Summary of <MACHINE LEARNING AS A TOOL FOR HYPOTHESIS GENERATION>

Zeli Chen

May 25, 2024

1. What are the research questions?

- Can and how ML algorithms generate novel and testable hypotheses?

2. Why are the research questions interesting?

- **Science is curiously asymmetric.**
  - How ideas are generated versus tested is noteworthy.
  - Idea generation by human is an empirical activity.
- **Two developments.**
  - ML algorithms can find patterns people might not notice.
  - Data on human behavior is exploding
- **Challenges:** Interpretable, and "black boxes" prediction of ML algorithms.

3. What is the paper's contribution?

- Contribute to the Literature that exploring hypothesis generation.
  - **Prior studies** use previous models and theories to generate hypothesis.
  - **This study** use ML to generate hypothesis.
- Contribute to the literature that exploring ML in prediction.
  - **Existing studies** build new models or measures.
  - **This study** focus on ML itself to generate hypothesis.

4. What hypotheses are tested in the paper?

- ***Hypothesis1*** ML algorithms can generate novel and interpretable hypotheses from high-dimensional data sets.
- ***Hypothesis2*** The generated hypotheses by ML improve the understanding and prediction of existing model.

a) These hypotheses answer the research question.

b) These hypotheses follow by a systematic procedure to generate hypotheses using ML algorithms.

5. Sample

- Using an example about judicial decision making in Mecklenburg county, North Carolina.

6. Dependent and independent variables

- Features of photos of the defendants captured by human and ML to make judge detain prediction, the comparison is correct.

7. Regression/prediction model specification

- Using GBDT to predict judge decisions using the structured administrative data, which is specific.

8. What difficulties arise in drawing inferences from the empirical work?

- There may be other results when using another ML model (replicability).

9. Describe at least one publishable and feasible extension of this research.

- Using ML to generate hypotheses to predict assert price. (stock price using

annual reports)

- Using ML to generate hypotheses to predict managers' behavior. (IPOs, M&As, mutual funds' holdings/returns)
- Using ML to generate hypotheses to predict macroeconomic.