Sydney Lockwood

The Relationship Between Sacks and Turnovers on Win Probability in the NFL

Data Overview

This project examines how sacks and defensive turnovers contribute to a team's win probability in the NFL. The first data set I will be evaluating is from NFL Savant for all play-by-play data from 2018 to 2024. I can then filter to get a view on all defensive plays and provide a funneled view into sack rates and turnover rates. A second dataset from NFLVerse (GitHub) provides game-level outcomes, including wins and losses, for the same time period. Combining these datasets allows me to analyze how pressure-related plays correlate with game outcomes.

Data Source Summary

NFL Savant is a publicly available dataset that compiles official NFL play-by-play data using feeds straight from the NFL. Data is downloadable in a CSV file by season. The yearly play-by-play data was available after the NFL published the official records.

NFLVerse Github is maintained by an analytics community that compiles official game data for every NFL game since 1999. The GitHub page is dedicated to data and analysis of the National Football League and is regularly updated to ensure accuracy.

Data Structure & Content

NFL Savant data provides a wide range of analytical variables. The main ones I will use include GameId, Quarter, Minute, Second, OffenseTeam, DefenseTeam, Down, Formation, PlayType, and more as needed. The dataset is very extensive, so it will be important to decide what values I do and do not need.

GitHub NFLVerse dataset features variables like game_id, season, game_type, away_team, away_score, home_team, home_score, and result. This dataset is also extensive, but easier to filter, and will primarily be used to calculate win probability.

Together, these datasets can be merged by game_id and season to create a clean, singular dataset for analysis.

Data Completeness & Consistency

Both sources provide complete play-by-play coverage for all NFL games from 2018-2024. Minor inconsistencies may occur, such as different naming conventions or missing values. These can be corrected through data cleaning before merging both data sets.

Quality Issues & Potential Biases

Some quality issues that may arise are play tagging inconsistencies, reporting bias, or overlapping metrics. To mitigate these issues, I will filter for regular-season games only and filter for sacks in turnovers in the same play. This can help me identify a strip-sack, where the play would be recorded as a fumble turnover and not a sack.

Initial Cleaning & Preparation Plan

1. Import both datasets into Excel
2. Filter for regular-season games 2018-2024
3. Merge datasets using game_id, play_id, and season
4. Remove or address missing values and ensure consistent team labels across the sheet
5. Create new calculated values
    a. Sack rate: sacks/opponent pass attempts
    b. Turnover rate: turnovers/opponent possessions
6. Aggregate data by team-season and calculate the win percentage for modeling

Next Steps

1. Begin cleaning the data sets
    a. Identifying missing variables or incomplete data
    b. Verify consistent labeling
2. Add calculated metrics
3. Merge the datasets
    a. Perform exploratory checks
    b. Remove duplicate data
4. Plan the modeling approach