Breaking In:

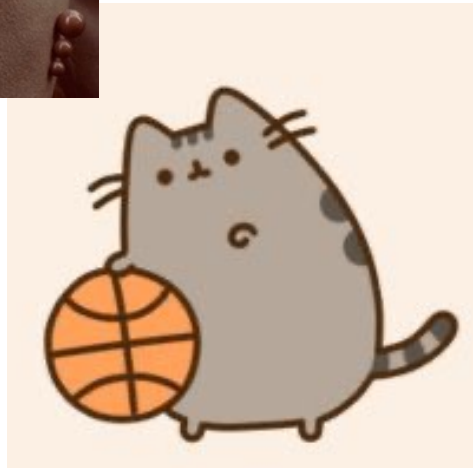
Advice and Example Problems for Professional Sports Interviews

Caitlan Krasinski
Tegan Bunsu Ashby
Josh Pohlkamp-Hartt
Ethan Douglas
Nick Wan

JSM 2023

Tegan Bunsu Ashby

Assistant Director, Software Engineering
Philadelphia Phillies



Full-Stack SWE Prompt

Using the provided dataset of game box scores from a single season (with player and team metadata), develop an application that displays the “Four Factors” for a given player (eFG%, TOV%, ORB%, DRB%, and FT%).

Your project should include…

- A frontend that displays:
  - A player profile
  - An interactive data visualization (e.g. D3, visx, react-vis, plotly)
- A backend with a database and an API
  - E.g. Postgres + FastAPI
- A README with clear instructions for how to run your project locally

Tips

- As is often the case, the data sometimes contains corrupted or malformed values. Do your best to handle these and feel free to ignore missing values, but make sure they don’t interfere with the calculation or presentation of the final result
- You’re welcome to enhance your app with any publicly available data, such as scraped play-by-play data

What makes this project a bummer

A lot of NBA teams are looking for some signal that you already know and like basketball analytics…so even if you're a really strong engineer, there's a certain amount of (sport-specific) statistical literacy that's required upfront. You're also interviewing for a relatively small R&D team, so you'll need to be a generalist (i.e. a full-stack developer data scientist analyst who can also rebound at practice in a pinch).

- It's a LOT of work
- It's pretty open-ended (half intentional, half lazy)
- Non-SWEs (technical and not) may be evaluating your project
- You'll be judged on both the technical and basketball merits of your project and it's unclear how both of those are weighted
- It doesn't test for parts of the SWE role that aren't obviously related to "basketball product" for external stakeholders, e.g. handling tracking and pose data, data pipelines + MLOps – even though these are some of the greatest technical needs for the future!

It’s why the WISD Hackathon exists

2023 SYMPOSIUM + DATA COMPETITION

WOMEN IN SPORTS DATA

HOOP THERE IT IS

Hackathon: NBA Tracking Data

This year, we're balling. Participants will work on projects featuring an NBA player tracking data set over a six-week virtual sprint period. During this time, they must build a functioning application prototype OR submit a data science project or notebook with recommendations geared towards a front office or coaching staff. Participants must use NBA tracking data provided by Sportradar, to develop their projects. Participants may also use any other *publicly available* data set to enhance their projects, such as biography data.

Suboptimal Solutions

- **It doesn't run** 😪
  - Your code should compile and be easily reproducible in the reviewer's local environment
  - Ideally you include a build script
  - Mega brownie points for a Docker container
- **You didn't control for data quality** 🧹
  - You'll likely have to clean the data – be suspicious!
- **Your code is poorly formatted** 💬
  - Flex your system design and component-driven development muscles and make sure your code is abstract, dry, and reusable
  - Write good comments!
- **Your code is buggy** 🐞
  - Tests are your friend

Optimal Solutions

What *I* really want to see

- Containerized application (and database!) in Docker
- API with query params
- Functional UI (component library)
- Error handling and tests
- Publicly deployed app*
    - You may need permission to do this depending on what data you're given (e.g. tracking data is proprietary and cannot be shared publicly), but if your dataset is generally available, like PBP data, you should take this opportunity to add a project to your portfolio!

Really nice to haves

- ORMs
- Additional advanced stats! TS%!
- Feature to filter and sort players by performance
- Pretty design (coaches love pretty design)

Phillies R&D Opportunities

🚀 Quantitative Analyst Associate (Spring/Summer 2024)
👾 Software Engineering Associate (Spring/Summer 2024)
🤖 Machine Learning Engineer

Coming Soon

🎨 UI Designer
💻 Software Engineer

Josh Pohlkamp-Hartt

Associate Director of Analytics

Boston Bruins

JSM 2023

Question : Who is going to win the Stanley Cup in 3 years? Why? Show your work and defend your decision.

Details

- No data provided ☹
- Expectations is for a report and code
- “want to see how you think”
- Non-technical staff reviewing report
- Will have in-person discussion of problem during on-site interview

Data Used

- Scraped/found data from a variety of sites:, Hockey Abstract, Corsica, Hockey Graphs, Man Games Lost, NHL API
- Merged and feature-engineered datasets for modeling playoff wins and player projections.

Metric	Description
Age	Average age of players on team
PTS%	% of points available that were won
ConfRank	End of season conference standings
STE	Special Teams Efficiency = PP% +PK%, a measure of a team's special teams puck luck
PDO	PDO= = 5v5 Shooting % + 5v5 Save %, a measure of a team's even strength puck luck
GSAA30_Team	Goals saved above average per 30 shots - team average
GSAA30_1G	Goals saved above average per 30 shots - Starting Goalie
GSAA30_2G	Goals saved above average per 30 shots - Backup Goalie
xGDA_Team	Expected goals scored differential above replacement per 60 minutes - team average
xGDA_T6F	Expected goals scored differential above replacement per 60 minutes - Top 6 Forwards
xGDA_T4D	Expected goals scored differential above replacement per 60 minutes - Top 4 Defense
xGDA_B6F	Expected goals scored differential above replacement per 60 minutes - Bottom 6 Forwards
xGDA_B3D	Expected goals scored differential above replacement per 60 minutes - Bottom 3 Defense
xGDA_RepF	Expected goals scored differential above replacement per 60 minutes - Replacement Forwards
xGDA_RepF	Expected goals scored differential above replacement per 60 minutes - Replacement Defense
xGDA_SRS	Strength of schedule adjusted expected goals scored above replacement per 60 mintues
GInj	Total number of man games lost to injury

What makes a good solution?

- Well Defined Statistical Approach
  - Target Metric Design
  - Statement of Assumptions
  - Testing and tuning
  - Measure of uncertainty

- Data manipulation
  - Collection, merging and cleaning
  - Feature Engineering

- Nice to haves
  - Domain expertise (roster management, positional considerations)
  - Temporal Effects

Attributes of an underwhelming solution:

- Unclear Analytical Process
  - No defined KPI
  - Lack of stated assumptions
  - Thin on methodology discussion in report
  - “Black Box” approach
  - Hard to “see how they think”

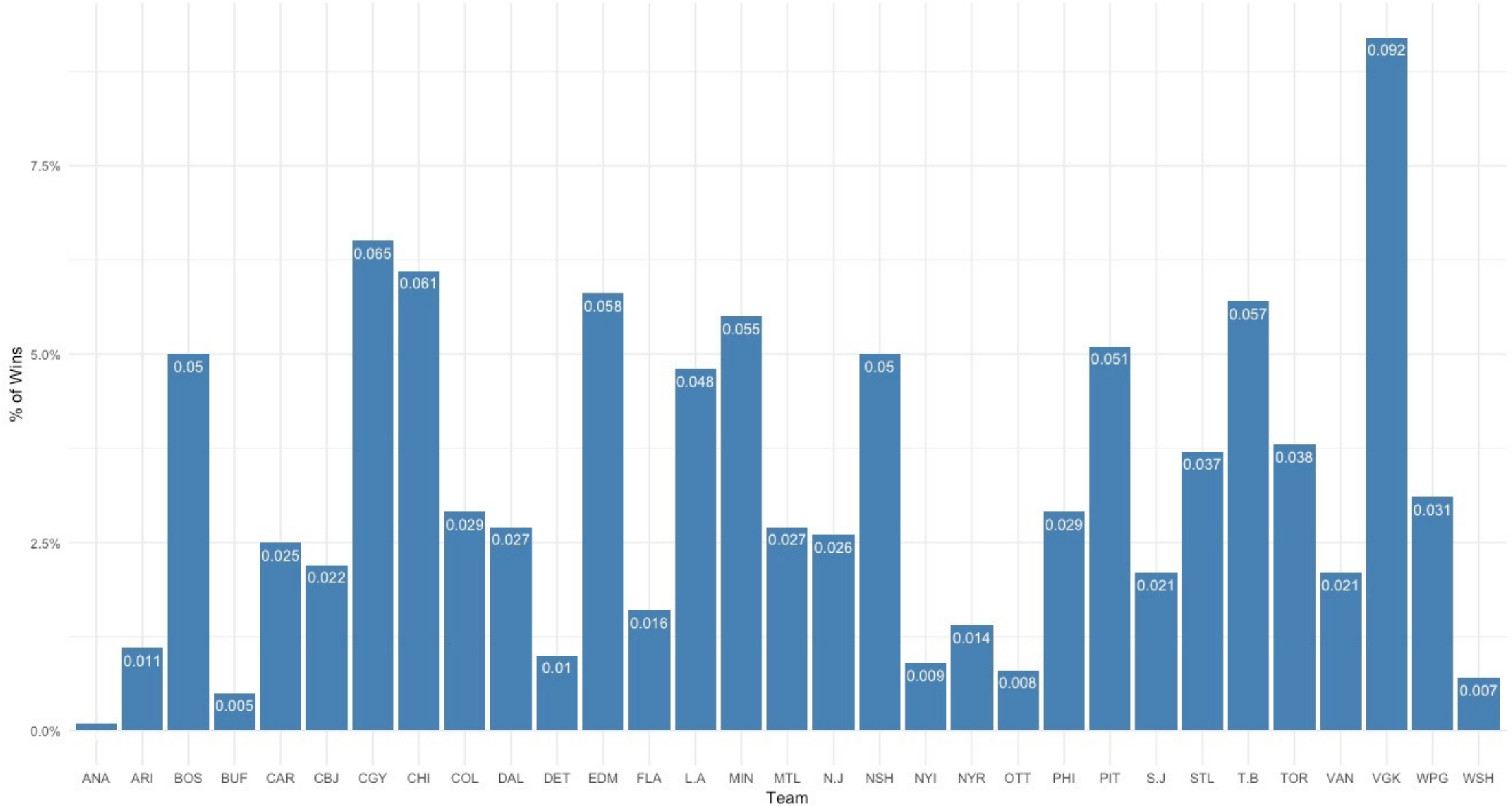
- Difficulty with data manipulation
  - Okay to ask for help with finding and collecting data
  - Needs to be able to merge and validate data

A Potential Solution:

- Multi-phase simulations
    1. Project Rosters in 3 years
        1. Age players
        2. Sign and graduate players
        3. Construct Game Rosters
    2. Assign Team Effects (Luck, Special Teams, Goalies)
    3. Predict Cup Winner
    4. Repeat 1-3 many times

- KPI: % Cup Wins = # Simulated Cup Wins/ # Simulations

- Prediction is team with highest %.


% of Wins
7.5%
5.0%
2.5%
0.0%
0.011
0.05
0.005
0.025
0.022
0.065
0.061
0.029
0.027
0.01
0.058
0.016
0.048
0.055
0.027
0.026
0.05
0.009
0.014
0.008
0.029
0.051
0.021
0.037
0.057
0.038
0.021
0.092
0.031
0.007
ANA
ARI
BOS
BUF
CAR
CBJ
CGY
CHI
COL
DAL
DET
EDM
FLA
L.A
MIN
MTL
N.J
NSH
NYI
NYR
OTT
PHI
PIT
S.J
STL
T.B
TOR
VAN
VGK
WPG
WSH
Team

JSM 2023

Ethan Douglas

Product Scientist

Zelus Analytics

ZELUS

Zelus Technical Assessment

Candidates are provided Cricket delivery outcome data from the Indian Premier League, asked to build a model to predict runs scored per delivery

Details

- Data is provided and explained
- Answers to be submitted as R Markdown or Jupyter Notebook
- Work is reviewed by senior data scientist
- Will discuss problem with candidate in followup virtual meeting

ZELUS

Question:

Because for all intents and purposes it's only possible to observe run outcomes of exactly 0, 1, 2, …, 6, one could consider using a classification model instead of a regression model in the prior question.
Without creating such a model, can you think of any advantages of this approach?
Could you describe how it could be possible that two players could have similar effects estimated using a regression model, but different effects estimated using a classification model?

ZELUS

Question: Evaluate tradeoffs of classification vs regression approach

Best solutions

- Clear understanding of technical concept
- Well communicated answer
  - Is it obvious what the candidate is trying to say?
- Demonstrate ability to think of potential model applications
  - What could the different approaches mean for how teams are able to use the model?
- Connect to sports-specific context
  - E.g. classification model could better capture different batting styles

ZELUS

Nick Wan

Director of Analytics
Cincinnati Reds

Question 1: Classify pitch types

Let's assume the data from `train.csv` is collected from a high tech scout. The scout is able to provide information on the speed of the pitch, the spin rate of the pitch, and the release point of the pitch (specifically the height of the pitch and the side of the pitch). The scout also provides information on pitch type for these particular pitches.

Using the information from our scout, please predict the pitch types thrown in `test.csv`. You can use any techniques you want; some statistical/machine learning techniques would be appropriate here. Comment your code thoroughly, so we can review your thought process on methodology for this question. The output file should have the same structure as `sample_solution.csv`.

Your solution will be evaluated by log loss.

Data Provided:

Pitcher Unique ID
Release speed (MPH)
Spin rate (RPM)
Release point (height and side, ft)
Pitch type (classification label, only in the training data)

A	B	C	D	E	F	G
UID	PITCHER_KEY	RELEASE_SPEED	SPIN_RATE_ABSOLUTE	RELEASE_HEIGHT	RELEASE_SIDE	PITCH_TYPE_TRACKED_KEY
361	483	92.17958832	1976.671631	6.226350784	-2.7887609	SI
362	483	93.35202789	2113.584229	6.384256363	-2.626076698	SI
350	864	93.10941315	2153.424561	5.912424088	-1.954880595	FB
353	928	93.71099854	2457.569092	5.913224697	-1.000528216	FB
359	483	92.90188599	1993.8302	6.303658009	-2.822147131	SI
360	483	93.19267273	2144.768799	6.342080116	-2.950623035	SI
356	405	87.104599	2213.919189	4.720386505	-2.217806339	FB
357	76	85.41158295	2186.283691	6.943728447	-1.369087338	CF
347	864	82.6007309	2840.050781	5.79373312	-2.023489475	CB
349	427	80.09098053	2574.014404	5.864082813	1.978016973	SL

Best solutions Included:

- Great statistics fundamentals
  - Estimating distributions vs mean/median aggregation
  - Normalization
  - Outlier consideration

- Modeling
  - Great research methodology
  - Testing trade offs between models
  - Hyperparameter tuning

- Nice to haves
  - Domain expertise
  - Feature engineering

Run-of-the-mill solutions

- Machine learning “kitchen sink” approaches
  - Lack of feature exploration prior to modeling
  - No methods on “raw” features vs normalized features

Find our Baseball Operations/Analytics postings at:

TeamWork Online

- **Under “Player Operations” and “Technical Services”**

First batch of job postings begin around September/October