Delving into Corsi: Building predictors for shot attempts in hockey

Patrick Rooney

November 6, 2017

Motivation

A player's Corsi-for percentage in hockey is a measure of how many shots his team attempts while he is on the ice:

$$CF\%(X) = \frac{\text{shot attempts for X's team while X on-ice}}{\text{shot attempts for both team while X on-ice}}$$

Note: Shot-attempt = shot on goal OR shot missed / blocked

Questions:

- Can we predict shot-attempts using the *CF%* statistic?
- Can we build a *better* predictor, using all shot attempt data?
- Most importantly: is the *CF%* a good measure of how a player influences shot attempts?

Our Data Set

Main data set:

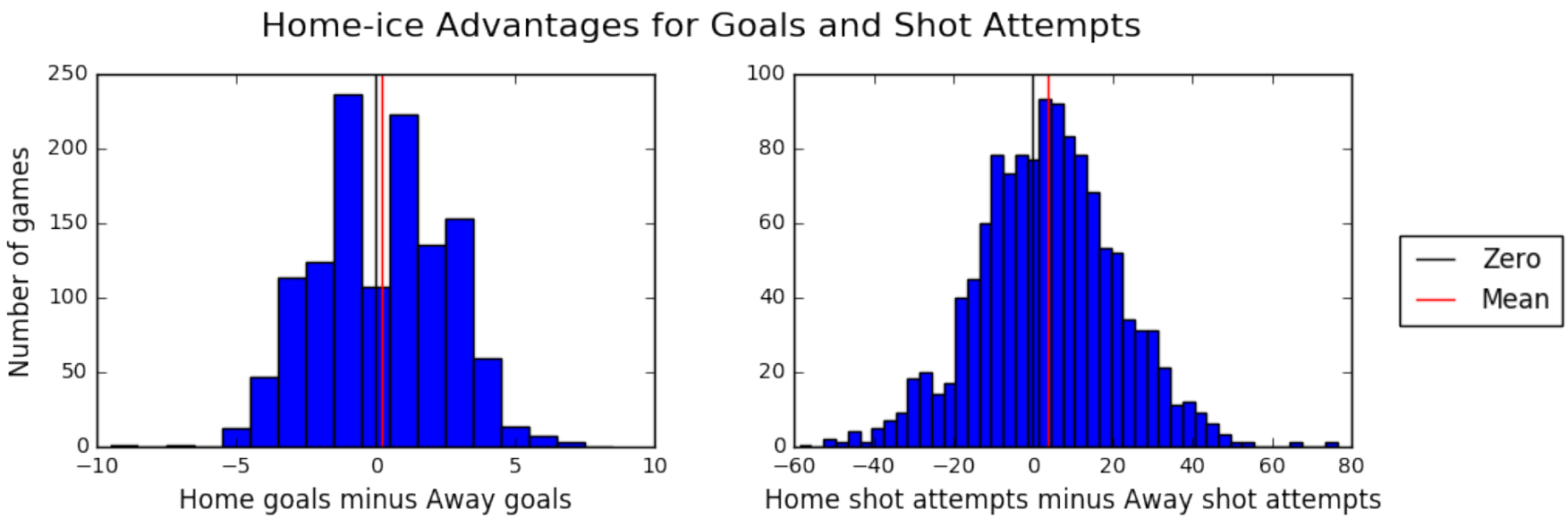
- All 1,230 regular-season games in the 2015-2016 NHL season
- 136,540 shot attempts (of which 66,601 were on goal, and 6,565 were goals)
- 900 players (no goalies). Player on home-ice treated differently then on away-ice (home-ice advantage is real!)

Additional data:

- Salaries of all players
- Average time-on-ice per game for all players

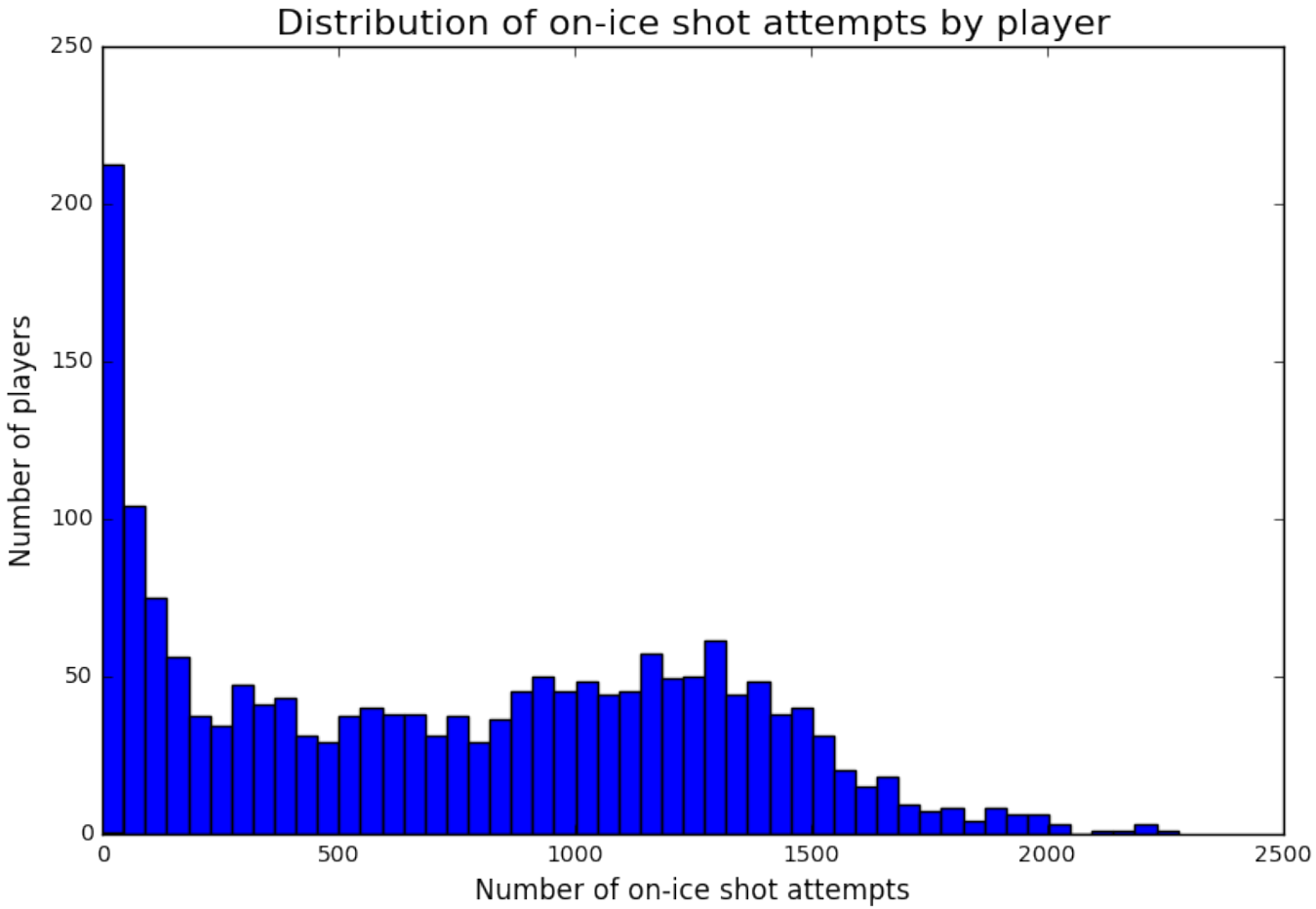
Home-ice advantage

Home-ice advantage is worth on average 0.198 goals/game, and 3.582 shot attempts/game.



Distribution of Shot-attempts

Studying shot-attempts instead of shots on goal gives us a bigger sample size. 87.2% of players attempted shots more than 50 times.



Predictors

We built three different predictors, with varying complexity:

1. A logistic regressor using only six features: the cumulative $CF\%$ of home and away players, the average salaries, and the average playing time.
2. A logistic regressor using 1,800 binary features. Each player has two features: whether he is on-ice at home, or on-ice away.
3. A random forest algorithm using the above 1,800 features.

The data set used for the second and third predictors are the same, but the random forest algorithm is more complex than a logistic regressor.

Crude Logistic Regressor

Six features:

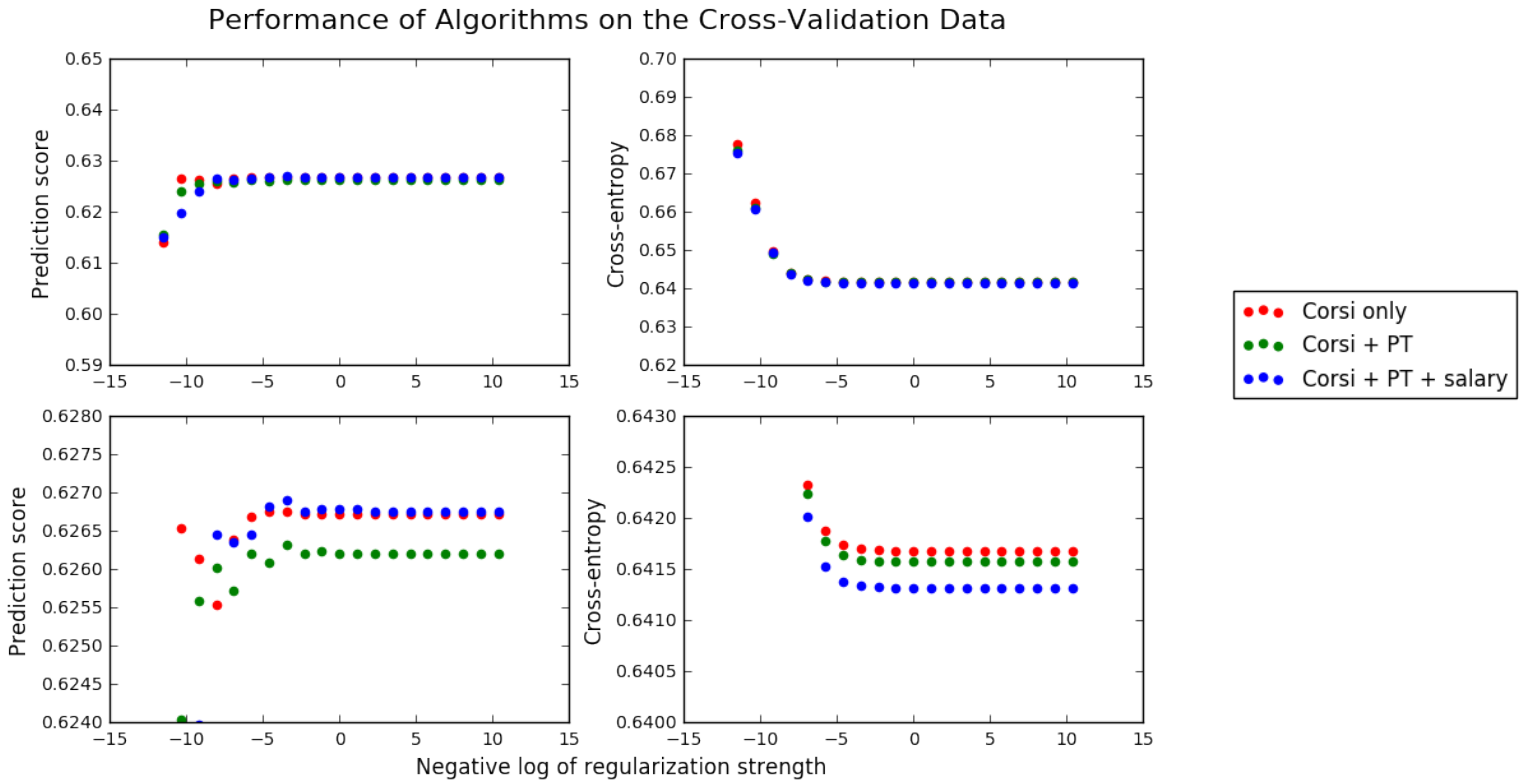
1. Sum of $\sum_i \log(CF\%(X_i))$ for either players X_i home, or away.
2. Average salary of home players, and away players.
3. Average playing-time of home players, and away players.

Notes:

- Sum of log-Corsis used instead of average, to account for power plays.
- Salary and play-time used as proxies of "player quality". It is harder to generate shot-attempts if your opponents are good! And it is easier if your linemates are good.
- Salary is a limited proxy, because young players are underpaid relative to ability.

Training the Crude Logistic Regressor

Result: including playing time and salary improved the cross-entropy very slightly on the cross-validation set. Best prediction score: 62.7%



Sensitive Logistic Regressor

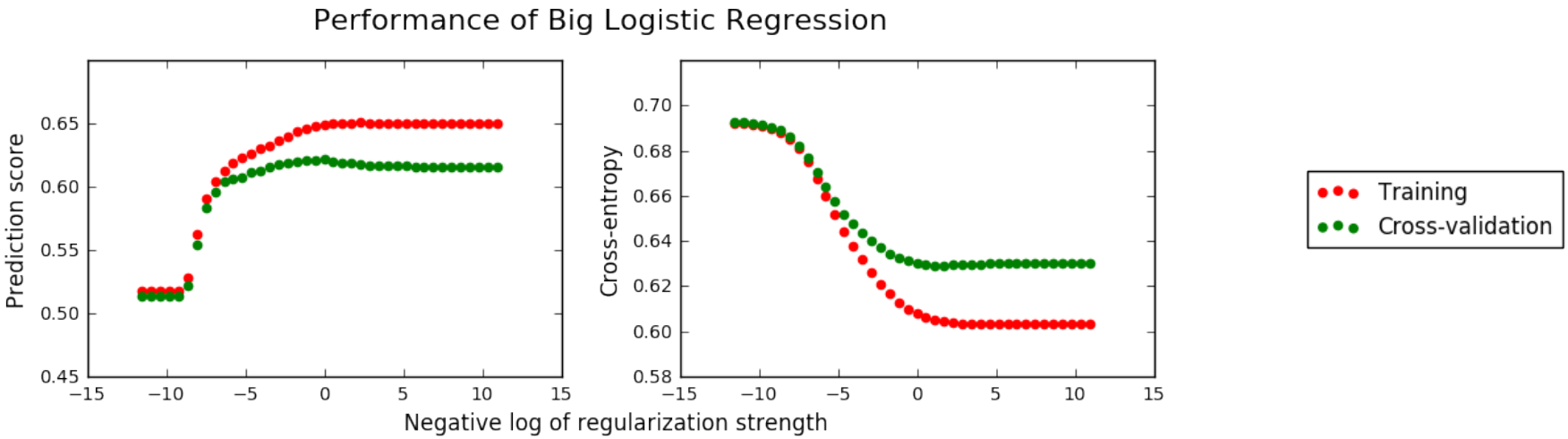
Now: forget playing-time and salary. Let the algorithm decide itself who the the better players are. Salary and PT are decided by coaches and general managers. Goal of data science is to see past possible biases.

Result: cross-entropy on the CV set has improved. But prediction is worse: 62.1%.

Conclusion: perhaps inherent randomness in system prevents us from getting better prediction? Or: algorithm is not good enough?

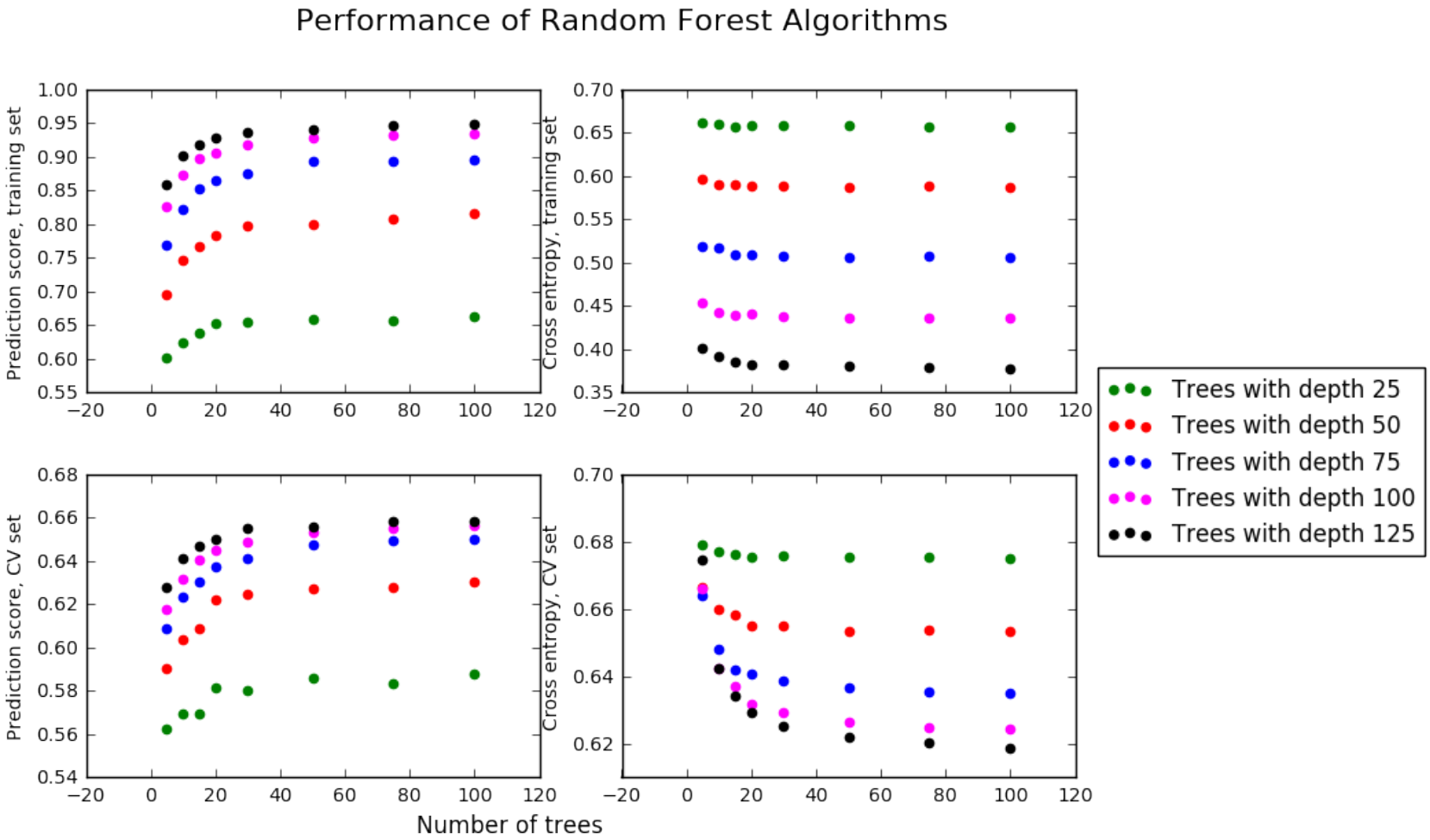
Training the Sensitive Logistic Regressor

Training prediction is better than crude model. But CV prediction slightly worse.



Random Forest Predictor

Let's use a random forest algorithm on the same data set.

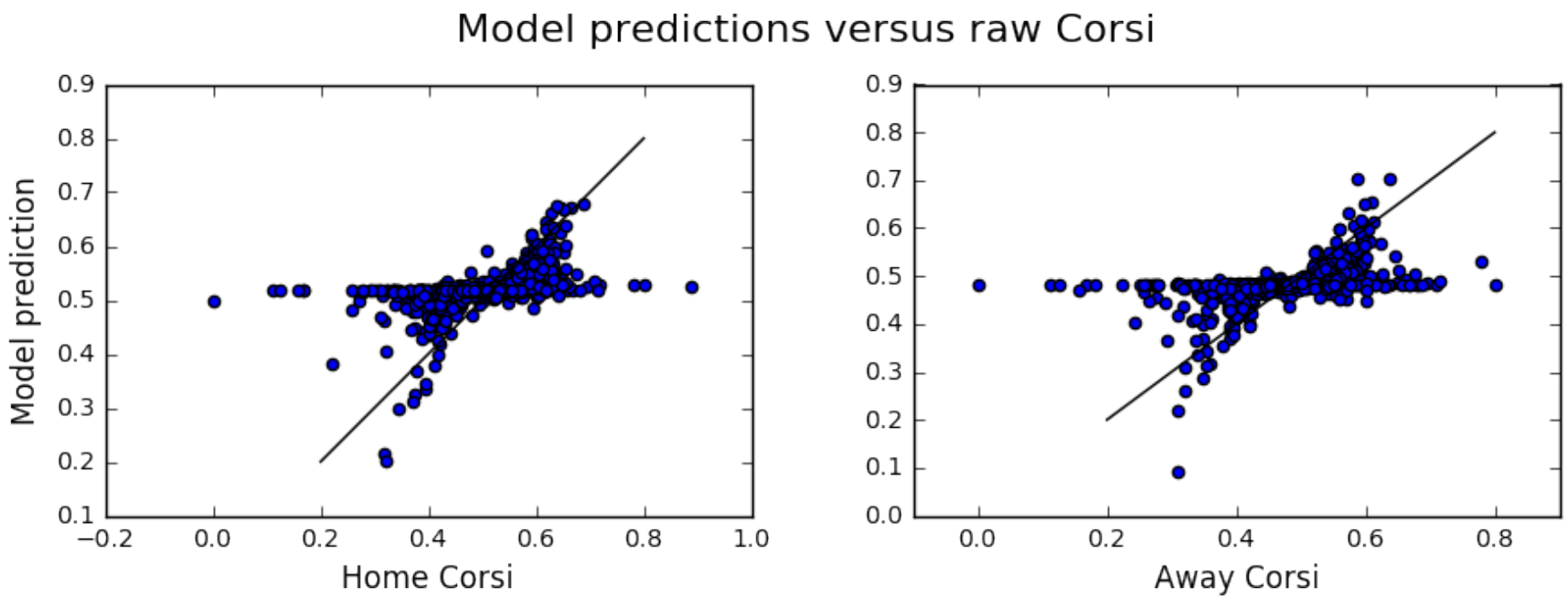


Analysis of Random Forest Predictor

- 125 trees with a depth of 100 seem to be sufficient to optimize training prediction.
- Large difference between training and CV cross-entropy.
- Prediction on CV set has improved considerably: 65.8%
- How to analyze the results and compare to Corsi? Tell the algorithm only player X is on the ice, and ask for a prediction. Artificial but objective.

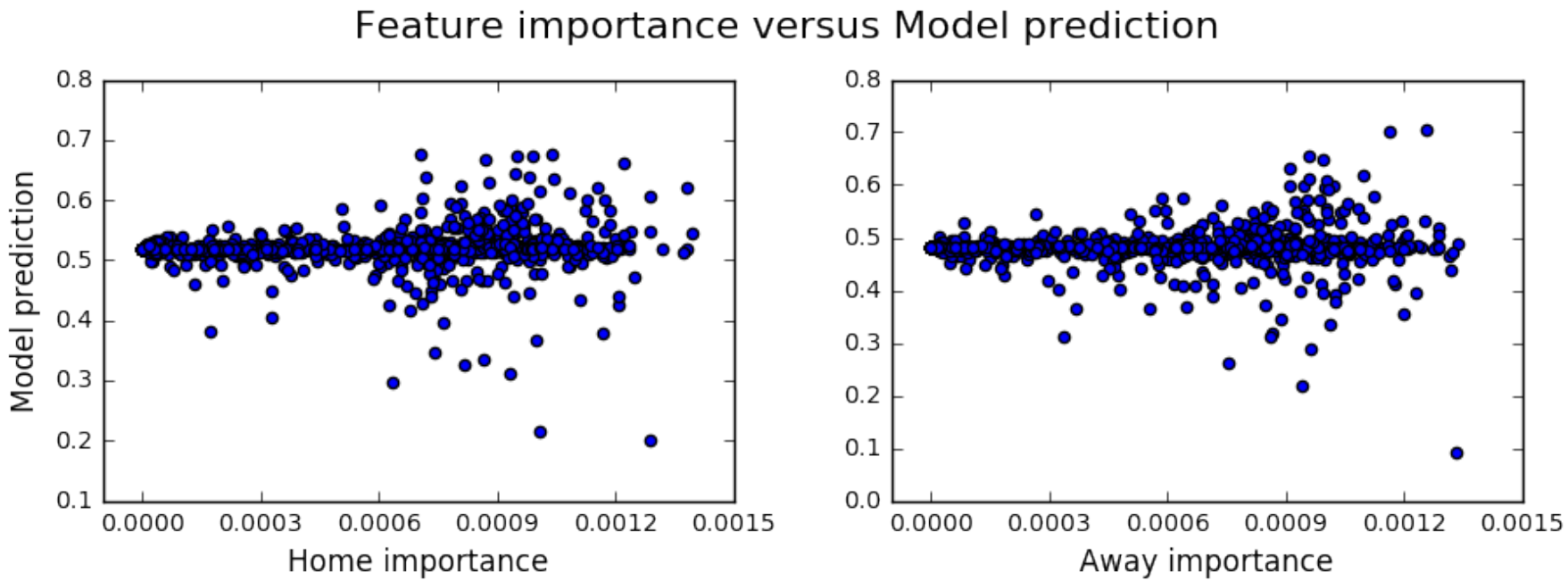
Model predictions vs. Corsi

When we compare model prediction to Corsi, we see that many players cluster around the 0.5 level, regardless of Corsi:



Model predictions vs. Feature importance

Perhaps these players have low feature importance? We can see this is only partially the reason. Extremity of model prediction correlates with importance, but still many important players in the center.



Categorizing players

We can conclude that for many players, their Corsi rating is misleading. A player with high/low Corsi may not actually affect the shot much. We can categorize players with "false" or "robust" Corsi:

- If the model rating deviates from the mean by less than 0.3 times that of his Corsi, we say that his Corsi is **false**. He is either not as good as, or not as bad as, his $CF\%$ suggests.
- If the model rating deviates from the mean by more than 0.7 times that of his Corsi, we say that his Corsi is **robust**. In these cases, the model agrees that a good/bad Corsi reflects the true ability of the player.

Results:

- 536 of 900 players had false Corsis both at home and away.
- Only 24 of 900 players had robust Corsis both at home and away.

Summary of Findings

- We can build a logistic regressor to predict shot attempts with a 62.1% prediction accuracy.
- A better predictor can be built using random forests. The prediction accuracy can be improved to 65.3%.
- To train a random forest model with such accuracy, one needs on the order of 100 trees, at a depth of 125. The predictions do not vary much when changing hyper-parameters slightly.
- The random forest model rates the vast majority of players as average relative to their Corsi. Almost 60% of players had their model rating deviate from the mean less than 0.3 that of their Corsi.

Summary of Findings

- Only 24 players had robust Corsi. The model indicated their Corsi rating was reflective of their actual ability. These 24 players are: Alex Ovechkin, Andy Greene, Calvin de Haan, Dan Boyle, Erik Karlsson, Gregory Campbell, Jarome Iginla, Jarret Stoll, Karl Alzner, Kris Versteeg, Kyle Palmieri, Luke Glendening, Matt Hunwick, Matt Niskanen, Nick Schultz, Nikolai Kuleimin, Oliver Ekman-Larsson, Paul Gaustad, Phil Kessel, Reid Boucher, Rob Scuderi, Ryan McDonagh, Tomas Tatar, and Wayne Simmonds.