Problem Set 2

Predictive Analytics for Business Strategy

Spring 2023

Due: Sunday, February 5th

McDermott

Team Member :

- **Jimmy Yang**
- **Youngsun Song**
- **Genadi Susilohadi**
- **Arun Thomas**
- **Yashada Hemant Nikam**

1. Wayfair has an idea for a new version of their website and they want to know what impact it would have on profits. They take a **random sample** of the users of their website and randomize who gets version A (old) and who gets version B (new).
    a. **Why does Wayfair do this?**
    To avoid bias in the result of comparing the new and previous version.
    b. **Is this experiment maximizing their profits? Explain why or why not.**
    It depends. If through this experiment, Wayfair can determine clearly which version has a better impact on their profit, then yes. But if they can't determine that (there is no difference in impact on the profit between the new and previous version) then this experiment is a sunk cost that lowers their profit.
    c. **What is the benefit of having a random sample?**
        - It lowers the risk of having a biased result
        - It is easier to find the sample than samples with specific characteristics
        - Ensure sample is representative of the whole population
        - Post experimental data analysis is easy compared to observational data
    d. **What is the benefit of randomization?**
        - Ensures ATE=ETT
        - Prevents selection bias commonly present in observational data

2. A cannabis company in the state of Washington, let's call them Rxx, wants to make sure they have reserved the optimal amount of storage for their product at different times of year, taking into account weather patterns, seasonal demand, etc. **Will they need to engage in active prediction or passive prediction to do this? Explain**.
    Passive Prediction. The company is trying to observe the impact of variables outside their direct control upon optimal amount of storage. Since the company cannot control the weather or consumer demand, they can only observe the effect those variables have on storage amount over time**.**

3. Suppose you have data on the daily number of taxi rides in NYC from 2000-2020. You can also see the number of Uber and Lyft rides. Uber's share has been increasing and the taxi rides have been decreasing.
   a. If you take a random sample of this data and regress number of taxi rides on the presence of Uber, **will this give you a good estimate of the causal impact of Uber's presence on taxi rides? Explain why** OR **why not**.
      No, it won't. Because it will only show the correlation between the number of taxi rides and the number of Uber rides. In order to estimate the causal impact, we need to observe other variables that affect the taxi rides, such as:
      - Price rate of each ride
      - Availability of each ride
      - Safety measures for each ride
      - Existing complementary benefit of each ride (*for example Uber has food & grocery deliveries for their customer*)
      - Convenience factor (wait time) for the customer

4. A company called Taegyo wishes to market products specifically for consumers who are expecting a baby, such as classical music recordings, etc. They claim their products cause the baby to have higher verbal skills. **Do you think their claims are likely valid? Explain why or why not?**
   It is likely to be invalid. Because a baby's verbal skill is likely to be affected by multiple variables which affect one another. Their claim might not be proven if other variables are not controlled as well. Other variables that might be related to the baby's verbal skill development are :
   - Nutrition Consumption
   - Type of Activity Engaged
   - Genetic Role
   - Age of Baby
   - Family history
   - Parent's availability to bond with children
   - Child going to day care or not (Y/N)

5. Given observational data for a randomly sampled selection of adults in the U.S. on Y = annual income and X = college degree (binary), answer the following questions:
   a. Following the approach from class, **what are some potential confounding factors?** (*State once what makes all of them potential confounding factors.*)
   b. For each potential confounding factor, **state whether it is a confounding factor**. **You can state once for all what makes something a potential confounding factor an actual confounding factor. **It is fine to do this as a table and add the separate statements requested*.

Potential Confounding Factor	Type	Reason
Family's wealth	Confounding Factor	Because these factors are : 1. Correlated/impact prediction of "income" 2. It is within the "U" of the model
Living area	Confounding Factor	
Individual intelligence (IQ)	Confounding Factor	
Study pursued in College (Major)	Confounding Factor	
Working Experience	Confounding Factor	

6. You are consulting a political campaign as a data scientist (after the election, so an ex post analysis). Given voter-level observational data on Y = voted for the campaign's candidate and X = targeted with an ad, answer the following questions:
   a. **Following the approach from class, what are some potential confounding factors?**
   b. **For each potential confounding factor, state whether it is a confounding factor.**

Potential Confounding Factor	Type	Reason
Number of Ads	Confounding Factor	Because these factors are : 1. Correlated with "targeted with an ad" 2. It is within the "U" of the model
The distribution area of the Ad	Confounding Factor	
Message in the Ad	Confounding Factor	
Type of Ad	Confounding Factor	

7. If you wish to make a passive prediction, **explain what kind of data you would need to have and why the other would not work**.
   We would need observational data (historical data) of variable that we believe to be related to the event that we are trying to predict. Experimental data would not work in this case, because experimental data requires us to actively change one or more variables to measure the impact on the outcome, doing so would be active prediction and not passive prediction.