CS 254: Machine Learning
Project Short Proposal
Peter Larsen

Introduction

When it comes to sports in America, teams change and attempt to get better by one of two ways: free agency and drafting. Free agency is relatively easy, as once a player is successful, it can be seen that they are successful, and thus teams will attempt to bid for having the player come play with them. However, drafting is less certain. Often, in the sports world, the terms “bust” and “steal” are thrown around, meaning players who either didn’t live up to expectations, or exceeded expectations. In hockey, this is especially true, as a player’s success at lower levels (Junior Leagues, college, European Leagues) is never a sure thing to base their NHL success on. Consequently, there seems to be a lack of determinability, and more of a luck of the draw in determining how a drafted player works out.

Problem Definition and Algorithm

Thus, the question is relatively simple: can Machine Learning aid in the selection of players such that it starts to pick the correct player for each team by comparing incoming players to established players in the league? The inputs would be data on players currently in the league, and players in the upcoming draft class. The output would be a draft order of players based on numerical analysis and not “gut feelings”, as so often seems to be the case. The idea would be to run the algorithm multiple times for successive years and refine it each time (essentially telling it yes or no on each of its selections).

Dataset

The dataset would be a hand-constructed .csv file based on multiple different .csv files from Hockey Reference.com, a statistics site based around collecting all types of data that have to do with hockey. The dataset has features such as: Name, Height, Weight, Age, Position, Goals (Current Year), Goals (Career), Assists, Save Percentage, Blocks, Corsi For, All Star etc. All of these would potentially have the ability to factor into selection, especially as different positions use different statistics. The dataset is labeled, but disjoint, and requires time to be assembled. However, a solid portion is already assembled (2010-2011, 2011-2012, 2012-2013, 2013-2014, 2014-2015) due to work I’ve already put in. After constructing the last couple of seasons (2015-2016, 2016-2017, 2017-2018), I should be able to run comparisons, similarly to the first day of class in comparing receivers to linemen. Grouping by position and then by success would allow us to see how incoming prospects (weighting their statistics based on the level of play) could potentially fare.

Related Work

This type of work has already been undertaken in the NFL and the NBA, but no one has bothered to do so in the NHL. In doing this work for the other leagues, others used many less features in their analysis, but this is also due to the relative simplicity of the sport and its statistics. A fast runner will do well at many positions in the NFL, but simply being a fast skater does nothing to get around a goalie, or to help you aim. Therefore, the way to improve their methods is to utilize more data with more features, thus avoiding overfitting.

Bibliography

Data analytics in sports: improving the accuracy of NFL draft selection using supervised learning c2015; [accessed 2018 Sep 20] https://mospace.umsystem.edu/xmlui/handle/10355/47027

Hockey Reference. c 2007; [accessed 2018 Sep 10]. https://www.hockey-reference.com/