ETL:

All data for MLB starting pitchers was downloaded from FanGraphs for seasons 2018 – 2023.

1. **Data Cleaning and Manipulation**:
   * Filtered out rows with the season 2020 due to the pandemic-shortened year.
   * Null values in the DataFrame are filled with zeros.
   * Columns containing redundant stats were dropped from the DataFrame.
   * The data was filtered for specific relevant seasons (2018, 2019, 2021, 2022).
   * Data for the year 2023 is grouped by player ID.
   * Average statistics for ERA, FIP, and WHIP are calculated.
   * The average stats were merged with the 2023 stats for each player.
   * The pitcher names were merged with the results DataFrame.
2. **Z-Score Calculation**:
   * Z-scores were calculated for ERA, FIP, and WHIP for both the average values and the values for the year 2023.
   * The differences between z-scores were computed.
3. **Adding Z-Scores and Differences**:
   * The calculated z-scores and differences were added as new columns to the DataFrame.
4. **Output and Saving Data**:
   * The script outputs the merged DataFrame with z-scores and differences.
   * Intermediate values, standard deviations, and some specific columns are displayed.
5. **Creating Learning Datasets**:
   * Created separate datasets for learning ERA, FIP, and WHIP.
   * Specific columns were dropped from each dataset.
   * The modified datasets were saved as separate CSV files.

Analysis and ML:

Classification Results:

Logistic regression and XGBoost was used in training, evaluating, and visualizing the results of classification models. The goal was to predict performance metrics (ERA, FIP, WHIP) for pitchers.

1. \*\*Data Preprocessing:\*\*

- Imported necessary libraries and reading data from CSV files into separate DataFrames for ERA, FIP, and WHIP.

- Missing values in these DataFrames were imputed using the median strategy.

2. \*\*Logistic Regression Model for ERA, FIP, and WHIP:\*\*

- The `train\_and\_evaluate\_logistic\_regression` function used for a logistic regression model for (ERA, FIP, WHIP).

- The target columns were transformed into binary classes (0 or 1) based on whether the value is greater than 0.

- The data was split into training and testing sets, followed by applying SMOTE for oversampling to address class imbalances.

- Standardization using StandardScaler was applied to the features.

- Logistic regression model was trained, predictions were made, and accuracy calculated.

- A SHAP explainer was created to analyze feature importance.

3. \*\*Classification Model with XGBoost for ERA, FIP, and WHIP:\*\*

- The `train\_and\_evaluate\_classification\_model` function was evaluated using XGBoost classifier for each specific target column.

- Missing columns in the DataFrame were dropped, and the target column transformed similarly to the logistic regression case.

- Data was split, scaled, and a pipeline with SMOTE and XGBoost was created.

- Cross-validation scores were computed, the pipeline trained, and predictions made.

- SHAP values were calculated for feature importance.

4. \*\*Creating Predictions DataFrame:\*\*

- A function `create\_predictions\_dataframe` was defined to create a DataFrame containing predictions, actual values, and correctness (whether the prediction matches the actual value).

5. \*\*Running the Models for ERA, FIP, and WHIP:\*\*

- The code ran the logistic regression and classification models for ERA, FIP, and WHIP separately.

- The results, included accuracy and cross-validation scores, and were printed for each model.

- SHAP summary plots were generated to visualize feature importance.

6. \*\*Merging and Organizing Results:\*\*

- The prediction correctness data from ERA, FIP, and WHIP models were merged based on pitcher names.

- A DataFrame named `pitching\_verdict\_df` was created containing pitcher names and their correctness for ERA, FIP, and WHIP predictions.

7. \*\*Final Verdict DataFrame:\*\*

- The `pitching\_verdict\_df` DataFrame provides a consolidated view of correctness for ERA, FIP, and WHIP predictions for each pitcher.

Overall, the code performed classification tasks using logistic regression and XGBoost to predict the performance metrics of pitchers. It emphasizes accuracy, cross-validation scores, and SHAP values for feature importance analysis. The final DataFrame `pitching\_verdict\_df` summarizes the correctness of predictions for each performance metric and pitcher. The results can be used to assess the models' performance and provide insights into which features are most influential in making accurate predictions.

ERA Classification scores:

Cross Value Scores: [0.94814815 0.93333333 0.94814815 0.95522388 0.99253731]

Accuracy: 0.93

Recall: 0.93

ERA Shap graph

A screen shot of a graph

Description automatically generated

FIP training and evaluation results

Cross Value Scores: [0.93333333 0.94074074 0.96296296 0.95522388 0.97761194]

Accuracy: 0.95

Recall: 0.95

FIP SHAP Graph

A graph of different colored shapes

Description automatically generated

WHIP training and evaluation scores:

Cross Value Scores: [0.92592593 0.95555556 0.96296296 0.94029851 0.97761194]

Accuracy: 0.92

Recall: 0.89

WHIP SHAP Graph

A graph of different colored shapes

Description automatically generated

CONCLUSIONS:

Overall, the models seemed to work pretty well. Reviewing several sports websites were the current top pitchers in the MLB are mentioned showed that the model predicted those named players fairly accurately.