**SCTPDSAI-M3-DS2-Coaching-MoneyBall Analytics**

**Submit Prediction**

**Overview**

Welcome to the MLB Season Wins Prediction Challenge! In this exercise, you'll predict the number of games a Major League Baseball team will win in a season based on their performance statistics. This is a regression problem where you'll use historical MLB team data to build a model that can accurately forecast seasonal win totals.

**Start**

a day ago

**Close**

12 days to go

**Description**

**The Data**

The dataset contains comprehensive team statistics from the 2016 Lahman Baseball Database, including:

Batting statistics: Runs, hits, home runs, strikeouts, etc.  
Pitching statistics: Earned run average, saves, strikeouts, etc.  
Fielding statistics: Errors, double plays, fielding percentage  
Team information: Year, team name, franchise ID  
Game outcomes: Wins, losses, championships

Refer to the Data tab for more details on the data description

**Evaluation**

**Evaluation**

Submissions are evaluated using Mean Absolute Error (MAE), which measures the average absolute difference between the predicted wins and actual wins. Lower scores indicate better performance, with a perfect score being 0. The MAE is calculated as the mean of the absolute values of the differences between predicted and actual wins across all teams.

**Submission File**

For each team-season in the test set, you must predict the number of wins (W) as an integer. The file should contain a header and have the following format:

ID,W

2,85

5,92

6,78

etc.

**Dataset Description**

**MLB Season Wins Prediction: Data Description**

**Files**

* **train.csv** - Historical MLB team seasons with all features and the target variable (W) - 1698 samples
* **test.csv** - MLB team seasons for prediction submission (without the W column) - 567 samples
* **sample\_submission.csv** - A sample submission file in the correct format

**Analysis Support Fields**

These fields are included to support data exploration and analysis but use caution when including them as features in model training.

* yearID - The season year (integer, e.g., 2019)
* year\_label - Categorical label for historical baseball eras (1-8):
  + 1: Pre-1920 (Dead-ball era)
  + 2: 1920-1941 (Live-ball era)
  + 3: 1942-1945 (WWII era)
  + 4: 1946-1962 (Post-war era)
  + 5: 1963-1976 (Pitcher's era)
  + 6: 1977-1992 (Free agency era)
  + 7: 1993-2009 (Steroid era)
  + 8: 2010-present (Post-steroid/analytics era)
* decade\_label - The starting year of the decade (e.g., 2010 for 2010-2019)
* win\_bins - Categorical binning of win totals:
  + 0: < 50 wins
  + 1: 50-69 wins
  + 2: 70-89 wins
  + 3: 90-109 wins
  + 4: 110+ wins

**Note for Model Training:** The above fields are only available in the training data but are not included in test.csv as they would cause data leakage or are derived from the target variable

**Target Variable**

* W - Number of wins in the season (integer, range ~40-120)

**Features**

**Basic Statistics**

* G - Games played (integer)

**Batting Statistics**

* R - Runs scored (integer)
* AB - At bats (integer)
* H - Hits (integer)
* 2B - Doubles (integer)
* 3B - Triples (integer)
* HR - Home runs (integer)
* BB - Walks (integer)
* SO - Strikeouts (integer)
* SB - Stolen bases (integer)

**Pitching/Defense Statistics**

* RA - Runs allowed (integer)
* ER - Earned runs allowed (integer)
* ERA - Earned run average (float)
* CG - Complete games (integer)
* SHO - Shutouts (integer)
* SV - Saves (integer)
* IPouts - Outs pitched (innings pitched × 3) (integer)
* HA - Hits allowed (integer)
* HRA - Home runs allowed (integer)
* BBA - Walks allowed (integer)
* SOA - Strikeouts by pitchers (integer)
* E - Errors (integer)
* DP - Double plays (integer)
* FP - Fielding percentage (float)

**Derived Features**

* mlb\_rpg - MLB average runs per game for the season (float)

**Era Indicators**

Binary indicators for different historical MLB eras:

* era\_1 - Pre-1920: Dead-ball era
* era\_2 - 1920-1941: Live-ball era
* era\_3 - 1942-1945: WWII era
* era\_4 - 1946-1962: Post-war era
* era\_5 - 1963-1976: Pitcher's era
* era\_6 - 1977-1992: Free agency era
* era\_7 - 1993-2009: Steroid era
* era\_8 - 2010-present: Post-steroid/analytics era

**Decade Indicators**

Binary indicators for each decade (1910s-2010s):

* decade\_1910 - 1910s
* decade\_1920 - 1920s
* decade\_1930 - 1930s
* decade\_1940 - 1940s
* decade\_1950 - 1950s
* decade\_1960 - 1960s
* decade\_1970 - 1970s
* decade\_1980 - 1980s
* decade\_1990 - 1990s
* decade\_2000 - 2000s
* decade\_2010 - 2010s

**Dataset Notes**

1. All teams in the dataset played at least 150 games in their respective seasons
2. The data spans from 1901 to 2016
3. Some statistics may be missing for older seasons due to incomplete record-keeping
4. The target variable (W) typically ranges from about 40 to 120 wins per season, with most teams winning between 65 and 95 games

**Competition Objective**

Predict the number of games a team will win in a season based on their performance statistics. Your model should be able to generalize across different eras of baseball.

**Evaluation Metric**

Submissions are evaluated using **Mean Absolute Error (MAE)**, which measures the average absolute difference between the predicted wins and actual wins. The lower the MAE, the better the model's performance.

**Submission Format**

Your submission file should contain a header and have the following format:

ID,W

0,85

1,92

2,78

etc.

Where:

* ID is the row index from the test set (0-based)
* W is your predicted number of wins (must be an integer)