Analyzing MLS Club Performance

Christopher Lee and Harshal Rukhaiyar

Overview: This project aims to examine the key factors in determining how well a club (team) performs in the Major League Soccer (MLS) league – specifically, the number of goals a club scores within a given season. We employ various machine learning techniques such as principal component analysis (PCA) and neural networks. Data was gathered from the official MLS club statistics website.

Methods: We used the H2O package in R to implement PCA. To determine the number of principal components we should use, we employed the proportion of variance explained (PVE) and scree plot criterions, resulting in 13 principal components. To improve interpretability in a linear regression model, we extracted the most important statistics within the 13 components by taking statistics with higher eigenvalues. Following PCA, we inputted the resulting 20 statistics into a linear regression model to predict the number of goals a team would score in a season. Finally, to see if a non-linear approach would benefit our analysis, we created a neural network using the Keras package. This neural network contained two hidden layers with ReLu activation functions.

Results: Our linear regression model explained approximately 86% of the variability in the number of goals a club would score in a season. Significant statistics included losses, goals allowed, wins, pass percentage, ties, save percentage, number of clean sheets, shot percentage, yellow cards, and fouls. Our neural network, however, did not yield the best results. Upon multiple runs, the root mean squared error of the neural network ranged from 20 to 45.

Conclusion: Given the large error and wide range of results from our neural network, we conclude that our linear regression model is the best tool to predict the number of goals a team will score in a season. Had we had more time to tune the neural network, determine the optimal number of hidden layers and nodes, and fully understand and test different activation functions, our results may be different. However, it seems that when the output variable is numerical, linear regression or other, more simple, machine learning techniques may be better-suited for our project.