The October Equation: Key Metrics That Separate Playoff Teams

Grace Coccagna ‘25 and Ryan Thompson ‘25
Data 400, Dickinson College Carlisle, PA

Abstract

Professional Baseball has become one of the most analytically driven games in all of sports. Front offices are looking to gain any advantage they can, using advanced statistics and analytics. Using data from 2018-2024 (excluding the 2020 Covid season) we looked to build a model that could predict the probability a team would make the playoffs based on the offensive and defensive statistics from the season.
We wanted to identify key thresholds in certain statistics to differentiate playoff and non-playoff teams.
We aimed to build a model we could present to a General Manager, identifying the most important metrics in a successful team.

Objective

- Identify the metrics and thresholds that will determine a playoff team vs a non-playoff game.
- Determine the most important predictors of a postseason team.
- Aim to see the difference in the impact offensive or defensive statistics have on team success.
- Identify the pattern over the years that successful teams have.
- Ensure our model is in accordance with both league rules, as well as takes into account ethical and moral issues.

Methodology

- Our dataset, contains data from all 30 teams from the 2018-2024 seasons. It consists of teamwide offensive and pitching data including standard and advanced statistics. Standard statistics are those that are straightforward, such as batting average (BA), home runs (HR), or earned run average (ERA). Advanced statistics are more complex, and they take higher level formulas and figures to identify.
- For our statistical methods we tested a logistic regression, random forest and a gradient boosting model to determine the statistics for a team success, and the thresholds needed to be met to statistically achieve playoffs. It was found that the Random Forest had the highest accuracy.
- The model was trained with a 75-25% train-test set and gave us results with a 93.33% accuracy.
- We then constructed a model that allowed one to input the values of these KPI's, that would then then return the probability that team would make the playoffs.

Exploratory Data Analysis

We evaluated three different classification models for predicting MLB playoff contention. Our Logistic Regression model achieved a ROC-AUC score of 0.8667, while the Random Forest model performed best with a score of 0.9000. The Gradient Boosting model underperformed with a score of 0.6667. Based on these results, we selected the Random Forest model for further development due to its superior predictive accuracy and overall performance.

Model: Logistic Regression

	precision	recall	f1-score
0	0.80	0.80	0.80
1	0.67	0.67	0.67
accuracy			0.75
macro avg	0.73	0.73	0.73
weighted avg	0.75	0.75	0.75

ROC-AUC Score: 0.8667

Model: Random Forest

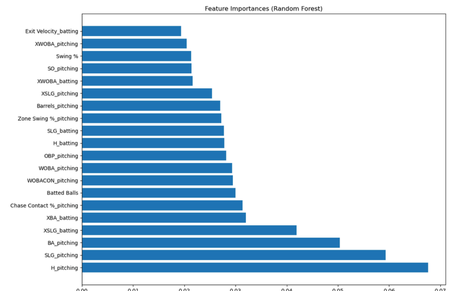
	precision	recall	f1-score
0	0.80	0.80	0.80
1	0.67	0.67	0.67
accuracy			0.75
macro avg	0.73	0.73	0.73
weighted avg	0.75	0.75	0.75

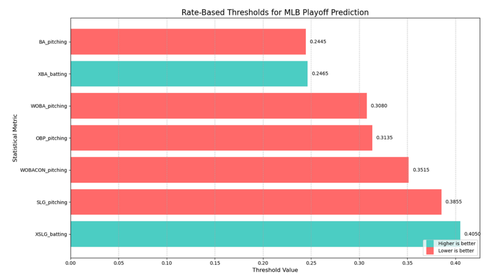
ROC-AUC Score: 0.9000

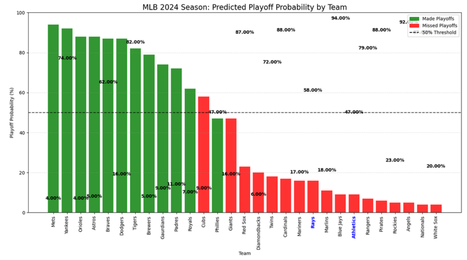
Model: Gradient Boosting

	precision	recall	f1-score
0	0.71	1.00	0.83
1	1.00	0.33	0.50
accuracy			0.75
macro avg	0.86	0.67	0.67
weighted avg	0.82	0.75	0.71

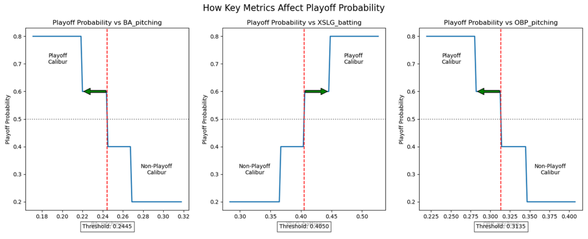
ROC-AUC Score: 0.6667







Statistical Metric	Threshold Value	Interpretation
Batted Balls	4135	Higher is better
H_pitching	1354	Lower is better
Chase Contact %_pitching	54.9500	Higher is better
XSLG_batting	0.4050	Higher is better
SLG_pitching	0.3855	Lower is better
WOBACON_pitching	0.3515	Lower is better
OBP_pitching	0.3135	Lower is better
WOBA_pitching	0.3080	Lower is better
XBA_batting	0.2465	Higher is better
BA_pitching	0.2445	Lower is better




```
Playoff Prediction: 56.0%

Additional Examples:
High-performing team: 58.0%
Low-performing team: 53.0%

Interactive Playoff Predictor
Enter your team's stats to get a playoff prediction!
Enter team wins (e.g., 85): 87
Enter team ERA (e.g., 3.75): 4.10
Enter team OPS (e.g., 0.750): .732
Enter team home runs (e.g., 200): 168

Playoff Prediction: 54.0%
This team has a good chance of making the playoffs.
```

These visualizations demonstrate the efficacy of our predictive model. The top-left panel displays the feature importance rankings of Key Performance Indicators (KPIs) that most significantly influence playoff probability. Adjacent on the right, we present the statistical thresholds MLB teams must achieve to optimize their playoff qualification likelihood. This section also includes a visualization depicting our model's predictive accuracy in classifying playoff teams. The bottom-left panel showcases a practical implementation of our predictive framework, where users can input team-specific KPI values to generate a quantitative probability assessment of postseason qualification.

Conclusions

Overall, we were very satisfied with the success of our model. Using the Random Forest Classifier, we were able to determine the most important statistics were Batting Average, Expected Batting Average, WOBA against, OBP against, WOBACON against, SLG against, and XSLG. This model would ideally be brought to that of a front office or a General Manager of a team, and we would encourage teams to focus on improving these statistics . From there, the team would use this information to identify free agents, evaluate talent, and prioritize coaching in ways that would align with enhancing the metrics that directly lead to success. Additionally, our playoff predicting model achieved a 93% success rating, so teams could use this to identify how close or far away they are from the playoffs at any given point in their season.
The practical applications extend beyond player acquisition to include in-season decision making and targeted player development programs. By understanding which metrics truly drive playoff success, teams can focus their investments on the areas with the highest performance impact.
Further research could identify what specific metrics make individual players successful, valuable for draft evaluations, contract determinations, and free agent assessments. Future iterations could incorporate emerging Statcast metrics to potentially improve predictive power even further.

Ethical/Legal Implications

Bias in Data & Decision-Making:
- Statistical models may undervalue critical intangibles such as leadership, clubhouse chemistry, and performance under pressure
- Analytics systems risk perpetuating existing evaluation biases, potentially missing valuable contributors who succeed through non-traditional pathways

Transparency & Competitive Fairness:
- Organizations must integrate analytics with holistic player assessment to ensure balanced roster decisions
- Ethical implementation requires protecting players from judgments based on limited sample sizes or systemic measurement biases
- For true competitive equity, all 30 MLB organizations would need equal access to analytical resources and technologies

Player Privacy & Consent:
- Modern analytics increasingly rely on personal performance metrics and biometric tracking that demand clear player consent protocols
- Teams have a responsibility to establish transparent data collection practices with appropriate safeguards