**Date: 08/06/2024**

**Problem Statement**

The primary goal of the research is to investigate the impact of robot umpires on player performance and strategies in baseball, using modern machine learning methods. Specifically, the focus is on understanding heterogeneous treatment effects of robot umpires and detecting changes in batting and pitching metrics across treatment and control groups.

**Previous Work**

The prior work by Jerry W. Kim and Brayden G. King, titled "Seeing Stars: Matthew Effects and Status Bias in Major League Baseball Umpiring," explores status bias in MLB umpiring decisions. The study found that umpires are more likely to favor high-status pitchers (e.g., those with All-Star selections) by expanding the strike zone or avoiding calling strikes against them. This bias was shown to impact actual game outcomes, providing high-status pitchers with a performance advantage.

**Key findings from the previous work include:**

1. Status Bias: Umpires are more likely to make favorable calls for high-status pitchers.

2. Ambiguity and Reputation: The ambiguity of the pitch and the pitcher's reputation as a control pitcher influence the degree of status bias.

3. Performance Impact: Biases in umpiring decisions translate into real performance advantages for high-status pitchers, affecting game outcomes.

**Expected Outcomes**

For the current research project involving robot umpires, the expected outcomes include:

1. Reduced Human Bias: Robot umpires are expected to reduce the status bias present in human umpiring decisions.

2. Performance Metrics: Analysis of performance metrics such as batting average, on-base percentage, strikeout rates, and pitching statistics to detect changes when robot umpires are used.

3. Player Strategies: Insights into how players (both batters and pitchers) adjust their strategies in response to the more consistent and unbiased calls of robot umpires.

**Research Questions**

1. Impact on Performance: How does the introduction of robot umpires affect player performance metrics?

2. Heterogeneous Treatment Effects: Are there heterogeneous effects of robot umpires on different types of players (e.g., high-status vs. low-status players)?

3. Strategic Adjustments: How do players alter their strategies in response to the use of robot umpires?

**Methodology**

1. Causal Forests: To uncover heterogeneous treatment effects, the research will use causal forests, a machine learning method designed to estimate the impact of interventions across different subgroups within the data. This technique will help identify which players benefit the most from robot umpires.

2. Classification Models: Various machine learning classification algorithms will be employed to detect changes in batting and pitching metrics. The goal is to classify the outcomes based on whether robot or human umpires were used, thereby identifying shifts in player behavior and performance.

**1. Data Collection and Preparation**

**Data Sources:**

- MLB Pitch Location Data: Historical data on pitch locations, outcomes, and player performance.

- Minor League Baseball Data: Data from leagues where robot umpires are used, including pitch locations, outcomes, and player performance.

- Player Statistics: Comprehensive stats for players, including batting average, on-base percentage, strikeout rates, etc.

**Data Preparation:**

- Data Cleaning: Remove any erroneous or incomplete data entries.

- Feature Engineering: Create new features from the raw data, such as pitch type, pitch location, game context (e.g., inning, score), and player status.

- Labeling: Label the data to indicate whether the umpire is a human or a robot.

**2. Exploratory Data Analysis (EDA)**

- Descriptive Statistics: Calculate basic statistics for all features to understand their distributions.

- Visualizations: Use plots (e.g., histograms, box plots, scatter plots) to visualize the data and detect patterns or anomalies.

- Correlation Analysis: Compute correlations between features to identify relationships that may inform the model.

**3. Causal Inference with Causal Forests**

Understanding Causal Forests:

Causal forests are a type of machine learning model used to estimate heterogeneous treatment effects, i.e., how different subgroups in the data respond to a treatment (in this case, the introduction of robot umpires).

**Steps to Implement Causal Forests:**

*Define Treatment and Outcome:*

- Treatment: Use of robot umpires.

- Outcome: Player performance metrics (e.g., batting average, strikeout rate).

- Covariates: Include relevant player and game features (e.g., player status, pitch location, game context).

*Model Training:*

- Split the data into training and test sets.

- Use a causal forest algorithm to estimate the treatment effect of robot umpires on different subgroups of players.

- Model Interpretation:

- Analyze the results to identify which subgroups benefit the most or least from the use of robot umpires.

- Visualize the heterogeneous treatment effects using tools like partial dependence plots.

**4. Classification of Batting and Pitching Metrics**

*Classification Models:*

Different machine learning classification algorithms will be used to detect changes in batting and pitching metrics:

- Logistic Regression

- Random Forests

- Gradient Boosting Machines (GBMs)

- Support Vector Machines (SVMs)

- Neural Networks

*Steps to Implement Classification Models:*

- Define Features and Labels:

- Features: Player statistics, pitch data, game context, etc.

- Labels: Indicator of whether robot or human umpires were used.

- Data Splitting:

- Split the data into training and test sets.

- Model Training:

- Train multiple classification models to predict the label.

- Use cross-validation to tune hyperparameters and prevent overfitting.

- Model Evaluation:

- Evaluate the models using metrics such as accuracy, precision, recall, F1 score, and ROC-AUC.

- Compare the performance of different models to select the best one.

- Feature Importance:

- Analyze feature importance scores to understand which features are most influential in distinguishing between robot and human umpire scenarios.

**5. Strategy Analysis**

- Player Strategy Adjustments: Analyze how players alter their strategies based on umpire type.

- Pitching: Changes in pitch selection, pitch location, and frequency.

- Batting: Adjustments in swing decisions, batting stance, and approach to different pitch types.

- Performance Metrics Comparison:

- Compare traditional performance metrics (e.g., batting average, strikeout rates) across games with human umpires vs. robot umpires.

- Use statistical tests (e.g., t-tests, chi-square tests) to determine if the differences are significant.

**6. Reporting and Visualization**

- Results Summary: Summarize the findings from the causal forest analysis and classification models.

- Visualizations: Create visualizations (e.g., bar charts, line graphs, heatmaps) to present the results clearly.

- Discussion: Interpret the results in the context of the research questions, discussing the implications for player performance and strategies.

- Recommendations: Provide recommendations for stakeholders (e.g., baseball leagues, teams) based on the findings.

**Tools and Libraries**

- Python/R: For data analysis and machine learning.

- Libraries:

- Scikit-learn: Machine learning algorithms.

- Pandas/Numpy: Data manipulation and analysis.

- Matplotlib/Seaborn: Data visualization.

- EconML: Causal inference methods, including causal forests.

- Statsmodels: Statistical tests and models.

Date – 09/06/2024

1. **Game Context Data:**
   * game\_date: Date of the game.
   * game\_type: Type of game (e.g., regular season, postseason).
   * inning: Inning of the game.
   * inning\_topbot: Top or bottom of the inning.
   * home\_team: Abbreviation of the home team.
   * away\_team: Abbreviation of the away team.
   * home\_score: Home team score before the pitch.
   * away\_score: Away team score before the pitch.
   * post\_home\_score: Home team score after the pitch.
   * post\_away\_score: Away team score after the pitch.
   * outs\_when\_up: Number of outs when the batter is up.
2. **Pitch Data:**
   * pitch\_type: Type of pitch.
   * release\_speed: Speed of the pitch.
   * release\_pos\_x, release\_pos\_z: Horizontal and vertical release positions.
   * pfx\_x, pfx\_z: Horizontal and vertical movement of the pitch.
   * plate\_x, plate\_z: Horizontal and vertical position of the pitch as it crosses the plate.
   * zone: Zone location of the pitch.
   * type: Result of the pitch (e.g., ball, strike, in play).
   * events: Outcome of the plate appearance (e.g., single, double, home run).
   * description: Description of the pitch result.
   * spin\_rate: Spin rate of the pitch.
   * release\_extension: Release extension of the pitch.
3. **Batter Information:**
   * batter: MLB Player ID of the batter.
   * stand: Batter's stance (left or right).
   * batting\_avg: Batting average.
   * on\_base\_percentage: On-base percentage.
   * strikeout\_rate: Strikeout rate.
4. **Batted Ball Data:**
   * launch\_speed: Exit velocity of the batted ball.
   * launch\_angle: Launch angle of the batted ball.
   * hit\_distance: Projected distance of the batted ball.
   * hit\_location: Position of the first fielder to touch the ball.
   * bb\_type: Type of batted ball (ground\_ball, line\_drive, fly\_ball, popup).

 **Feature Engineering**: Creating new features from raw data, such as the difference in speed for pitches, game context variables, and derived performance metrics.

 **Label Encoding**: Convert categorical variables (like player names, pitch types) to numeric values for analysis.

 **Date Features**: Extract useful features from date columns (year, month, day).

#### Data Acquisition

1. **Historical Pitch Data**:
   * **Pitch Location**: X, Y coordinates of the pitch as it crosses the plate.
   * **Pitch Type**: Fastball, curveball, slider, etc.
   * **Pitch Speed**: Start and end speed of the pitch.
   * **Outcome**: Ball, strike, hit, out, etc.
   * **Game Context**: Inning, score, count, etc.
2. **Player Performance Data**:
   * **Batting Metrics**: Batting average, on-base percentage, slugging percentage, strikeout rate, etc.
   * **Pitching Metrics**: ERA (Earned Run Average), WHIP (Walks and Hits per Inning Pitched), strikeout rate, etc.
   * **Player Status**: High-status player (e.g., All-Star) or not.
3. **Game Information**:
   * **Umpire Type**: Whether the umpire is human or robot.
   * **Date and Venue**: Date and location of the game.

#### Types of Data Needed

1. **Pitch Data**:
   * Detailed information on each pitch thrown, including type, speed, location, and result.
   * Information about the context in which the pitch was thrown (inning, score, count).
2. **Player Data**:
   * Comprehensive statistics for both batters and pitchers, including performance metrics like batting average and ERA.
   * Information about the player’s status (high-status or not).
3. **Game Data**:
   * Contextual information about each game, such as the type of umpire (human or robot), the venue, and the date.