NFL ETL Pipeline Project

Max Munits

This passion project is very much a work in progress. The initial purpose of this is not to have accurate models predicting the upcoming NFL season, but rather to showcase my knowledge of data architecture and ETL pipeline construction. As the groundwork is laid with implementation of CI/CD and ETL best practices, I will begin to hone in more on developing meaningful and insightful analyses and visualizations.

Initial Goal

To develop a linear regression model that will predict win totals for all 32 NFL teams in the 2023 season. The idea is to take stats from the previous two to three seasons and use them to predict the teams performance and overall win total. This will be done for the previous five NFL seasons to show how accurately the model would have performed over those years and allow room for tweaking to ensure further accuracy.

Technologies

Extraction: All data scraping from https://www.pro-football-reference.com/ was done using the BeautifulSoup4 Python package.

Transformation: Pandas dataframes were used to transform the scraped data into a suitable form for storage in PostgreSQL, which was chosen for its convenience of being open source and flexible hosting options.

Data Modelling: As dbt (data build tool) has become the industry standard for building data models, I thought this would be a good opportunity to use it and build an entire data stack from the ground up. The ScikitLearn Python package will be used for regression analysis. Various options and BI tools will be explored later on for data visualization

Virtualization: Docker is used for establishing connections between Python and PostgreSQL, and will have further usage for other DevOps purposes later on

Code

nfl_etl/extract.py: Scraping and reformatting of all required data for building dbt models later on

nfl_etl/ nfl_etl/models/season_projection_stats.sql:

Framework for our first data model, which is in the works. Contains the stats required for performing linear regression to predict win totals for the upcoming season based on stats of previous seasons. This will be an ongoing process for a little bit and be tweaked with various advanced metrics to ensure accurate results