Orange Hoops Data Science Challenge

Team: Data Wizards

- Rushikesh Shinde
- Sagarika Shinde
- Sejal Sardal
- Srushti Shobhane

Problem Statement

Objective: Predict player injuries using various attributes from three provided datasets.

Loading and Inspecting the Dataset

- **Datasets Loaded:**
  - Player Sessions: Contains details of each player session.
  - Muscle Imbalance: Includes data on player muscle imbalances.
  - Injury History: Records of past injuries for each player.
- **Initial Inspection Steps:**
  - Used “info()” to view data structure, data types, and null values.
  - Displayed the first few rows with “head()” to get an initial look at each dataset.

Data Cleaning and Preprocessing

Missing Values:

Checked each dataset for missing values.

Plan: Address missing values after merging the datasets.

Date Conversion:

Converted “Session_Date” column in “player_sessions” to datetime format for easier time-based analysis.

Cleaning Merged Data

- **Merge Process:**
  - Merged “injury_history” with “player_sessions” on Player.ID.
  - Further merged with muscle_imbalance on Player.ID.
- **Post-Merge Adjustments:**
  - Identified redundant columns (Name, Player Name, Group.Id) created during merging.
  - Dropped extra columns to retain only unique identifiers.
  - Handled missing values present in the “Side” and “Severity” column.

Before Cleaning:

Player.ID	0
Name_x	0
Group.Id_x	0
Injury Type	0
Body Part	0
Side	12259
Injury Date	0
Severity	22074
Recovery Time (days)	0
Additional Notes	0
Group.name	0
League.ID	0
Session.ID	0
Session_Date	0
Position	0
Distance..mi.	0
Distance...min..mi.	0
Duration..s.	0
Steps	0
Speed....of.max......	0
Speed..max....mph.	0
Speed..?ò...mph.	0
Time..s.	0
Accumulated.Acceleration.Load	0
Anaerobic.Activity..distance...mi.	0
...	
HamstringImbalance Percent	0
Calf Imbalance Percent	0
Groin Imbalance Percent	0

After Cleaning:

Player.ID	0
Name_x	0
Group.Id_x	0
Injury Type	0
Body Part	0
Side	0
Injury Date	0
Severity	0
Recovery Time (days)	0
Additional Notes	0
Group.name	0
League.ID	0
Session.ID	0
Session_Date	0
Position	0
Distance..mi.	0
Distance...min..mi.	0
Duration..s.	0
Steps	0
Speed....of.max......	0
Speed..max....mph.	0
Speed..?ò...mph.	0
Time..s.	0
Accumulated.Acceleration.Load	0
Anaerobic.Activity..distance...mi.	0
...	
HamstringImbalance Percent	0
Calf Imbalance Percent	0
Groin Imbalance Percent	0

Feature Engineering

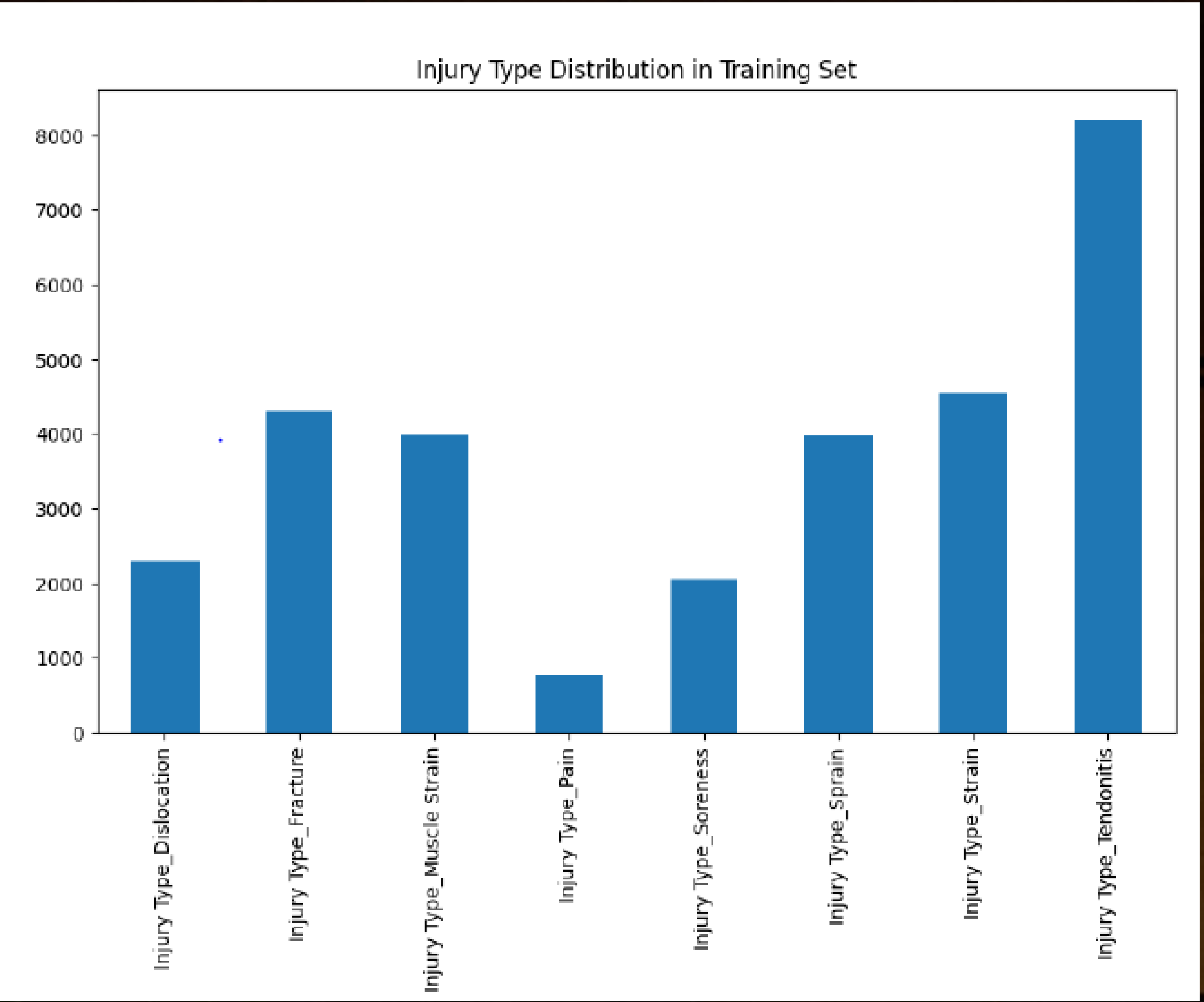
- **Binary Flags:** Created flags for missing Severity and Side columns.
- **One-Hot Encoding:** Converted categorical columns (Position, Body Part, Side, Injury Type) to numeric.
- **Date Conversion:** Converted "Injury Date" to datetime format.
- **Scaling:** Applied "StandardScaler" to numerical features for better model performance.
- **Ordinal Encoding:** Mapped Severity to ordinal values (e.g., Grade 1 = 1, Grade 2 = 2).
- **New Features:** Created "Duration_per_mile" (time per distance) , giving an idea of time taken to cover a distance.
- **Target Creation:** Created binary Injury_Flag for injury occurrence.
- **Data Splitting:** Split data into 80% training and 20% testing.

Model Selection and Training

- **Model Chosen:**
  - Gradient Boosting Classifier wrapped with “MultiOutputClassifier” for multi-label classification.
- **Training:**
  - Trained the model using X_train and y_train, then made predictions on the test set (X_test).
- **Evaluation:**
  - **Accuracy:** Achieved accuracy of **87.79%** on the test set.
  - **Cross-validation**: Performed 5-fold cross-validation with an average score close to 0.90.
  - **Final Choice**: Gradient Boosting selected for strong recall, F1 scores, and predictive power after testing other models (e.g., Random Forest, Logistic Regression, SVM, KNN).
  - **Practical Implications:** High recall ensures at-risk players are flagged for preventive care. Feature importance insights (e.g., muscle imbalance, playing position) guide injury prevention strategies.

Injury Type Distribution in Training Set

- **Objective:**
  - Analyzed the distribution of injury types in the training set to identify the most common injury.
- **Observation:**
  - **"Tendonitis"** is the most frequent injury type, highlighting its importance for preventive measures.
- **Visualization:**
  - Bar chart showing the distribution of injury types in the training set.
- **Key Insight:**
  - Focusing on preventive measures for Tendonitis could reduce the most common injury occurrence.



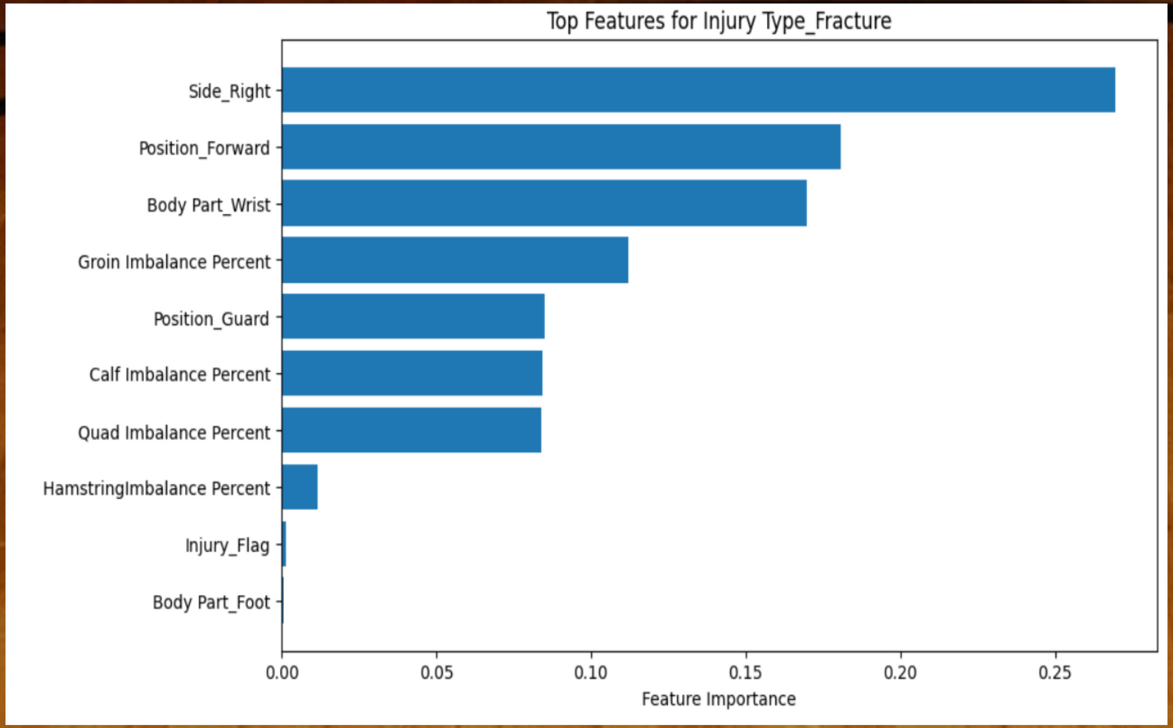
Visualization

Feature Importance

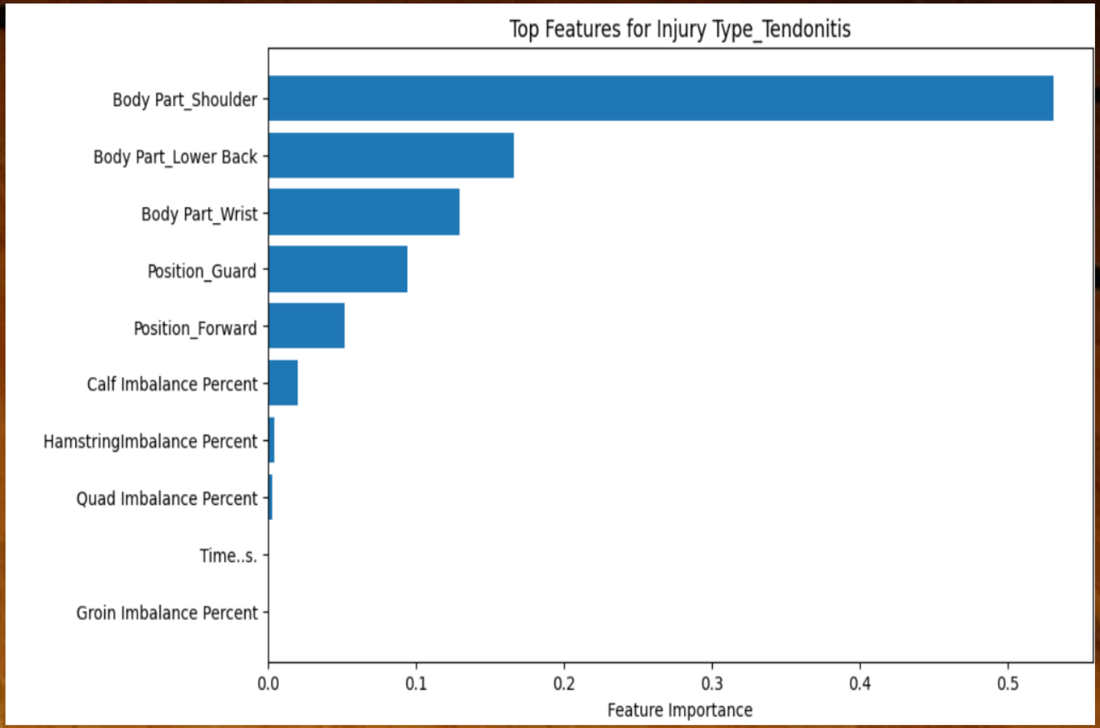
- **Objective**:
  - Identify key features influencing each injury type using the Gradient Boosting model.
- **Method**:
  - Extracted feature importances for each injury type.
  - Displayed top 10 features impacting the injury predictions.
- **Key Insight:**
  - Understanding which body parts or other features contribute most to injuries like Tendonitis helps predict future injuries.
  - This analysis supports better injury prevention and management strategies.
- **Visualization**:
  - Bar charts displaying top features for each injury type.

Visualization

Top features impacting “Injury Type_Fracture”



Top features impacting “Injury Type_Tendonitis”



Conclusion

- **Model Performance:**
  - Achieved 88% accuracy and 90% cross-validation score.
  - Model effectively predicts injury types based on training data.
- **Key Insights:**
  - Dislocation is the most common injury type, highlighted by the injury type distribution.
  - Identified key features influencing injury occurrence, helping predict future injuries.
- **Impact:**
  - Helps in injury prevention and player management.
  - Informs team selection decisions, which can impact game outcomes.


THANK YOU