We'll Do It Live!

Open-Sourcing the Sports Analytics Hiring Process

Sam Ventura
VP of Hockey Strategy & Research
Buffalo Sabres

October 30, 2022

github.com/sventura/Hiring-For-Sports-Analytics

Goals and Disclaimers

Goals:

- **Resource**: To provide a resource, from the team's perspective, for how to conduct the sports analytics hiring process, and to set baseline expectations for sports analytics job applicants
- **Open-Source**: To be as open as legally allowed about this *`github.com/sventura/Hiring-For-Sports-Analytics`*
- **Honesty**: Some things worked well, some things didn't; I hope others can learn from and improve on this process

Disclaimers:

- I am not an expert
- I am looking for feedback
- There is a limit to what I can share here
- Circumstances influenced my process

General Principles of Hiring Process

Rigor: Prepare targeted questions, ask difficult follow-ups

(examples to follow / on GitHub)

Communication: Set the scene, state my purpose, be flexible

(examples to follow / on GitHub)

Respect, Empathy: Offer specific feedback, be cognizant of the stakes

Sports Analytics Interview Framework

1. **Resume Review** (300+ candidates):
   - Wide net approach, better to overshoot than miss a star candidate
   - Looking for: programming experience, modeling experience, sports experience a plus

2. **Technical Assessment** (30-50 candidates):
   - Essentially a take-home exam
   - Data Scientist: candidates completed a “Data Analysis Challenge”
   - Data Engineer: completed a “Data Engineering Assessment”

3. **Technical Interview** (6-15 candidates): More on this soon

4. **Final Round Interviews** (2-5 candidates)
   - Second technical interview
   - Hockey-specific interview
   - Standard behavioral interview
   - Misc. extra conversations

Candidate Diversity

I am not an expert:

- Relied heavily on our HR department for guidance
- Systematic review of resumes to reduce risk of unconscious bias

HR posted the positions to these job boards:

- Minorities in Sports Business Network
- Women in Sports Tech
- Black Sports Professionals
- HBCU job boards (via Handshake)

Resume Review Checklist (Data Scientist)

Candidates were given a score for each of the following categories:

1. **Research Experience**: experience working on long-term, research-oriented projects, ideally with complex data
2. **Open-Source**: publicly available research, GitHub presence, can evaluate technical expertise via GitHub, personal website, etc; experience with reproducible research; production-level coding
3. **Programming**: data analysis w/ code, many languages/tools/etc
4. **Sports**: experience working in sports, experience playing sports, strong hockey/sports knowledge, etc
5. **Education**: Degrees or certifications are relevant for the position
6. **Recommendations**: Colleagues reached out w/ recommendation

Goal: Checklist will help reduce risk of unconscious bias

Data Analysis Challenge – Data & Questions

Data: ≈ **80,000 shot attempts** with detailed shot information
(e.g. location, shooter, goalie, shot result, etc)

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)?

Data Analysis Challenge – Assessment Criteria

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.

Data Analysis Challenge – Include In Answer

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.

Data Analysis Challenge – What I Was Looking For

First, **their analysis had to be reproducible, with code submitted**

Second, I'm looking for "**statistical thinking**" in their answers, i.e.:
an attempt to **account for context** and (ideally) **quantify uncertainty**.

Among those with reproducible code, three levels of solutions:

- **Simple arithmetical analysis** of shooting/save percentage
  (most submissions did this, one passed)
- **Controlling for contextual factors with statistical model**
  (e.g. by modeling goal probability with logistic regression)
  (10-ish submissions did this, all passed)
- **Advanced modeling**, e.g. mixed effects models, fitting shooter and goalie effects, posterior distributions of shooter talent, etc
  (two submissions did this, both passed, both were finalists)

Bonus for creative approaches, exceptional writing/data viz, etc

Data Analysis Challenge – Example Answers

```r
## Add in some smooth terms for distance and angle
shots_mod1 <- gam(isGoal ~ 1 + shot_type + off_wing + rush_situation +
                    player_strength + RebAngVel + s(dist) + s(angle),
                  data = dat, family = "binomial", method = "REML")

## Interact distance and angle (preserves their main smooth terms as well). Use
# a tensor product because they are on different scales
shots_mod2 <- gam(isGoal ~ 1 + shot_type + off_wing + rush_situation +
                    player_strength + RebAngVel + te(dist, angle),
                  data = dat, family = "binomial", method = "REML")

## Remove off_wing
shots_mod3 <- gam(isGoal ~ 1 + shot_type + rush_situation +
                    player_strength + RebAngVel + s(dist) + s(angle),
                  data = dat, family = "binomial", method = "REML")

## Compare AIC
AIC(linear_mod, shots_mod1, shots_mod2, shots_mod3)
```

Data Analysis Challenge – Example Answers

1.2.3 Incorporating Shooter Effects

To account for shooter identity in our xG model, we could use the `gamm4::` package and simply add a random effect term for shooter identity to the model formula used in `shots_mod3`. This works, but it can take some time to run the model. For this analysis, I will use the logit transformation of our predicted probabilities from `shots_mod3` as an offset in a mixed model and then incorporate shooter identity as an intercept-varying random effect post-hoc, without the need to fit everything simultaneously. It's important to note that by doing this, we are assuming independence between the shooter effects and the other variables in the model. (A benefit of this shortcut, beyond tractability under a 24 hour deadline, is that it grants the option of constructing the underlying xG model with your algorithm of choice)

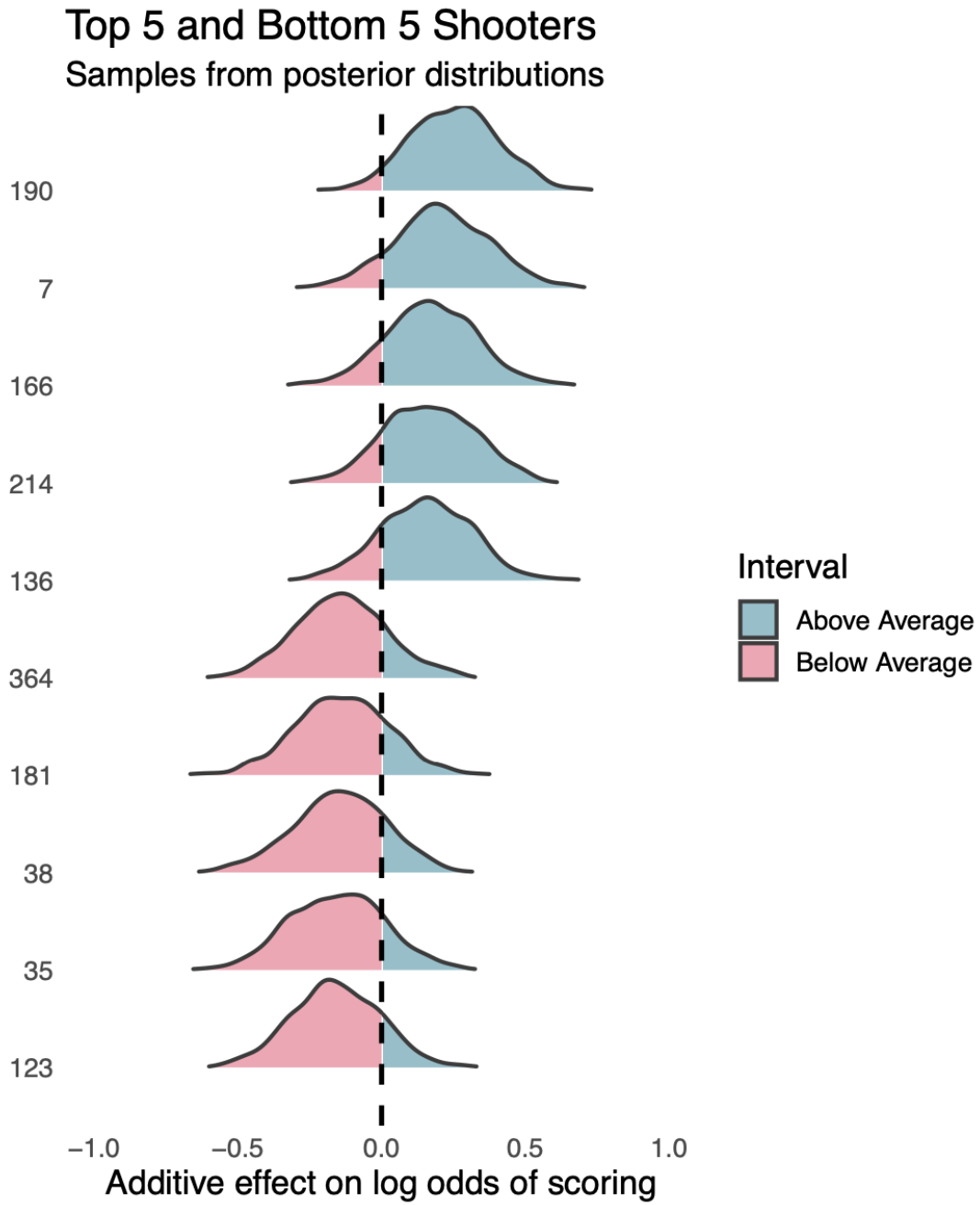
```
## Add predicted log-odds of scoring to the data
dat$shots_odds <- arm::logit(fitted(shots_mod3))

## Fit a GLMM with lme4
# I won't use their estimate here, but I'm also adding goaltenders as a control
shots_mer_mod <- glmer(data = dat, isGoal ~ 1 + offset(shots_odds) +
  (1|shooter_id) + (1|goalie_id), family = "binomial")
```

Extracting the shooter effects and plotting the top 5 and bottom 5

```
## Extract random effect estimates
shooters <- ranef(shots_mer_mod)$shooter_id %>%
  rename(eff = `(Intercept)`)
```

Data Analysis Challenge – Example Answers



Technical and Final Round Interview Demo

Data Engineering Assessment – Problem Set Up

API Documentation: Pointed them to public documentation for the Elite Hockey Prospects API, and asked targeted questions:

1. How would you **gather, store, & update** data from the API?
2. Pick a few tables and **describe a schema** for them
3. How would you deal w/ **name-standardization / record linkage** issues across Elite Prospects and other data sources?
4. How would you handle the **streaming nature of hockey data**?
5. **What issues to you anticipate encountering** along the way?

Data Engineering Assessment – What I'm Looking For

Reframe non-technical instructions into data engineering language: I communicated the problem using non-technical language, but I am looking for them to reframe it with technical details in their answer.

Pipeline-oriented thinking and focus on automation

Focus on testing: Anticipation of failure points, testing frameowrk in place, anticipating issues, etc

The best submissions used diagrams to represent their proposed schema; paid attention to data types; discussed pipelines, automation, testing, timing, etc; discussed potential issues they may run into, how to solve, etc; gave sensible answers to the record linkage question

(example submissions on GitHub)