PREDICTING NFL GAME MARGINS USING LINEAR REGRESSION

Elliot Wilens

Github.com/edubu2/metis-project2

PREDICTING NFL GAME MARGINS USING LINEAR REGRESSION

Agenda

1. Data gathering
2. Feature engineering
3. Feature selection
4. Results
5. Next Steps

But first, a question…

Why do people like sports?

Data Sources and Tools

Tech Stack

- Python3
- pandas
- NumPy
- BeautifulSoup
- scikit-learn
- NumPy
- StatsModels
- matplotlib
- Seaborn

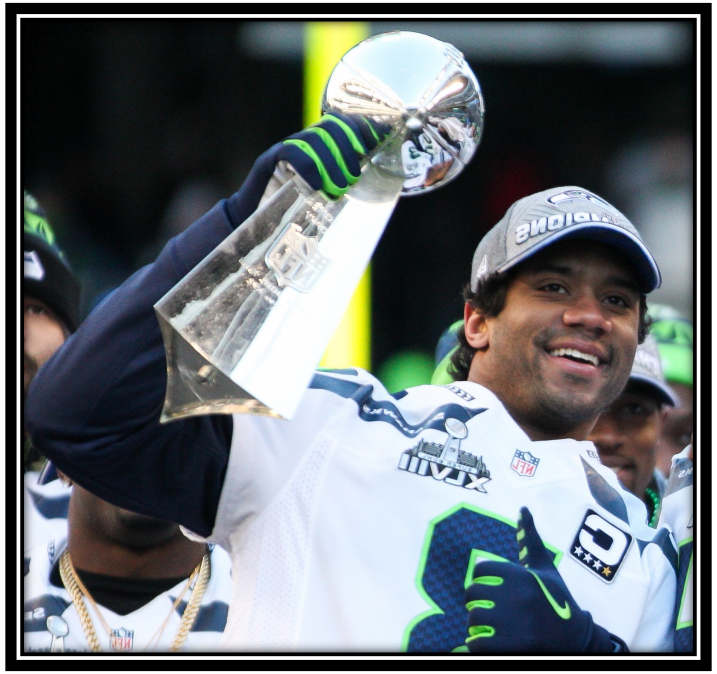
All stats gathered from **pro-football-reference.com** via web scraping with BeautifulSoup

- 11,948 data points (filtered down from 20500)

Feature Engineering

Objective: Implement features that will help answer the question:

How is this team playing right now?



- EWMA vs Standard Rolling Averages
  - Exponentially Weighted Moving Average
  - Weights most recent week stronger than past weeks
  - 4-week EWMA vs. 19-week EWMA
    - Subtracting these helps answer above question

Feature Selection

200 available features → 20

Used Lasso regularization to simplify model in terms of num. features (with K-Fold CV)

Top Ten features in terms of Lasso Coefficients (abs. value)

1. EWMA Margin
2. Home Game
3. EWMA (10-week) Margin
4. Season Total Margin (opponent)
5. EWMA Margin (opponent)
6. Season Total Wins
7. Third Down Conversion Percent (opponent)
8. Third Down Conversion Percent
9. EWMA Total Yards Allowed
10. EWMA (10-week) Wins (opponent)

Model Selection & Results

Standard Linear Regression Model with Standardized Features

- Limitation: cannot predict games in weeks 1-3

Test R^2 : 0.151

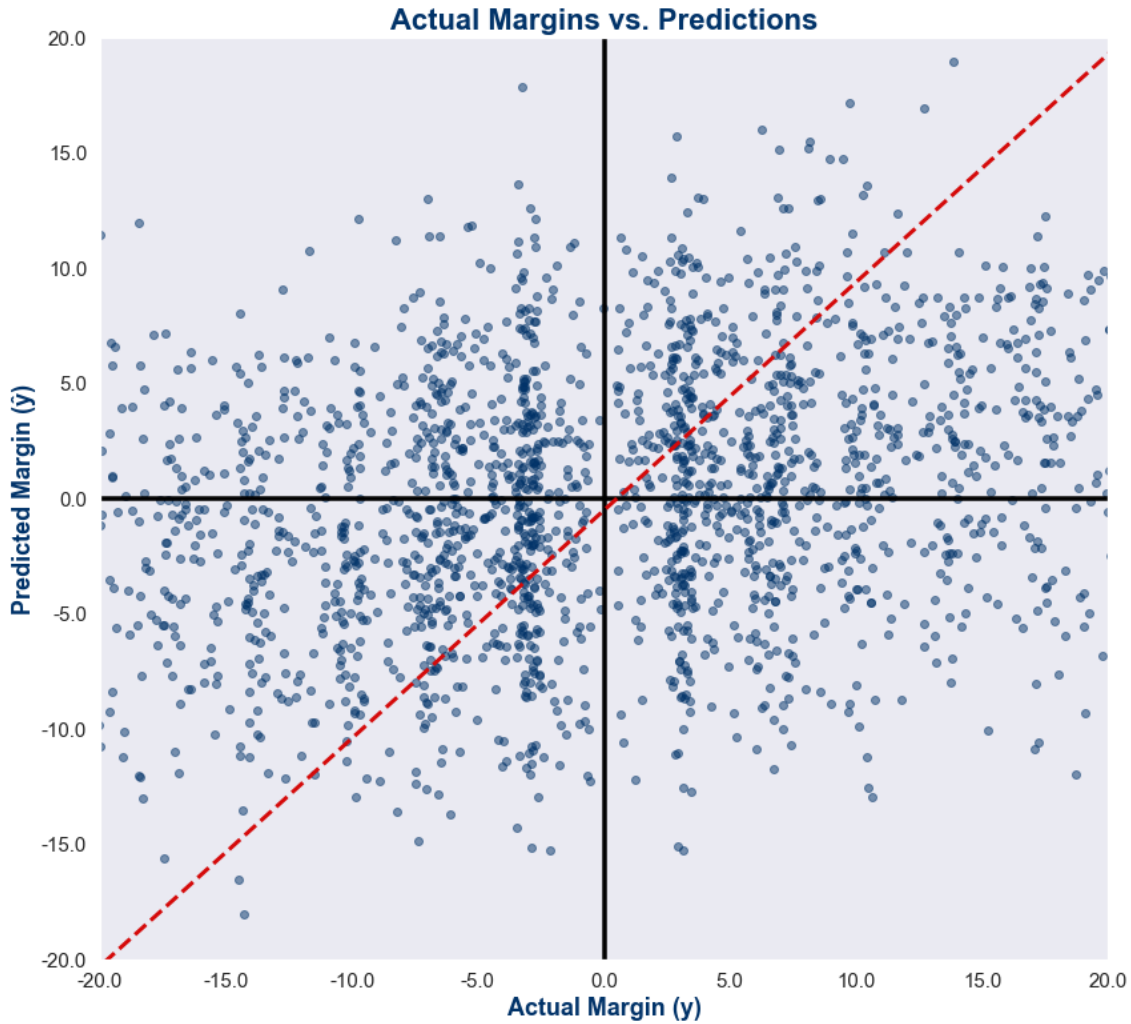
Predicted W/L outcome successfully 64% of the time in test set

Standard Deviation of Residuals: 13.58 pts

- can expect margin to be within 13.58 pts approximately 68% of the time, since the distribution of points margin is relatively normal
  - (since each game adds another +x, -x to the distribution)

Simplified Linear Model R^2 : 0.118

- Using: Home Game, Season Total Margin & Wins



Top right/ Bottom left: Correctly predicted (W/L)

Top left/ Bottom right: Incorrectly predicted (W/L)

Next Steps

- Further explore model results by team & by year
- Model did not seem to recognize change in gameplay over time
  - Explore time-series modeling
- Calculate ELO Ratings for each team (fivethirtyeight.com)
- Incorporate team rankings in certain stat categories into algorithm
- Include features that indicate whether the team is at/near ”full strength” (i.e. injured QB, WR, star DE)
- Capture weather data during web-scraping process

Appendix

- Dataset
- Data Gathering

- Feature Engineering
  - Home Game * EWMA Margin