Machine Learning Project Proposal(CS-584)

Date: 03/01/2024
Team members:
Tirumalesh Nagothi (A20520569)
Tarun Sai Varanasi (A20526965)
Mohana uma sushmanth (A20525576)

Goal: Analysis to forecast the likelihood of hitting a home run based on a batter's different attributes.

Research Question:

Can our group create a machine learning model that reliably forecasts, given a player, pitch, and environmental combination, whether a batted ball will be a home run?

Hypothesis:

There are some factors (such as launch angle, speed, pitcher handedness, and ballpark) that will significantly impact the probability of a home run. Our group ```thinks these intricate linkages can be efficiently learned by a machine learning model.

Methodology

Dataset Acquisition and Preprocessing:

Collaboratively, our team will locate baseball data with the designated columns. We are considering: https://www.kaggle.com/code/jcraggy/baseball-hr-prediction-xgboost-0-08-log-loss/input

Data cleaning will be a collaborative effort:

Handling outliers, missing values, and discrepancies.

Exploratory Data Analysis (EDA):

As a team, we will make use of visualizations (box plots, scatterplots, and histograms) to comprehend distributions and possible correlations between factors and the occurrence of home runs.

Together, we will compute statistical summaries to determine means, medians, correlations, and other information.

Feature Engineering:

The development of new features with predictive potential will be fueled by brainstorming sessions: Terms used in interactions (such as launch speed * launch angle), player-specific data (such as the home run rate of a hitter), and park factors (which account for ballpark dimensions)

Feature Selection:

In order to determine the most significant features, we will jointly assess feature importance utilizing methods like correlation analysis, dimensionality reduction (e.g., PCA), or feature importance from tree-based models.

Model Training and Selection:

Trying out various categorization algorithms as a group:

- The Logistic Regression
- Random forests or decision trees
- Vector machines for support (SVM)

If the dataset is huge, take into account possible neural networks.

The task of dividing the data into training and testing sets will be divided among us.

Utilize cross-validation in conjunction with hyperparameter optimization to prevent overfitting.

Evaluation:

Discuss about and decide which metrics are best for classification:

Accuracy:

- Accuracy (critical for reducing false positives)
- Recall (critical to reducing false negatives)
- The harmonic mean of recall and precision, or F1-score

Confusion Matrix:

Collaboratively examine the ROC curve and determine the AUC.

Interpretation and Visualization:

Understanding which variables influence the model's predictions will be the main topic of group discussions. To visualize model performance and decision boundaries (if any), we will work together.

Expected Outcomes:

- A machine learning model that can predict home runs with a respectable degree of accuracy.
- The team's understanding of the main variables influencing home runs.
- Possible uses in game analysis, fantasy sports, and player scouting.

Project Tools:

- Python: Main programming language
- pandas: Data manipulation and analysis
- numpy: Numerical computation
- scikit-learn: Machine learning algorithms
- matplotlib / Plotly: Visualizations

References:

https://pandas.pydata.org/pandas-docs/stable/index.html
https://numpy.org
https://www.python.org
https://scikit-learn.org/stable/
https://matplotlib.org
https://plotly.com