Capstone Project 2
Predicting NHL player performance using non-game metrics with machine learning

Kevin Hill

Statement of Purpose

- The goal of my capstone project 2 is to predict a National Hockey League (NHL) player's performance using non-game metrics such as height and weight.
- The primary clients I envisioned when planning this project are mainly NHL teams, and other non-NHL hockey teams.
- This project was inspired partly by my brother-in-law who was told he was 2 inches too short to play in the NHL. I am also a huge hockey fan and wanted to join my passion for the sport with my growing knowledge of data analytics.

Problem Formulation and Data Origin

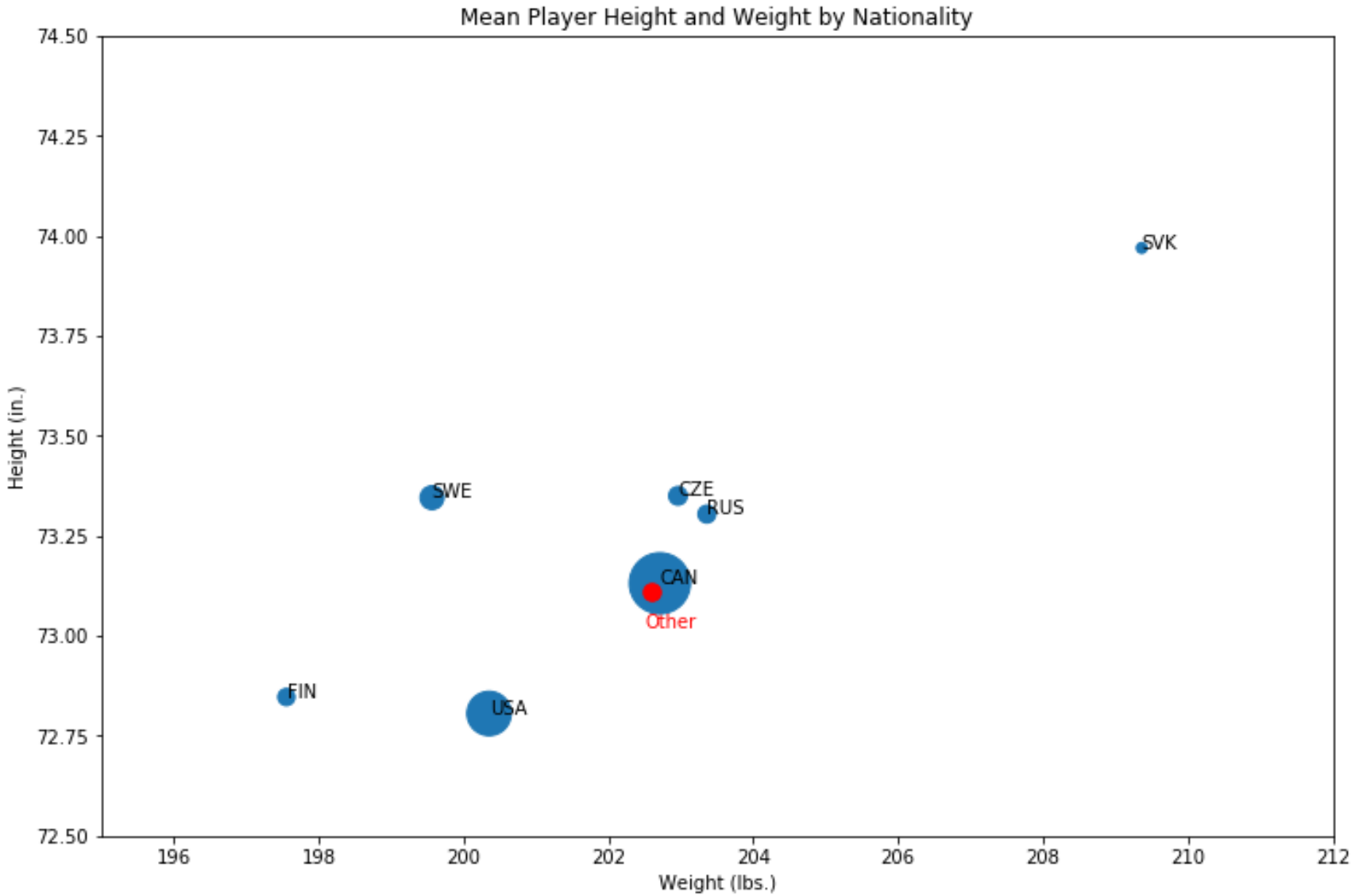
- The main problem to solve is: How many points will a player score? If points are not predictable, how about other stats such as corsi, aka puck possession?
- An additional problem to solve is: What metrics contribute most to a player's performance?
- The data was collected from NaturalStatTrick.com. It contains various statistics ranging from basic player information to game stats attributed to players and teams. The same data can be obtained online from such websites as NHL.com or Corsica.hockey.

Data Wrangling

- The data had to be combined from 3 different data sets, removing duplicated information, such as player name and team. Also, I removed several stats that I didn't feel I needed for this study, for instance I kept ratio stats like goals scored for percentage, instead of the numbers used to generate those stats.
- Additionally, I calculated neutral zone start percentages, because the data only included offensive and defensive zone starts.

Exploratory Analysis

- Player height and weight haven’t changed much over the past 10 seasons.
- The next slide shows mean player height and weight by nationality with the size of the dot representing number of players.
- My report details some interesting finds specific to player nationality by height and weight.
- Players appear to have the most productive year for points around age 29, generally getting better before that and worse after that.


Mean Player Height and Weight by Nationality
Height (in.)
Weight (lbs.)
74.50
74.25
74.00
73.75
73.50
73.25
73.00
72.75
72.50
196
198
200
202
204
206
208
210
212
SVK
SWE
CZE
RUS
CAN
Other
FIN
USA

Inferential Statistics

- I wanted to know if there was any correlation between height and weight and the various metrics I was trying to predict. On the right, notice the top stats that seem to correlate positively with height and weight are penalty minutes and hits, not points or corsi, which actually appear to be negatively correlated to height and weight.

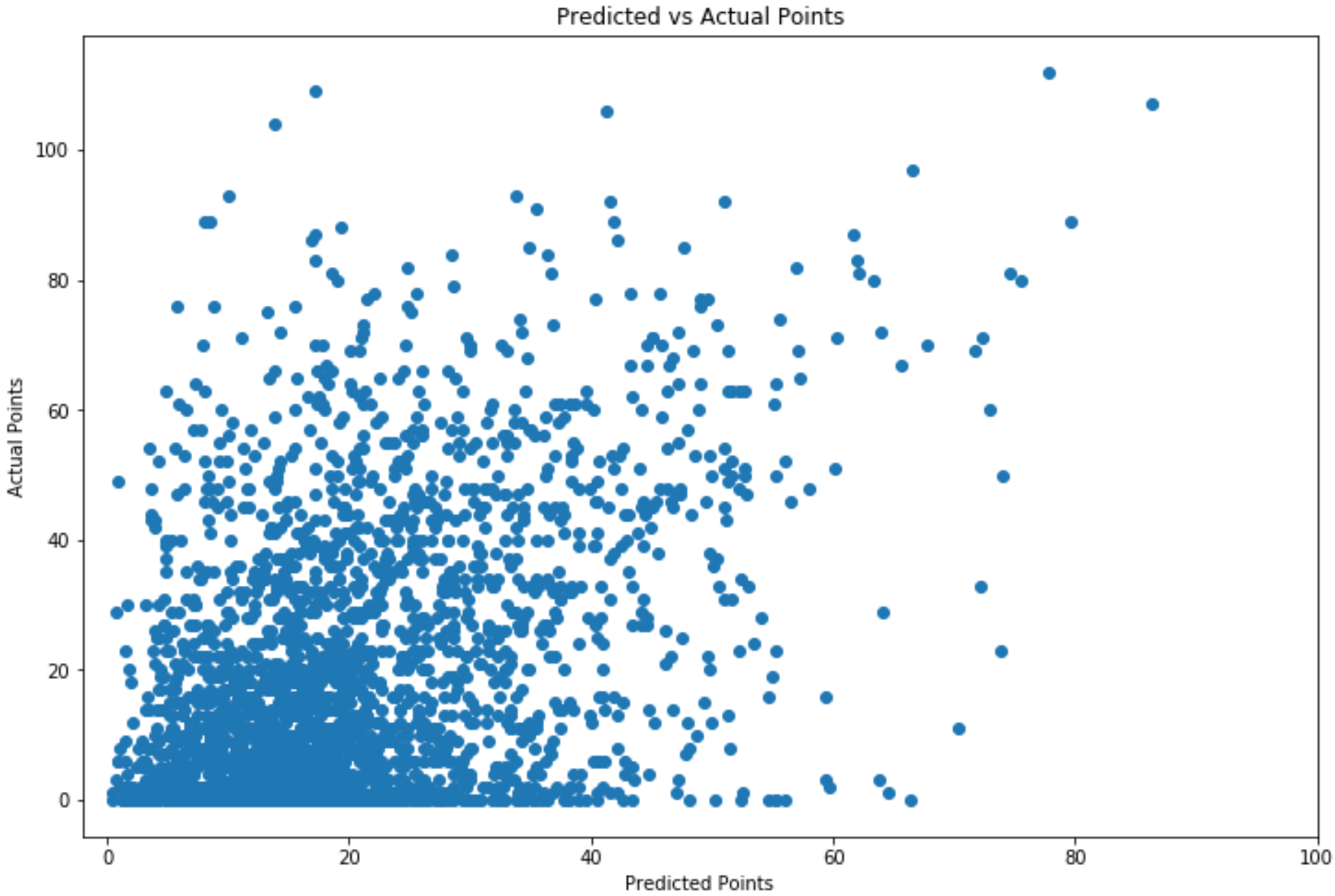
Height (in)	1.000000
Weight (lbs)	0.723978
PIM	0.220726
Hits	0.208112
Shots Blocked	0.139598
Def Zone Faceoff %	0.122515
Giveaways	0.054116
Neu Zone Faceoff %	0.041121
Age	-0.015806
PDO	-0.032167
Penalties Drawn	-0.068917
HDGF%	-0.092398
Takeaways	-0.108198
GF%	-0.109339
Rush Attempts	-0.111391
Total Points	-0.128907
Off Zone Faceoff %	-0.131483
CF%	-0.140692
Goals	-0.150659

Weight (lbs)	1.000000
Height (in)	0.723978
Hits	0.352725
PIM	0.327954
Shots Blocked	0.131205
Def Zone Faceoff %	0.115979
Neu Zone Faceoff %	0.079200
Age	0.076493
Giveaways	0.033424
Penalties Drawn	0.014032
PDO	-0.045023
Rush Attempts	-0.099929
HDGF%	-0.117117
GF%	-0.122696
Takeaways	-0.143154
Total Points	-0.143407
Off Zone Faceoff %	-0.145556
Goals	-0.154019
CF%	-0.156968

- I also found Czech players seem to be better defensively than other players, so calculated players' position breakdown by country, but found nothing of note.
- Russian players appeared to have higher mean stats in many categories, and there may be statistically significant evidence to support their advantage in point production and corsi. However, there may not be evidence to support an advantage in goal scoring or shooting percentage.

In-Depth Analysis

- I first chose to build a linear regression model to solve the problem. Using hyperparameter tuning, I would select the best performing classifier.
- However, very poor r-squared scores for models for different stats made me choose another classifier, the random forest. Again, hyperparameter tuning would provide me the best model. Results were better, but still low.
- The next slide shows the graph of predicted points versus actual points, showing no pattern to suggest a successful model.


Predicted vs Actual Points
Actual Points
Predicted Points
0
20
40
60
80
100

Final Results

- The best results came from the random forest regressor, but none of the models can be considered successful.
- The scores shown here demonstrate the lack of good results. An r-squared score varies from 0 to 1, with 1 being an excellent model. An RMSE score is 0 or greater, with close to 0 suggesting the model doesn't result in much error. Training scores come from the model building and test scores from data unknown to the models.

	R2 score test	R2 score train	RMSE score test	RMSE score train
base linear regression	0.047656	0.040776	414.148139	386.141300
ridge regression	0.047656	0.040776	414.148143	386.141300
lasso regression	0.047656	0.040776	414.148163	386.141300
random forest	0.124061	0.719230	380.921655	113.025847

Recommendations

- Although point prediction is extremely poor at best, I can suggest which statistics are most affected by non-game metrics.
- Height and weight both seem to have a positive impact on hits and shots blocked.
- Age appears to have a beneficial effect on defensive zone starts, and roughly outlines a player's best years for several stats, including point production.
- Nationality has perhaps minor influences to specific stats.
- In any case, players apparently should be judged by their in-game performance, not by their appearance or background, but non-game metrics may help target specific roles on a team for players to fill.

Future Considerations

- Goalie data was not available from the website, so that is an obvious next step.
- Studying individual NHL teams' mean height, weight and age may reveal trends in scoring or defense according to this study.
- Also, collecting data from future seasons to incorporate with this existing data may track changes over time.