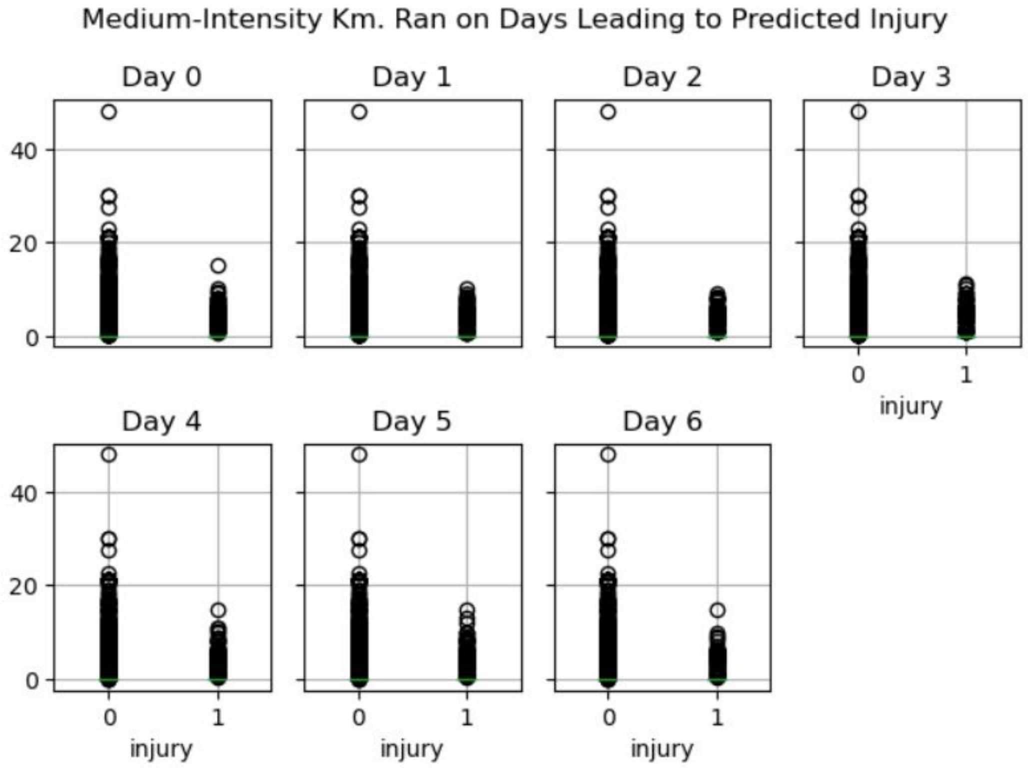
✓ 0.1s

```
1.0    27103
0.0    11476
2.0     4187
Name: nr. sessions, dtype: int64
```



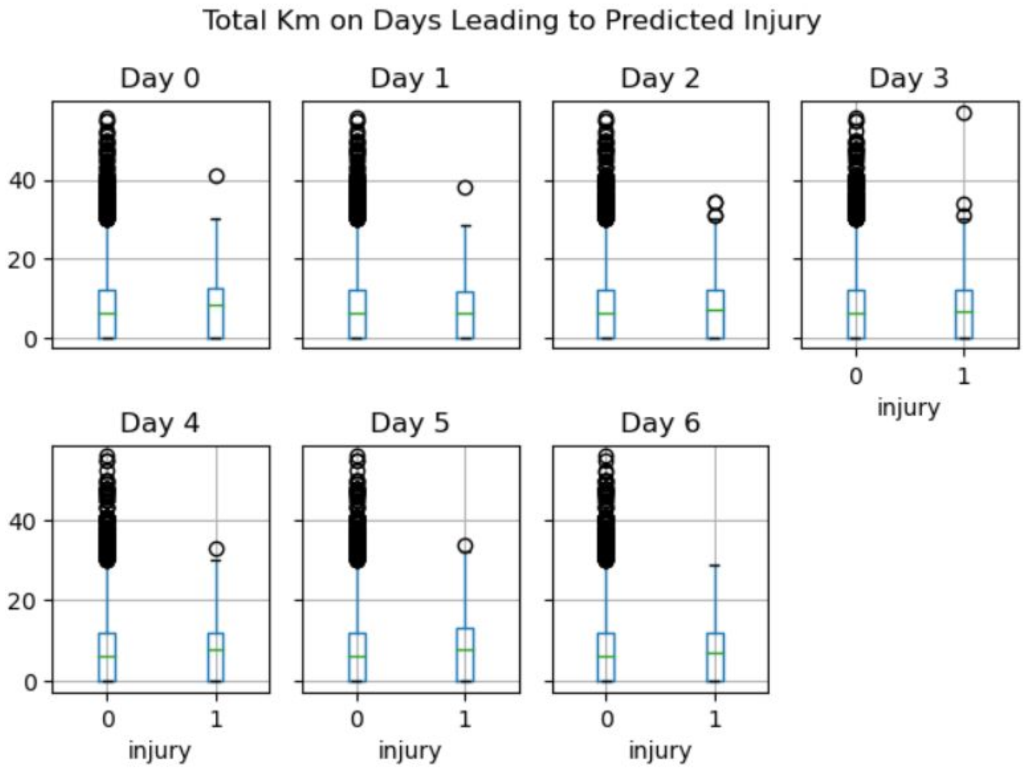
Results: Plots of Most Predictive Features

- Consistent throughout day logs
- All outliers
- Surprisingly, high km ran = no injuries



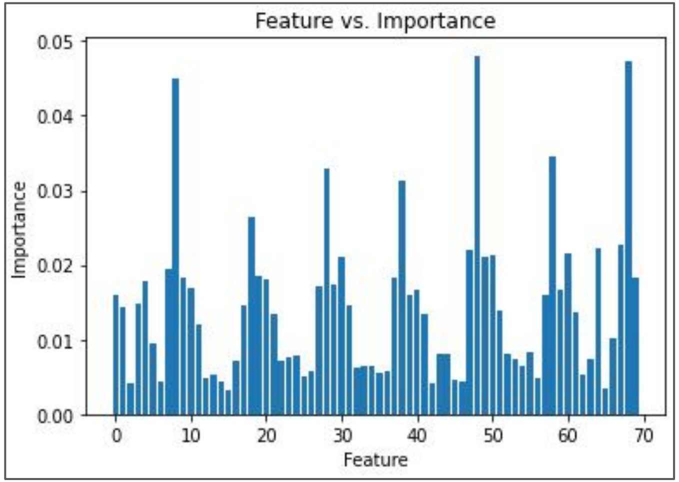
Results: Plots of Most Predictive Features

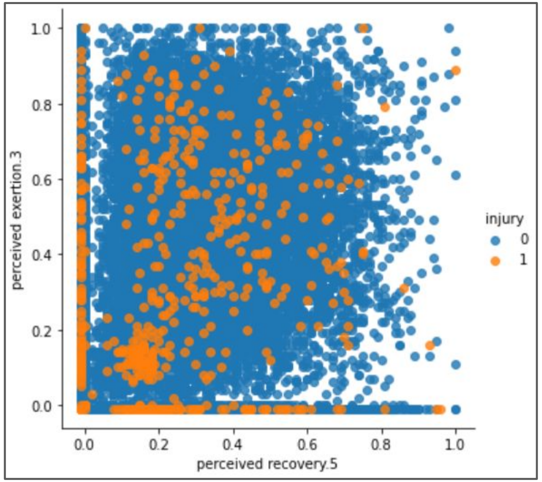
- Consistent throughout day logs
- Many outliers
- No obvious separation besides outliers for non-injuries



Model Results

- Random Forest Model Top 3 Most Predictive:
    - Perceived training Success
    - Perceived exertion
    - Perceived recovery
- Metrics for RF Model w/ Feature Selection:
    - AUC Score: .83
    - F1 Score: .88





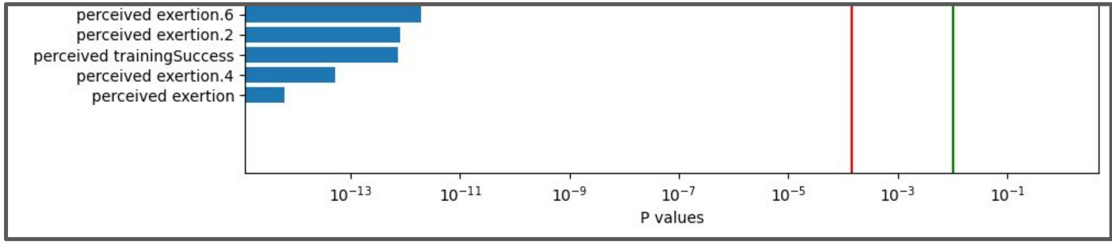
Discussion: Model

- Hypothesis and research questions are supported by our model
- Future Uses
- Different Imbalanced Classification Techniques (i.e. bagging)

Discussion: Hypothesis

- The greater the perceived exertion an athlete reports, the higher the probability that later training sessions result in injury.
- Perceived exertion = lowest p-value of all features
- Positive Pearson correlation

- Hypothesis: Accepted!



	perceived exertion	injury
perceived exertion	1.000000	0.039748
injury	0.039748	1.000000

Discussion: Features

- **Research Question:** What features are most predictive of injuries in long distance runners?
- **Answer:** Perceived exertion, perceived training success, perceived recovery
  - Many athletic pursuits can incorporate these

- Unbalanced data set resulted in difficult EDA
- Team aspect likely skewed data

- How would these features hold up for other teams and/or sports?

Questions?


ATHLETIC TRAINERS
DAILY SCHEDULE
8 AM. – TAPE ANKLES
9 A.M. – TAPE ANKLES
10 A.M. – TAPE ANKLES
11 A.M. – TAPE ANKLES
LUNCH – SAVE THE WORLD
1 P.M. – TAPE ANKLES
2 P.M. – TAPE ANKLES
3 P.M. – TAPE ANKLES
4 P.M. – TAPE ANKLES
5 P.M. – TAPE ANKLES

Source(s):

- Lovdal, S., den Hartigh, R., & Azzopardi, G. (2021). Injury Prediction in Competitive Runners with Machine Learning. International journal of sports physiology and performance, 16(10), 1522–1531. https://doi.org/10.1123/ijspp.2020-0518