Gradient Boosting in Opioid Distribution

Joshua Gabino, Amrit Gill, Alex Jacobs, Riley Rizzo, Regan Schulman, and Burke St. Claire
Texas A&M University
April 29, 2024

Abstract

When dealing with large amounts of high dimensional data, machine learning methods such as gradient boosting, help to go through large amounts of data in a effective and efficient way. More efficient machine learning techniques being integrated into studies done in the past can lead us to discerning more information.

Introduction

- Independent pharmacies distributed 39.1 percent more opioids than chain pharmacies, highlighting significant differences in drug distribution patterns based on ownership type.
- Gradient boosting was chosen due to feature importance and its ability to handle complex nonlinear interactions between variables.
- Machine Learning is introduced to provide a more nuanced and accurate model.

Literature Review

- Gradient Boosting is efficient in processing large volumes of pharmacy data (Paperspace; Analytics Vidhya).
- Limited research on ML systems for immediate intervention in drug abuse
- Addressing gaps can help the opioid epidemic in pharmacies

Methodology

- Original study uses a difference-in-difference model that tracks independent pharmacies' transition into a chain pharmacy
- Pharmacy fixed effects, time fixed effects, geographic fixed effects, and year-month fixed effects
- Using scikit learn's Gradient Boosting Regressor algorithm, we evaluated supply and demand features

Findings

Table: Regression Results Summary

Variable	Original	Replication
DPre	5.099	3.12
Dpost	-9.303	-8.824
Chain	-8.362	-6.229
Constant	32.036	29.86
Observations	5,055,761	5,071,787
R-squared	0.003	0.0018

Table: Error Comparison

Regression	MSE	ME
OLS	73.873	5457.15
OLS+	73.479	5399.20
Boosted	74.643	5571.525
Boosted+	73.755	5439.734

- Number of Observations does not match as original authors hand-picked several pharmacies from the cleaned data set
- The model with the best fit is the Gradient Boosted Regression with Labor Force and Unemployment Rate added

Discussion

The original paper used a non-traditional DiD design that may have skewed the results and the ML algorithim's accuracy. Despite that, the Gradient Boosting Regressor was still provided reliable results for feature determination.

Conclusions

Using Gradient Boosting, researchers can limit intensive analysis of multiple regressors by using Feature Importance to determine the most relevant influences on the outcome. Gradient Boosting is particulary useful for feature determination vs. precition in non-traditional DiD designs.

Appendix and References

Figure: Citations listed in QR Code