J. POHLKAMP-HARTT

EXPECTED GOALS MODELING IN HOCKEY: MODEL SELECTION'S EFFECT ON GEOMETRIC ASSUMPTIONS

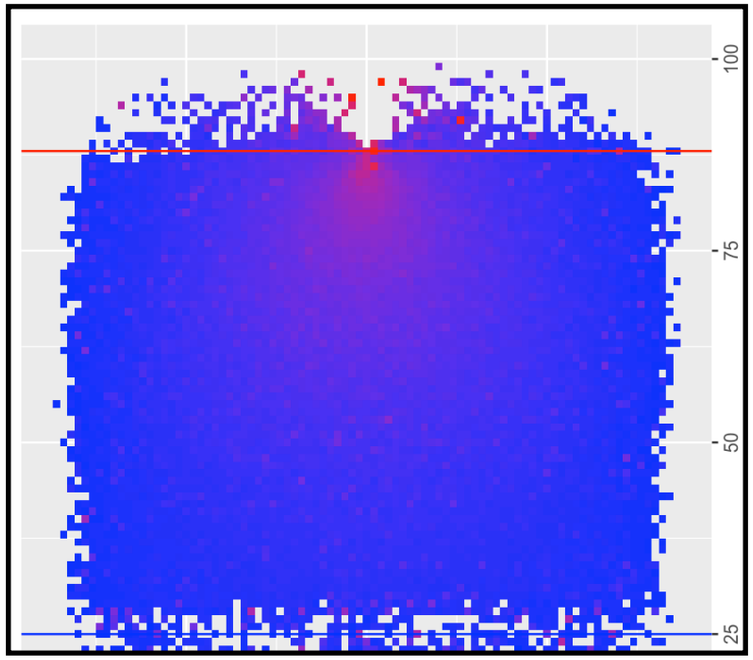
BACKGROUND

- The objective of hockey is to score more goals than your opponent.
- Goals are scored when players shoot the puck on net and it is not stopped by the opposing team.
- Goal scoring is a random process with many factors affecting the probability of success.
- Goals occur infrequently (~5 per game) and do not make a good measure of game level performance.

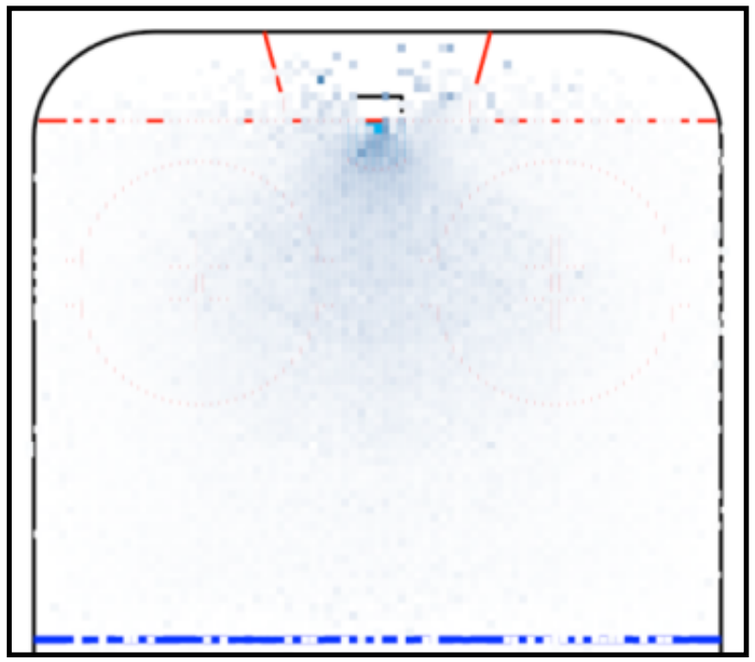
BACKGROUND CONT.

- Shots on net are a poor measure for performance, not all shots have same probability of scoring.
- Expected Goals is the preferred measure. The most common definition of Expected Goals is:
  XPG = P(score|shot on net)*P(on net|shot)
- For simplicity's sake we will focus on P(score|shot on net) today.

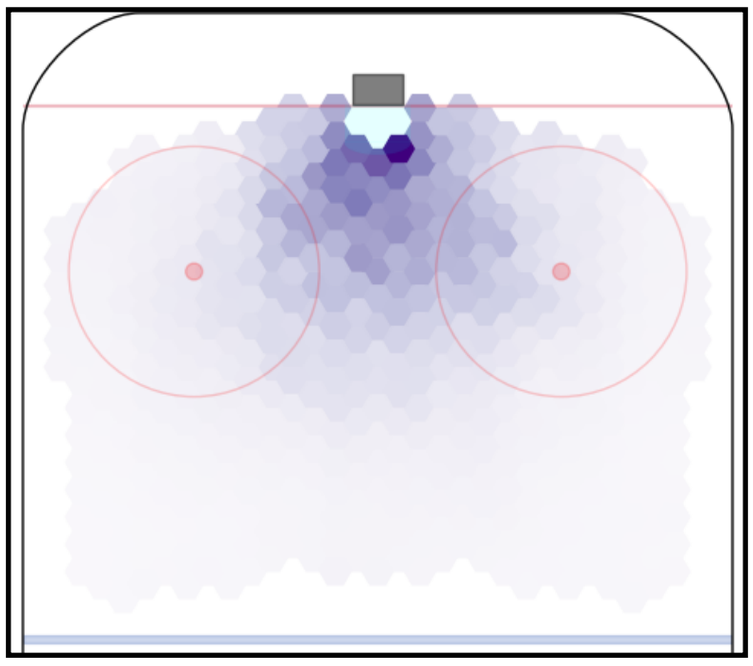
MOTIVATING EXAMPLES



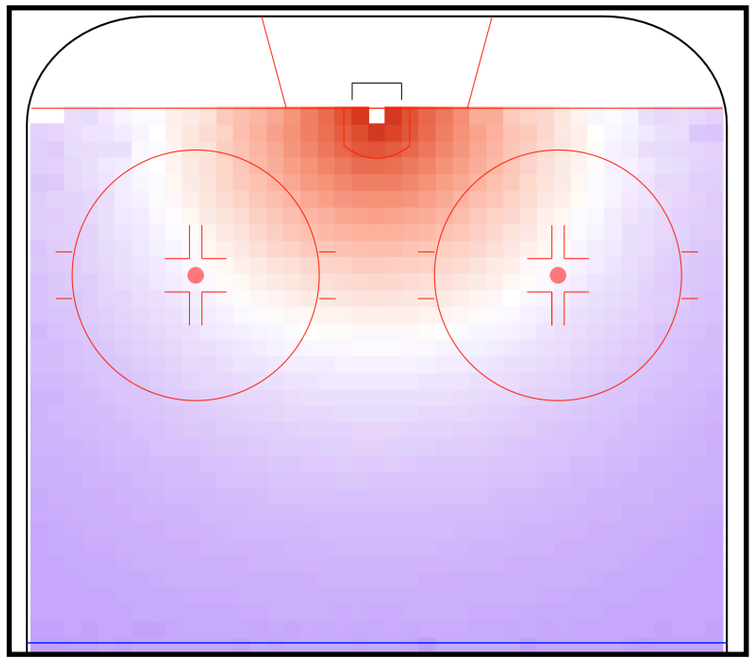
1. Polynomial, red high - blue low



2. XGBoost, blue high - white low



3. Non-linear Regression, blue high - white low



4.Linear, red high - blue low

OUR DATASET

- Proprietary data provided by a third party company, Sportlogiq.
- Data is produced through the use of computer vision (Yolo, etc.).
- Data created is rows of on-ice events with some contextual data.
- Example events are: Shots, Puck Recoveries, Passes
- As an alternative, the NHL currently provides Shot related data through their API.

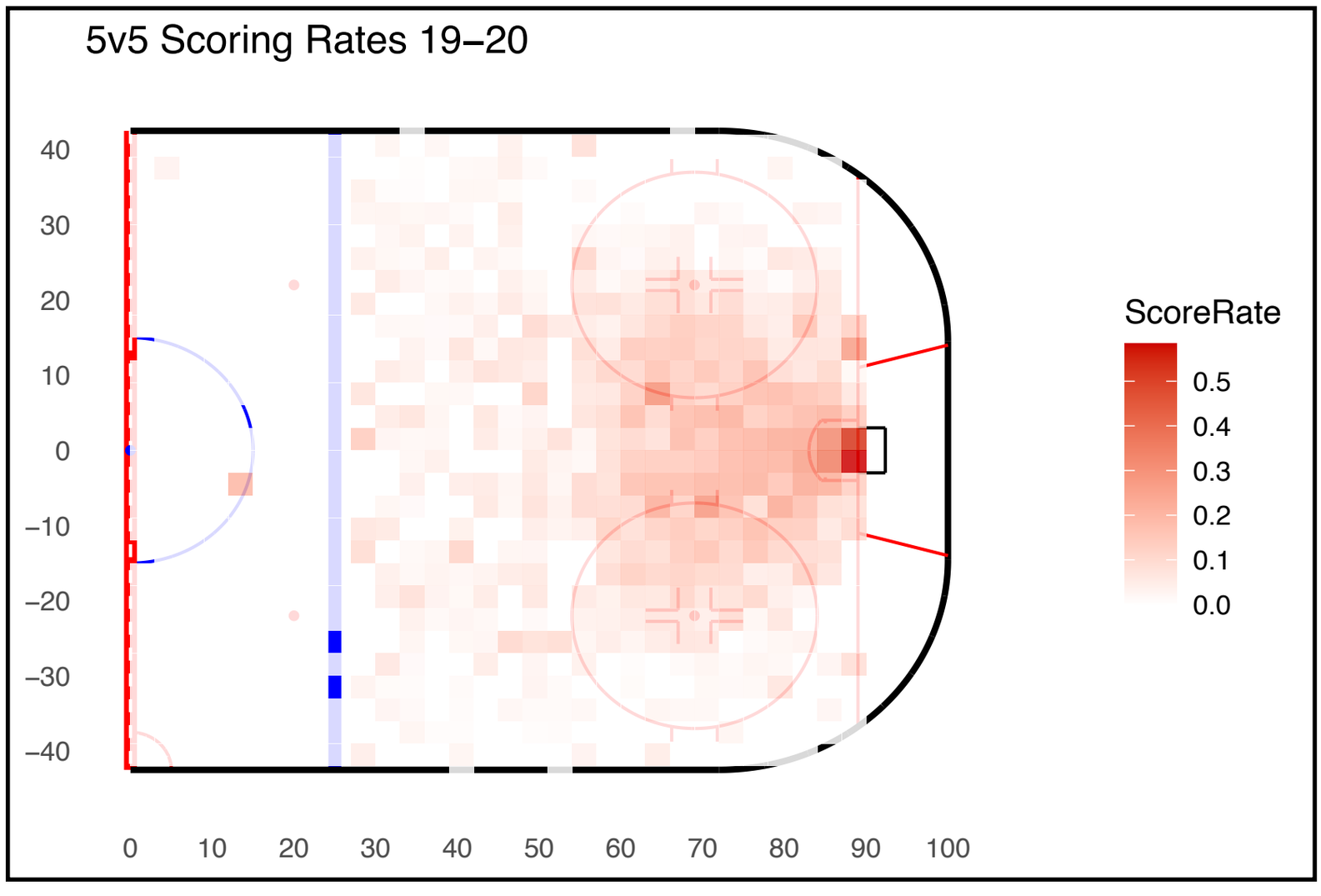
OUR DATASET CONT.

- We will use on-net 5v5 shots data from 2019-2020, n = 43614.
- Our focus is on shot location data. There are an additional 10+ variables commonly used we will not discuss here.

Event	is_goal	x_coord	y_coord
1	0	70	20
2	1	82	-10
3	1	65	2
...	...	...	...
n	0	71	31

OUR DATASET CONT.

- Let's take a look at scoring rates for 19-20:



MODEL SETUP

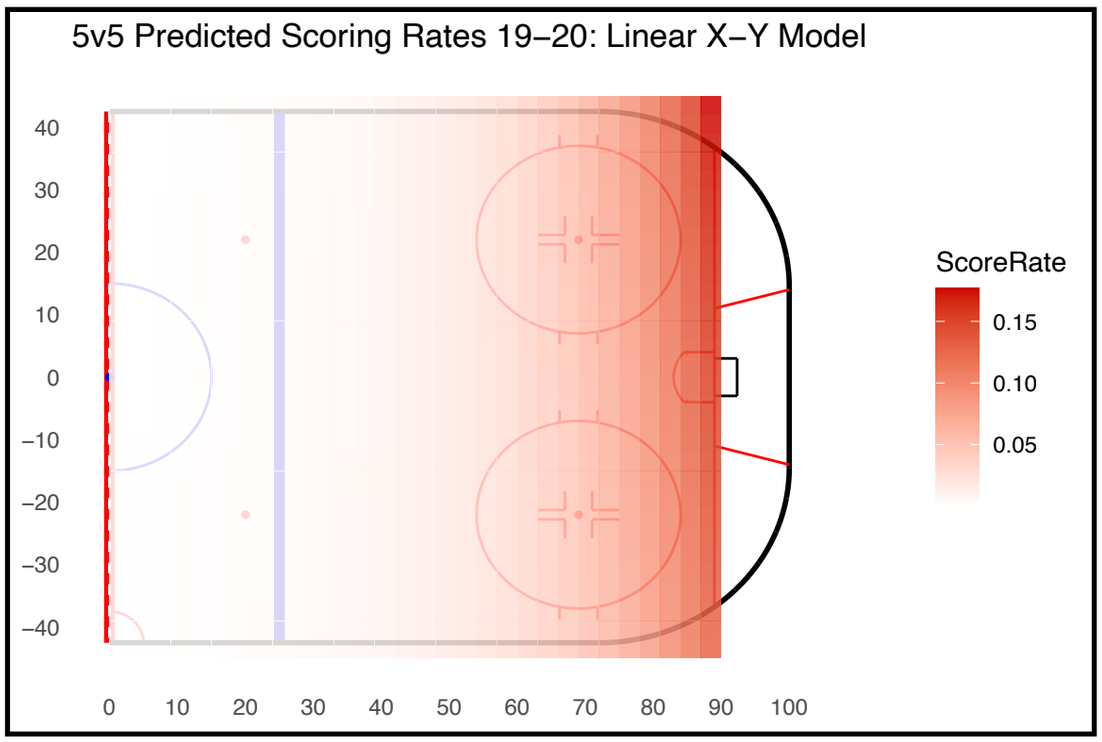
- We want to model the probability of a goal being scored on a given shot location.
  P(Goal | shot location)
- We will treat this as a binary classification setup where we are using location as the independent variables and whether there is a goal as the dependent.

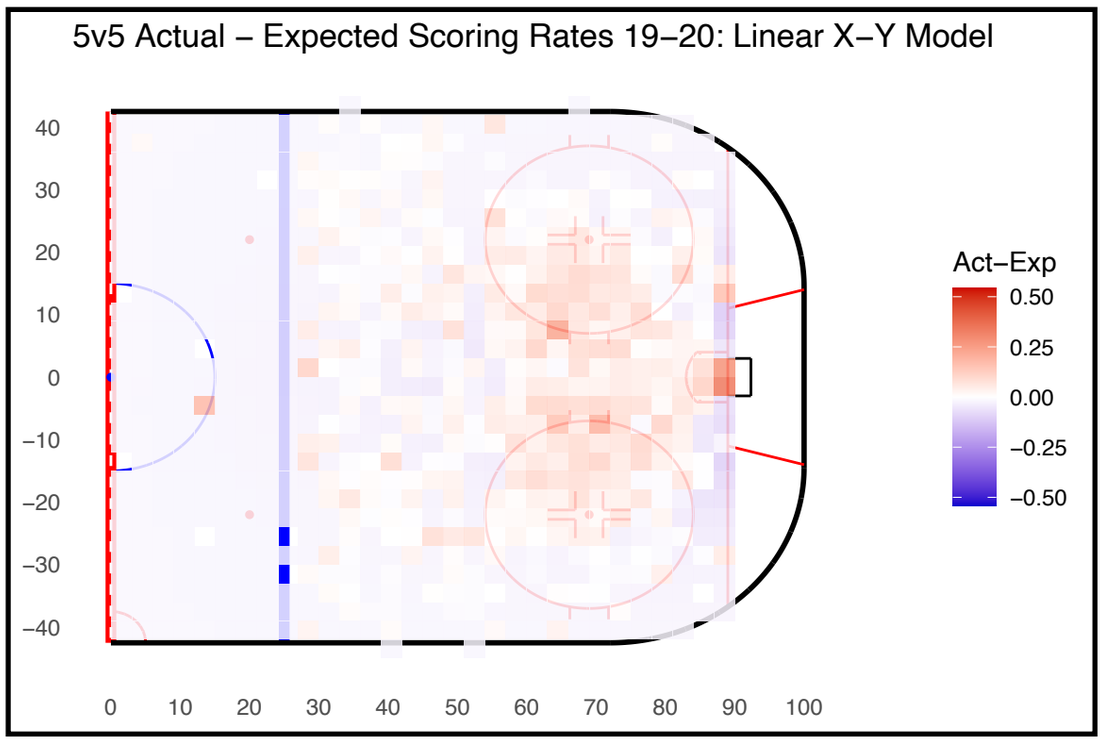
COORDINATE SYSTEM?

- Should we use the raw X,Y data or does it make more sense to use a radial coordinate system focused on the net?
- Does left-right side of the net (angle sign) matter? Or is side unimportant?

NAIVE PARAMETRIC MODELS

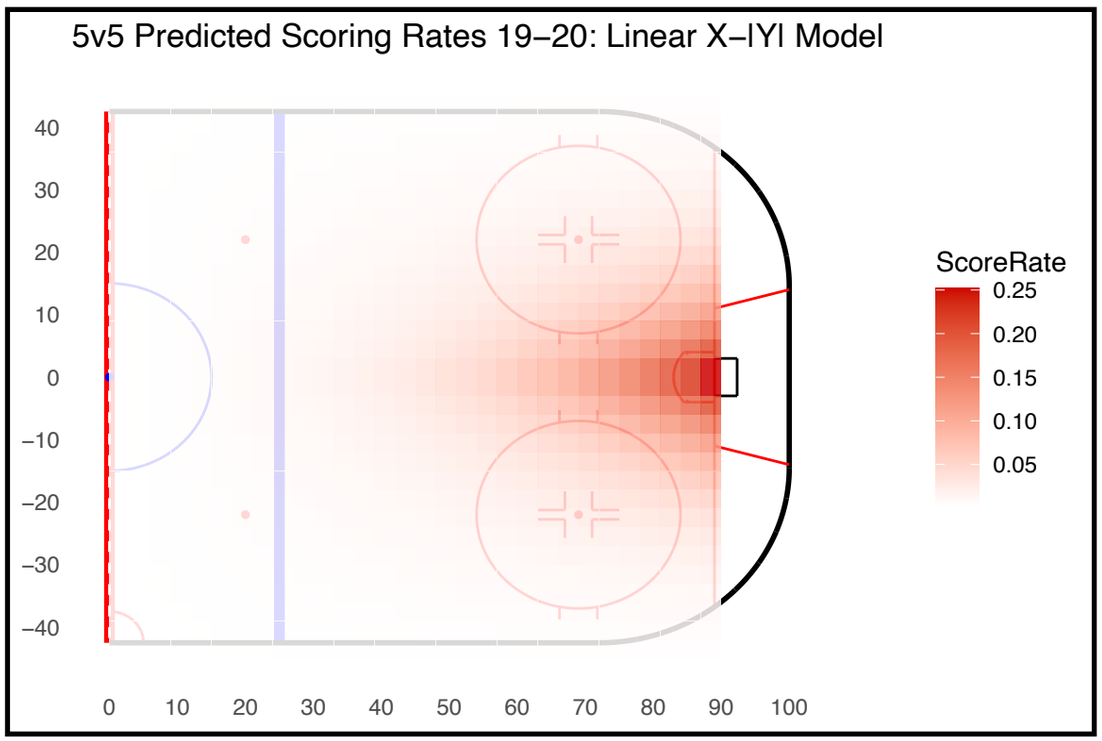
- Start by assuming X and Y have linear effects on scoring probability.
- This produces a decent (AUC = 0.754) fitting model but there is an obvious over simplification.

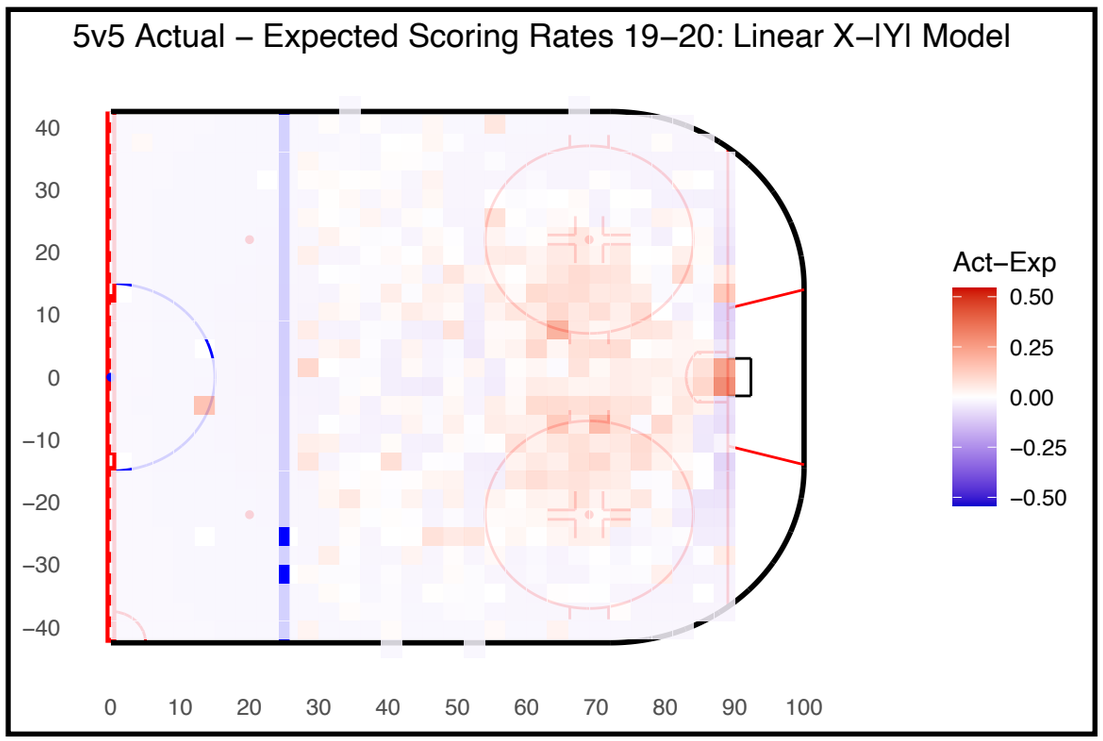




NAIVE PARAMETRIC MODELS CONT.

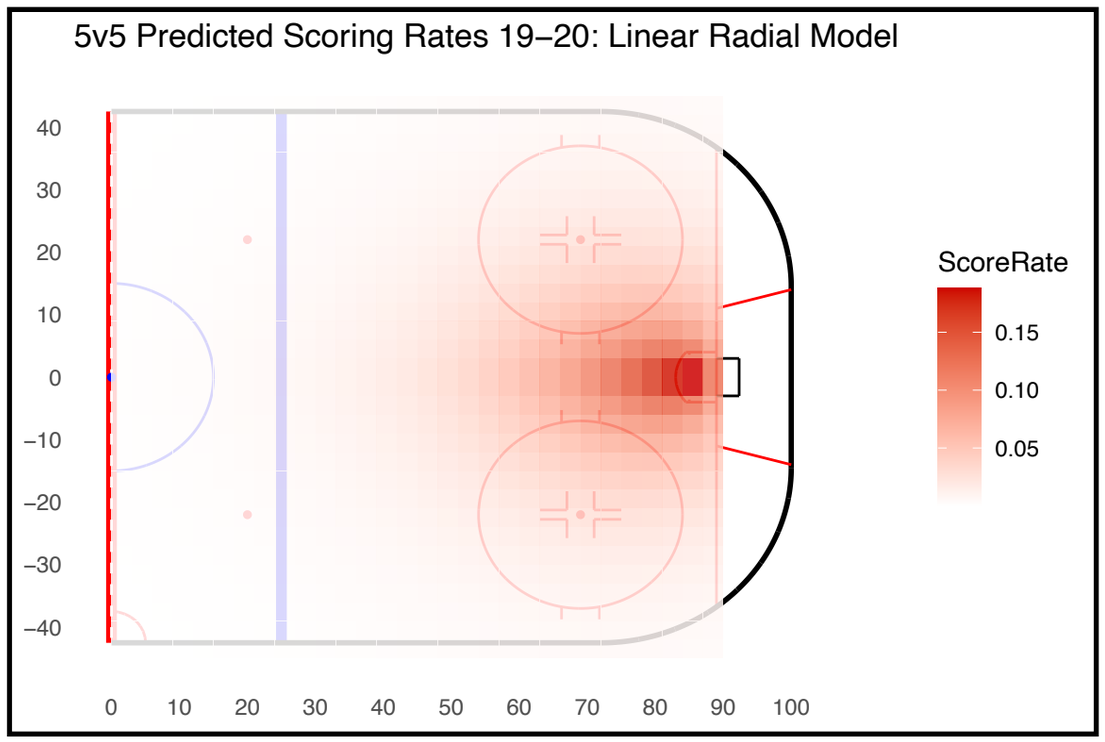
- Now what happens if we use the |Y| rather than Y?
- Improved results (AUC = .807)!

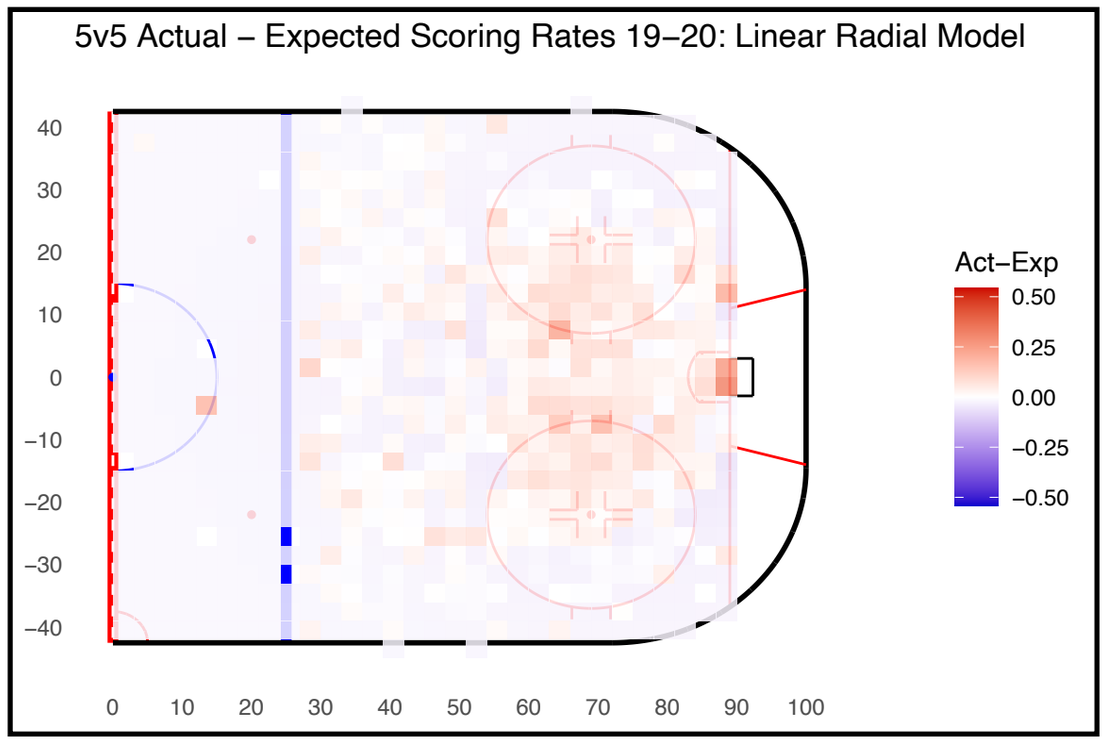




NAIVE PARAMETRIC MODELS CONT.

- Now let's try with radial coordinates and absolute angle.
- Even better (AUC = .813).



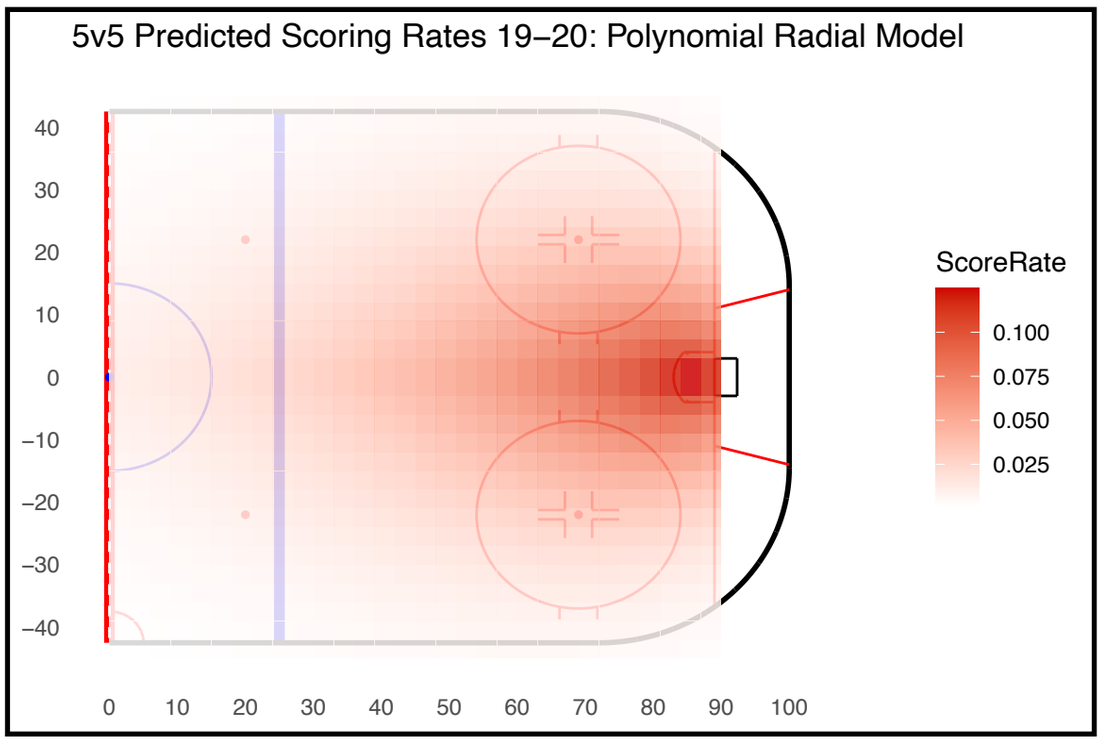


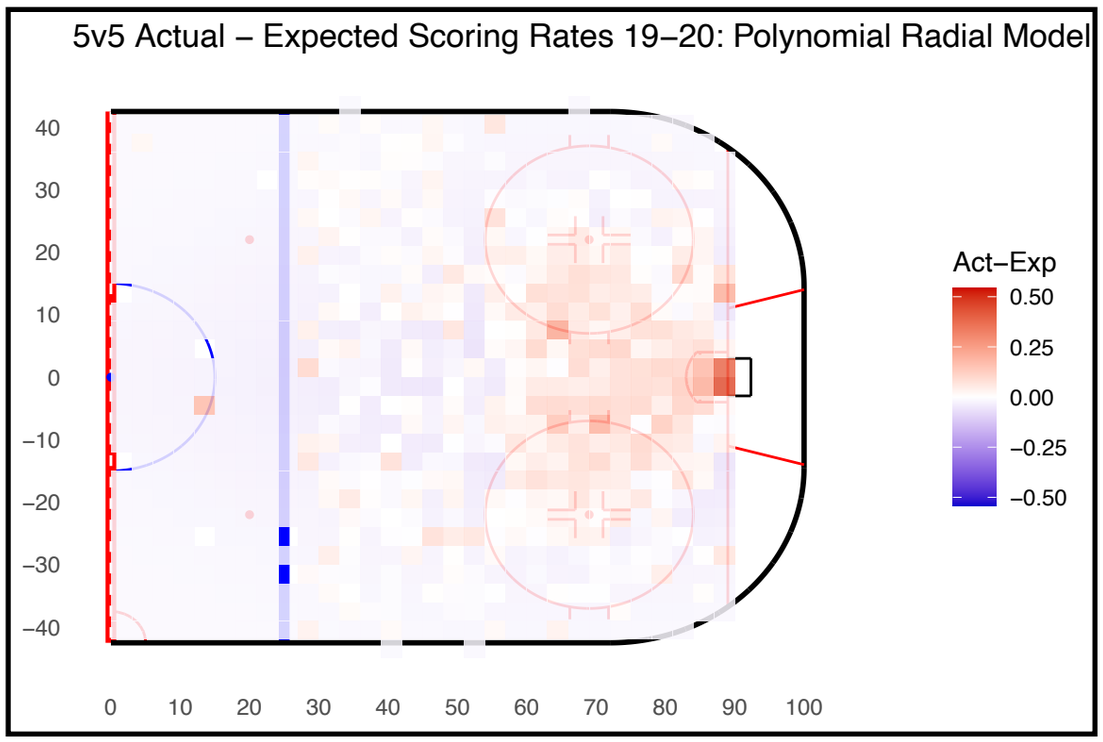
NOTE ON REGULARIZATION

- To avoid some of the overfitting as we introduce more complicated regression models, we use regularization. Elastic-Net Regression is a common tool in sports analytics for this sort of task.
- This method constricts the coefficients of the independent variables and doesn't allow for the coefficients to tune aggressively to training data.
- We use this method for our next two types of models.

POLYNOMIAL PARAMETRIC MODEL

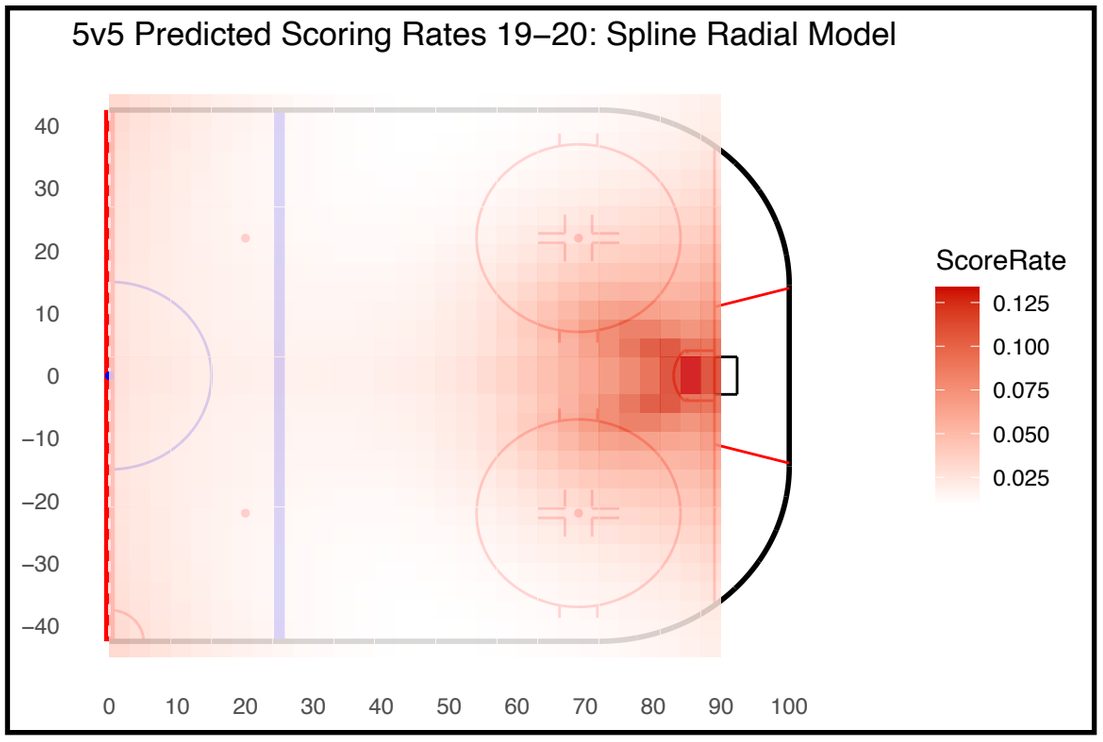
- Continuing with radial coordinates, what if we assume the effects of location follow a 2nd degree polynomial?
- Polynomial effects do not improve our model. (AUC = .811)

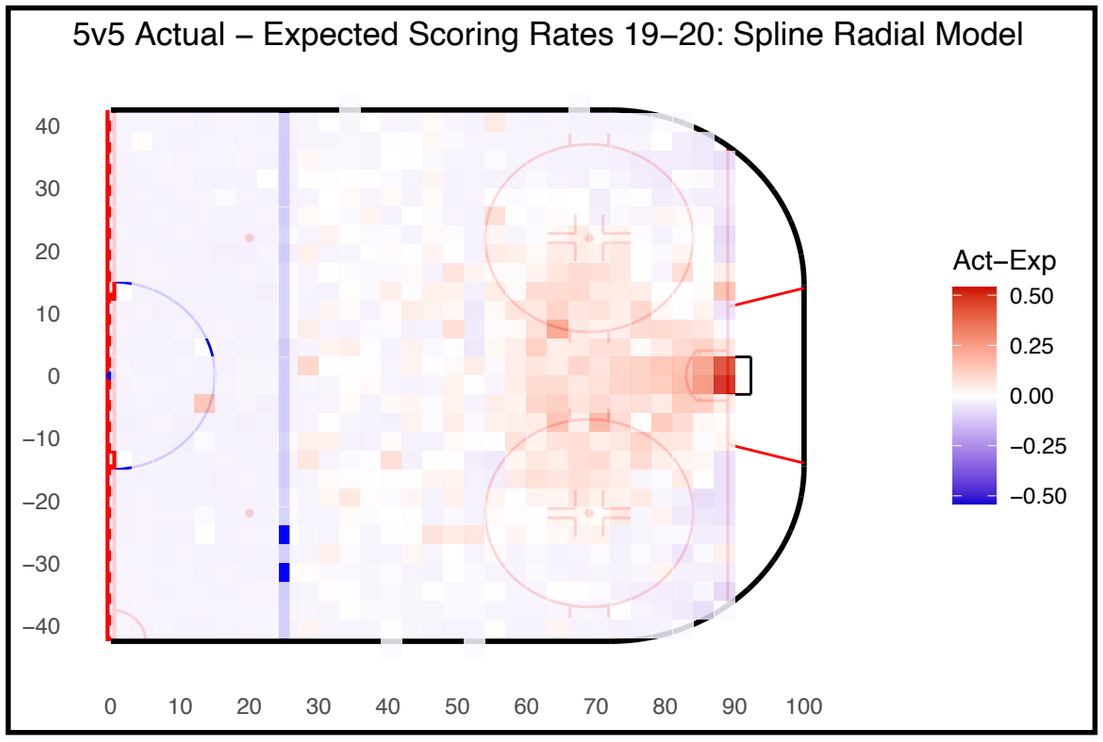




SPLINE PARAMETRIC MODEL

- Now what happens if we use splines? Let's try with 5 knots.
- Slightly worse performance (AUC = .808).



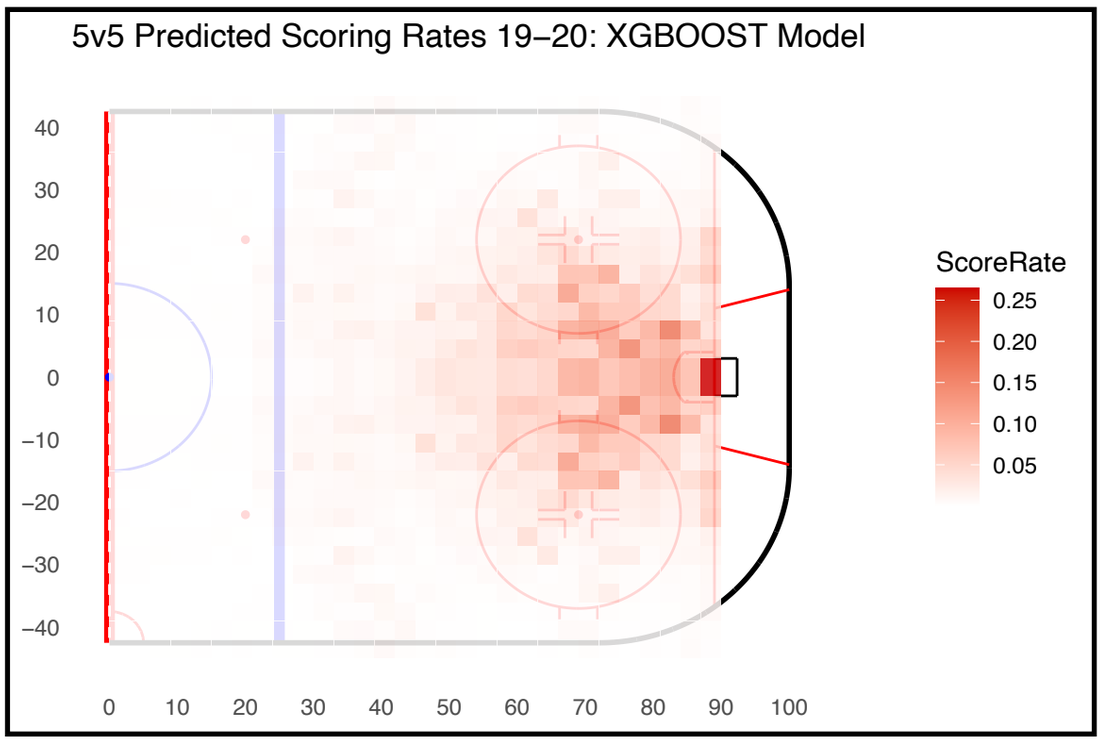


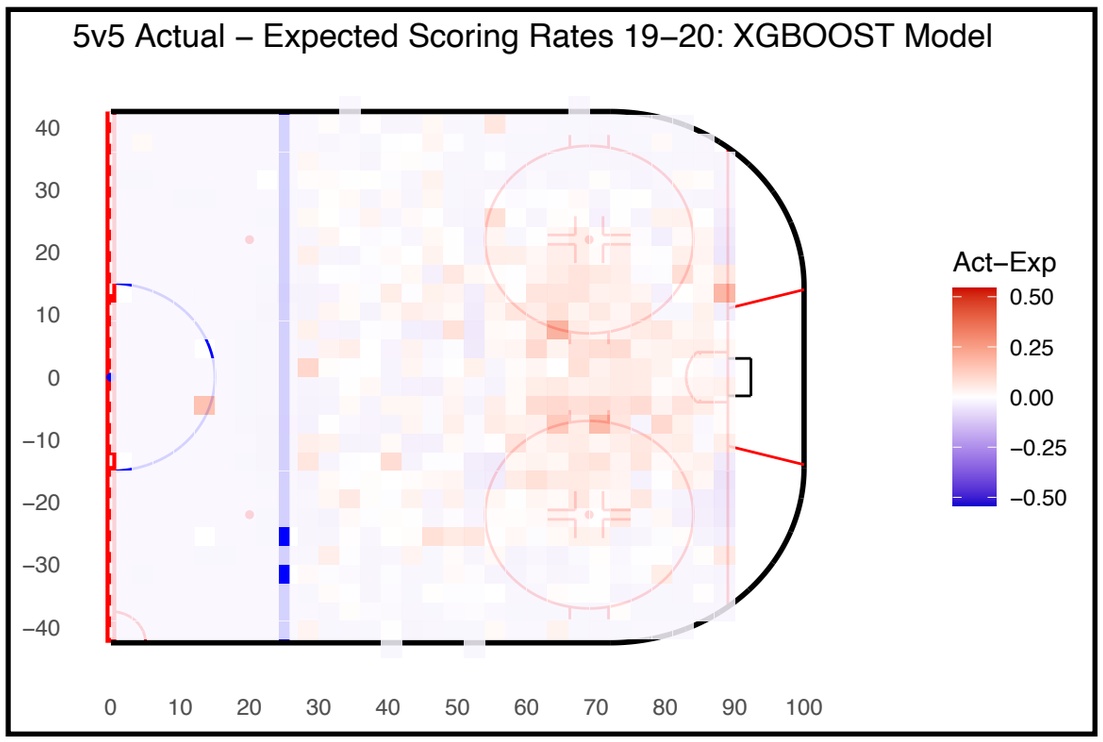
NON-PARAMETRIC MODELING

- There are a variety of non-parametric models that are commonly used, they include: Decision Trees, Random Forests, XGBoost/LightGBM, Nearest Neighbors, Support Vector Machines.
- Non-parametric models assume no shape or interaction between the data, increasing flexibility at the cost of time to train/build.
- We will focus today on XGBoost. XGBoost is basically really fast and flexible Random Forests.

NON-PARAMETRIC MODEL

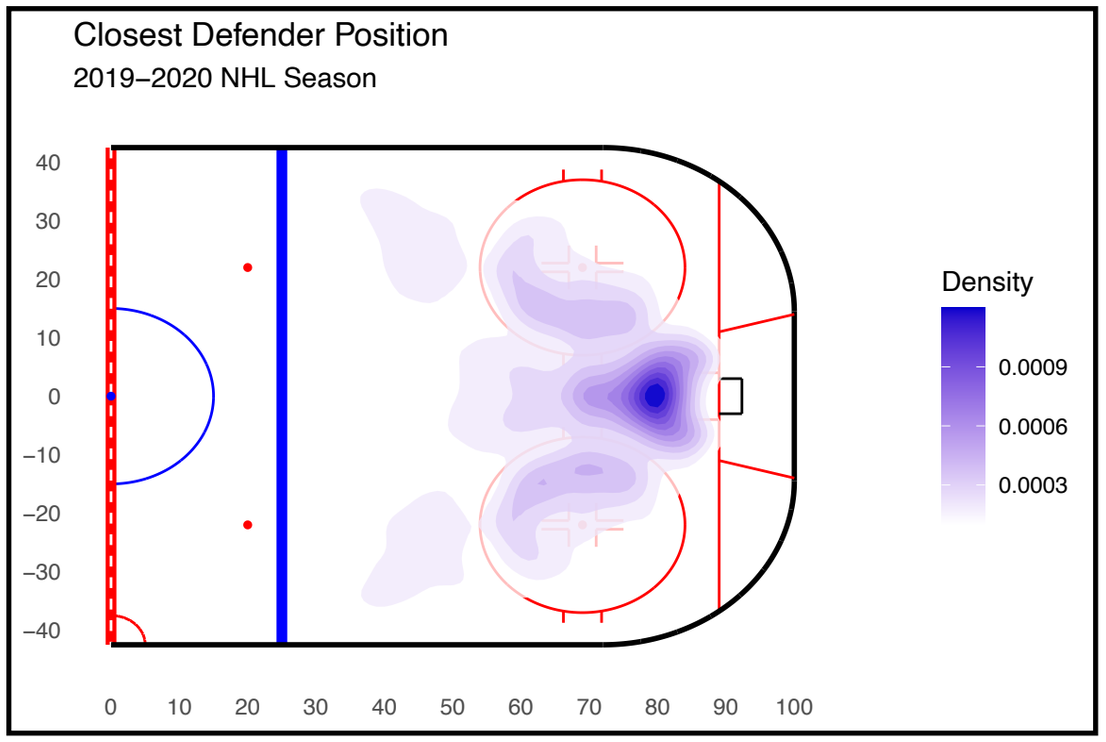
- Using an XGBoost model with |Angle| and Distance as variables, how is our performance?
- Much closer to the home plate pattern, model performance is similar to our past models (AUC = .812).





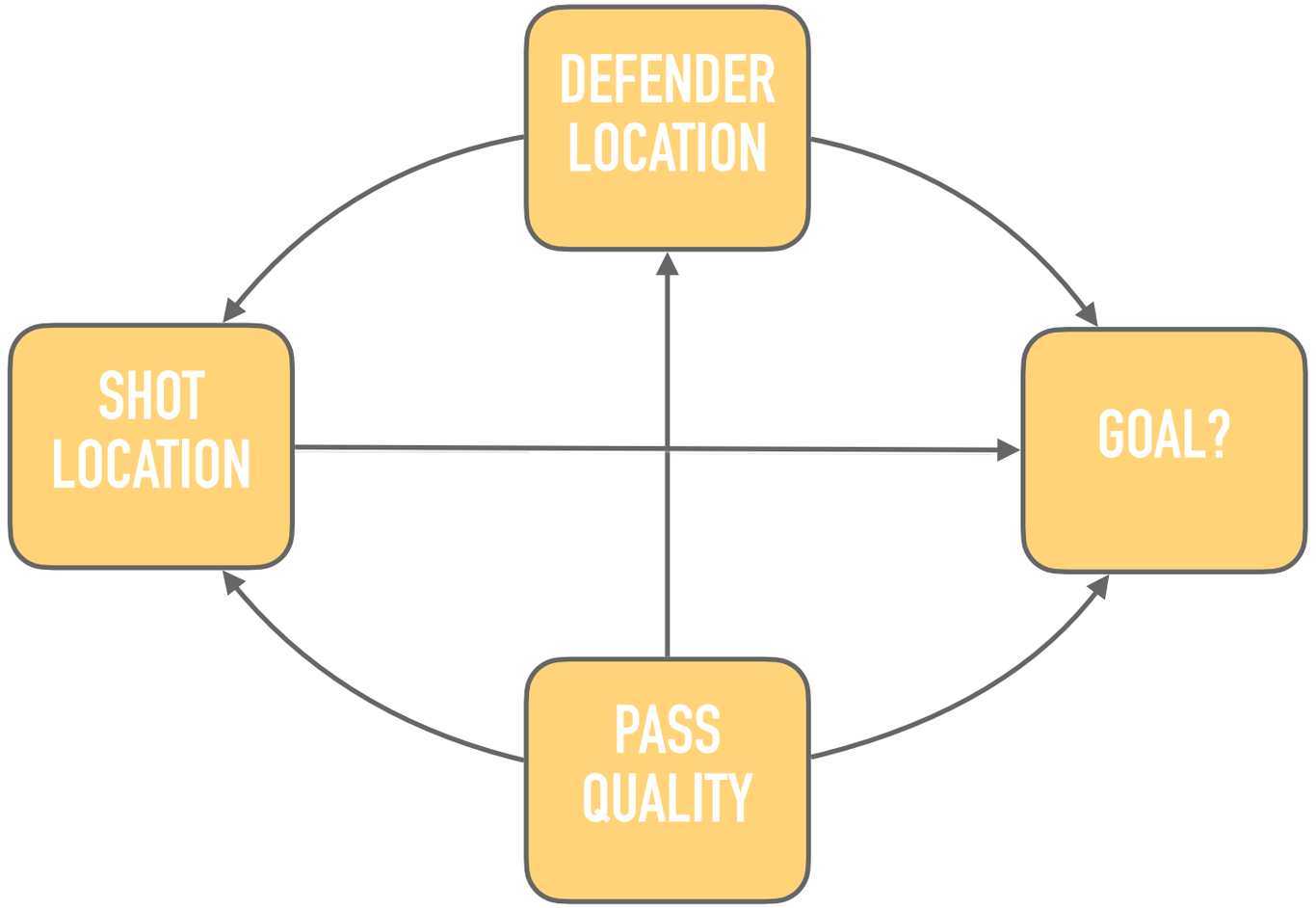
POTENTIAL CONFOUNDING EFFECTS

- Using only location is naive of us. There are a lot of other factors in scoring goals. These may be masked in our location data.
- Getting to better scoring areas is dependent on defender positioning.



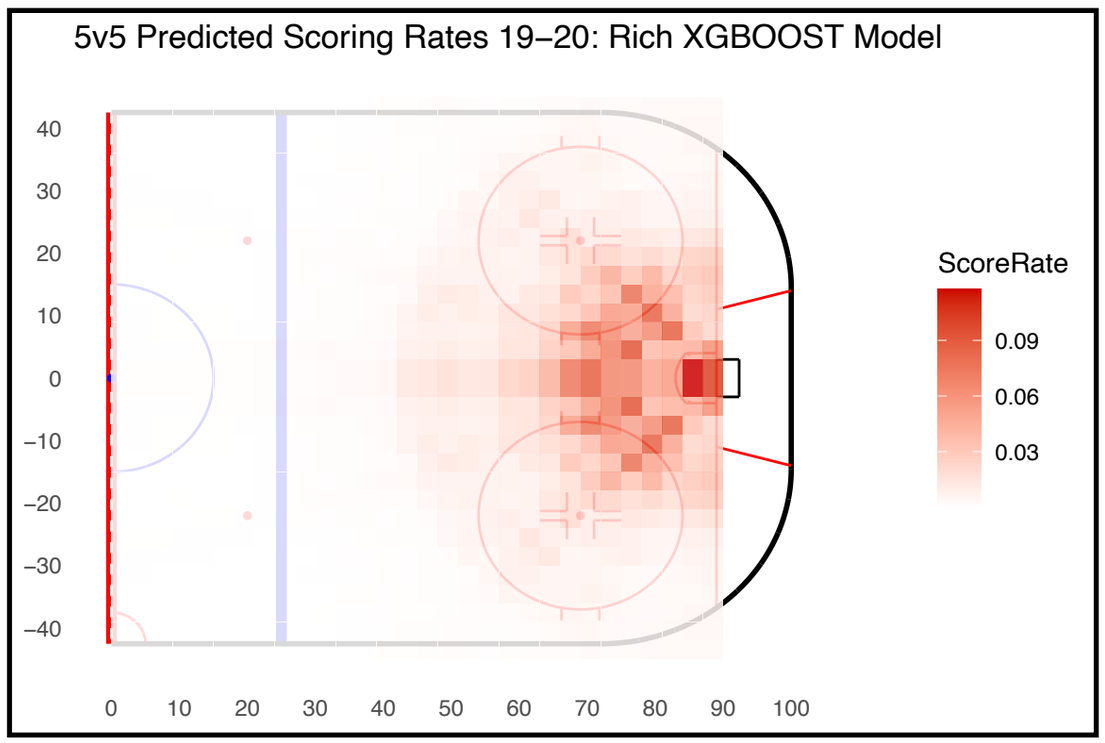
POTENTIAL CONFOUNDING EFFECTS CONT.

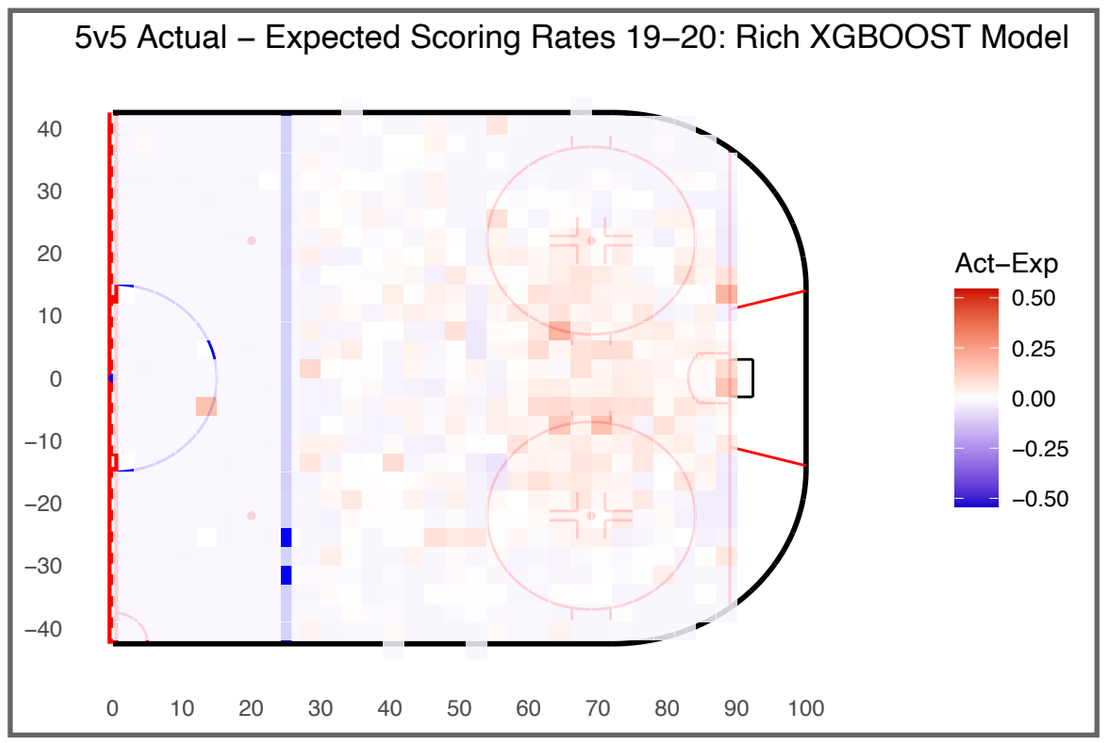
- Another factor that can impact shot quality is passing.
- Directed Graph of the possible confounding relationship:



RICH NON-PARAMETRIC MODEL

- Final model is XGBoost with location, preceding pass, closest defender.
- Gained definition to the home plate pattern, model performance is better than our past models (AUC = .826).





REMARKS

- Lots of room to still improve XPG Models.
- Your choice of model will impact the geometric effects and overall performance.
- Watch out for confounding variables.
- Check your assumptions.
- Get better data, when available.

ACKNOWLEDGEMENTS

- Thank you for virtually attending and Queen's for hosting
- Data provided by Sportlogiq with permission from the Boston Bruins
- Motivating figures pulled from Twitter accounts of public hockey analysts: Micah McCurdy, Josh and Luke Younggren , Matt Barlowe, Alex Novet

- Further Questions? Email me - Pohlkamp.Hartt@gmail.com