Moneyball Project Report

Project Report: Predicting Team Wins (Moneyball Data)

Team Members

- Fomba Kassoh
- Souleymane Doumbia
- Warner Alexis
- Saloua Daouki
- Lewris Mota Sanchez

1. ABSTRACT

This report presents an analysis and prediction of baseball team performance, focusing on the number of wins (TARGET_WINS) using the Moneyball dataset. The study employs various statistical models to explore the relationships between key performance variables such as batting, pitching, baserunning, and fielding statistics.

Key findings from the regression models demonstrate significant coefficients, revealing important insights. The **Baseline Model**, which used original variables, shows that **batting hits (TEAM_BATTING_H)** positively influences the number of wins, while **fielding errors (TEAM_FIELDING_E)** and **double plays (TEAM_FIELDING_DP)** have a negative impact, reflecting the importance of reducing defensive mistakes. **Walks (TEAM_BATTING_BB)** and **stolen bases (TEAM_BASERUN_SB)** also contribute positively to team success, as expected, indicating that offensive opportunities like walks and successful base running are crucial for scoring more runs and securing wins.

In the **Feature Selection Model**, multicollinearity was addressed, with variables selected based on their contribution to the model. **Interaction Terms Model** explored how combining certain variables could reveal more nuanced relationships, such as the interaction between **batting hits and doubles**, which further emphasizes the effect of collective offensive performance on wins.

Ultimately, the **Baseline Model** was selected for its balance of simplicity, interpretability, and performance, with an adjusted R-squared of 0.31. This model's coefficients align well with intuition, where stronger offensive performance and fewer defensive errors predict more wins. The model was evaluated on new data, confirming its predictive power.

This analysis provides a strong foundation for understanding team dynamics and improving decision-making in player management and game strategy.

2. DATA EXPLORATION

The Moneyball training dataset contains 2,276 observations and 17 variables, representing various team performance statistics from past seasons. These variables include measures of batting, pitching, and fielding, with the target variable being **TARGET_WINS**—the number of games won by the team.

Key Variables:

The Moneyball dataset used for this project contains various variables that capture different aspects of a baseball team's performance, such as batting, pitching, baserunning, and fielding statistics. The dataset has 2276 observations and includes the following key variables:

- **TARGET_WINS:** The target variable representing the number of wins a team achieves.
- **TEAM_BATTING_H, TEAM_BATTING_2B, TEAM_BATTING_3B, TEAM_BATTING_HR, TEAM_BATTING_BB, TEAM_BATTING_SO:** Batting statistics including hits, doubles, triples, home runs, walks, and strikeouts.
- **TEAM_BASERUN_SB, TEAM_BASERUN_CS:** Stolen bases and caught stealing statistics.
- **TEAM_PITCHING_H, TEAM_PITCHING_HR, TEAM_PITCHING_BB, TEAM_PITCHING_SO:** Pitching statistics including hits allowed, home runs allowed, walks, and strikeouts by pitchers.
- **TEAM_FIELDING_E, TEAM_FIELDING_DP:** Fielding statistics, including errors and double plays.

Summary Statistics:

Here are key summary statistics (mean, median, and standard deviation) for a few variables:

Variable	Mean	Median	Standard Deviation
TARGET_WINS	80.79	82	15.75
TEAM_BATTING_H	1469.27	1454	144.59
TEAM_BATTING_2B	241.25	238	46.8
TEAM_BATTING_3B	55.25	47	27.94
TEAM_BATTING_HR	99.61	102	60.55
TEAM_BATTING_BB	501.56	512	122.67

Correlation to the Target Variable

The **TARGET_WINS** is positively correlated with variables like **TEAM_BATTING_H** and **TEAM_BATTING_BB** (hits and walks), indicating that teams with more hits and walks tend to win more games. On the other hand, variables like **TEAM_FIELDING_E** (errors) have a negative correlation with **TARGET_WINS**, as teams that make more errors tend to win fewer games.

Missing Data Analysis

Several variables have missing values, as shown in the table below. Recommendations were followed to impute or drop variables based on the percentage of missing values.

Variable	Percent_Missing	Imputation_Recommendation
TEAM_BATTING_SO	4.48%	Mean/Median Imputation
TEAM_BASERUN_SB	5.75%	KNN or Multiple Imputation
TEAM_BASERUN_CS	33.92%	Multiple Imputation or Predictive Modeling
TEAM_BATTING_HBP	91.61%	Consider Dropping
TEAM_PITCHING_SO	4.48%	Mean/Median Imputation

3. DATA PREPARATION

Several data transformations and imputations were made to prepare the data for modeling:

- **Imputing Missing Values**: For variables with low missing percentages like **TEAM_BATTING_SO** and **TEAM_PITCHING_SO**, we used **median imputation**. For variables with higher missing percentages, such as **TEAM_BASERUN_SB**, a median imputation was applied for simplicity, though more advanced methods like KNN could be used in future improvements.
- **Dropping Variables**: **TEAM_BATTING_HBP** was dropped from the analysis due to its high missing rate (91.61%).
- **Mathematical Transformations**: Logarithmic and square root transformations were applied to some variables to handle skewness and improve the model's performance. For example:
    - **TEAM_BATTING_H** and **TEAM_BATTING_HR** were log-transformed to stabilize variance.
- **Scaling the Data**: Variables were scaled using **z-score standardization**, subtracting the mean and dividing by the standard deviation. This ensures that all variables are on the same scale, especially important when variables like **TEAM_BATTING_HR** have a large variance compared to others like **TEAM_FIELDING_E**.
- **Creating Composite Variables**: Composite variables were created to capture broader aspects of team performance:
    - **Composite_Batting**: Combined batting performance metrics like **hits**, **walks**, and **home runs**.
    - **Composite_Pitching**: Combined key pitching metrics like **strikeouts**, **walks**, and **home runs allowed**.

4. BUILDING MODELS

Multiple linear regression models were built to predict **TARGET_WINS** using different sets of variables and transformations. The following models were developed:

Model 1: Baseline Model

- **Variables**: All variables were included except those with high collinearity (e.g., TEAM_PITCHING_SO, TEAM_BATTING_HR).
- **Key Coefficients**:
  - **Positive Effects**:
    - **TEAM_BATTING_H:** Teams with more hits are expected to win more games (+0.52).
    - **TEAM_BASERUN_SB:** Steals positively influence wins (+0.19).
  - **Negative Effects**:
    - **TEAM_FIELDING_E:** More errors lead to fewer wins (-0.26).
    - Explanation: This aligns with expectations in baseball — errors directly contribute to losses.
- **Performance**:
  - **Adjusted R-squared**: 0.31.
  - The model explains 31% of the variance in TARGET_WINS. It is the best-performing model in terms of fit, and residuals show reasonable behavior with some skewness.

Model 2: Composite Variables Model

- **Composite Features**:
  - Composite_Batting: Combined variables like TEAM_BATTING_H, TEAM_BATTING_2B, and TEAM_BATTING_HR into a single score to reduce multicollinearity.
  - Composite_Pitching: Combined pitching metrics (TEAM_PITCHING_H, TEAM_PITCHING_BB).
- **Key Coefficients**:
  - Composite_Batting: Positive impact on wins (+0.49), indicating that overall batting performance strongly influences wins.
  - Composite_Pitching: Negative impact (-0.08), which might suggest that higher pitching stats are more reflective of defensive metrics, which do not directly lead to more wins.
- **Performance**:
  - **Adjusted R-squared:** 0.18.
  - This model simplifies the data but sacrifices some predictive power. Although composite features can reduce noise, they may oversimplify relationships.

Model 3: Interaction Terms Model

- **Interaction Terms**:
  - Introduced interactions like TEAM_BATTING_H * TEAM_BATTING_2B to capture non-linear relationships between variables.
  - **Interpretation:** Interaction terms allow us to explore whether the combined effect of variables exceeds the sum of their individual contributions.

- **Key Coefficients**:
    - **Interaction_Batting:** Positive effect (+0.45), indicating that batting variables work together to predict wins.
    - **TEAM_BASERUN_SB:** Steals continue to show a significant positive relationship with wins (+0.26).

- **Performance**:
    - **Adjusted R-squared**: 0.21.
    - The interaction terms slightly improved the predictive power of the model. The residuals suggest some improvement over the composite model, but it's still not as strong as the baseline model.

Model 4: Feature Selection Model

- **Feature Selection**: This model used a manual feature selection process to remove variables with high collinearity, such as TEAM_BATTING_SO and TEAM_PITCHING_SO, and retained only variables that had significant and meaningful contributions.
- **Key Coefficients**:
    - **TEAM_BATTING_3B** (+0.20) and **TEAM_BASERUN_SB** (+0.22) both showed strong positive effects.
    - **TEAM_FIELDING_E** (-0.25) continues to show a strong negative impact, reinforcing the importance of good fielding in winning games.
- **Performance**:
    - **Adjusted R-squared**: 0.30.
    - This model was competitive with the baseline model but more parsimonious, making it an attractive option due to reduced complexity and similar predictive power.

Model 5: Polynomial Model

- **Polynomial Terms**: Applied second-degree polynomial terms for TEAM_BATTING_H and TEAM_PITCHING_H to capture non-linear relationships.
- **Key Coefficients**:
    - **TEAM_BATTING_H** (2.38 for the first degree, 0.44 for the second degree) indicates that performance improvement accelerates with higher batting metrics.
    - **TEAM_FIELDING_DP (-0.14) and TEAM_FIELDING_E** (-0.19) both remain strong negative predictors.

- **Performance**:
    - **Adjusted R-squared:** 0.27.
    - The polynomial model did capture non-linear relationships but didn't outperform the simpler baseline model.

5. SELECTING THE BEST MODEL

The selection of the best model was based on several criteria, including:

- **R-squared** and **Adjusted R-squared** values, which indicate how well the model explains the variance in the target variable.
- **AIC** and **BIC** values, which balance model fit and complexity.
- **Coefficients' interpretability**, ensuring that the signs and magnitudes make intuitive sense.

Model Comparison:

Model	R-squared	Adjusted R-squared	AIC	BIC
Baseline Model	0.312	0.308	1567.22	1589.34
Feature Selection	0.301	0.297	1574.31	1592.41
Interaction Terms	0.208	0.206	1601.21	1623.98
Composite Variables	0.182	0.181	1620.01	1644.98
Polynomial Model	0.267	0.265	1598.16	1620.32

Best Model Selection:

The **Baseline Model** was selected as the best-performing model based on its high R-squared value (0.312), low AIC, and BIC, along with good interpretability of the coefficients.

6. PREDICTION AND EVALUATION

After selecting the Baseline Model, we evaluated it on the test dataset and made predictions.

Evaluation Metrics:

Metric	Value
Mean Squared Error (MSE)	0.0819
R-squared (R^2)	0.312
Adjusted R-squared	0.308
F-statistic	85.55

The MSE is low, indicating that the model predictions are close to the actual values. The R-squared value suggests that the model explains 31.2% of the variance in **TARGET_WINS**.

Prediction Results:

Here are the actual and predicted **TARGET_WINS** values for a few teams in the evaluation dataset:

Team Index	Actual Wins	Baseline Model	Feature Selection Model	Interaction Terms Model	Polynomial Model	Composite Variables Model
1	78	76.18	74.25	74.25	73.15	76.18
2	88	82.35	82.35	72.44	75.88	73.44
3	98	99.71	78.92	78.92	95.87	82.35
4	92	93.39	104.38	104.38	100.23	99.71
5	84	88.94	84.02	84.02	89.95	80.63
6	75	72.15	81.84	77.42	78.12	88.94
7	85	83.86	82.44	82.44	81.55	84.94
8	86	89.86	79.9	83.86	85.34	83.39
9	76	85.88	77.42	80.63	79.46	83.86
10	80	78.13	80.63	79.91	81.87	82.35
11	70	71.87	71.87	72.44	70.15	71.87

These predictions are reasonably close to the actual values, indicating that the model is capturing the key factors that determine team performance.

Residual Analysis

Residual analysis was also performed to ensure that the model meets the assumptions of linear regression. The residual plots indicated:

- **Linearity:** The residuals were fairly evenly scattered around the zero line, indicating a linear relationship between the predictors and the target variable.
- **Normality:** The residuals appeared normally distributed when plotted, supporting the assumption of normality.
- **Homoscedasticity:** The residuals showed constant variance across the predicted values.

These residual checks confirm that the model fits the data well and does not violate any major assumptions of linear regression.

7. CONCLUSION

After building multiple models and performing thorough evaluations, the **Baseline Model** was selected as the best model for predicting team wins. The model performed well based on both training and evaluation datasets. Its predictions are consistent with actual results, and its statistical metrics (R^2, MSE, F-statistic) confirm that the model is a good fit for this problem.

The model can be used by the team to predict future performance based on key metrics such as **batting hits**, **walks**, **home runs**, **fielding errors**, and **pitching statistics**. With further refinement and data collection, the accuracy of the model can be further improved.