GA-SEA-DAT2

Course Project Initial Presentation
Dan Johnston

Starting Point

Context:

Professional sports have seen a rise in analytics to support coaching and scouting decisions. The National Hockey League lags behind other sports leagues in the adoption of these techniques.

Project Question(s):

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

Data Sources

Two sources

1. http://www.nicetimeonice.com/api - independently run NHL statistic sites.
    - API provides easy reference for IDs used by NHL.com API
2. http://statsapi.web.nhl.com/api - Official API used by NHL.
    - No documentation is available to the public.
    - Provides game level data in JSON format.

The Data Collection

1. Collected game level data for the 2012-2013 season
   - Excluded postseason games
   - (30 teams * 82 game season) / 2 teams per game = 1,230 API calls
2. For each game, extract player level summary details
   - Excluded goalies
   - Desired details were buried deep in the JSON
   - game['liveData']['boxscore']['teams'][‘home’]['players']['playerID’]['stats']['skaterStats']
   - 18 players per team * 30 teams * 82 games =~ 44,280 player/game entries
3. For each player, create season summary statistics
   - Season summary statistics serves as basis for analysis
     - ~880 player records for the season
     - 16 features per player

Getting a Feel for the Data

- Used KNN to attempt to predict known positions
- 4 Positions: Defense, Left Wing, Right Wing, Center
- Tested n_neighbor = 1
- Scaled data. Time based features have much larger scales than others
- Results vary widely based on if/how positions are grouped
    - No Grouping - 0.568
    - Group Wing positions - 0.653
    - Group Wing and Center - 0.959

Digging Deeper

- Used Cross Validation to determine the optimal set of features and number of neighbors
- Using Group Wing positions as outcome
  - Not enough room for improvement when grouping Wing and Center Positions
- Lesson learned - estimate the number of iterations first
  - 1-25 Neighbors, 1-16 features = 1.64 million models, i.e. never finished
  - Ran for ~16 hours. 35% Complete
- Best fit:
  - Features: ('assists', 'blocked', 'evenTimeOnIce_s', 'giveaways', 'goals', 'penaltyMinutes_s', 'plusMinus', 'powerPlayAssists', 'shortHandedTimeOnIce_s', 'shots', 'takeaways')
  - Neighbors = 9
  - Cross Validation Accuracy Mean: 0.741054510623

Next Steps

1. Use clustering to determine player typologies beyond position
    a. K-means
2. Estimate effectiveness of combinations of clusters for team makeup
    a. Use Plus/Minus as outcome metric.
    b. Precise method TBD; likely linear regression