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

Jens Ludwig, Sendhil Mullainathan(Working Paper, 2023)

2024.05.28 YuQingYan

1. **What are the research questions?**
   - How to generate hypotheses that are both novel and testable using machine learning algorithms?
2. **Why are the research questions interesting?**
   - Human cognition is no longer the only way to notice patterns. Machine learning algorithms can also notice patterns, including patterns people might not notice themselves.
   - Data on human behavior is exploding, and this kind of information researchers once relied on for inspiration is now machine readable.
3. **What is the paper's contribution?**
   - Contribute to the literature on hypothesis generation.
     - Existing literature: Human researchers generate hypothesis based on existing theory or economic intuition.
     - In contrast: apply a systematic procedure to generate hypothesis using machine learning.
   - Contribute to the solution to the black box problem.
     - Existing literature: when the data is rich and high-dimensional, the machine learning algorithms is usually not inspectable.
     - Extension: procedure is not fully automatic and will be shaped and constrained by people. New concepts that humans do not yet understand cannot be produced.
   - Contribute to the literature on machine learning in economic research.
     - Existing literature: apply new measures and new models to researches.
     - Extension: apply data-driven machine learning algorithms to a novel field.
4. **What hypotheses are tested in the paper?**
   - Judge decisions are made based on the important features human labeled.
   - Features chosen by machine learning procedure can expain judge decisions.

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

   - The hypotheses are proposed under the context of judge decisions, the process of analyzing this problem is the answer to the research question.

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

   - Yes, the hypotheses are generally common sense.
5. **Sample: comment on the appropriateness of the sample selection procedures.**
   - The procedure this paper proposed is applied to the US criminal justice setting. This application includes clear decision making, also large amount and high-dimensional data can be used for analysis.
6. **Dependent and independent variables: comment on the appropriateness of variable definition and measurement.**

- The dependent variable is judge detain decision. The independent variables are features of defendants chosen that might influence judge decisions.

7. **Regression/prediction model specification: comment on the appropriateness of the regress/predict model specification.**

   - The regression explains how well the features can predict judge decisions.

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

   - The problem is whether this procedure overcomes the black boxes problem as ihe paper reported.

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

   - How can this procedure used for hypothesis generation applied to the generation of stock market factors?