# Combating the Opioid Crisis: Insights from Machine Learning

Joshua Gabino, Amrit Gill, Alex Jacobs, Riley Rizzo, Regan Schulman, and Burke St. Claire
M. Jahangir Alam - ECON 470: Data Science for Economic and Social Issues - Spring 2024
Texas A&M University
May 4, 2024

## Abstract

Machine learning methods, such as gradient boosting, help to go through large amounts of data in a effective and efficient way. Integrating machine learning techniques into past studies 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 | Boosted+ |
|---|---|---|---|
| DPre | 5.009 | 3.110* | 2.544* |
| Dpost | -9.303* | -8.824* | -10.860* |
| Chain | -8.362* | -6.229* | -6.345* |
| Constant | 32.036* | 29.860* | 14.766* |
| LFPR | | | -2.41e-6* |
| Unemployment Rate | | | 2.079* |
| Observations | 5,055,761 | 5,071,787 | 4,695,416 |
| Mean Outcome | 27.14 | 26.79 | 25.80 |
| Root MSE | | 73.873 | 73.479 |
| R-squared | 0.003 | 0.0018 | 0.0093 |

- 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