

Data Analysis Challenge

Buffalo Sabres

September 15, 2021

1 Introduction

As a data scientist in the NHL, you will often be asked to answer ambiguous, open-ended questions with data – with very limited time to get your response into the hands of a decision-maker.

This data analysis challenge asks a few seemingly simple questions about shooters and goalies in hockey. However, there are many ways to interpret these questions, and many approaches you might take to answer them. The questions are intentionally ambiguous, because we are more interested in your approach to answering the questions than your answers themselves.

2 Assessment

With ambiguous questions, how will you be assessed? We are interested in the following:

- Your technical ability/knowledge to answer a question, solve a problem, etc
- Your ability to clearly explain your answer to a non-technical audience
- Your ability to clearly describe what you did to a technical audience
- How you think about and approach common problems in hockey analytics

Again, we are less interested in your actual answer, and more interested in how you come to your answer, how you present your answer, etc.

3 Data

The file `shot-attempts.csv` contains 81496 rows and 14 columns. Each row is a single shot attempt taken in an imaginary hockey league. Columns include:

- `season`: 1 is the first season, 2 is the second season, and so on
- `game_id`: 1 is the first game, 2 is the second game, and so on
- `game_time`: the game time, in seconds. 0-1200 is the first period, 1200-2400 is the second period, etc.
- `y_coord`: the y-coordinate, where $y = 0$ is the line going from the center of one goal line to the center of another goal line, and $-42.5 \leq y \leq 42.5$; $y > 0$ means that the shooter is on their left-wing side of the ice
- `x_coord`: the y-coordinate, where $x = 0$ is the center-ice red line, and $-100 \leq x \leq 100$; $x > 0$ means that the shooter is closer to the opponent's goal than their own goal
- `shooter_id`: self-explanatory
- `goalie_id`: self-explanatory
- `shooter_hand`: self-explanatory
- `goalie_glove`: the goalie's glove hand
- `rush_situation`: the rush situation, with respect to the shooting team; "Odd" means that the shooter is on a 1-on-0, 2-on-0, 2-on-1, (etc) rush; "Even" means a 1-on-1, 2-on-2, (etc) rush; "Trap" means a 1-on-2, 1-on-3, 2-on-3, (etc) rush; "None" means that the shot was not taken in transition
- `shot_type`: seven self-explanatory categories
- `location_on_net`: the on-goal location of the shot, from the perspective of the shooter; seven self-explanatory categories (e.g. "High Right", "Five Hole", etc)
- `player_strength`: the game-situation, with respect to the shooting team (e.g. "5v4" means that the shooter is on the power play (5 skaters, 1 goalie), and the goalie is on the penalty kill (4 skaters, 1 goalie)).
- `shot_result`: goal, save, miss, block

4 Your Answers

How To Format Your Answers: However you see fit.

What To Include In Your Answer: For each question below, include the following in your answers:

1. **Non-Technical Answer:** Provide your answer as if it was being shared directly with a GM, assistant GM, or head scout (i.e. to an intelligent but non-technical audience). Assume that your audience is numerate (can read and understand stats, tables, charts, etc), but does not have a background in data analysis (e.g. does know what regression is, does not write code, etc).
2. **Technical Description:** Provide a detailed, technical explanation of how you came to your answer (i.e. where the audience has a background in data analysis, e.g. another hockey data scientist). Including well-documented code is recommended.
3. **More Time?** Describe what else, if anything, you would do if you had more time to answer the question.
4. **More Data?** Describe what, if any, additional data you would want in order to better answer the question.

5 Questions

1. Who are the best (and worst) shooters?
How confident are you in your answer(s)?
2. Who are the best (and worst) goalies?
How confident are you in your answer(s)?
3. Which goalies will perform best next season (i.e. in season 3)?
How confident are you in your answer(s)?

A Bonus Questions

Below are three optional questions that you can answer if you have time. You can answer these however you want (i.e. you do not need to follow the directions in Section 4).

Note on assessment: Not answering these questions will not hurt you. Answering them well will help you. Answering them poorly will hurt you.

- a. What does it mean to be a “good shooter”? **Discuss.**
- b. What is the difference between questions of the form “who performed best in the past” vs. “who do you predict will perform best in the future”? From a technical perspective, how does your method for answering these questions change, if at all? **Discuss.**
- c. Suppose you had this data, but for 20 seasons instead of 2. **Describe how would you answer the following question:** *Does a goalie’s ability change over time?*