Identifying Underpaid and Overpaid MLB Hitters: Rubric

DS 4002 – Spring 2023 – Terence Moriarty
Submission Format: Source Code File and PDF Documents Uploaded to Canvas

Group Assignment

General Description: Create a model that can identify underpaid and overpaid MLB hitters. Write a defense of your model and its results.

Why am I doing this? The goal of this assignment is to experience and work through crucial parts of the data science lifecycle. While the objective and datasets have been provided, you must make decisions about what type of model works best for your goal and which features of the datasets are most significant to your model. Additionally, you will provide a written explanation of your model and its results, which will ensure your careful consideration throughout the model-building process. The skills developed through these tasks will be critical as you continue to move through the UVA Data Science curriculum and into a potential career in Data Science.

What am I going to do? Develop a model that anlayzes the statistics of every hitter from the 2022 MLB season, and qualifies their overall level of contribution. The model should also include a comparison of level of contribution with salary. From here, identify the ten most underpaid and ten most overpaid hitters in the MLB for the season. Two datasets are provided, one with statistics for every hitter in the 2022 season, and one with every MLB player’s 2022 salary. You will be graded based upon the consideration you put into your model, your choice of model, and your written explanation of the model and its results. You will not be assessed on the statistics and salary of the players chosen. Deliverables include:

- Well-documented source code file.
- A PDF document containing your written explanation of your model and its results.
- A PDF document featuring the names of the chosen players.

All of these items will be submitted electronically through Canvas.

Tips for success:

- Don’t be afraid to get creative. There are essentially unlimited possibilities for how you can build your model.
- Challenge yourself. While not required, consider innovative ways to make your model most useful to an MLB team. As an example, perhaps include a games played or at bats minimum, so small sample sizes do not carry too much weight.
- Don’t stress about results. The explanation of your thought process and decisions you made are far more important than which players your model identifies.
- Talk to the professor, the TA, and other students. This is a creative assignment, and you are allowed to show ideas to people for comment.

How will I know I have succeeded? You will meet expectations on Identifying Underpaid and Overpaid MLB Hitters when you follow the criteria in the rubric below.

Spec Category	**Spec Details**
Source Code	**Format** ● .R, .rmd, or .py file **Details** ● Well-documented source code file used to create your model. ● Should also include: ○ Any necessary data cleaning ○ Exploratory data analysis ○ Analysis of the model's results that led you to choose certain players
Written Explanation	**Format** ● PDF file **Details** ● A written explanation of your model and its results. ● Include the following sections: ○ Original Thoughts (1 paragraph) ■ After (or even during) exploratory data analysis, what was your original plan for how to create your model? ■ Did any aspects of the datasets stand out to you? ○ Model Building (1-2 paragraphs) ■ What decisions arose while creating your model? What did you chose to do and why? ■ Did you change model types at any point? If so, why? ■ Which statistics did you choose to focus on when evaluating player's contributions? Why? ○ Model Justification (1-2 paragraphs) ■ What about your model make it the best way for you to complete your objectives in this assignment? ■ Note: Don't be afraid to brag, you should be proud of your work. ○ Explanation of Results (1 paragraph, additional figure(s)) ■ What aspects of your model and its output led to your choice of those 20 players? ■ Identify one underpaid and one overpaid hitter and describe how their statistics and salary reflect this. ■ Include at least one figure or plot generated throughout your project that would support your explanation. Several plots or figures are welcome and encouraged. ● Note: Lengths of sections are not strict constraints, and should function as guidelines as you complete your written explanation.

Identified Players	**Format** ● PDF file **Details** ● Two lists, one identifying the ten most underpaid players and the other identifying the ten most overpaid players. ● Clearly indicate which list is which. ● Lists may be ordered by severity of under/overpay, but this is not a requirement.

Acknowledgements: This rubric was created based on the rubrics used by Loreto Alonzi, Assistant Professor of Data Science at the University of Virginia. This structure is pulled direction from Streifer & Palmer (2020).