

GA-SEA-DAT2

Course Project Initial Presentation
Dan Johnston

Where We Left Off

Project Question(s): Using 2013-2014 regular season player statistics...

1. Can players be segmented into typologies based on individual season performance statistics
2. If a team's mix of player typologies predicts a team's success

Status

- Data collection complete
- Data exploration [mostly] complete
- Project Questions: unanswered

Transforming the Data

- Move away from total summaries
 - shots, goals, blocks, etc.
- Move towards measures of efficiency
 - goals per shot, blocks per time on ice, etc.
- Reasoning: players' time on ice varies widely based on coaching decisions. I wanted to focus on what players do with the time they have

Feature Selection and Clustering Exploration

- Initial clustering lead to clusters with very small (<5) base sizes
 - Found that players with few games were driving small clusters
- Special Teams features were removed
 - Short Handed and Power Play
- Tested $\sim 70k$ combinations of number of clusters and features. Not all made sense, so I filtered results
 - Smallest cluster ≥ 30
 - 5+ clusters
 - 3+ features
 - $\sim 9k$ models remaining

Using Clusters to Predict Team Success

- From the 9k clustering models, picked the 500 models with top Silhouette Coefficients
 - 0.24 - 0.38
- Player clusters were matched to each game record
- Used clusters as features in Linear Models to predict Plus/Minus
 - Goals Made - Goals Against
- Used Cross Validation (RSME) to Evaluate - best RSME is 2.29
 - **Features:** shots_per_min, assists_per_min, faceOff_wins_per_attempt
 - **Clusters:** 8

Thoughts on Outcomes

- Results are only okay. Best RSME is 2.29
- $\pm \sim 2$ covers the middle 50% of cases
- Player statistics from NHL feel incomplete

Max	+8
75%	+2
50%	0
25%	-2
Min	-8