NFL Moneyball: Maximizing Player Efficiency and Team Success

Introduction

The "Moneyball" approach revolutionized baseball analytics, and this project explores its application to NFL football. The goal is to analyze player performance metrics alongside salary and team success to uncover undervalued players, identify key contributors to team success, and explore the importance of player versatility. Given the multidimensional nature of NFL player contributions, a visual storytelling approach is well suited to convey insights in a clear and interactive way. This project provides tools and visualizations to support analysts, coaches, and even fantasy football enthusiasts in making data driven decisions.

Dataset

Overview

The dataset features NFL player performance metrics from the 2023 regular season, primarily sourced from Pro-Football-Reference. Unfortunately, Pro-Football-Reference does not include salary data in any way, so this was sourced from OverTheCap. Key columns include rushing, passing, and receiving yards, total touchdowns, player salaries, team W-L%, and Salary to Team Success Ratio.

Data Quality and Cleaning

The dataset required significant preprocessing:

- Incomplete Salary Data: Salary data was missing for approximately 25% of players. This affected the value analysis for these players and led to their exclusion from these calculations.

- Potential Season Mismatch: Performance and salary data might not align perfectly seasonally. While we're certain the player performance metrics were all from the 2023 regular season, the data source for the salary data did not specify a season, so there is a chance they are not aligned, introducing some uncertainty. Thankfully, salary data does not necessarily change between seasons, and if it does, it is often within a similar range. Given this, we hope if there is a season mismatch between player and salary data, it did not significantly affect the analysis.
- Player Name Standardization: Merging datasets required cleaning and standardizing player names to ensure accuracy. Because the salary data was from a different source and did not share the name standardization of Pro-Football-Reference, additional steps were needed to accurately merge salary data. There were some cases where salary data was not merged due to name standardization differences, these were manually corrected when identified.
- Team Abbreviation Inclusion: The data set including team win-loss percentage (W-L%) only featured the teams' full name, not the standard abbreviation that was included in other datasets. This was manually added to the team data, with some resentment, so that W-L% could accurately transfer to the comprehensive player data.

Data Abstraction

- Dataset Type: Multivariate, including numerical and categorical attributes.
- Item Type: Each row corresponds to a unique player-season entry.
- Attributes:
    - Categorical: Player Name, Team, Position
    - Quantitative: Passing Yards, Rushing Yards, Receiving Yards, Passing TDs, Rushing TDs, Receiving TDs, Salary, Team W-L%

   - Derived: Total Yards, Total TDs, Yards per Dollar, TDs per Dollar, Normalized Metrics (Yards, TDs, Salary, Yards per Dollar, TDs per Dollar), and Player Value Ratios such as Salary to Team Success Ratio

Tasks

This project supports several key tasks:

1. Assessing Player Contributions: Examine how individual player metrics correlate with team success metrics like win percentage (W-L%).
2. Identifying Undervalued Players: Compare player performance against salary to identify value outliers. Users can then optimize their salary cap, fantasy football team, or other applications by trading for or drafting players with a high impact to salary ratio.
3. Evaluating Versatility: Highlight players who contribute significantly across multiple offensive dimensions (passing, rushing, and receiving).

Visual Designs

1. Scatter Plot: Player Contributions vs. Team Success

- Design: Scatter plot with team win percentage on the x axis and yardage metrics on the y axis. A dropdown parameter allows users to toggle between total, passing, rushing, and receiving yards. Color indicates player position, and size relates to salary.
- Interactions: Filters for player positions, selection parameter for type of yard contribution, hover interaction to show player name, position, selected yard contribution, normalized salary, team W-L%, and raw salary.
- Insights: Highlights correlations between yardage contributions and team success, allowing identification of impactful players and positions. Filtering by position and yard contribution allows a more narrowed search for specific players while retaining salary context.

- Justification: A scatter plot is an easy and accessible starting point for exploring correlations and relationships. The user friendly parameter and filter adjustment provides the flexibility to highlight specific contributions or positions while retaining the familiar W-L% context.

2. Salary to Performance Bar Chart

- Design: Dual axis bar chart comparing player salaries to derived salary-to-team-success ratios (STSR). Bars use color to show the magnitude of difference between salary and STSR. STSR is also visualized separately with its own data point for comparison.
- Interactions: Sorting options and filters for player positions as well as a salary range selection filter allowing a search to fit any given budget. Hover interactions show player name, position, team W-L%, Difference (salary vs salary-to-team-success ratio), and raw salary.
- Insights: Identifies players offering high value relative to their salary and those potentially overpaid. By using STSR and comparing to raw salary, the user can get an impression of how the player is valued by their current team. In the visualization, this was calculated using Difference, and was included in the color coding. To expand on this relationship, if a player's salary is greater than their STSR, then they are likely overpaid or less impactful. If raw salary is about equal to STSR, the player's salary is likely aligned with the team performance. If salary is less than STSR, this player may be undervalued relative to their impact. Using the visualization, the darker green bars indicate a favorable difference between salary and STSR, and identify players that might be undervalued and worth pursuing.
- Justification: The dual axis bar chart is a great way of comparing similar values, like a players raw salary and their derived STSR. Visualizing both metrics on the same chart

allows user to quickly identify disparities between the two. Similarly, bars are color coded based on the magnitude of difference between salary and STSR.

3. Offensive Versatility Heatmap

- Design: Heatmap showing normalized metrics for passing, rushing, and receiving yards.
- Interactions: Filtering by position and sorting by performance metrics.
- Insights: Highlights versatile players who excel across multiple categories, as well as specialized contributors.
- Note on Implementation and Results: As mentioned in the video presentation, this heatmap implementation did not show versatility the way I had envisioned. Essentially, it showed minimal player versatility, the majority of players had normalized values of 0 or 1, with little variability. I didn't initially expect this, but it might actually be expected behavior for the data. NFL players are generally highly specialized, so the extremes (0 or 1) seen so frequently in the normalized values might be valid. That said, I also considered alternate explanations and have some ideas about why/how this happened, and if it is valid or a consequence of how I analyzed the data. Unfortunately, as of the time of writing, I haven't been able to successfully implement any changes, so this is theoretical. While the heatmap didn't show the expected versatility, or much versatility at all, it did highlight some standout contributions. I'm mostly curious if the lack of versatility is due to my implementation or the data itself.
    - For a potential cause, most players only contribute significantly in one category (rushing, receiving, passing yards) due to their position. So if we normalize this data, these contributions could be scaled to 1 for the category they contribute most significantly to, and 0 for the others. So in this dataset with specialized roles/contributions, maybe versatility is more rare than I expected.

        - There could also be a small sample size for certain metrics. Given the specialization mentioned, there might only be a handful of players with non-zero values in certain categories, like passing yards for non-quarterbacks. In this case, normalization could exaggerate the differences, leading to extreme values (1 or 0) dominating the heatmap, as observed in my visualization.
        - For this visualization, I normalized using a max value calculation, so each value was divided by the max value for that metric. Since most players have contributions far below the max value I used, the normalized values will cluster at 0, with some outliers at 1. This could also explain the lack of versatility seen in the visualization.
            - To improve this implementation, I believe a z-score normalization or other alternative statistical scaling methods could be used to show more nuanced differences in versatility.
- Justification: The heatmap is designed to showcase a players' versatility by normalizing passing, rushing, and receiving yards and visualizing them side by side. Using a heatmap, players who contribute across multiple offensive dimensions standout, but players who are more specialized and contribute more to a single metric are still noticeable. Normalization attempts to ensure that contributions are compared on the same scale, so standout players are emphasized regardless of their raw totals.

Results

This project effectively demonstrated the power of visual storytelling in NFL analytics. Key findings include:

- Undervalued Players: The scatter plot and bar chart revealed players who significantly outperformed their salary, presenting potential trade or draft opportunities. This was generally identified using the scatter plot, which made it easy to find players with a low salary, high yard contributions, and high team success metrics. The multiple filtering options allow a narrower scope where appropriate.

  - Notable undervalued players identified:
    - Quarterbacks: Brock Purdy, Patrick Mahomes (surprisingly)
    - Running backs: Gus Edwards, James Cook
    - Wide receivers: Amon-Ra St. Brown, Puka Nacua, Brandon Aiyuk

- Versatility vs. Specialization: While the heatmap showed an unexpected lack of versatility, the results encouraged additional thought, research, and analysis of the data, as well as data normalization methods. Whether the lack of versatility is due to the implementation or the data itself, the heatmap did highlight both versatile players and highly specialized players.
  - Versatile:
    - Justin Fields - QB with a normalized metric value of 0.256 for rushing yards and 1.0 for passing yards.
    - Derrick Henry - RB with a normalized metric value of 0.183 for receiving yards and 1.0 for rushing yards.
  - Specialized:
    - Tua Tagovailoa - QB with a normalized metric value of 0.016 for rushing yards, 1.0 for passing yards.
    - Stefon Diggs - WR with a normalized metric value of 0.004 for rushing yards and 1.0 for receiving yards.
- Salary Disparities: The bar chart showed notable discrepancies in player salaries relative to their contributions and team success metrics, suggesting potential inefficiencies in team spending.
  - Notable disparities identified:

        - Justin Herbert (QB), Kyler Murray (QB), Keenan Allen (WR), Curtis Samuel (WR)

Process and Tools

Tools Used

- Google Sheets: Initial data collection, cleaning, and preprocessing.
- Tableau: Visualization and interaction design, some data processing.

Process

1. Data Integration: Merged performance metrics with salary and team success data.
2. Derived Metrics: Created normalized and ratio metrics for deeper insights.
3. Visualization Development: Designed and improved on visualizations to ensure clarity and functionality. User interactions were also developed during this stage.
4. Presentation: Showcased visualizations and interactions, explained usage and challenges in development.