HOCKEY TEAM SUCCESS: A SUM OF ITS PARTS

MOLLY MCNAMARA

SPRINGBOARD FOUNDATIONS OF DATA SCIENCE

CAPSTONE PROJECT

THE NATIONAL HOCKEY LEAGUE

- Building a successful team is the greatest challenge for every General Manager (GM)
  - Can player characteristics and statistics predict team performance?
  - How well can future player performance be predicted at the time of the draft?
- Can data science help a GM to make more informed decisions in player selection?

THE DATA

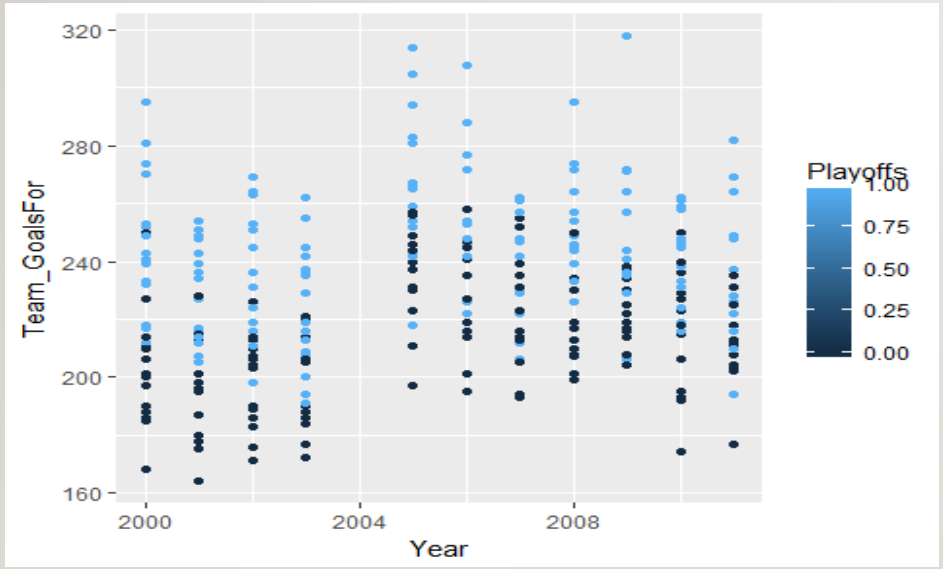
- NHL player and team statistics for years 2000-2011 pulled from:
  - Master (basic demographic information on all NHL players from 1908 to 2011)
  - Scoring (scoring statistics for all NHL players from 1908 to 2011)
  - Teams (team statistics and playoff outcomes from 1908 to 2011)
  - Goalies (defensive statistics for all NHL goaltenders from 1908 to 2011)
- NHL draft data from:
  - Draft (player draft information from 1979 to 2010)

THE DATA (CONTINUED)

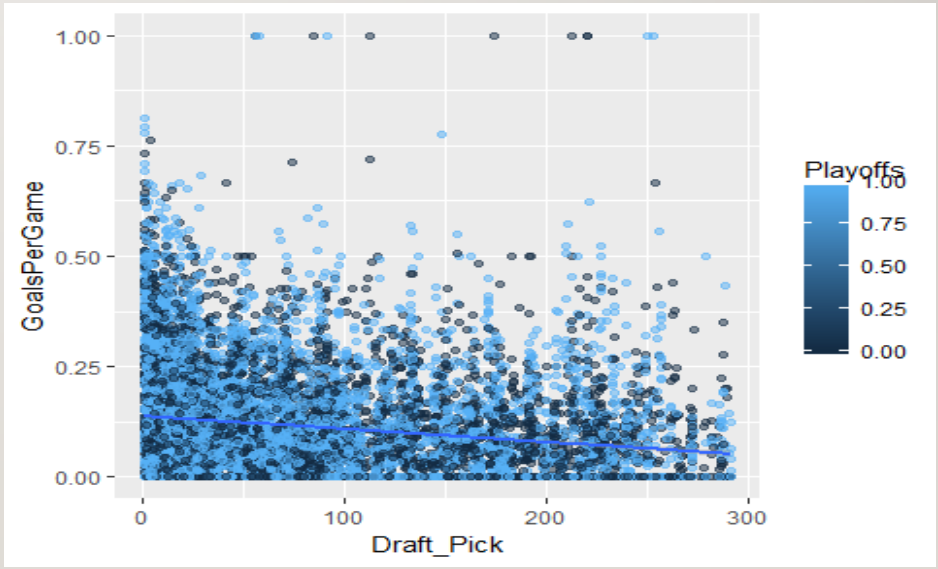
- Data was wrangled from 5 different files
  - Files joined/merged
  - Duplicative variables removed
  - New calculated scoring variables created
  - NA values handled as appropriate by variable
  - Dataset subset to years of interest
  - Final cleaned data file saved and exported

PLAYER CONTRIBUTION TO TEAMS: SCORING

- Teams that score more goals in a year are more likely to go to the playoffs

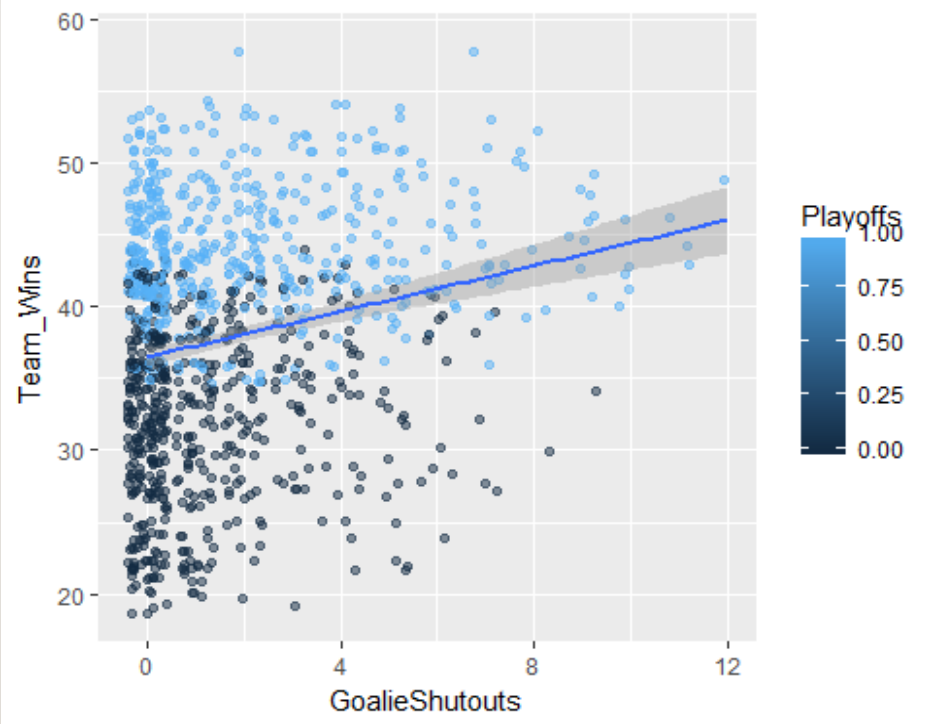


- Players drafted in higher rounds appear to score more goals



PLAYER CONTRIBUTION TO TEAMS: DEFENSE

- Goaltenders with a higher rate of shutouts (where the other team is prevented from scoring entirely) appear to contribute to their teams winning more games and making the playoffs

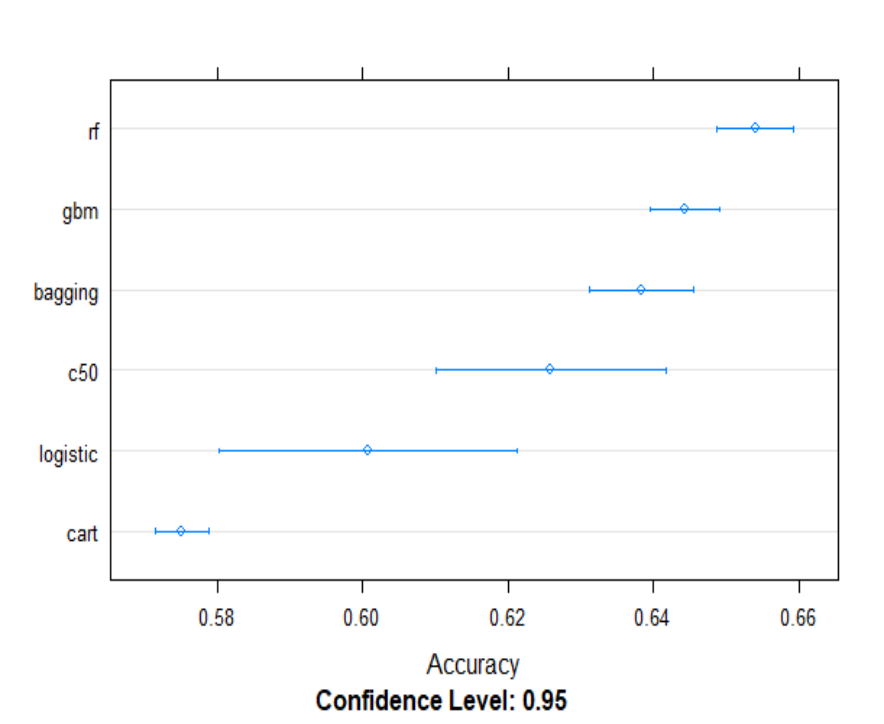


THE ANALYSIS APPROACH

- The primary question was evaluated using several different types of algorithms to build a model that allows for less than linear relationships between the variables; model performance was assessed to determine what type of algorithm produces the best model.
- The secondary question was evaluated using logistic regression to determine how the variables may impact the final outcome of player performance.

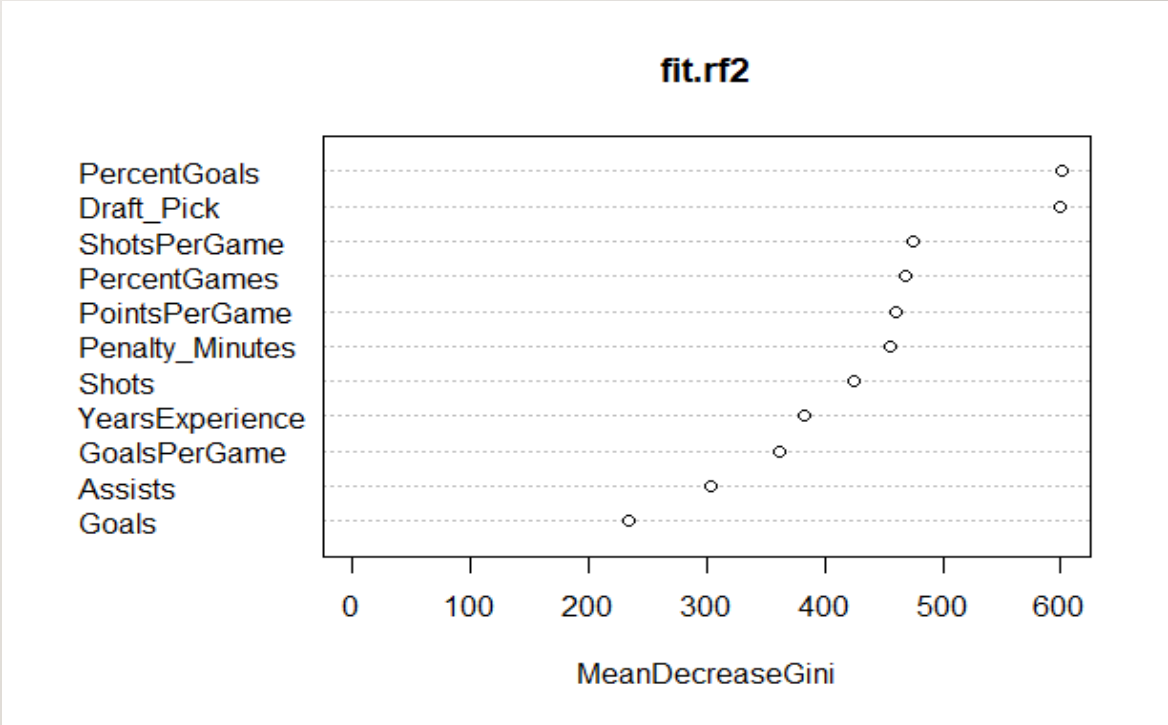
ANALYSIS (1)

- Repeated k-fold cross-validation was used to model team performance as a function of player features.
- The random forest model provided the most accurate prediction of if teams would make the playoffs.



ANALYSIS (2)

- A followup model using the variables of greatest importance from the first model was able to predict success with nearly 70% accuracy.



ANALYSIS (3)

- To address the question of predicting future player performance at the time of the draft, a subset of underachieving (high draft pick, scoring less than average goals) and overachieving (low draft pick, scoring more than average goals) players was used to determine what factors known at the time of draft can predict future performance.
- Logistic regression provided a fairly accurate model with the largest predictor variables being draft age, position played, height, amateur league, and drafting team.

```
Generalized Linear Model
2325 samples
   7 predictor
   2 classes: '0', '1'
Pre-processing: centered (85), scaled (85)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 2718, 2718, 2718, 2717, 2719,
2718, ...
Resampling results:
  Accuracy   Kappa
  0.9333757  0.4589219
```

RECOMMENDATIONS

- Look for an experienced player selected in a higher draft position who is able to play a high percentage of the team's games (perhaps indicative of a lower injury rate), and who has a strong scoring history and/or high shot rate
- Consider players who take penalty minutes (this may be an indicator of willingness to take risks to make plays).
- Player position, draft age, height, and coming from certain amateur leagues may give an indication as to how successful a lower or higher drafted player will perform longer-term. If the chance arises to take a player that meets the criteria (for example a tall Russian League forward) in a lower draft round, that saves the team money and earlier-round draft picks.

PROPOSED NEXT STEPS

- Expand dataset to enhance predictions with more years of NHL history
- Addition of pre-draft play statistics would help build an even better model of future performance
- Draft combine results would add another layer of data for predicting player performance

THANK YOU

- Thank you to my mentor Branko Kovač for all of the advice and support
- Data generously available from:
  - The Hockey Database by Doug Reynolds, via Open Source Sports (player and team statistics)
  - Sports Reference LLC. "NHL Entry And Amateur Draft History." Hockey-Reference.com - Hockey Statistics and History (draft data)