Derrick Robinson

DSC 550

Final Project Report: Predicting Athlete Recovery Time

---

1. Introduction

Problem Statement

The objective of this project is to develop a predictive model that estimates the recovery time of athletes following injuries. This model uses features such as player height, weight, training intensity, and injury history to make predictions. Accurate predictions of recovery time are valuable to sports teams, athletic trainers, and medical professionals as they aid in managing player workloads, optimizing rehabilitation schedules, and enhancing overall team performance.

Importance and Usefulness

Injury recovery is a critical aspect of sports performance and team strategy. Incorrect estimates can result in premature returns to play, risking re-injury, or overly cautious delays, impacting team success. A data-driven recovery prediction model enables:

- Improved decision-making in rehabilitation planning
- Efficient resource allocation by medical staff
- Enhanced injury prevention strategies

- Competitive advantage through optimal player availability

Stakeholder Pitch

"Imagine a tool that provides your medical team with quick, reliable estimates of recovery time based on real data. With this predictive model, coaches and trainers can better plan team rotations, monitor training loads, and reduce costly injuries. This is more than a model; it's your new assistant in strategic player management."

Data Source

The dataset used for this project was obtained from Kaggle's Injury Prediction Dataset. It includes anonymized records of athletes, featuring demographics, training variables, previous injury history, and actual recovery times.

---

2. Milestone 1: EDA and Problem Framing

Target Variable and Problem Type

The target variable for this project is Recovery_Time, a continuous variable representing the number of days it takes an athlete to return to play. This defines the problem as a **regression task**.

Key Exploratory Data Analysis (EDA) Insights

1. **Scatter Plot of Height vs. Recovery Time**: No clear linear trend was observed.

2. **Box Plot of Previous Injuries vs. Recovery Time**: A positive relationship was found—athletes with more past injuries tend to have longer recovery times.
3. **Histogram of Recovery Time**: The distribution is right-skewed, with most recovery times clustered below 10 days.
4. **Training Intensity vs. Recovery Time**: Indications of a weak non-linear relationship, suggesting more intense training may slightly increase risk.

These findings informed decisions on feature engineering and model complexity in later phases.

---

3. Milestone 2: Data Preparation

Dropped Features

Columns with unclear predictive value or high correlation to other fields were dropped. For example, Likelihood_of_Injury was removed due to redundancy and potential data leakage.

Handling Missing Data

Missing numeric values were imputed using median imputation, which is robust to outliers. This preserved the dataset's size without skewing the distribution.

Feature Engineering

- **BMI** was calculated using player height and weight to provide a better proxy for body composition.
- **Scaled Training Intensity** using Min-Max scaling to normalize the input range.

Dummy Variables

Categorical variables were encoded using one-hot encoding, though few categorical features were included.

The dataset was cleaned, enhanced, and reshaped into a form suitable for modeling.

---

4. Milestone 3: Model Building and Evaluation

Models Used

1. **Linear Regression**: Served as the baseline model due to its simplicity and interpretability.
2. **Random Forest Regressor**: Selected to capture non-linear relationships and interactions between features.

Evaluation Metrics

- **Mean Absolute Error (MAE)**: Measures average prediction error
- **Root Mean Squared Error (RMSE)**: Penalizes larger errors
- **R-Squared (R^2)**: Indicates proportion of variance explained by the model

Model Performance

Model	**MAE**	**RMSE**	**R^2**
Linear Regression	4.32	5.86	0.47

Model	MAE	RMSE	R^2
Random Forest Regressor	3.15	4.20	0.71

The Random Forest model demonstrated superior performance, confirming the presence of non-linear relationships within the dataset.

---

5. Conclusion

Insights from the Model

The modeling process highlighted the predictive importance of features like previous injuries, training intensity, and BMI. The baseline linear regression provided interpretability but was limited in accuracy. The Random Forest model, by contrast, better captured the complexity of the data and delivered significantly improved performance.

Deployment Readiness

While the Random Forest model shows strong potential, deployment would benefit from:

- More diverse and comprehensive data sources
- Integration with real-time tracking systems for up-to-date predictions
- Validation on new injury cases to ensure generalizability

Recommendations

- Use the model for internal analytics and medical planning

- Explore advanced models (e.g., XGBoost, neural networks)
- Collaborate with sports teams for live trials and feedback loops

Challenges and Future Opportunities

- The limited sample size may impact generalizability
- The model could be expanded to classify injury severity or predict injury occurrence
- Additional features like injury location, type, and sport played could further improve accuracy

This project lays a strong foundation for future predictive analytics in sports medicine and team management.