Predicting Soccer Injuries with Machine Learning

Module 7: Final Project Presentation

Jennifer Lawless
DASC 9311: Data Science Project
December 10, 2025

Problem and Solution Overview

Core Problem	• Predict whether a soccer player will suffer an injury (requiring at least 7 consecutive days off) in their next season
Why This Matters	• Player safety and career longevity • Identifying key injury risk factors • Preventative healthcare • Training program optimization
Goals	• **Primary Goal:** >85% accuracy with strong generalization, high recall • **Secondary Goal:** Outperform existing Kaggle solutions • **Practical Goal:** Provide actionable insights for injury prevention
Solution Achieved	• **95.1%** accuracy • **95.2%** recall • Beat Kaggle benchmarks • Identified key risk factors

Dataset Overview

Samples

800 Chinese university soccer players aged 18-24 participating in collegiate and provincial leagues

Target

Injury_Next_Season: perfectly balanced binary classification value

Data Quality

No missing values, multi-source verification (medical records, coach reports, student surveys)

18 Features Across 5 Categories:

Physical Characteristics

Age, Height, Weight, BMI

Physical Fitness Assessments

Knee Strength, Hamstring Flexibility, Balance, Sprint Speed, Agility, Reaction Time

Soccer-Specific Metrics

Position, Training Hours, Matches Played, Previous Injury Count

Lifestyle Factors

Sleep Hours, Stress Level, Nutrition Quality

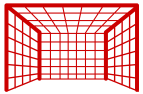
Training Compliance

Warmup Routine Adherence

Link to Kaggle: https://www.kaggle.com/datasets/yuanchunhong/university-football-injury-prediction-dataset/data

Solution Strategy





Comprehensive EDA	Feature Engineering	Model Development	Cross-Validation	Multi-Metric Evaluation
• Performed an EDA using distributions, injury patterns, correlation analysis, and feature importances	• Created composite scores (Fitness Score, Wellness Score) and risk indicators	• Built Baseline to Advanced to Ensemble models with iterative refinement, total of 8 models	• Used nested cross-validation on models to prevent overfitting	• Evaluated models based on Accuracy Rate, Recall, Precision, F1-Score, and ROC-AUC

Key EDA and Feature Engineering Findings

Methods Used

Feature Distribution Analysis (histograms, bar plots)

Injury Pattern Analysis (statistical tests, box plots)

Correlation Analysis (correlation matrix)

Feature Importance Tests (ANOVA, Mutual Information)

Strong Predictors

Engineered composite scores (Fitness Score, Wellness Score)

Physical fitness assessments (balance, reaction time)

Lifestyle factors (sleep, stress, nutrition)

Previous injury count

Weak Predictors

Demographics (age, height, weight, BMI)

Playing position

Training statistics (training hours per week, matches played)

Project History and Evolution

Weeks 1-2 (9/1-9/14)

- Dataset acquisition, Kaggle benchmark, review of relevant literature

Weeks 3-4 (9/15-9/28)

- EDA, feature analysis and engineering, data preprocessing

Weeks 5-7.5 (9/29-10/22)

- Baseline model development (LR, RF, XGB), initial underperformance

Weeks 7.5-10.5 (10/23-11/5)

- Refinement of baseline models, advanced model development (LightGBM, SVM)

Weeks 10.5-13 (11/6-11/23)

- Ensemble model development (Soft Voting, Stacking, BMA)

Weeks 14-15 (11/24-12/7)

- Final analysis, documentation, ethical review

Key Design Changes

Feature Set Refinement

Split into df_raw (tree models) vs. full engineered set (linear models) after discovering engineered features caused noise in Random Forest/XGBoost

Hyperparameter Strategy

Moved from broad grid searches to focused ranges based on preliminary results

Model Persistence

Added joblib to save models, enabling ensemble methods to combine earlier base models

Final Model Performance

Rank	Model Group	Model	Accuracy	Precision	Recall	F1	ROC-AUC	Key Strengths/Weaknesses
1	Ensemble	Model 7: Stacking Ensemble	0.951	0.951	0.952	0.951	0.990	Ranked #1 overall by Accuracy and Recall, best overall balance of metrics
2	Baseline	Model 1: Logistic Regression	0.951	0.953	0.950	0.951	0.991	Highest ROC-AUC among models, simple, fast, and highly interpretable
3	Ensemble	Model 6: Soft Voting Classifier	0.950	0.953	0.948	0.950	0.990	Benefits from model diversity but is computationally heavier, very robust
4	Ensemble	Model 8: Bayesian Model Averaging	0.950	0.953	0.948	0.950	NaN	Probabilistic framework offers robust results, yet still worst-performing of the ensembles
5	Baseline	Model 3: XGBoost	0.949	0.946	0.953	0.949	0.988	Strong recall and fast training, slightly lower precision and interpretability
6	Advanced	Model 4: LightGBM	0.948	0.940	0.957	0.948	0.988	Highest recall, efficient on large datasets but sensitive to hyperparameter tuning
7	Advanced	Model 5: Support Vector Machine	0.946	0.948	0.945	0.946	0.988	Solid precision, good for high-dimensional data but slower and harder to scale
8	Baseline	Model 2: Random Forest	0.945	0.952	0.938	0.945	0.985	Worst performing model, slightly lower recall and less efficient on large datasets

Challenges Encountered

Kaggle Benchmark Replication Issues

Kaggle Solution 2 Discrepancy:

- Reported: XGBoost accuracy 0.969, recall 1.000
- Replicated: accuracy 0.943-0.945
- Suggests potential inflation or unreported preprocessing

Dataset Limitations

- Only 800 samples – did careful cross-validation to prevent overfitting
- Perfectly balanced target (50-50) may not reflect real-world injury rates (typically lower)

1

Feature Engineering Complexity

Engineered features (Fitness Score, Wellness Score) highly predictive for linear models but introduced noise for tree-based algorithms. Solution: Model-specific feature sets

2

3

Computational Resources

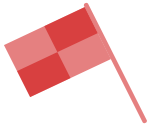
Large hyperparameter grids required RandomizedSearchCV and GPU acceleration for LightGBM to reduce training times

4

5

Beating Benchmarks

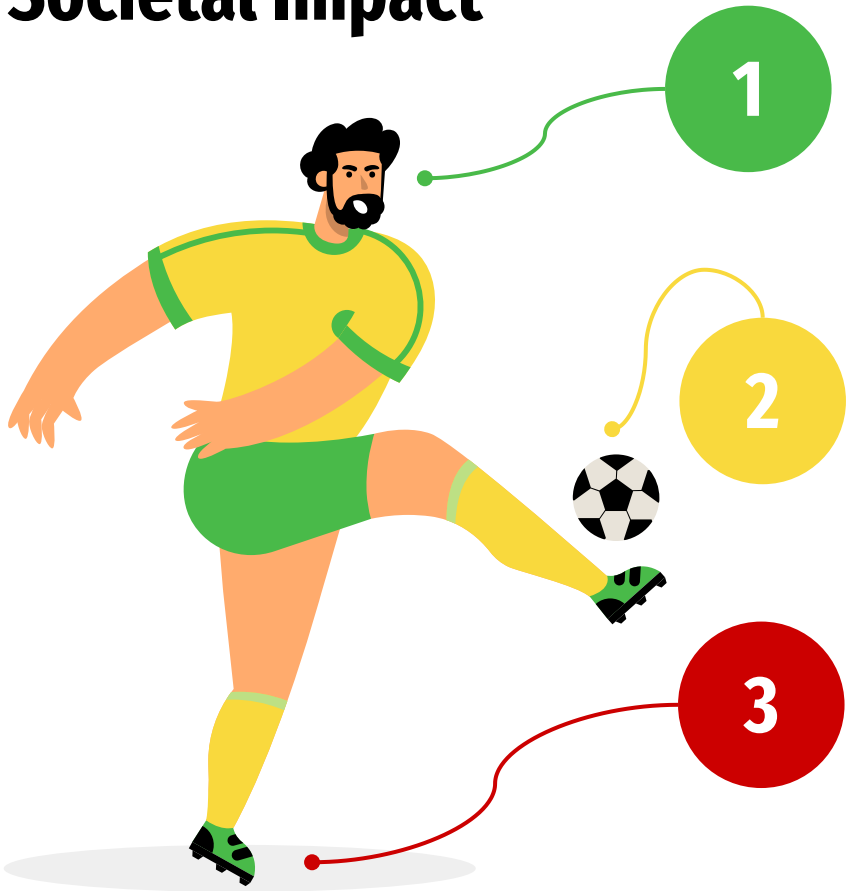
Initial baseline models underperformed and required multiple iterations of feature refinement and ensemble development to beat existing solutions



Outstanding Issues and Future Directions

Open Questions		Proposed Solutions	
?	Do predictions remain stable across multiple seasons?	✓	Longitudinal tracking across seasons
?	How do real-time physiological changes affect risk?	✓	Incorporate wearable sensor data (HR variability, load monitoring)
?	Can deep learning find additional patterns?	✓	Explore neural network architectures
?	How should coaches interpret and act on predictions?	✓	Develop LIME explanations for individual players
?	Does the model generalize to other leagues/sports?	✓	Validate on diverse datasets (different countries, sports)

Societal Impact



Athlete Health and Safety

- Identify high-risk players before injuries occur
- Implement preventative programs (modified training loads, recovery time, strength programs)
- Reduce injury rates and extend athlete careers

Economic Impact

- Reduce costs of medical treatment and rehabilitation
- Minimize lost playing time
- Universities and organizations save on injury-related expenses

Evidence-Based Training

- Model emphasizes modifiable risk factors: sleep, nutrition, stress, warmup adherence
- Athletes empowered to take active roles in their health
- Educational programs can promote healthier lifestyle choices

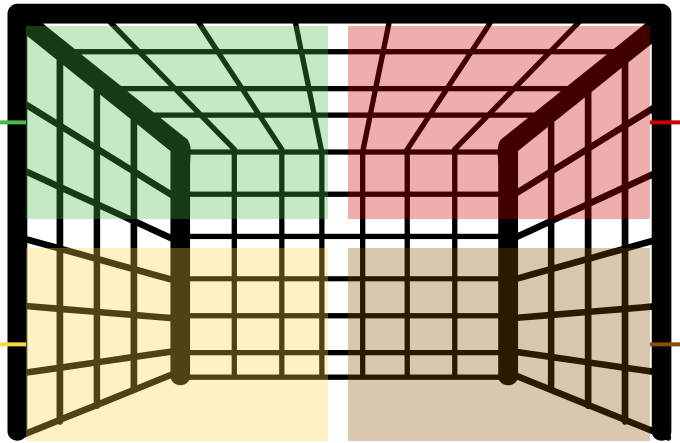
Ethical Implications

Privacy and Data Security

Model uses sensitive health and physiological data. Requires secure storage, explicit consent, and clear communication. Athletes must retain the right to opt out without penalty.

Model Accuracy and False Predictions

False positives may cause stress, false negatives risk preventable injury. Communicate clearly that no model is 100% accurate.



Discrimination and Misuse Risks

Predictions must not influence recruitment or playing time. High-risk labels could unfairly impact athletes. Solution: use predictions only for preventative care, not selection decisions.

Bias and Fairness

Training data is limited to Chinese university athletes (18-24). May not generalize to other groups. Requires bias audits, diverse validation, and fairness safeguards.

Conclusion

Key Insights

Dynamic Factors > Demographics

Fitness, wellness, and recovery are far more predictive than age, height, or BMI

1

Model-Specific Feature Engineering

2

Tailored feature sets are crucial for optimal performance

Ensemble Methods Win

3

Combining diverse models achieved most robust predictions

Ethics Are Essential

4

Privacy, bias auditing, and transparency are vital

Project Success

Achieved 95.1% accuracy and 95.2% recall, exceeding 85% threshold and outperforming the Kaggle benchmarks through Stacking Ensemble approach

GitHub Code Repository

https://github.com/lawlesje/predicting-soccer-injuries-with-machine-learning

Thank You!

Questions?