UDS Midterm Study Guide

Yabei Zeng

Table of contents

1 UDS Midterm Study Guide

1.1 Intro Chapter

1.1.1 Solving Problems with Data: No Order

- **Step (1) Specifying the Problem**
- **Step (2) Solving Problems through Answering Questions**
  - Types of Questions:
    - **Exploratory Questions**: Questions about general patterns in the data
      - Useful for understanding the problem space better and prioritizing subsequent efforts
      - E.g.: how many job applicants are you receiving when you post a job?
    - **Passive Prediction Questions**: Questions about likely outcomes for individual observations or entities
      - Useful for targeting individuals for additional attention or interventions being considered
      - E.g.: given the symptoms of this patient and their test results, how likely are they to develop complications after surgery?
    - **Causal Questions**: Questions about the effect of actions or interventions being considered
      - Useful for deciding on appropriate courses of action
      - E.g.: what is the effect of an action X on an outcome Y? If I do X, how will Y change?
      - Fundamental problem of causal inference: we can never perfectly know what the value of our outcome Y would be in both a world we do X and one where we don't do X

1.1.2 Problem Refinement & Stakeholder Management

- **Step (0) Recognize your role**
  - Helping your stakeholder better understand their problem is a core part of the job
  - Data science is about pairing domain expertise with computational methods and quantitative insights
- **Step (1) Don't assume your stakeholder knows what they need**
- **Step (2) Abstract the problem**
- **Step (3) Ask questions (Especially Quantitative Ones)**
  - Questions about Success

  - Questions about the Problem
- **Step (4) Propose questions you might answer**
  - Make your questions specific and actionable
- **Step (5) Iterate**
  - Bring your work back to your stakeholder as often as possible

1.2 Proscriptive vs. Descriptive Questions

- Descriptive Questions: Questions about the **state of the world** and about objective reality (have right or wrong answers) include the following examples:
  - "What kinds of users are clicking our ads?
  - "Do high-income and low-income countries emit similar amounts of carbon dioxide?"
- Proscriptive Questions: Questions about **how the world should be** , don't have right or wrong answers because answers to proscriptive questions require evaluating the desirability of possible outcomes, include the following examples:
  - "Should higher income and low income countries be expected to meet the same carbon emission reduction standards?"
  - "Do high-income countries have a moral obligation to provide tuberculosis drugs to developing countries for free (or at cost)?"

1.3 Exploratory Questions

- Using Exploratory Questions: questions about elicit information, questions about broader patterns in the world, could be answered by simple summary statistics or plots, answering the exploratory questions makes you understand the contours of the problem you seek to solve, include the following examples:
  - What type of buildings (industrial, residential, commercial) consume the most power in the US?
  - In what region of the US are buildings consuming the most power?
  - Is there a region of the US where buildings are generating the most CO2?
  - Does the average energy use per building vary by region or building type?
  - In what season is most building energy consumed? Is more energy consumed by heating or AC needs, or do the two use similar amounts of power?

1.3.1 Internal versus External Validty

- Internal validity: how well you have analyzed the data you have
- External validity: how well you expect the answer you generated from that data to generalize to your stakeholder's context
- Internal validity concerns
  - Relate to your ability to properly characterize the causes of severe accidents in this data, accidents causes you generate from this data are meaning and faithful representations of the patterns in the data
  - Concerns over the accuracy with which things are measured
- External Validty Concerns
  - New features rolls out, whether there was anything exceptional about the data generating process
- The more control one has over a study context, the more likely one is to have good internal validity, but that control can often create an artificially to limits external validity, Thus internal and external validity should not necessarily be thought of as things to try and simultaneously maximize at all costs, rather, they are best thought of as distinct features of analysis that should always be considered

1.4 Passive Prediction Questions

- Passive prediction questions: about **the future or potential** outcomes of individuals entities, including examples:
  - "How likely is patient X to experience a heart attack in the next two years?"
  - "How likely is it that Mortgage Holder Y will fail to make their mortgage payment next month?"
- Passive Prediction don't usually have "an answer", need by considering the feasibility of developing a model to give individual-level answers to a Passive Prediction Question
- Differentiating between exploratory and passive prediction questions
  - Passive prediction questions focus on the values that get spit out of a model for each entity in the data, the only thing we care about in the passive prediction is **the quality of these predictions**
  - With exploratory question, our interest is in improving our understanding of the problem space
- Passive prediction and causal questions
  - Both questions are trying to predict future outcome

  - The difference is that passive prediction are about make accurate prediction, not causal relationships that we can directly manipulate to shape outcomes (causal questions)
  - Correlation does not necessarily imply causation

- **Internal validity and external validity** in answering passive prediction questions
  - Internal validity: measure of how well a model captures the meaningful variation in the data we already have
  - External validity: measure of how well we think that our model is likely to perform when faced with new data
  - Some of the factors that influence the External Validity to Passive Prediction Questions are the same as those that shape the External Validity of Exploratory Question
    * Population represented in the data
    * The time period in the questions
  - Two external validity concern (different from exploratory questions):
    * Overfitting
    * Adversarial users: attempt to subvert a statistical or machine learning model

- Not Using Black Box:
  - Problem with black box machine learning models: these models are used in high-stakes areas like `healthcare` and `criminal justice` but lack transparency, leading to severe consequences due to their inaccessibility and complexity
  - Explanation towards the black box could be problematic, lacking fidelity and can be misleading
  - Well-structured data with meaningful features can be equally well-predicted by simpler, interpretable models

- pyGAM
  - Generalized Additive Models(GAMs)
    * $g(\mathbb{E}[y|X]) = \beta_0 + f_1(X_1) + f_2(X_2, X_3) + ... + f_M(X_N)$
    * $X_1, X_2, ..., X_N$ are independent variables, and y is the dependent variable
    * $g()$ is the link function

1.5 Causal Questions

- Causal questions: designed to help us predict the consequences of our actions
- Stakeholder will ask a causal question when they want to know whether an action may be beneficial
- An answer to causal questions will involve designing a study that not only measures the effect of treatment but also knows any measured effect will generalize to the stakeholder's context.

- Internal and External Validity
  - Internal: the study itself in the context of studying
  - External: The context in which the stakeholder wishes to generalize the result
- Answering causal questions:
  - An action X on an outcome Y
  - `Fundamental Problem of Causal Inference`: directly measure the causal effect of X on Y for a given entity
  - `Randomized experiment`: most familiar tool for answering causal questions
    - Randomized Control Trials (RCTs) or A/B Tests

1.5.1 Potential Outcomes Framework

- T is the treatment, a binary treatment, meaning it can only take on two values : {0,1}
- Y_i is the potential outcome variable for i individual, ${Y_i}^1$ is the treated potential outcome variable for i individual
- D is the observed receipt of either treatment or control, takes two values {0, 1}
- Given value of D, we will only be able to observe
  - Y_i^0 if $D_i = 0$ or Y_i^1 if $D_i = 1$
- Defining the causal effect:
  - $\delta_i = Y_i^1 - Y_i^0$
  - The difference between the individual potential outcome
  - This quantity is not able to observe
  - For all individuals, the value is call ATE
    - $ATE = \frac{1}{N} \sum_{i \in 1,2,3...N} \delta_i$
    - $ATE = \frac{1}{N} \sum_{i \in 1,2,3...N} Y_i^1 - Y_i^0$
    - If we assume our data is a random sample from the population, then we can write the ATE in the following form
      - $ATE = E(\delta_i)$
      - $ATE = E(Y_i^1 - Y_i^0)$
      - $ATE = E(Y_i^1) - E(Y_i^0)$
    - The binary treatment
      - $E(Y_i^1 | D_i = 1) - E(Y_i^0 | D_i = 0)$
      - This is a exact quantity you get from a linear regression of Y on D
- Correlation and Causation
  - $\hat{ATE} = E(Y_i^1 | D_i = 1) - E(Y_i^0 | D_i = 0) + E(Y_i^0 | D_i = 1) - E(Y_i^0 | D_i = 1)$
  - because $E(Y_i^0 | D_i = 1) - E(Y_i^0 | D_i = 1) = 0$, therefore $\hat{ATE} = E(Y_i^1 | D_i = 1) - E(Y_i^0 | D_i = 0)$

  - We can also write $\hat{ATE}$ as
    * $\hat{ATE} = E(Y_i^1|D_i = 1) - E(Y_i^0|D_i = 1) + E(Y_i^0|D_i = 1) - E(Y_i^0|D_i = 0)$
    * where $E(Y_i^1|D_i = 1) - E(Y_i^0|D_i = 1)$ is ATT
    * and $E(Y_i^0|D_i = 1) - E(Y_i^0|D_i = 0)$ is the baseline difference

- Average Treatment Effect on the Treated (ATT)
  - $E(Y_i^1|D_i = 1) - E(Y_i^0|D_i = 1)$
  - $ATE = E(Y_i^1) - E(Y_i^0)$
  - $= \lambda \left(E(Y_i^1|D_i = 1) - E(Y_i^0|D_i = 1)\right) + (1 - \lambda) \left(E(Y_i^1|D_i = 0) - E(Y_i^0|D_i = 0)\right)$
  - where $E(Y_i^1|D_i = 1) - E(Y_i^0|D_i = 1)$ is the ATT (Average treatment effect on the treated)
  - where $E(Y_i^1|D_i = 0) - E(Y_i^0|D_i = 0)$ is the average treatment effect on the untrated
  - ATT = ATE when $E(Y_i^1|D_i = 1) - E(Y_i^0|D_i = 1) = E(Y_i^1|D_i = 0) - E(Y_i^0|D_i = 0)$, meaning no differential treatment effects
  - neither $E(Y_i^0|D_i = 1)$ nor $E(Y_i^1|D_i = 0)$ are observable

WHAT is `SUTVA`?

- The Stable Unit Treatment Value Assumption (SUTVA) is a fundamental principle in causal inference that assumes `the treatment given to one unit does not affect the outcomes of other units.` This assumption comprises three main ideas:
  - `homogeneity of treatment effects`: each unit receives the same treatment effect
  - `No spillovers or externalities`: The treatment of one unit does not influence the outcomes of another unit.
  - `No general equilibrium effects`: The assumption that the treatment does not lead to broader systemic changes that could affect the outcomes.

1.5.2 Indicator Variables

`Indicator variables` :sometimes referred to as dummy variables, takes values of 0 and 1, used to indicate whether a given observation belongs to a discrete category in a way that can be used in statistical models.

- the coefficient on an indicator variable is an estimate of the average **DIFFERENCE** in the dependent variable for the group identified by the indicator variable
- the **REFERRENCE GROUP**, is the set of observations for which the indicator variable is always zero.
- the coefficient on an indicator variable is an estimate of a **DIFFERENCE** with respect to a **REFERENCE GROUP**
- For categorical variables with more than 2 categories, use `one-hot encoding`

1.5.3 Experiment

A/B testing

A/B testing: two versions, and see which one performs better, includes the following steps:

- Identify the goal
- Create variants: two versions to be tested
- Split the audience: one group sees version A, and the other sees version B
- Collect the data
- Analyze results
- Make decisions

Twyman's Law: interesting or different results are often mistaken and emphasizes the need for skepticism and thorough verification of experiment data

- implications in Data Analysis:
  - error checking
  - skepticism
  - verification
- application in research and experimentation
  - design robustness
  - reproducibility
  - result interpretation

Overall Evaluation Criteria (OEC)

OEC: metrics used to measuer the success of an experiment and determine whether the tested changed have achieved the desired outcome

- **Selection of OEC**: The right OEC involves the followings
  - considerinf the goals of the experiment
  - the expected impact of the changes
  - how success will be measured
- **Balance between short-term and long-term goals**
  - the balance should consider how changes might affect user behavior or business metrics over time

A/A Testing

Purpose/Definition: running two identical versions of a product or feature against each other to check the testing infrastructure's accuracy and to detect any biases or errors before conducting actual A/B tests

Through AA testing, a baseline for normal performance variation without experimental changes is established. The baseline is used for comparision in sebsequent A/B tests to measure the impact of changes accurately

Internal Validity

Internal Validity: an experiment accurately establishes a causal relationship between the independent and dependent variabkes, ensuring the results are due to the manipulated variables and not other factors

Threats to internal validity

- selection bias
- history effect (external events affecting outcomes)
- maturation (participants changing over time)
- instrumentation changes (differences in measurement tools or procedures)

Strategies for improving internal validity

- random assignment to control and treatment groups
- controlling external variables
- using blinding methods
- ensuring consistent measurement techniques throughout the experiment

External Validity

external validity: how well the results of an experiment can be generalized beyond the specific conditions of the study

Factors affecting external validity

- sample characteristics
- experimental setting
- the context of which the experiment is conducted

Threats to external validity

- selection biases
- situational factors

- primary effects versus novelty effects (new features release, and users need time to adapt to it(primary), new features may attract new customers in short term (novelty))
- interaction effects between experimental treatments and participants

Strategies for improving external validity

- using representative samples
- plot over time (primary and novelty)
- ensure the experimental setting closely mirros the real-world context
- replicating studies in different settings

1.5.4 Power Calculations

- **Statistical Power and Type II Errors**: defines statistical power as the probability of correctly rejecting a false null hypothesis and explains Type II errors (false negatives)
- **calculating sample size**: discusses how to compute the sample size needed to achieve desired power and control type I (false positives) and Type II error rates
- **Errors in Hypothesis Testing**:
  - True Positive (Correct Rejection of Null Hypothesis): The test correctly rejects the null hypothesis when it is false. This means that there is a real effect, and the test has successfully detected it.
  - True Negative (Correct Acceptance of Null Hypothesis): The test correctly fails to reject the null hypothesis when it is true. This means that there is no real effect, and the test correctly identifies that there is no significant difference or effect.
  - Type I Error (False Positive): The test incorrectly rejects the null hypothesis when it is actually true. This means the test indicates there is an effect when there isn't one, leading to a false alarm.
  - Type II Error (False Negative): The test incorrectly fails to reject the null hypothesis when it is actually false. This means the test fails to detect a real effect, missing the discovery of an actual difference or effect.
- **Type II Error Rate (β)**: the probability of not detecting a true effect when it exists, decreases with larger sample sizes or greater effect sizes
- **Trade-offs in Testings**: the inverse relationship between Type I and Type II errors and the necessity of balancing sensitivity (power) against the risk of false positives.
- **Effect Size**: statistical power depends on the specified effect size, larger effects are easier to detct
- **Sample Size**: influenced by desired power, siginificance level (Type I error rate), and the minimum effect szie of interest
- **Minimum Effect of Interest (MEI)**: drive the design of an experiment based on the smallest practical effect size that would be meaningful to detect

2 Quiz

2.1 Intro Chapter

1. In the view of the author (me) of the chapter you read, what is the most important task for you, a data scientist, hoping to help your stakeholder?

Choice 1of 4: Properly articulating the problem you are trying to solve.

Choice 2 of 4: Providing results that are more accurate than what the stakeholder had before you were you hired.

Choice 3 of 4: Leaving the stakeholder with a model that they know how to maintain and use effectively after you leave.

Choice 4 of 4: None of these are the most important thing.

ANSWER : Properly articulating the problem you are trying to solve.

2. In the view of the author (me) of the chapter you read, what is the main/most common purpose of Exploratory Questions?

Choice 1 of 4: To help you understand how variables in your dataset are coded and what variables are in the data.

Choice 2 of 4: To help improve your understanding of the problem space and thus better prioritize subsequent efforts.

Choice 3 of 4: To identify new markets for a business wishing to expand.

Choice 4 of 4: None of these are what the author deems the main/most important purpose of Exploratory Questions.

ANSWER: To help improve your understanding of the problem space and thus better prioritize subsequent efforts.

3. Why is it critical you always progress from problem articulation to answering an Exploratory Question, then a Passive Prediction Question, then a Causal Question?

Choice 1 of 4: It's critical that one start with problem articulation and answering an Exploratory Question, but you may sometimes skip answering a Passive Prediction Question before proceeding to answering a Causal Question.

Choice 2 of 4: If you don't follow this order, your analysis will be harder to explain to stakeholders. This flow helps to ensure the "legibility" of your analysis to third parties.

Choice 3 of 4: This order ensures you always work from the easiest task to the hardest, ensuring you don't waste time unnecessarily.

Choice 4 of 4: It's not critical you proceed in that order. While problem articulation should always be the first thing you do, and there is sometimes a natural flow from Exploratory to Passive Prediction to Causal Questions, in practice data science projects will almost never flow mechanically from one to the other.

ANSWER: It's not critical you proceed in that order. While problem articulation should always be the first thing you do, and there is sometimes a natural flow from Exploratory to Passive Prediction to Causal Questions, in practice data science projects will almost never flow mechanically from one to the other.

4. What is the difference between a Passive Prediction Question and a Causal Question?

Choice 1 of 4: A Passive Prediction Question is a question about the future, while a Causal Question is about what caused something to happen in the past.

Choice 2 of 4: They are different terms for the same thing.

Choice 3 of 4: A Causal Question is a question about how the world would be different if a certain action took place (usually an action that you or your stakeholder is thinking about undertaking). A Passive Prediction Question is a question about how things are likely to proceed if the status quo prevails.

Choice 4 of 4: Causal Questions are always harder to answer than Passive Prediction Questions.

ANSWER: A Causal Question is a question about how the world would be different if a certain action took place (usually an action that you or your stakeholder is thinking about undertaking). A Passive Prediction Question is a question about how things are likely to proceed if the status quo prevails.

2.2 Exploratory and Descriptive / Prescriptive

1. Why is it critical you always progress from problem articulation to answering an Exploratory Question, then a Passive Prediction Question, then a Causal Question?

Choice 1 of 4: This order ensures you always work from the easiest task to the hardest, ensuring you don't waste time unnecessarily.

Choice 2 of 4: If you don't follow this order, your analysis will be harder to explain to stakeholders. This flow helps to ensure the "legibility" of your analysis to third parties.

Choice 3 of 4: It's critical that one start with problem articulation and answering an Exploratory Question, but you may sometimes skip answering a Passive Prediction Question before proceeding to answering a Causal Question.

Choice 4 of 4: It's not critical you proceed in that order. While problem articulation should always be the first thing you do, and there is sometimes a natural flow from Exploratory to

Passive Prediction to Causal Questions, in practice data science projects will almost never flow mechanically from one to the other.

ANSWER: It's not critical you proceed in that order. While problem articulation should always be the first thing you do, and there is sometimes a natural flow from Exploratory to Passive Prediction to Causal Questions, in practice data science projects will almost never flow mechanically from one to the other.

2. Which of the following is a Prescriptive Question about marijuana (also known, in various circles, as weed, pot, cannabis, ganja, Mary Jane, and of course, best of all, the Devil's Lettuce)?

Choice 1 of 5: Is marijuana legalization likely to generate tax revenue for local governments?

Choice 2 of 5: Is it a good idea to legalize recreational marijuana?

Choice 3 of 5: Does marijuana use lead to use of other drugs like heroin or cocaine?

Choice 4 of 5: None of these are Prescriptive Questions.

Choice 5 of 5: Does recreational marijuana legalization cause violent crime?

ANSWER: Is it a good idea to legalize recreational marijuana?

3. According to the author of the reading (me!), what's wrong with advertisements reporting Aimovig "caused episodic migraine sufferers to experience 3-4 fewer days of migraines a month"?

Choice 1 of 4: 3-4 days was the average treatment effect, but almost no patient experiences that effect. For most patients, the effect of Aimovig is much less or much greater.

Choice 2 of 4: Aimovig wasn't actually designed to treat migraines. The clinical trial was designed to measure it's effect on regular headaches, but it didn't work, so the Pharmaceutical company dug through the data till they found an outcome — migraines — for which it had an effect, then called that their target.

Choice 3 of 4: Aimovig's claims aren't based on an FDA approved clinical study. Amgen conducted an observational study to get that number, not a controlled experiment, and you can't make causal medical claims without a randomized clinical trial.

Choice 4 of 4: They didn't report the standard errors associated with that estimate, and the standard errors are huge, making it virtually meaningless.

ANSWER 3-4 days was the average treatment effect, but almost no patient experiences that effect. For most patients, the effect of Aimovig is much less or much greater.

4. Which of the following is NOT one of the author's concerns about the term EDA?

Choice 1 of 4: Its prevalence creates the misguided impression among students they need to invest in really "getting to know their data" before fitting a model, where students should really be focused on post-modeling diagnostics (e.g., residual plots) to catch errors.

Choice 2 of 4: It devalues analyses not based on sophisticated statistical modeling.

Choice 3 of 4: It conflates several distinct activities under a single name — learning the structure of your data, validating your data, and answering questions about the world without formal modeling.

Choice 4 of 4: It creates the impression one can learn a lot about the world by "doing EDA" without first articulating a clear question on seeks to answer.

ANSWER: Its prevalence creates the misguided impression among students they need to invest in really "getting to know their data" before fitting a model, where students should really be focused on post-modeling diagnostics (e.g., residual plots) to catch errors.

5. Select all of the following questions that are Exploratory Questions.

Choice 1 of 5: Using past performance indicators, market trends, and companies's financial reports, is it possible to forecast next year's profitability for different technology companies? Choice 2 of 5: Examining data from various schools, what patterns emerge in terms of student performance, extracurricular involvement, and socioeconomic background? Choice 3 of 5: Given data on an individual's lifestyle choices and genetic predispositions, can we anticipate the onset of a particular health issue accurately? Choice 4 of 5: Will the introduction of a new digital learning platform cause improvement in student achievement? Choice 5 of 5: Are employment rates better in US counties with younger populations?

ANSWER:

Examining data from various schools, what patterns emerge in terms of student performance, extracurricular involvement, and socioeconomic background?

Are employment rates better in US counties with younger populations?

2.3 Exploratory and Descriptive/ Prescriptive Redux (only include the different one)

1. How do Internal and External Validity differ (according to the author)?

Choice 1 of 4: None of the above.

Choice 2 of 4: Internal Validity is the accuracy of an analysis conducted by a team that also collected the data. External Validity is the accuracy of an analysis conducted by a team that did not collect the data themselves but rather got it from another source.

Choice 3 of 4: A given analysis can generally be said to have a certain level of Internal Validity, but there is no such thing as "a study's External Validity." External Validity is only defined with respect to a specific context.

Choice 4 of 4: Because Internal Validity is more important than External Validity, analyses should be designed to maximize Internal Validity first, and then look for ways to increase External Validity second.

ANSWER: A given analysis can generally be said to have a certain level of Internal Validity, but there is no such thing as "a study's External Validity." External Validity is only defined with respect to a specific context.

2.4 Interpretable ML

1. How do Internal and External Validity differ (according to the author)?

Choice 1 of 4: Because Internal Validity is more important than External Validity, analyses should be designed to maximize Internal Validity first, and then look for ways to increase External Validity second.

Choice 2 of 4: A given analysis can generally be said to have a certain level of Internal Validity, but there is no such thing as "a study's External Validity." External Validity is only defined with respect to a specific context.

Choice 3 of 4: Internal Validity is the accuracy of an analysis conducted by a team that also collected the data. External Validity is the accuracy of an analysis conducted by a team that did not collect the data themselves but rather got it from another source.

Choice 4 of 4: None of the above.

ANSWER: A given analysis can generally be said to have a certain level of Internal Validity, but there is no such thing as "a study's External Validity." External Validity is only defined with respect to a specific context.

2. In recent years, many companies have turned to AI tools to screen resumés to identify which job applicants to advance to interviews (UGH!). Not long after companies started rolling out these tools, many applicants realized they could game the system by adding keywords the models might be looking for in tiny font written in white (so a human wouldn't see them, but an NLP algorithm would).

Note: this trick is known and doesn't work today.

From the perspective of the data scientists who developed these tools, this behavior is an example of (select all that apply):

Choice 1 of 5: Adverse selection. Choice 2 of 5: An external validity concern. Choice 3 of 5: None of the above. Choice 4 of 5: Adversarial users. Choice 5 of 5: An internal validity concern.

ANSWER

An external validity concern.

Adversarial users

---

3. Which of the following is NOT true of an explainable black box model?

Choice 1 of 4: These are all true of Black Boxes.

Choice 2 of 4: Black box models usually outperform interpretable models, according to standard evaluation metrics. But interpretability should also be considered when choosing a metric, and so with you also take into account interpretability, interpretable models are often better.

Choice 3 of 4: the explanations generated for the black box model must be wrong. They cannot have perfect fidelity with respect to the original model. If the explanation was completely faithful to what the original model computes, the explanation would equal the original model, and one would not need the original model in the first place, only the explanation. (In other words, this is a case where the original model would be interpretable.) This leads to the danger that any explanation method for a black box model can be an inaccurate representation of the original model in parts of the feature space.

Choice 4 of 4: Explainability doesn't help address problems that arise when there are problems with input data, like when a typographical error results in an inmate getting an inaccurate risk assessment score

ANSWER: Black box models usually outperform interpretable models, according to standard evaluation metrics. But interpretability should also be considered when choosing a metric, and so with you also take into account interpretability, interpretable models are often better.

4. Rudin characterizes the COMPAS risk model as a black box model. In what way is the COMPAS risk model a black box?

Choice 1 of 4: The model applies a neural network model to defendents' full criminal records, which can be designed in a manner that allows for some interpretability, but the specific framework used by the creators of COMPAS did not make use of these methods.

Choice 2 of 4: The company that created it refuses to share the model because they argue it is a trade secret.

Choice 3 of 4: None of the above — Rudin actually presents COMPAS precisely because it is a successfully deployed interpretable model.

Choice 4 of 4: The company that created COMPAS considered interpretable models, but showed that best results were obtained from a boosted decision tree, and so have chosen to rely on that model instead.

ANSWER: The company that created it refuses to share the model because they argue it is a trade secret.

2.5 Online AB Internal and External Validity

Q1 OEC

According to Kohavi, Tang and Xu, which of the following are desirable properties of an OEC?

- ☐ It should be **sensitive** (I might say responsive) enough that your treatment will cause measurable changes in the outcome.
- ☐ Changes in your OEC should be expected to causally impact your organization's achievement of its long-term objectives (e.g., you are confident an increase or decrease in OEC will contribute to your organization's long-term goals).
- ☐ The measure should be **measurable** (in the context of where the experiment is being run) and **attributable** (you can readily associate treatment assignment with outcomes).
- ☐ The measure should be **timely**, meaning it should respond to changes (like your treatment intervention) quickly enough that you don't need to wait forever to measure the effect you care about and make a decision.

ANSWER: all of them

Q2 OEC

Kohavi, Tang and Xu discuss Bing search's problems with using the number of distinct queries per search as an OEC (where a "search" is a distinct effort by a user to find something they want).

Which of the following best characterizes the problem they identified?

- ( ) Bing didn't have a way of figuring out how many distinct queries should be attributed to each "search."
- ( ) Queries per search doesn't tell Bing anything about how much **revenue** they are making from searches. As an ad-revenue based company, they should be focused on ad revenue per search.
- ( ) What the hell is Bing? (this option was funnier before chatGPT was integrated into Bing).
- () When the quality of search results goes **down**, then in the short run people actually end up using the service more because they aren't getting good results (even though in the long run it drove people away from Bing). In other words, the OEC was causally related to the long-term goals of Bing.

ANSWERS: When the quality of search results goes **down**, then in the short run people actually end up using the service more because they aren't getting good results (even though in the long run it drove people away from Bing). In other words, the OEC was causally related to the long-term goals of Bing.

Q3 AA Testing

A/A Testing is:

- ( ) A critical tool for validating the internal validity of any a/b study in which you run an experiment without actually exposing anyone to your treatment and then look for differences across treatment arms.
- ( ) A useful tool, but making sure there aren't differences in whatever observable characteristics you can measure across treatment and control conditions accomplishes the same thing.
- ( ) A critical tool for validating the external validity of any a/b study in which expose users to the same treatment (a) twice to see if how they respond changes between the first and second exposure.
- ( ) None of the above.

ANSWER: A critical tool for validating the internal validity of any a/b study in which you run an experiment without actually exposing anyone to your treatment and then look for differences across treatment arms.

Q4 Primacy and Novelty Effects

What are the impacts of primacy and novelty effects on A/B experiments?

- ( ) Primacy ("first" users) and novelty (new-feature-seeking users) effects are the distortions caused by the fact that when a new feature is released, a certain type of user is likely to find it and use it first, making the first people who use a new feature unrepresentative.
- ( ) Primacy and novelty effects are types of SUTVA violations that undermine a study's internal validity.
- ( ) Primacy effects (at the start of an experiment, users are accustomed ("primed") to the old version of a site/app and need time to learn the new site) will result in smaller treatment effects in the short run. Novelty effects (people interact with something just because it's new) will result in larger treatment effects in the short run than in the long run.
- ( ) None of the above.

ANSWER: Primacy effects (at the start of an experiment, users are accustomed ("primed") to the old version of a site/app and need time to learn the new site) will result in smaller treatment effects in the short run. Novelty effects (people interact with something just because it's new) will result in larger treatment effects in the short run than in the long run.