Abstract
The goal of this project was to determine what features an NFL team should consider when negotiating player salaries. Through web scraping, I collected roster data from all 32 NFL teams for the 2020 season. I then created different regression models and found that the best was a polynomial regression. The features of the model included the number of years in the NFL, games started, games played, body mass index, and the team's record. These features indicate a good starting point in determining a player's salary.

Design
One of the most important decisions when running an NFL team is deciding how much to pay players. If players are paid too much, the team will not be able to afford many high-quality players. Conversely, if players are not paid enough, the team will not be able to recruit or keep high-quality players. This project provides insight on features associated with NFL player salary that will improve a team's ability to negotiate an appropriate player salary.

Data
The data for this analysis comes from web scraping data from pro-football-reference.com. Data were collected from the full rosters of all 32 teams for the year 2020. When including only those with a value for the target (salary), there were 1,935 players in the dataset. The features provided on the website included: age, years in the NFL, games played, games started, weight, height, position, calculated player "value", and team record.

Algorithms
Feature Engineering:

- Created features: general position (offense/defense/special teams), position group (eg, QB, defensive line), height/weight compared to others within the same position (low or avg/high), starter (no/sometimes/yes), and body mass index.
- Log-transformed the target due to a positive skew in the distribution.
- Categorical features were converted to binary dummy variables.

Models and Evaluation:
Created OLS, polynomial, ridge, and Lasso regressions. First, data were split into 80% training and 20% testing. Cross validation was then performed on the training dataset to determine the best alpha to use in the ridge and Lasso regressions. Models were then trained and the R^2 and root mean squared were determined for each model although the mean absolute error was primarily used to select the ideal model. Features that were highly correlated with each other were not included together in the models (eg, player age and number of years in the NFL). The ideal model was then tested on the holdout data.

Due to the number of outliers for the target, I also experimented with creating models using a dataset free from outliers (Z-score < 4). However, when tested on the holdout data *with* outliers, these models did not perform better than the above models.

Tools

Data was collected via web scraping through the use of Python's BeautifulSoup and then converted to a Pandas dataframe. Data cleaning and exploratory data analysis occurred using Numpy and Pandas. Regression models were created and cross-validated using SKlearn. Data and model visualizations were created using Matplotlib, Seaborn, and SciPy.

Communication
Slides and visuals were presented and posted on my personal GitHub page.