UNIVERSITY OF
SAN FRANCISCO

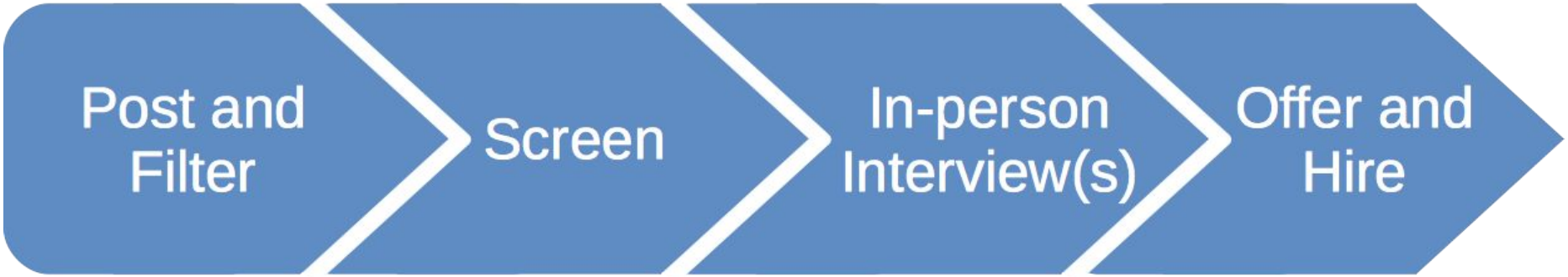
Master of Science
in Analytics

(Statistics &) Machine Learning

Interview Skills

The Interview Process

- Reminder (seen from the interviewer's POV):



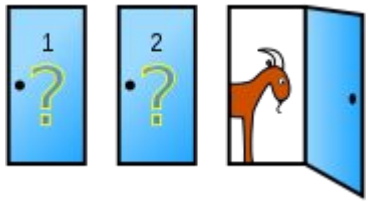
- Machine learning questions could occur during screening or during in-person interviews
- Different questions — types and formats — occur at different stages

The Mind of the Interviewer

- The interviewer's motivation
  - Does the candidate know what is necessary? (Does the candidate have the right training?)
  - Can the candidate apply what is known to what is not known? (Is the candidate creative?)
  - Sometimes: is the candidate able to get along with others?
  - How quickly can these answers be determined?
- Interviewer's questions: machine learning
  - Statistical measure
  - Applied to a business metric
  - Executed at scale (i.e. on computers)
- The candidate's dilemma
  - Show you know what you need to know
  - Show that you can be creative in the application of your knowledge
  - Do this as quickly as possible
  - Oh, and be personable while doing that

Let’s Make a Deal

- Monty Hall problem
  - Problem detailed in 1975; also on wikipedia
  - Scenario:



    Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens one door you did not pick — say No. 3 — which has a goat. He then says to you, "Do you want to stay with your choice or switch to the other door?"

    Question: Should you stay or switch?

  - The behaviour of the host is important:
    - The host must always reveal a door with a goat not picked by the contestant
    - The host must offer you the chance to switch doors
- Strategy:
  - Avoid using your intuition
  - Explain the steps you took to get to your answer
  - If possible, enumerate the choices and assign probabilities to each

Basic Questions from Statistics

- What is a type I error? A type II error?
- What is sampling? What sampling methods have you used?
- What is selection bias? What's the problem outcome of selection bias? How do you prevent it?
- You don't know a true population's mean or variance in some measure (eg. height/weight). How can you find it?

Metrics

- Statistical metrics
  - What is a P-value?
  - What is a coefficient?
  - How do you interpret regression coefficients for linear relationships?
  - What is an R-squared (R^2) value?
- Metrics for customer service / advertising
  - Churn (churn rate) — the rate at which customers discontinue (i.e. actively leave) or abandon (i.e. passively leave) a system
  - Conversion / growth / adoption rate — the rate at which customers subscribe to a system
    - Businesses care about growth / churn in order to determine whether their customer base is shrinking or expanding
    - Growth / churn can be examined for a business' units (eg. "Midwest sales") to determine where the productivity lies
- Metrics for advertisement
  - CPM (cost per impression) — the user has seen an ad
  - CPC (cost per click) — the user has seen an ad and has clicked on it
  - CPA (cost per acquisition) — the user has seen an ad, clicked on it and subscribed to the service
- Research the company to determine metrics important to them

Testing

- A/B Testing
  - Comparing two versions of something (a web page?)
  - Variant pages are shown to similar site visitors at the same time
  - Variant with the better conversion rate is chosen
  - Also called “split testing”
- Open questions:
  - Can “similar visitors” be chosen? (Can you prove they are similar? Can you ensure they remain in their respective splits?)
  - Implementation should be with temporary redirects (HTTP “301” redirect)
  - How long should the test run?
- Process:
  - Study the site / application
  - Observe user behaviour
  - Construct a hypothesis
  - Test hypothesis by dividing the population into two groups
  - Analyse data and draw conclusions

Types of ML Questions

Question Category	Example Question
Generic	What ML techniques have you used?
Theoretical / Applied	Suggest a few ways to determine the optimal “k” in KMeans.
Coding	Write one implementation of {k-Means, kNN, random forest}.
Screening	(Questions are unique to each company.)

Generic Questions

- Example: What ML techniques have you used?
- Approach
  - This is an opportunity to showcase your knowledge and experience
  - Discuss anything you have used in practicum
  - Consider other approaches you've studied and know well
- How to win
  - Explain the problem you were addressing
  - Compare and contrast different approaches and different algorithms

Theoretical Questions

- Example: Suggest a few ways to determine the optimal “k” in k-means
- Approach
  - Understand that your interviewer is looking for how well you understand the theory — i.e. there may be no optimal answer, but you understand the various tradeoffs
  - Talk briefly about the problems; use an example if you have one
  - Answer the question as directly as possible
- How to win
  - Ensure your list is diverse — eg. “use BIC or start with large k and continually remove centroids until it no longer reduces the description”
  - Detail more than one method — eg. “use Bayesian Information Criterion (BIC), which is…”
  - Briefly discuss options to the questions (eg. X-Means)

Coding Questions

- Example: Develop kNN from scratch
- Approach
  - Clarify the question
  - Give an example / use sample inputs and outputs
  - Write the algo
- How to win
  - Treat this as a whiteboarding question, even if it's not
  - Talk through the question — eg. "given a feature vector, find the euclidean distance to every other known vector... take the class that is the majority with the closest k vectors"
  - Manually derive an answer before coding

Screening Challenge

- Challenge / homework / take-home assignment
  - Screens out candidates who are technically weak, aren't committed to the opportunity
  - May pose specific questions or open-ended questions
  - Requires (deep?) analysis on a specific data set
  - Expectation is to be able to tell a story of what you did, insights you gained, etc.
  - Could be an opportunity to understand what the company a little better
- What to do
  - Reply within time allotted and follow the instructions
  - Answer questions posed by the challenge
  - Submit the source code (R, python) you used
  - Prepare narrative / discussion around it
- How to win
  - EDA is your friend: check the data and make sure it makes sense; cleanse the data
  - Stick with methods you know
  - Implement basic features; build (basic) models
  - Document your code well
  - Pick out specific examples germane to the company to write about / talk about

Example: Baby Names

Acquire and analyze the dataset on baby name popularity provided by the Social Security Administration.

A) Descriptive analysis

1. Describe the format of the data files. Can you identify any limitations or distortions of the data?
2. What is the most popular name of all time? (Of either gender.)
3. What is the most gender ambiguous name in 2013? 1945?
4. Of the names represented in the data, find the name that has had the largest percentage increase in popularity since 1980. Largest decrease?
5. Can you identify names that may have had an even larger increase or decrease in popularity?

B) Onward to Insight!

1. What insight can you extract from this dataset? Feel free to combine the baby names data with other publicly available datasets or APIs, but be sure to include code for accessing any alternative data that you use.

Answering “Baby Names”

- My plan (may not be your plan)
- Review questions
  - Answers to A) can all be answered with EDA
  - Answer to B) is open-ended — probably less important; name diversity?
- Wrangle the data
  - Unzip the data; create subset (2000+?) for dev — two smallest states? WY.TXT and ND.TXT
  - Read the StateReadMe.pdf file
- Start answer document
  - Question A1) can be answered almost immediately
  - Decide on how to handle output (graphs) — directly via Matplotlib? On output with OOo?
- Pilot the implementation
  - Create a function to read files files — read all files matching pattern (??.TXT) in directory
  - Store to one or more data structures — initial thoughts: list? dictionary... with key = NAME? NAME_YEAR?
  - Call from “main”: a series of functions, one for each question {A2|A3|A4|A5} — pass data structure for output; comment each function (Pydoc?) before writing function
  - Clean up; look for additional data, interesting trends (to answer B1)
  - End with function to download files from web and unzip

My Baby Names, BK's

- My implementation has limitations
  - Because of the data I chose for dev, I would have missed some signals, example: names "Baby", "Babyboy", "Babygirl"
  - Because I risk getting lazy and not looking for additional data
  - Because I might get lost in what may be in the data — US names getting more diverse?
- Implementation from Brian Kui (MSAN alumnus) is better
  - Not sure whether we have permission to share widely, but let's review & discuss
  - Critique of A4):
    - Question: find the name that has had the largest percentage increase in popularity since 1980
    - Answer compares two years (1980 vs. 2014) — may not be the right answer
  - Critique of A5):
    - Question: identify names that may have had an even larger increase / decrease
    - Answer seems vague
  - Part B answer takes advantage of generational data / trends vs. popularity of names

Submission for “Baby Names”

- Prepare a document
  - Follow the guidelines in writing this document
  - Where no guidelines exist, consider:
    - Introduction (purpose)
    - Input (data description)
    - Process (what you did)
    - Output (what you found)
    - Comment (implications of what you found, tailored to the company’s benefit)
  - Ensure your name and contact information appear at the top of this document
- Prepare a slide deck
  - Follow guidelines; if none exist, use 5 - 10 slides
  - Make slide titles similar to document; re-use graphs, wording, etc.
  - Add page numbers so audience can refer to slide number with questions
  - Ensure your name and contact information are clearly marked on each page

Example: Synthetic Dataset

Two synthetic datasets were generated using the same underlying data model. Your goal is to build a predictive model using the data in the training dataset to predict the withheld target values from the test set.

You may use any tools available to you for this task. Ultimately, we will assess predictive accuracy on the test set using the mean squared error metric.

You should return to us the following:

- An N x 1 text file containing 1 prediction per line for each record in the test dataset.
- A brief writeup describing the techniques you used to generate the predictions. Details such as important features and your estimates of predictive performance are helpful here, though not strictly necessary.

Answering Synthetic Dataset

- Review question
  - The challenge only cares about prediction error
  - Will my answer be judged by prediction error?
- Wrangle the data
  - Unzip the data file: two tab-separated files (sample train & test); all headers meaningless
  - Target is real number; need a regression model which reduces MSE
  - Most features are real (float / numeric); some features are categorical
  - Further split train file into (model + validation) ⇒ cross-validation?
- Start answer document
- Pilot the implementation
  - Create baseline model: use all features vs. linear regression
  - Missing values in some instances; impute… with mean / median?
  - Output error message as a reminder to use all train file and output predictions
  - Drive improvements over baseline:
    - Consider fewer features — ridge or lasso regularisation
    - Consider zero-normalising features
    - Explore other models — curvilinear, kNN, random forest?

After the Screening Challenge

- Depends on the company
  - May get a technical phone screen for one or more of: mathematics, statistics or coding
  - Pass the screening challenge and you'll get one or more in-person interviews
- Future peers (in-person interviews)
- Hiring manager (in-person interview)
  - Manager's motivation
    - Cares generally about how you communicate and fit with the team
    - Cares about your preparation, motivation, etc.
    - May have an idea of the "perfect" candidate and your delta (if any) from that idea
    - Already has feedback from your earlier interviews
  - How to interview well
    - Be confident and clear
    - Apply what you know from challenge and earlier interviews
    - Be prepared for the hiring manager with technical background!
- Executive (in-person interview)
  - Pass the in-person interviews and you may get an in-person interview with an executive
  - This is probably the final stage of interviews
  - Show that you understand the company, its problems and the ways you can contribute

After the Interviews

- Follow up
  - Self evaluate: do you still want to work there?
  - Be clear on who the decision maker is — the hiring manager?
  - Assuming you want to stay in touch, ask how to do it
- Say thanks?
  - This is not necessary for most companies or individuals
  - If you send a “thank you” note, you want to be remembered:
    - DO NOT send a boilerplate message
    - Build on some specific portion of the interview
    - Consider showing them you understand their company’s pain… and can help
- Get ready to send references
- Wait?
  - Process may take months depending on the companies
  - Follow up, if you asked how to… but don’t come across as desperate
  - Check in with people in your network periodically for status
  - Apply to the next company
  - If you get another offer while a decision is being made, reach out

Resources

- A/B Testing videos
  - Data Science in AirBnB
  - A/B Testing Done Right
  - A/B Testing by Google — via Udacity
- There's also A/A testing
- This and more in *Design of Experiments*