Professional Sports Injury Analysis

Risk Management Project Report

Executive Summary

- This report presents a comprehensive analysis of sports injury data collected from 30 athletes over a two-year period (May 2016 to April 2018).
- The analysis focuses on identifying injury risk factors, creating individual athlete risk profiles, and establishing relationships between workload and injury occurrence.

Project Overview

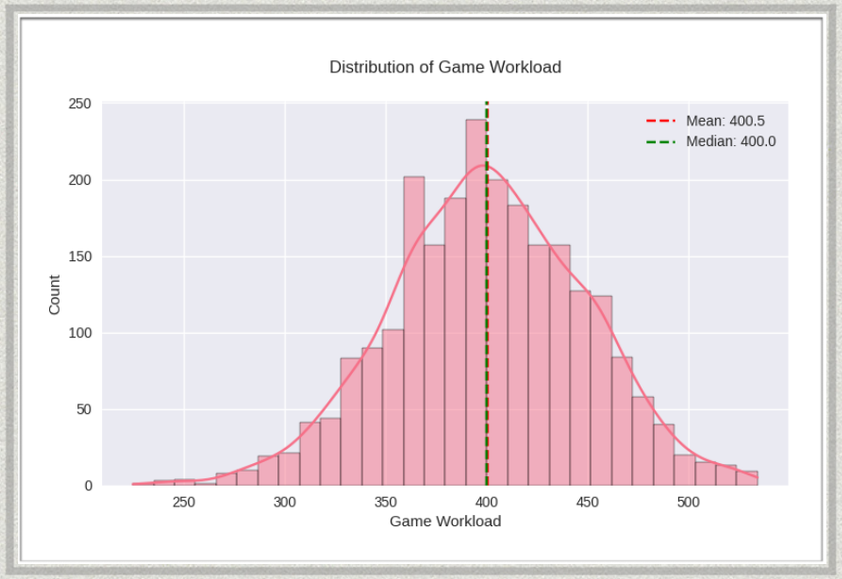
OBJECTİVES

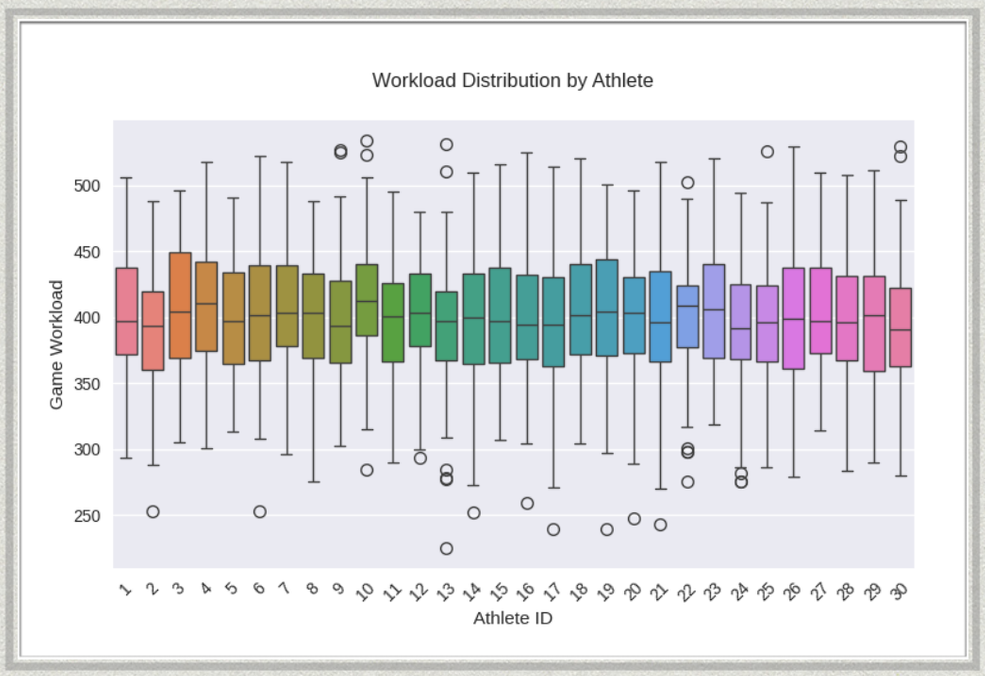
- Development of individual athlete injury risk profiles
- Early identification of high-risk athletes
- Analysis of the relationship between workload and injury occurrence
- Creation of a predictive model for injury risk assessment

DATA SOURCES

- Metrics Dataset - Hip mobility and groin squeeze measurements for 30 athletes (2016-2018)
- Workload Dataset - Game-specific performance and exertion metrics
- Injuries Dataset - Complete injury tracking system

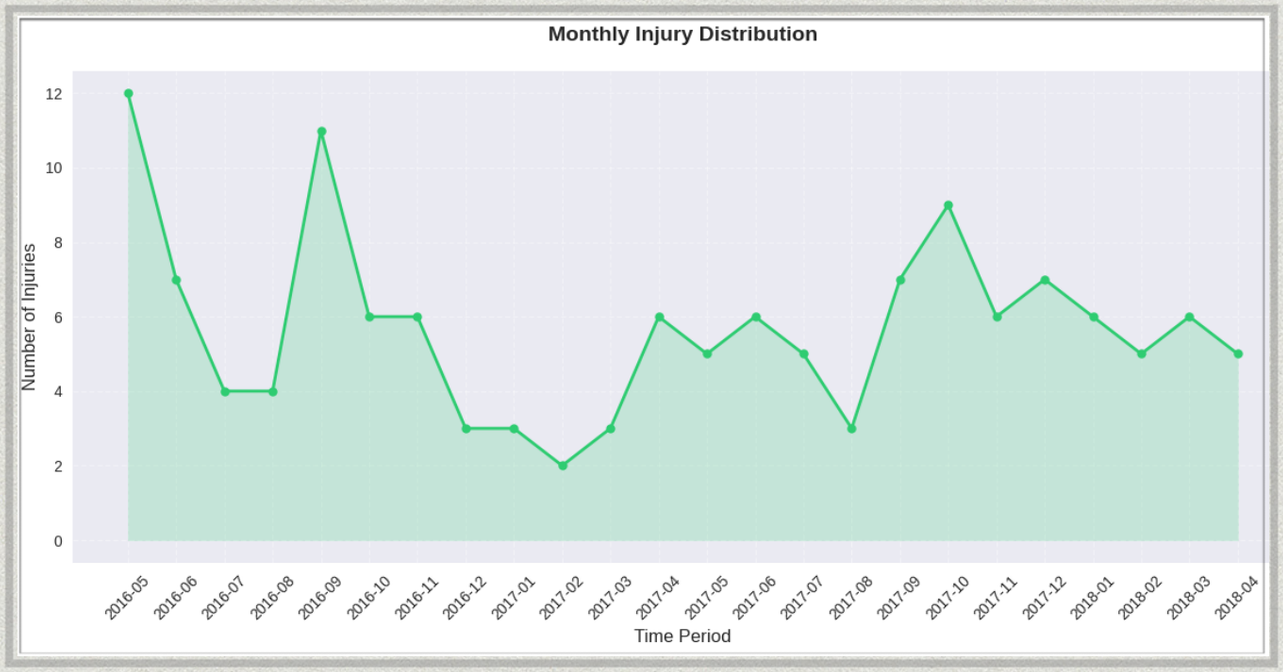
- Shows a normal-like distribution centered around 400 units
- The distribution is slightly right-skewed
- Range appears to be between ~250-500 units

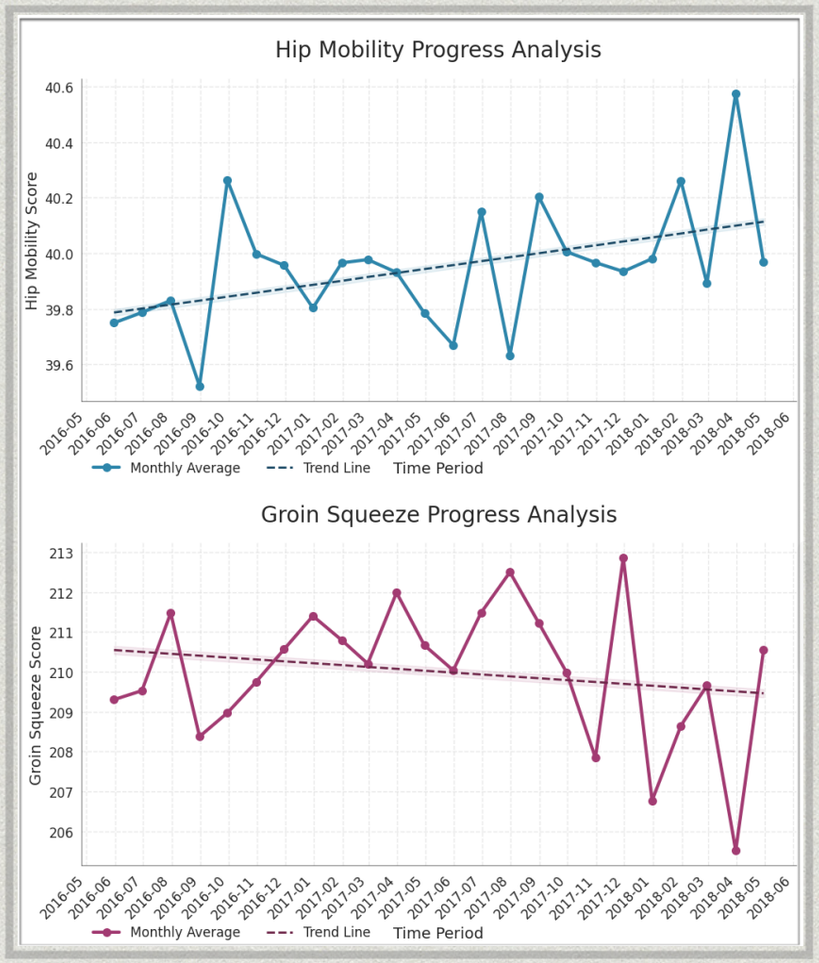




- Median workloads generally fall between 350-450 units
- Several outliers present, particularly in lower ranges
- Relatively consistent interquartile ranges across athletes

- Peak of 12 injuries in one month
- Minimum of 2 injuries in another month
- Average of 5.7 injuries per month





- Hip Mobility: Positive trend (slope: 0.0142), significant (p: 0.0312), weak correlation (R^2: 0.1941).
- Groin Squeeze: Negative trend (slope: -0.0471), not significant (p: 0.3699), very low correlation (R^2: 0.0367).
- High variability; hip mobility shows fluctuations, groin squeeze ranges between 206-213 units.

Methodology

Data Integration and Preprocessing

DATASET MERGING

- Merging DataFrames: Combined metrics_df and game_workload_df using an outer join on athlete_id and date. Merged injuries_df with an "injured" status.
- Handling NaN Values: Filled NaN values in injuries_status with "non_injured" and in game_workload with 0.
- Pivoting Data: Created a pivot table to consolidate measurements by athlete_id, date, game_workload, and injuries_status, transforming metrics into columns.

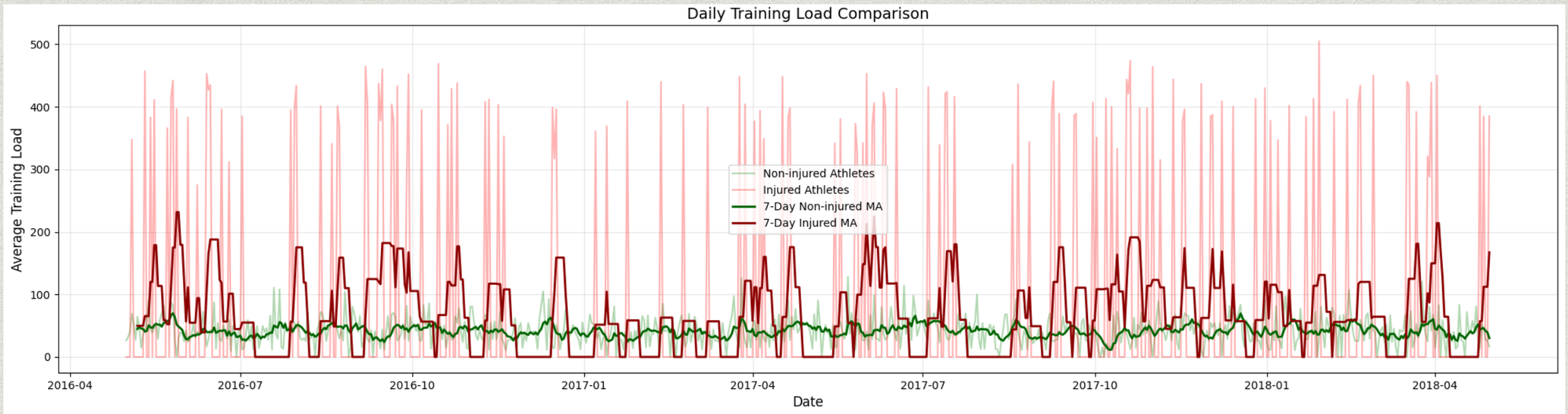
FEATURE ENGINEERING

- Injury Status: The injuries_status is converted to 0 (non-injured) and 1 (injured).workload_7d: Cumulative 7-day workload
- Resting Days: The resting feature tracks non-training days (0: resting, 1: active).
- Workload Features:
  - workload_7d: The total workload over the last 7 days.
  - acwr: The ratio of short-term (7 days) to long-term (28 days) workload.
  - workload_change: Daily percentage change in workload.
- Mobility & Injury Features:
  - hip_trend, groin_trend: Trends in mobility metrics.
  - injuries_30d, days_since_injury: Number of injuries in the last 30 days and days since the last injury.
- Risk Scores:
  - workload_risk: Risk associated with the current workload.
  - overall_risk: Overall injury risk, calculated using workload, mobility, and rest data. (Target feature)

Analysis Techniques

STATISTICAL ANALYSIS

- Correlation analysis between metrics and injury occurrence
- Time series analysis of workload patterns
- Distribution analysis of injury frequencies



Athletes with injuries have higher training loads and messy, and accordingly, their injury rates are higher.

LSTM (LONG SHORT-TERM MEMORY) ARCHITECTURE

LSTMModel Class

Defines the LSTM model with an LSTM layer, followed by two fully connected layers, batch normalization, and dropout to enhance model capacity and generalization.

EarlyStopping Class

Prevents overfitting by stopping training if the validation loss doesn't improve beyond a certain threshold (delta).

Individual Model

Trains a separate model for each athlete by preprocessing data (scaling, time series preparation), training the model, and evaluating performance using metrics like MSE, RMSE, MAE, and R-squared.

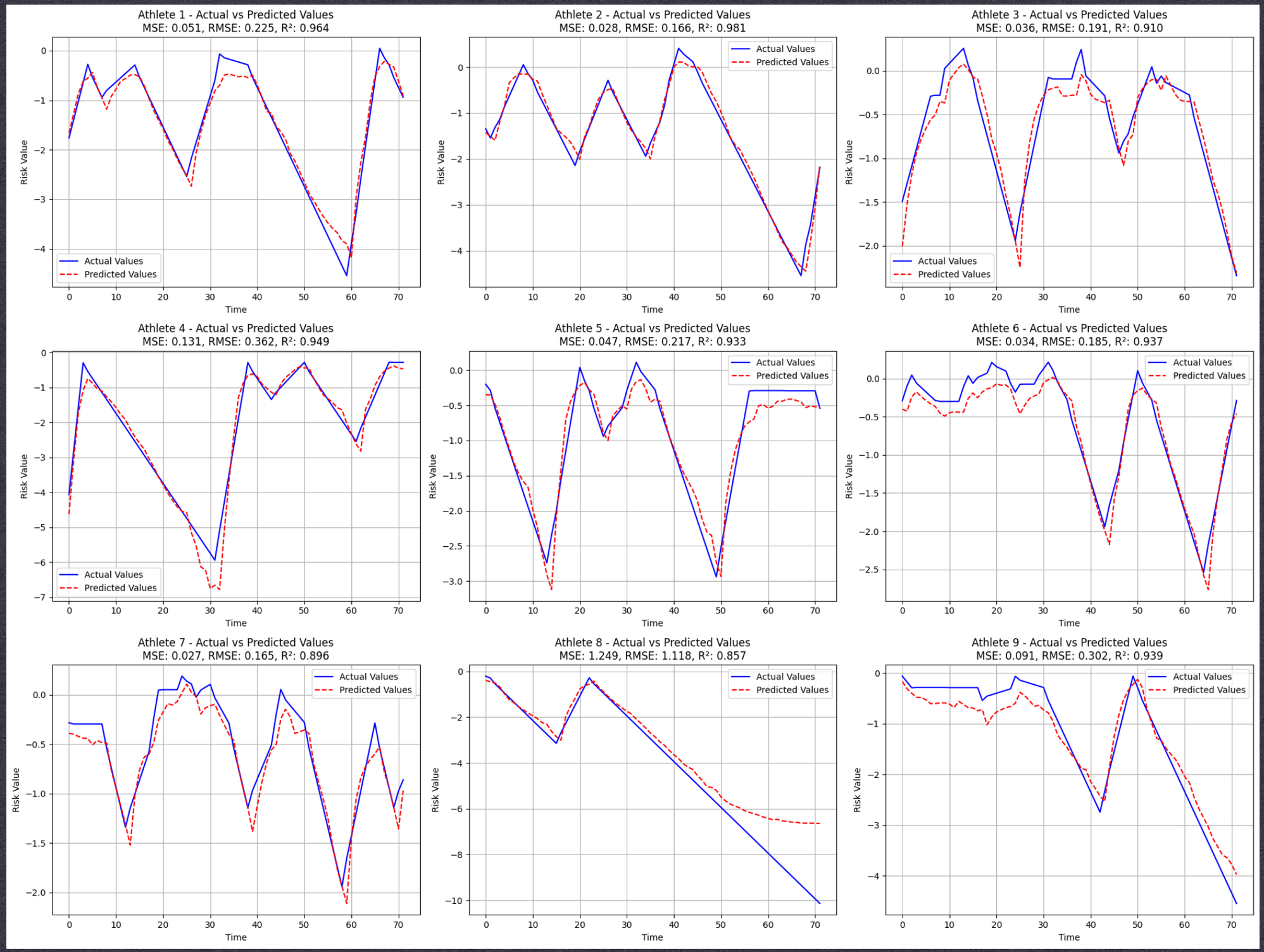
Sequences Function

Prepares input and output data for the time series model.

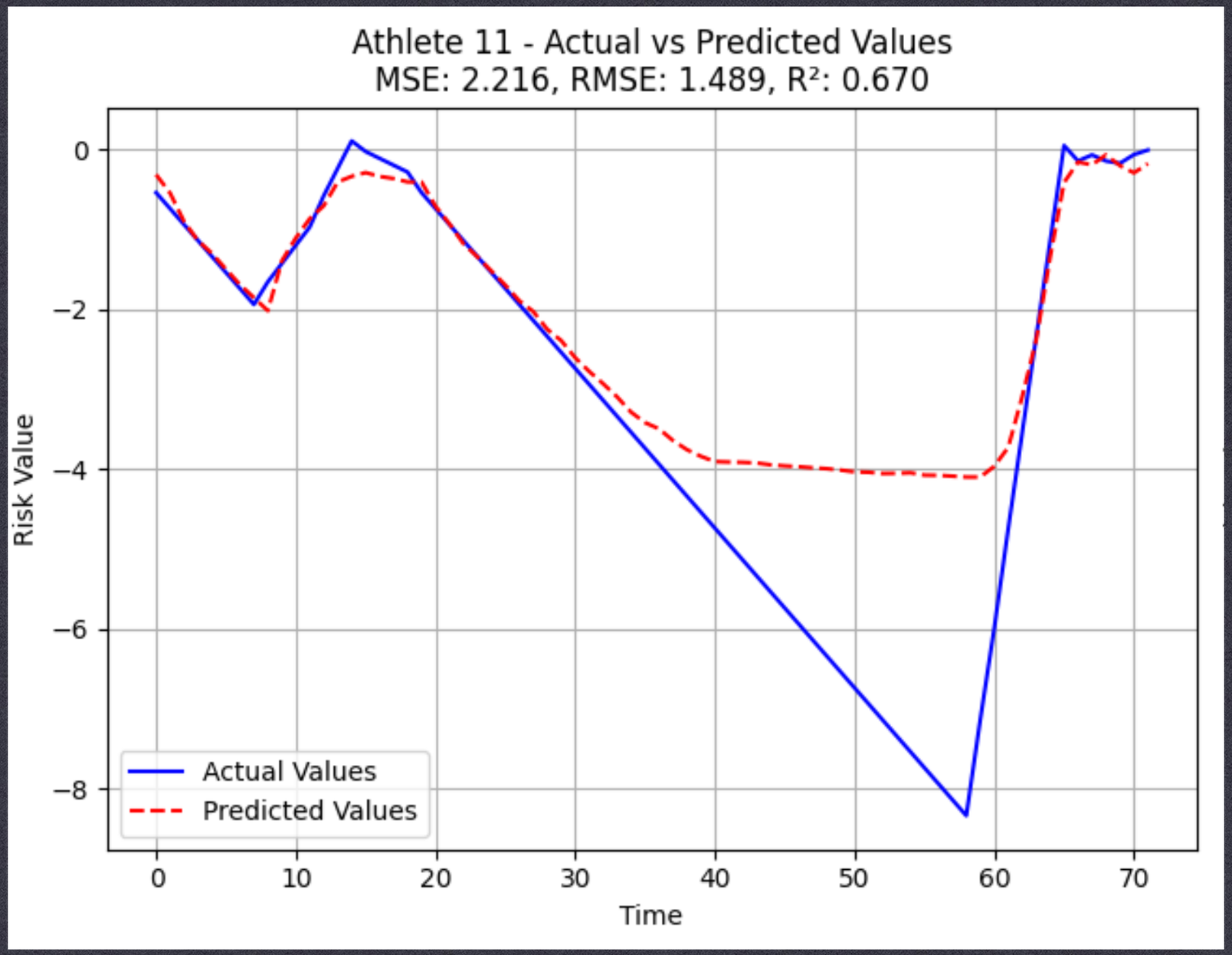
Hyperparameters

Lookback (sequence length), epochs, batch size, and patience (early stopping tolerance) control the training process.

The project trains individual LSTM models for each athlete and reports performance metrics both individually and overall. The code needs to be packaged for integration, with automated training, evaluation, and clear reporting of results.



For athletes with good alignment (e.g. Athlete 2, Athlete 5), the model appears to be capturing the trends and fluctuations in the actual data quite well. This suggests the model is performing well for predicting the performance of these athletes.



However, there are several athletes (e.g. Athlete 6, Athlete 8, Athlete 11) where the predicted values deviate quite substantially from the actual values, especially in certain time periods. This indicates the model is struggling to accurately forecast the performance for these athletes.

- The data emphasizes the need for individualized monitoring of athletes' injury risks. Athletes like **Athlete 8**, who show extreme variations, may benefit from additional rest or changes in their training regime.

- For athletes with consistently high injury risks, like **Athlete 11**, targeted interventions (e.g., adjustments in training load, mobility work) might be necessary to prevent injuries.

- These risk patterns underscore the importance of managing workload, rest, and recovery to minimize the likelihood of injury, particularly for athletes showing rising risks over time.

Athlete	Month 1	Month 2	Month 3
Athlete 1	17.30%	24.84%	50.90%
Athlete 2	20.45%	15.49%	61.24%
Athlete 3	23.72%	21.60%	30.38%
Athlete 4	28.69%	40.01%	11.56%
Athlete 5	38.80%	28.30%	30.22%
Athlete 6	10.22%	25.86%	41.41%
Athlete 7	26.36%	19.43%	42.96%
Athlete 8	18.40%	38.46%	90.44%
Athlete 9	12.22%	29.05%	48.31%
Athlete 10	11.03%	19.78%	53.70%
Athlete 11	21.31%	76.90%	68.16%
Athlete 12	30.52%	10.34%	33.28%
Athlete 13	35.02%	14.55%	67.53%
Athlete 14	63.58%	20.72%	10.05%
Athlete 15	11.35%	47.99%	8.93%
Athlete 16	30.05%	40.18%	15.37%
Athlete 17	46.36%	23.67%	73.34%
Athlete 18	7.99%	52.06%	31.45%
Athlete 19	36.53%	41.00%	18.07%
Athlete 20	51.53%	12.52%	17.03%
Athlete 21	29.94%	22.59%	31.30%
Athlete 22	14.59%	19.44%	35.45%
Athlete 23	26.04%	31.48%	10.45%
Athlete 24	15.69%	44.18%	10.77%
Athlete 25	19.57%	48.96%	36.12%
Athlete 26	24.00%	20.86%	45.26%
Athlete 27	27.07%	27.44%	24.82%
Athlete 28	18.96%	34.61%	36.32%
Athlete 29	37.51%	34.28%	9.42%
Athlete 30	11.23%	56.51%	31.61%

Note: This project has currently been developed solely for injury risk prediction. However, it can be examined from different perspectives, and the business problem can be altered to build models for various objectives. Further development on this topic will follow.

Hilal Alpak