Dataset Partition Strategy

For this project, I implemented an 80/20 split for the dataset partitioning, where 80% of the data from 2019 to 2022 was allocated for training and the remaining 20% for testing. This partitioning strategy was chosen to ensure a balanced approach that allowed the model to learn patterns from most of the data while still reserving a reasonable portion for evaluating its performance on unseen data. The 80/20 split is a widely adopted convention in machine learning as it provides sufficient data to train the model while keeping enough data for robust evaluation.

By using this split, I ensured that the model would generalize well when predicting 2023 fantasy football performance. The larger training dataset enabled the model to capture intricate relationships in the data, such as how player stats (e.g., Yards from Scrimmage, Passing Yards, Touchdowns) contribute to Fantasy Points. Meanwhile, the 20% testing set serves as an objective way to assess the model's performance, revealing potential overfitting or underfitting issues. This partition strategy is effective given the amount of historical data, and it aligns well with my goal of predicting the top 200 players for 2023.

Additionally, this partitioning approach helps in mitigating data leakage, where information from the test set might inadvertently influence the model. By carefully separating the train and test sets, I ensured that the model only learned from past seasons, making its predictions for the 2023 season more reliable. The approach also aligns with best practices in machine learning, where the aim is to evaluate the model's ability to generalize to unseen data, which closely mirrors real-world conditions.

EDA Analysis and Insights

In my exploratory data analysis (EDA), I focused on understanding key patterns in the data and identifying potential relationships between performance statistics and fantasy points. I ran several types of EDA, including descriptive statistics (mean, median, standard deviation), missing value checks, and visualizations such as histograms, scatter plots, and correlation heatmaps. These steps gave me a clear picture of the data distribution and allowed me to confirm the integrity of my variables.

The visualizations provided important insights. For example, the correlation heatmap revealed strong relationships between variables like Yards from Scrimmage, Total Yards, and Fantasy Points, which are critical for predicting player performance. The scatter plots showed how certain features, like Passing Yards, affected Fantasy Points distribution, with clear outliers for

positions like QBs. The distribution of Fantasy Points was also insightful, revealing that the majority of players clustered at the lower end, while a few high performers (like star QBs or RBs) dominated the upper range of scores.

Through EDA, I also identified certain trends that helped refine my problem statement. For instance, the data confirmed that Yards from Scrimmage and Passing Yards are strong indicators of a player's fantasy football value. I also recognized that the variability in performance across different positions suggests that position-based features will be crucial in improving model accuracy. This insight will guide future model iterations, as I plan to include advanced features that account for these positional differences.

EDA Insights and Problem Definition

The insights gained from the EDA helped me refine my approach to predicting the top 200 players for the 2023 season. First, it confirmed that Yards from Scrimmage and Passing Yards are key predictors for fantasy football performance, aligning well with my problem statement. The distribution plots and correlations supported the idea that players who accumulate more total yards (both rushing and receiving) tend to score higher Fantasy Points, which provides a strong foundation for model development.

One of the challenges highlighted by the EDA is the variability in performance across different positions (e.g., RBs, WRs, and QBs). Players in certain positions generate fantasy points in different ways, which suggests that I may need to treat positions differently during model training. For example, QBs typically generate fantasy points from passing yards and touchdowns, while RBs rely more on rushing stats. This insight emphasizes the importance of including position-based adjustments or additional features in the model.

Moreover, EDA also revealed possible outliers and inconsistencies, such as certain players whose performance varied significantly across seasons. These outliers may present challenges in model accuracy, but they also offer opportunities for advanced modeling techniques like weighted averages or rolling statistics, which could capture player trends over multiple seasons. These techniques can improve the robustness of the model, particularly for players with inconsistent performance histories.

Data Problems and Preprocessing Recommendations

Throughout the EDA, I identified several data problems. First, I encountered duplicate player names (e.g., "A.J. Brown" with symbols like "*+"), which needed to be cleaned to avoid misalignment when aggregating player statistics across years. I handled this by removing special characters from player names and ensuring consistency across all datasets. Additionally, I noticed missing values in the "Position" column, which required attention, as position-based features are critical for fantasy scoring predictions.

My recommendations for data preprocessing include consistently cleaning player names and inputting missing position data to avoid gaps in model training. Moreover, aggregating statistics for players appearing in multiple years was crucial to ensure that players' performance over time is captured correctly. For example, summing Yards from Scrimmage and averaging key metrics like touchdowns per game helped provide a more accurate picture of each player's contributions. These preprocessing steps will ensure the dataset is ready for accurate prediction modeling.

I also recommend creating derived features that capture season-over-season improvements or declines, as these trends could significantly influence a player's 2023 performance. For example, calculating average touchdowns per game or yards per game could provide a more granular understanding of each player's efficiency, helping the model account for rising stars or players nearing the end of their peak performance.