Predicting Playing Time for Women’s Soccer

Nancy Binowski

**Purpose:** In this project, I scraped, analyzed, cleaned, modeled and performed model improvements to determine the predictability of minutes of play for women soccer players in the US National Women’s Soccer League. The fundamental question trying to be answered is: are the right players playing?

**Data:** Data was scraped from FBref.com and the following features were pulled, screened, cleaned and refined: age, games, game starts, minutes, goals, assists, penalty shots made, penalty shots attempted, yellow cards received, red cards received, and various features/game statistics. I chose to use data from 2013, the first year of the league’s existence, through 2019, avoiding data altering effects of stoppage of play due to the COVID pandemic in 2020 and 2021. Goal keeper data was removed from the set because most of the statistics scraped did not apply to them. Some cleaning of data was required for missing data elements. Missing numeric data was set to 0 and missing categorical data was set to None. Most missing numeric data was for players who had very minimal or no playing time at all. Some of the numeric data also had to be processed to convert its text representation into integer format. The data was tested for outliers and none were determined.

**Tools Used:** A standard set of linear regression tools were utilized including Python, Jupyter Notebooks, BeautifulSoup, requests, pandas, re, numpy, seaborn, matplotlib, statsmodels, and sklearn.

**Preliminary Analysis:** A correlation study was performed and determined that there was sufficient correlation of minutes of play with several other features. Plotting showed that the relationships were linear in nature, thus making it a good candidate for a linear regression study.

**Feature Engineering:** Feature engineering was performed and multi-collinearity was reduced for features highly correlated like penalties and penalties/90, for example. The resulting feature set included: age, games, goals, assists, penalty shots made, yellow cards received, goals and penalties per game, and goals, assists and penalties per game. The target selected was minutes per game (minutes90). This feature set resulted in an OLS model R-squared value of 0.857 and all p-values less than the 0.05 threshold.

**Cross-Validation:** Cross validation testing was done on a simple linear regression model and its R-squared value was 0.854/0.866/0.849 for the training/validation/testing sets.

**Regularization:** Regularization was performed using the Ridge, RidgeCV, Lasso, LassoCV, ElasticNet and ElasticNetCV models. The ElasticNetCV model performed the best with R-squared values of 0.955/0.973 on the train/test sets.

**Future Study:** It would be interesting to run the Olympic Player data through the NWSL model and see if it performs the same. It also would be useful to segment the data by position to see if there were different prediction levels for different positions.