Baltimore Orioles Cedric Mullins (31) hit right-handed before making the change, only to hit left-handed. (AP Photo/Kathy Willens)

SOURCE: AP Photo/Kathy Willens

The Switch-Hitting Model

Building a model to predict whether switch hitters should only hit from one side.

*By: Greg Halperin & Patrick Holsey*

*MSDS456 – Sports Performance Analytics | August 22, 2024*

# Introduction

Matchup splits have always been an important topic when crafting a baseball lineup. For example, when facing a left-handed pitcher, the manager might prioritize right-handed hitters to take advantage of these splits. As detailed in this report, splits refer to the differences in a batter's performance against right-handed and left-handed pitchers. This has become increasingly important as analytics-inclined organizations use splits to construct their lineups to gain an advantage. Managers may choose a lineup of predominantly left-handed, predominantly right-handed, or alternate left and right-handed hitters depending on matchup splits. This creates value for hitters who do not have a significant difference between their matchup splits, particularly switch-hitters, as they can hit anywhere in the lineup and play every day without worrying about the handedness of the opposing pitcher. Most of the time, switch-hitters hit from the opposing side of the pitcher to gain an advantage.  This allows managers to feel comfortable playing them against any pitcher.

# Background

            In contrast, there have been instances in MLB history where players perform poorly while switch-hitting and improve drastically when they focus on their stronger side. An example of this is Cedric Mullins. Mullins, a center fielder for the Baltimore Orioles, came up to the Majors in 2018 as a switch-hitter and generated only six hits in his first 64 at-bats, resulting in a demotion back to the Minors in 2019. In the Minors, the Orioles had Mullins focus on hitting left-handed only since his performance against left-handed pitchers was already abysmal. After making this change in 2021, Mullins had 17 hits in his first nine games and began performing better against left-handed pitchers.[[1]](#footnote-1)

            For switch-hitters with poor splits from one side, it may make sense to stop switch-hitting, like Mullins. As stated above, the main benefit of having a switch-hitter on a team is their ability to hit against both types of pitchers – left and right-handed. If they fail to do so, are they providing much value? The other major factor is the practice time required to maintain two different swings. If a switch-hitter is underperforming, it may be more time-efficient for a player’s development to focus on one side.

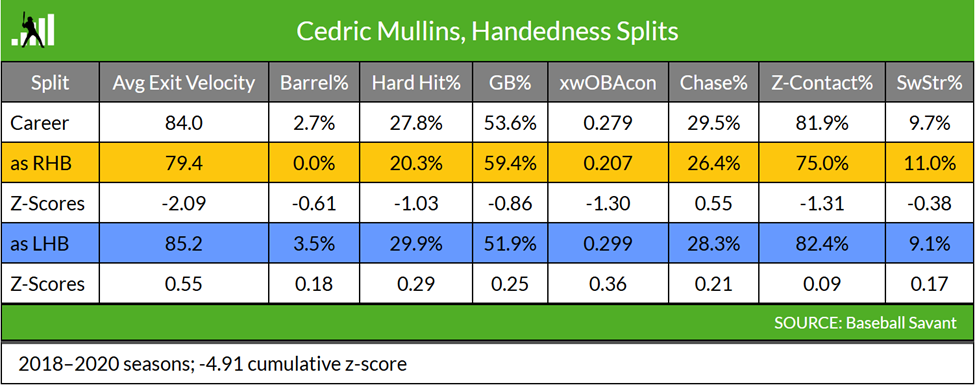
            This idea has been explored before. One resource used as a foundation for this study was Jake Mailhot’s article on *Fangraphs*, “Finding Switch-Hitters Who Should Stop Switch Hitting”[[2]](#footnote-2). In this article, Mailhot breaks down Cedric Mullins's splits and uses a linear regression model to determine the difference in z-scores between his right and left-handed hitting metrics. His results for Cedric Mullins are below:

Figure 1Mailhot model for Mullins

Based on this data, Mailhot analyzed five other hitters he believed had large enough splits to justify making changes. This led to the question: what type of impact on performance could be expected of these hitters if they made the change?

# Problem Statement

            Switch-hitting may not increase a player’s value in all cases. Organizations are always looking for ways to improve their teams. While having a switch-hitter in the lineup might be appealing for matchups at first glance, a team might get more value from players hitting exclusively from their stronger side, like in the Cedric Mullins scenario.

            Previous studies focus on the variability of a player’s splits and showcase that difference using a model. This information helps create lineups and consider how each player interacts with different pitcher handedness. It could also be used to determine specific training methods targeted for each of their swings. However, a missing component of these studies is predicting their performance if they stopped switch-hitting. This is important when looking at same-sided matchups, such as left-handed hitting against left-handed pitching or right-handed hitting against left-handed pitching. The sample size for hitter performance in these situations is minimal because they tend to hit from the opposite side. Therefore, this data does not provide an adequate sample size to be significant.

            This study intends to examine the relationships between switch-hitter splits and how they relate to left-handed-only and right-handed-only hitters’ splits. If a switch-hitter is stronger from the left side, he will be compared to other left-handed hitters, and vice versa for the right-hand side.

Hitters should not necessarily stop switch-hitting due to this model's results. However, they should evaluate the possibility and determine how their performance might be affected.

# Resources

This report used the following resources to extract data and build a foundation for this analysis. The complete list of citations is listed in the References section.

Malihot, J. *Finding Switch-Hitters Who Should Stop Switch-Hitting*, Fangraphs.com Blogs, 2022

Fangraphs.com served as the primary source for extracting data for this report and providing a starting point for previous research. In Malihot’s article, he notes the success of Cedric Mullins of the Baltimore Orioles after he decided to give up swinging from the right-hand side of the plate in 2021 and runs a regression model to estimate the performance of 25 switch-hitters. Focusing on five hitters with significant splits, Malihot then calculates z-scores based on league standard deviation for eight metrics, using the sum of z-scores to make his recommendations on performance from each side. In his conclusion, Malihot hesitates to recommend any changes for established players in the Major Leagues.

Lindbergh, B. *Overthinking It: Searching for Switch-Hitters Who Shouldn’t Switch-Hit,* BaseballProspectus.com, 2014.

Lindbergh uses a regression model based on True Average (TAv), comparing observed switch-hitters' observed platoon splits to his estimated split as a non-switch hitter. Lindbergh includes a “fudge factor” of one standard deviation, assuming a switch-hitter would succeed less than an average player hitting from only one side. By requiring 600 plate appearances vs left-handed pitchers, Lindbergh reduces his sample size to only five switch-hitters (one recently retired) and does not find enough statistical significance to recommend any player gives up switch-hitting.

Winston, W., Nestler, S., & Pelechrinis, K. *Mathletics*, Chapter 12: The Platoon Effect, 2022

Winston et al.’s *Mathletics* reviews managerial decisions to face a left- or right-handed pitcher based on the hitter’s platoon splits. They define platoon splits as managers starting left-handed hitters more often against right-handed pitchers and vice versa. This study guided research for this report as it found that only 35% of hitters against left-handed pitchers are left-handed and that 41% of hitters against right-handed pitchers are right-handed.

# Data Acquisition

The primary data source for this study was Fangraphs.com Major League Leaders Splits Statistics[[3]](#footnote-3). This site offers filters by year, position, and hitter or pitcher handedness, including switch-hitters. Data acquisition started by determining the minimum number of plate appearances against left-handed pitching required for a statistically significant sample. Based on *The Book: Playing the Percentages in Baseball* [[4]](#footnote-4), the recommended number of plate appearances for a switch-hitter’s platoon split to stabilize against left-handed pitching is 600 appearances. Using 600 plate appearances for switch-hitters against left-handed pitchers as a baseline, the next step in the data extraction process was filtering Fangraphs data for left-handed-only hitters against left-handed pitching. Leveraging *Mathletics* as a guide that only 35% of hitters against left-handed pitchers are left-handed themselves, it was evident that 13 seasons of data would be required for a sample size of 100 left-handed hitters to reach 600 plate appearances against left-handed pitchers. Since most of the current switch-hitters have not played that long, the minimum number of plate appearances against left-handed pitchers was reduced to 200, providing the time frame from the 2018 through 2023 seasons. Keeping 2018 through 2023 as the constant time frame, the following hitter data was extracted:

* A sample of 19 switch-hitters with 600 plate appearances against both left-handed and right-handed pitchers
* A sample of 123 left-handed hitters with a minimum of 200 plate appearances against left-handed pitchers and a minimum of 500 plate appearances against right-handed pitchers
* A sample of 230 right-handed hitters with a minimum of 250 plate appearances against left-handed pitchers and a minimum of 500 plate appearances against right-handed pitchers

Like Malihot’s study, Cedric Mullins’s switch-hitting plate appearances from 2018 to 2020 and left-handed-only plate appearances from 2021 to 2023 were extracted to compare to model predictions.

# Data Preparation

Fangraphs’ custom report feature was used to generate data sets. It was determined the following metrics provided a holistic view of a hitter’s performance: seasonal performance metrics such as weighted on-base average (wOBA) and on-base percentage + slugging percentage (OPS), along with batted-ball metrics such as ground ball percentage (GB%), line drive percentage (LD%), hard hit percentage (Hard%), walk percentage (BB%), and strikeout percentage (K%). There were limitations for the data available; for example, newer metrics used for Win Probabilities (ex., Wins Above Replacement, in Probability Added) were unavailable for custom reporting. The choice of wOBA and OPS was preferred over the standard metric of batting average as they include weights for the impact of a double, triple, or home run, whereas batting average treats all hits the same. Both wOBA and OPS also include walks and hit-by-pitches. With access to better data sets, the model will be more robust. However, these were the accessible metrics.

After importing the datasets, the next step was combining the left- and right-handed splits into data sets representing switch-, right-handed-only, and left-handed-only hitters. These data sets were filtered and adjusted to have the same variable names and only included the specific variables needed: name, wOBA, OPS, GB%, LD%, Hard%, BB%, and K%. Similar filtering, renaming, and cleaning processes were used to create the Cedric Mullins switch-hitting pre-2021 dataset.

# Exploratory Data Analysis

Scatterplots were created for current switch-hitter performance against left-handed and right-handed pitchers compared with the league splits averages for each metric. The purpose was to understand the switch-hitter sample better and identify players with significant discrepancies. The OPS scatterplot is shown below.

The vertical line identifies the right-handed hitter sample’s OPS against right-handed pitchers, and the horizontal line identifies the left-handed hitter sample’s OPS against left-handed pitchers. A main takeaway from the plot is that no switch-hitters in the sample have a worse OPS against left-handed pitching. The players identified as blue dots could be expected to appear in the model as players who might perform better if they decided to stop switch-hitting, as they are below the sample average against right-handed pitchers. The remaining scatterplot results for each metric are listed in Appendix A.A graph with numbers and lines

Description automatically generated

# Modeling

The approach to predicting switch-hitter performance began with creating two random forest algorithms. A random forest is a machine learning algorithm that combines multiple decision trees' outputs to reach a single result. This approach reduces the risk of overfitting, handles regression tasks accurately, and provides an easy path to evaluating variable importance[[5]](#footnote-5). Random Forest modeling also provides a root means square error (RMSE) that identifies the difference between the model’s predicted values and the actual values used to train it.

To create the left-handed random forest model, the left-handed hitter sample (LHH) was separated so that 80% of data was included in the training set, and the remaining 20% was reserved for testing the model. The LHH sample was used to predict how a switch-hitter would perform, hitting only left-handed. As previously mentioned, the metrics used in this model were OPS, wOBA, ground ball percentage (GB%), line drive percentage (LD%), hard hit percentage (Hard%), walk percentage (BB%), and strikeout percentage (K%). This model predicts a left-handed hitter’s OPS against left-handed pitchers using their metrics against right-handed pitchers. This is consistent with switch-hitters hitting left-handed against right-handed pitchers. The formula is included below:

This process was replicated to create a right-handed hitter model, using right-handed hitter metrics against left-handed pitchers to predict their performance against right-handed pitchers. The switch-hitter sample was run through both models to determine their projected OPSs.

# Results

A screenshot of a graph

Description automatically generated The predicted OPS results were compared to the switch-hitters’ actual OPS values. This determined whether the model predicted a performance improvement if they should change to only hitting from one side. The Root Mean Square Errors (RMSEs) of 0.005 for the left-handed model and 0.004 for the right-handed model were used as the margins of error. Any predicted OPS results within the RMSE margin of the actual values were not significant enough to make an informed decision. The players above the RMSE margin were predicted to improve by only hitting from that side, and those below this margin were predicted to perform worse. The prediction table is shown below:

The results are interpreted as follows:

LHH RMSE = 0.005, RHH RMSE = 0.004

* If a player has two green cells, they should continue switch-hitting as they are predicted to perform worse if they change to hitting from one side only.
* If a player has a red and a green cell, the model predicts they will improve if they hit only from the side that contains a red cell. For example, Tommy Edman is predicted to perform better if he hits exclusively right-handed.
* If a player has a yellow and green cell, their predictions for the yellow side fall within the margin of error (RMSE). In this case, the prediction is not statistically significant enough to determine whether a change is warranted.

Based on these results, the model predicts that five hitters (Santana, Albies, Grossman, Grandal, and Edman) would perform better if they only hit right-handed, and only one (Polanco) would perform better if he only hit left-handed. Another two (Profar and Candelario) were within the margin of error for right-handed hitters, and one (Happ) was within the margin of error for left-handed hitters. An important takeaway is that left-handed hitter splits tend to be more substantial. Therefore, predicting that players will perform better if they hit only left-handed is more challenging.

To test the model's viability, Cedric Mullins’ switch-hitting data before 2021 was run through the model and resulted in the following table:A screenshot of a computer

Description automatically generated

Based on these results, the model predicts that Mullins would have an OPS of 0.701 against left-handed pitching if he only hit left-handed. Mullins’ actual OPS against left-handed-pitching from 2021 to 2023 was 0.697, within 0.004 of the model’s prediction.

# Conclusion

Establishing a prediction model based on a random forest selection process provides greater versatility than previous regression models used by Malihot or Lindbergh. Factoring in the additional time spent by switch-hitters for training and maintaining a high level of performance from both sides of the plate, along with the prediction results, further emphasizes the need for this analysis. The model shows that switch-hitting does not always benefit a player’s performance. The recommendation is for the players identified in this report to re-evaluate hitting from their stronger side full-time. This prediction model can also be used to analyze switch-hitters of all levels, provided the relevant data. A focus for the next phase of this model is to discover Minor League switch-hitters who could benefit from these predictions, considering the limitation of the number of plate appearances required to train the model.

It is noteworthy to mention Diachronic Interpretation, which states that a hypothesis/prediction may change as new data becomes available. The potential for high variability in extracting data from a season in progress was why the 2024 season was not included in the prediction model.

The prediction model's next step would be to gain access to better, more complete data sets. A major limitation of this model was the percentage of variables explained. For the left-handed model, only 19.75% of variables were explained; for the right-handed model, only 34.32% were explained. Access to better data would provide better variables, such as Statcast metrics, which can improve the model. The goal would be to adjust the variables to account for the most significant percentage of data to maximize the model’s robustness.

# References

1. Baseball Savant, Pitch Tracking.<https://baseballsavant.mlb.com/>
2. Fangraphs, Major League Leaders (Custom Reports). <https://www.fangraphs.com/leaders/major-league?pos=all&hand=S&type=c%2C0%2C2%2C6%2C50%2C36%2C39%2C41&month=59&qual=600&startdate=&enddate=&season1=2018&season=2023&sortcol=9&sortdir=default&v_cr=202301>
3. Mailhot, Jake. *Finding Switch-Hitters Who Should Stop Switch-Hitting*. Fangraphs Blogs (2022). <https://blogs.fangraphs.com/finding-switch-hitters-who-should-stop-switch-hitting/>
4. Lindbergh, Ben. *Overthinking It: Searching for Switch-Hitters Who Shouldn’t Switch-Hit*
5. Baseball Prospectus (2014) <https://www.baseballprospectus.com/news/article/22731/overthinking-it-searching-for-switch-hitters-who-shouldnt-switch-hit/>
6. Lichtman, Michael. *The Book: Playing the Percentages in Baseball.* (2014)
7. Albert, Glickman, Swartz, and Koning. *Handbook of Statistical Methods and Analyses in Sports* - 1st edition. (2019)
8. Marchi,  Albert, and Baumer. *Analyzing Baseball Data with R* - 2nd edition. (2019)
9. Kuhn and Silge. *Tidy Modeling with R: A Framework for Modeling in the Tidyverse*. (2022)
10. Winston, Nestler, and Pelechrinis. *Mathletics: How Gamblers, Managers, and Fans Use Mathematics in Sports*, 2nd edition. (2022) Chapter 12: The Platoon Effect.
11. R-Bloggers. *Random Forest in R*. (2021). <https://www.r-bloggers.com/2021/04/random-forest-in-r/>
12. IBM.com. *What is Random Forest?* <https://www.ibm.com/topics/random-forest>
13. Gilbert, P. *Cedric Mullins makes changes resulting in hot start to season*. WBALtv.com. (2021) <https://www.wbaltv.com/article/cedric-mullins-baltimore-orioles-2021-season/36110875>

# Appendices

## Appendix A: Exploratory Data Analysis Plots

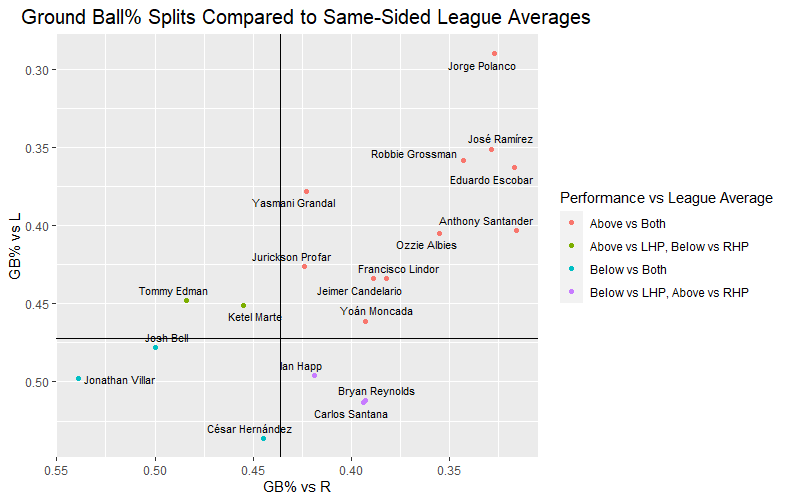
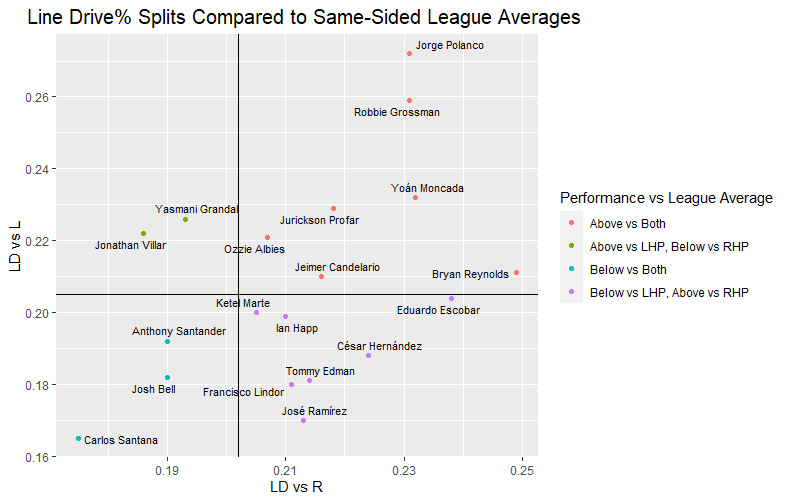
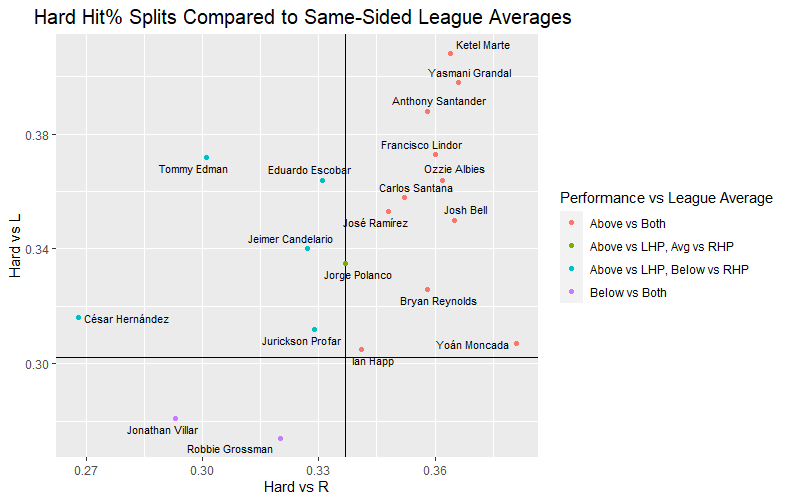
*Figure 2: wOBA Splits Compared to Same-Sided League Averages*

Figure 3: Ground Ball% Splits Compared to Same-Sided League Averages



*Figure 4: Line Drive% Splits Compared to Same-Sided League Averages*

*Figure 5: Hard Hit% Splits Compared to Same-Sided League Averages*

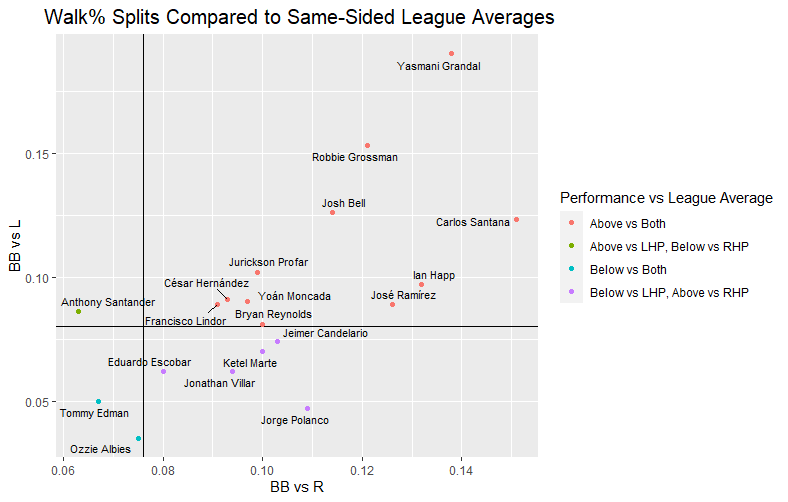
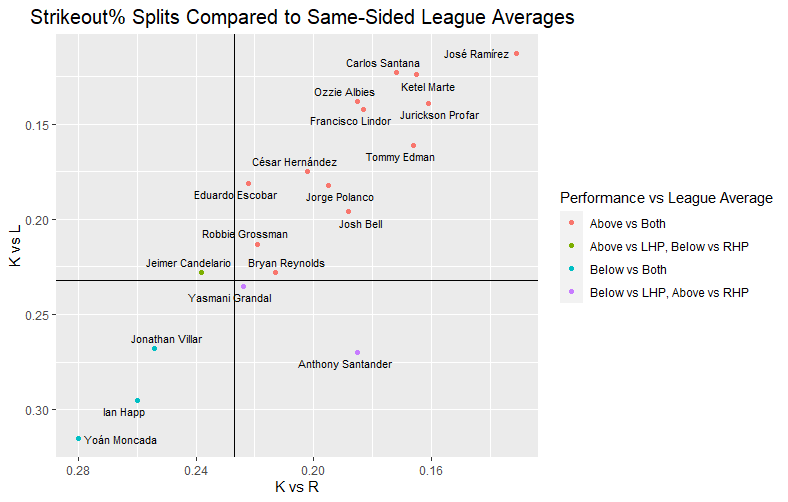


Figure 7: Strikeout% Splits Compared to Same-Sided League Averages

Figure 6: Walk% Splits Compared to Same-Sided League Averages

1. [Orioles centerfielder's changes result in hot start to season (wbaltv.com)](https://www.wbaltv.com/article/cedric-mullins-baltimore-orioles-2021-season/36110875) [↑](#footnote-ref-1)
2. [Finding Switch-Hitters Who Should Stop Switch-Hitting | FanGraphs Baseball](https://blogs.fangraphs.com/finding-switch-hitters-who-should-stop-switch-hitting/) [↑](#footnote-ref-2)
3. [Major League Leaderboards - 2018 to 2023 - Batting | FanGraphs Baseball](https://www.fangraphs.com/leaders/major-league?pos=all&hand=S&type=c%2C0%2C2%2C6%2C50%2C36%2C39%2C41&month=59&qual=600&startdate=&enddate=&season1=2018&season=2023&sortcol=9&sortdir=default&v_cr=202301) [↑](#footnote-ref-3)
4. Lichtman, Michael. *The Book: Playing the Percentages in Baseball.* (2014) [↑](#footnote-ref-4)
5. [What Is Random Forest? | IBM](https://www.ibm.com/topics/random-forest) [↑](#footnote-ref-5)