Patrick Quinn

ADAN 8888

6 October 2024

Week 5 Written Report

Data Processing and Feature Engineering for Fantasy Football Analysis

The primary goal of this week is to process and prepare a historical fantasy football dataset from the years 2019 to 2022. This week, I focus on ensuring the data is clean, well-structured, and ready for modeling and analysis. The dataset includes key features such as player performance metrics, including yards, touchdowns, receptions, and fantasy points. I perform several preprocessing steps on the data, including handling missing values, removing duplicates, treating outliers, and normalizing and transforming the data as necessary to make it suitable for machine learning models.

The first step involves cleaning the data by removing any noise and inconsistencies. I check the structure of the dataset to ensure that each dataset from 2019 to 2022 has a consistent format. Column standardization is essential, so I rename columns that have different names across datasets to ensure uniformity. This step is crucial when concatenating or merging datasets across different years. I also remove special characters from player names and other string-based columns to standardize the data.

To support accuracy, I find and address outliers in the dataset, particularly for performance metrics like fantasy points, total yards, and touchdowns. I use Winsorization to cap extreme outliers at the 1st and 99th percentiles, which preserves the overall data distribution while reducing the influence of extreme values. Additionally, I apply the Z-score method to flag any data points that deviate more than three standard deviations from the mean. Unrealistic values, such as excessively high fantasy points, are then treated to ensure the integrity of the dataset.

Addressing missing data is another critical part of the process. For columns with minor missing data, such as the "Position" column, I fill in missing values using the most frequent value or based on the player's position from previous years. In cases where rows have extensive missing data in critical fields, I drop those rows to avoid introducing bias into the analysis.

To prepare numerical features for modeling, I apply several transformations. Log transformations are used on skewed variables like fantasy points and total yards, helping to ensure these distributions are closer to normal. Additionally, I normalize numerical data using min-max scaling, which brings all values between 0 and 1. This transformation is particularly important for features with a wide range, such as total yards and touchdowns, to prevent any one feature from dominating the model. In some instances, I also standardize features by centering them at zero and scaling them to a standard deviation of one, which is beneficial for machine learning algorithms that perform better with standardized data.

To streamline the dataset, I remove several columns that are unnecessary for modeling. For instance, team information is dropped since it is less relevant for predicting individual player performance. Similarly, identifiers such as player IDs are taken out because they do not provide predictive power for fantasy points. I ensure that each player-season combination is unique by checking for and removing any duplicate rows. This step helps prevent skewed analysis and ensures that no player or statistic is overrepresented. I also remove any redundant columns that may have been duplicated during data preprocessing.

Although the dataset primarily contains numerical data, text columns like player names and positions must be processed to maintain consistency. I standardize all text columns by converting them to lowercase and clean these fields by removing any unnecessary punctuation or special characters. To convert categorical variables, such as player positions, into a format suitable for modeling, I utilize one-hot encoding. The "Position" column is transformed into binary columns for each position (e.g., QB, RB, WR, TE), allowing the model to better interpret categorical data.

At this stage of the project, I continue to refine the dataset through data cleaning, outlier treatment, handling missing values, and performing necessary transformations. Each of these preprocessing steps is essential to ensure that the dataset is robust and well-prepared for the next phase, where I will build predictive models to analyze player performance and generate insights for the 2023 fantasy football season.