Summary of *Machine Learning as a Tool for Hypothesis Generation*

The Quarterly Journal of Economics, 2024

Zhenting Hu

May 28, 2024

1. What are the research questions?

• How to generate novel hypotheses using machine learning algorithms?

2. Why are the research questions interesting?

• There is a significant asymmetry in the scientific process

  – Hypothesis testing is highly formalized but hypothesis generation remains informal.

• Machine learning offers the capability to detect patterns in large datasets

  – ML ethods can find novel patterns thanks to data exploding.

3. What is the paper's contribution?

• Contribution in novel hypotheses generation.

  – Prior studies generate hypotheses based on existing theories.

  – The paper develops a systematic procedure to generate hypo using ML.

• Contribution to ML in economic research.

  – **Prior**: Using ML for new measures, new models and casual inference tools.

  – **Extend**: Data-driven ML algorithms to a novel field.

4. What hypotheses are tested in the paper?

• H1: The algorithm can generate novel hypotheses about what facial features influence judicial decisions.

a) Do these hypotheses follow and answer the research questions?

• Yes.

b) Do these hypotheses follow from theory? Explain logic of the hypotheses

• Yes, the hypotheses are developed through empirical observation facilitated by machine learning algorithms, which found unknown patterns in mug shots

5. Sample: Comment on the appropriateness of the sample selection procedures.

• Large sample with high-dimension data can be found in financial studies too.

6. Dependent and Independent Variables.

• Detail the circumstances surrounding the current charge may help, such as whether it involved violence, use of weapons, or presence of victims.

7. Regression/prediction model specification.

• Use ensemble methods like Random Forest or XGBoost to combine the strengths of multiple models and improve predictive accuracy.

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

• Generating hypotheses cannot be totally independent from human?

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

• Apply the methods in feilds about voice data of Fed speak, explore speaking styles' influence on macro economy.