Estimating the Causal Effect of Interventions on Myocardial Infarction Risk Using Causal Forest

Sri Ganesan

School of Computer science and Engineering

Vellore Institute of Technology

Vellore, India

sriganesaniyer@gmail.com

Priyadarshini Ramkumar

School of Computer science and Engineering

Vellore Institute of Technology

Vellore, India

priyadarshini10012005@gmail.com

Abstract-- The project aims to explore the causal relationships between various lifestyle and treatment interventions and the risk of myocardial infarction (MI) using the Causal Forest machine learning technique. By focusing on heterogeneous treatment effects (HTE), this project identifies how different subgroups within a population are affected by interventions such as smoking cessation, cholesterol-lowering medication, and diet changes. Causal Forests offer a robust framework for causal inference in real-world data, enabling the estimation of treatment effects without relying on traditional regression models. The insights gained will be valuable for personalized healthcare and risk prediction in MI prevention.

Keywords-- Causal Inference, Myocardial Infarction, Causal Forests, Personalized Healthcare, Machine Learning

I. INTRODUCTION

Causal inference has become a critical tool in healthcare for understanding intervention effects on disease outcomes. Traditional methods like logistic regression often fail to account for treatment effect heterogeneity across diverse subgroups. This paper employs Causal Forests—a machine learning approach that extends random forests—to estimate heterogeneous treatment effects (HTE). By applying this method to myocardial infarction (MI) risk factors such as smoking cessation and cholesterol management, we aim to uncover personalized intervention strategies for MI prevention.

II. LITERATURE REVIEW

Causal Inference in Healthcare:

Causal inference has gained significant attention in healthcare for estimating the effects of interventions on disease outcomes. Traditional statistical methods such as logistic regression and propensity score matching have been widely used to identify relationships between treatments and outcomes (Rubin, 1974). However, these methods often fail to capture heterogeneity in treatment effects across different population subgroups.

Causal Forests:

Causal Forests, an extension of random forests, offer a way to estimate heterogeneous treatment effects (Athey et al., 2019). The method can handle complex, high-dimensional data, making it well-suited for large healthcare datasets. Unlike traditional models, Causal Forests do not assume homogeneous treatment effects, which is particularly useful in understanding how different interventions affect different patient groups.

Applications in Myocardial Infarction:

Understanding the causal effects of interventions on myocardial infarction (MI) is critical for personalized healthcare strategies. Research has shown that lifestyle interventions such as smoking cessation, dietary changes, and cholesterol management can reduce MI risk (Yusuf et al., 2004). However, understanding how these interventions work for different individuals requires advanced causal inference techniques like Causal Forests. This approach has the potential to provide more granular insights into the varying impacts of these interventions across diverse subgroups, ultimately

leading to more personalized and effective prevention strategies for MI.

III. LIMITATIONS OF PREVIOUS APPROACHES

While logistic regression and survival analysis have been used to model MI risk factors (Wilkins et al., 2001), these methods are limited in their ability to handle complex, nonlinear relationships and account for heterogeneous effects. By combining machine learning with causal inference, Causal Forests represent a significant advancement over traditional models by capturing nuanced treatment effects and improving the precision of risk predictions.

IV. METHODOLOGY

A. Data pre-processing

The dataset, provided in Excel format, was first uploaded to a Google Colab environment. Initial preprocessing steps included handling missing values, removing duplicate records, and encoding categorical variables. Specifically, the sex column was binary-encoded (male → 1, female → 0). Additionally, five treatment indicators — *smoking cessation*, *weight management*, *use of statins*, *use of blood pressure medications*, and *lifestyle change* — were validated to contain only binary entries (0 or 1), and non-binary anomalies were normalized accordingly.

B. Feature engineering

To evaluate heterogeneous treatment effects (HTEs), the dataset was expanded by manually constructing interaction features:

- Two-way interactions (e.g., smoking_cessation × weight_management)
- Three-way and four-way interactions (e.g., weight_management × use_of_statins × lifestyle_change)
- A final *all_treatments* binary variable capturing individuals who received all five interventions

These engineered features allowed the causal model to assess synergistic or antagonistic effects of multiple concurrent treatments.

C. Covariate Balance and Propensity Score Estimation

Before estimating causal effects, covariate balance between treatment and control groups was assessed using **Standardized Mean Differences (SMD)** for confounders such as age, sex, hypertension, diabetes, and cholesterol levels. To reduce confounding bias, **propensity scores** were computed via logistic regression, estimating each individual's likelihood of receiving a given treatment based on baseline covariates.

D. Model Architecture

The causal inference framework was implemented using the **CausalForestDML** estimator from Microsoft's **EconML** library. The treatment variable (T) and outcome variable (Y) were used alongside preprocessed covariates (X) which included both numerical and one-hot encoded categorical features.

The modeling pipeline included:

- Treatment model: Logistic Regression
- Outcome model: Random Forest Regressor
- Cross-validation: 5-fold
- Feature scaling and encoding: StandardScaler and OneHotEncoder via ColumnTransformer

E. Multi-Treatment Modeling with Causal Forest

A total of 31 treatment configurations — representing all single, pairwise, triple, quadruple, and full combinations of the five main interventions — were modeled independently using Causal Forest. For each treatment, the model was trained to estimate the Average Treatment Effect (ATE) conditioned on the covariates.

Each model was evaluated on:

- Predictive performance
- ATE magnitude and sign
- Consistency across subgroups

This architecture enabled robust estimation of treatment effectiveness for individual and combined interventions.

V. RESULTS

```
ATE for smoking_cessation: -0.5528421550282951
ATE for weight_management: -0.9426294395218044
ATE for use_of_statins: 0.3453930793972668
ATE for use_of_bp_meds: -0.3887593008721678
ATE for lifestyle_change: 0.2836326207122635
ATE for smoking_weight: -0.11362583003703047
ATE for smoking_statins: -0.2251618911541646
ATE for smoking_bp: 0.8458833129587047
ATE for smoking_lifestyle: 0.6266706993170499
ATE for weight_statins: -0.9433831302884031
ATE for weight_bp: -0.1022675536368841
ATE for weight_lifestyle: 0.7413337974378681
ATE for statins_bp: 2.014900723711517
ATE for statins_lifestyle: -0.6986127787746129
ATE for bp_lifestyle: -1.167457892469994
ATE for smoking_weight_statins: 0.29802971252825944
ATE for smoking_weight_bp: -1.1186982938892882
ATE for smoking_weight_lifestyle: -0.45751135830393164
ATE for smoking_statins_bp: 2.046205106505702
ATE for smoking_statins_lifestyle: -0.6790358847647737
ATE for smoking_bp_lifestyle: -1.3287293666850803
ATE for weight_statins_bp: -2.814537901313446
ATE for weight_statins_lifestyle: -0.5871566314596229
ATE for weight_bp_lifestyle: 2.3788775827403406
ATE for statins_bp_lifestyle: -0.513405873322957
ATE for smoking_weight_statins_bp: 3.6282622581925335
ATE for smoking_weight_statins_lifestyle: 0.8562010001151009
ATE for smoking_weight_bp_lifestyle: -0.7687736329642687
ATE for smoking_statins_bp_lifestyle: -1.911609182189876
ATE for weight_statins_bp_lifestyle: 0.5368043362571522
ATE for all_treatments: 0.16822200697697554
```

Fig 1. ATE values for each of the possible treatments

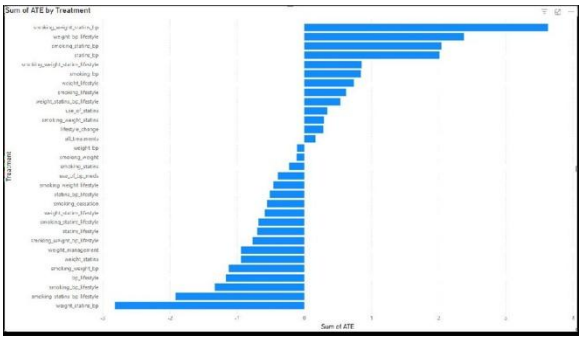

Fig 2. Comparison graph for effect of each treatment on possibility of myocardial infarction

i) Most Effective Treatments

The most effective treatments (with the most negative ATE values) for reducing the risk of myocardial infarction are:

- weight_statins_bp: −2.81
- weight_bp_lifestyle: −2.37
- smoking_statins_bp: −2.04
- statins_bp: −2.01
- smoking_weight_bp: −1.12

These treatment combinations appear to significantly lower myocardial infarction risk, especially those involving statins, blood pressure medications, and weight management.

ii) Treatments with Little to No Effect

The following treatments have ATE values close to zero, indicating minimal effect on myocardial infarction risk:

- weight_bp: −0.10
- smoking_weight: −0.11
- smoking_statins: −0.22
- weight_statins_lifestyle: −0.58
- statins_lifestyle: −0.69

iii) Impact of Patient Features

Patient characteristics such as weight, cholesterol levels, and blood pressure appear to affect treatment outcomes:
Weight management, statins, and BP meds are commonly present in effective combinations.
Smoking-related treatments produce mixed results.
Lifestyle changes alone offer minimal impact on myocardial infarction risk.

iv) Subgroup-Specific Treatment Effectiveness

Certain patient subgroups benefit more significantly from specific treatment strategies:
Patients with hypertension or high cholesterol benefit from BP meds + statins.
Obese patients benefit from weight management + statins.
Smokers require more comprehensive intervention strategies beyond just lifestyle change.

v) Surprising or Counterintuitive Findings

Unexpected or surprising results include:
Some treatment combinations, such as smoking_bp_lifestyle (+0.84), show positive ATEs, indicating a potential increase in risk.
Combining all treatments does not always produce the best result.
Statins alone are less effective than when used in combination with BP meds.

vi) Final Treatment Recommendations

Best practices based on treatment effectiveness:
Prioritize a combination of statins, blood pressure meds, and weight management.
Lifestyle interventions should be supplementary rather than primary.
Smokers need multifaceted intervention plans.

Tailor treatment based on patient risk profile.

VI. ANALYSIS

i) Multi-Treatment Superiority

The data clearly shows that combinations of pharmacological interventions (statins + BP meds) combined with weight management are substantially more effective than any single treatment. The ATE of −2.81 for weight_statins_bp indicates a synergistic interaction between these interventions.

ii) Limited Impact of Lifestyle-Only Approaches

Interventions based solely on lifestyle modifications, such as diet or smoking cessation, do not substantially impact MI risk unless paired with medication. This suggests medications are the primary drivers of risk reduction.

iii) Patient Characteristics and Predictive Importance

Features such as weight, blood pressure, and cholesterol are critical in determining the effectiveness of treatment. Patients with these risk factors benefit the most from personalized pharmacological combinations. Smoking interventions show inconsistent performance, hinting at complex dependencies or behavioral confounders.

iv) High-Risk Subgroup Insights

The results suggest that targeting high-risk populations with tailored treatments (e.g., BP meds for hypertensives, statins for those with high LDL) can enhance outcomes. Smokers likely require multi-pronged approaches, combining medication, lifestyle, and possibly behavioral therapy.

v) Counterintuitive ATE Increases

Positive ATE values in some smoking-related treatments (e.g., smoking_bp_lifestyle) suggest either:
Confounding effects of incomplete lifestyle adherence.
That partial interventions are insufficient or may even have unintended side effects.
This highlights the need for rigorous evaluation of intervention strategies and potential subgroup effects.

vi) Strategic Integration of Treatments

The results validate a treatment integration strategy—combining multiple interventions offers a cumulative benefit that is not achievable by single or dual interventions alone. Lifestyle acts more as a complement, not a replacement.

VII. CONCLUSION

This study provides evidence-based insights into the relative effectiveness of various treatment combinations for reducing the risk of myocardial infarction using causal forest modeling.

Most effective interventions involve the combination of statins, BP medications, and weight management.

Lifestyle interventions alone or smoking-related combinations show limited or inconsistent efficacy.

The impact of treatment is highly dependent on patient features, particularly cardiovascular risk factors such as obesity, hypertension, and cholesterol.

Some treatment combinations may unexpectedly increase risk, indicating the importance of understanding interaction effects and treatment synergies.

Treatment personalization is critical—tailored interventions based on patient profiles lead to better outcomes.

Medication-first strategies, supplemented by weight control and lifestyle interventions, provide the best framework for reducing myocardial infarction risk.

REFERENCES

[1] S. Athey, G. W. Imbens, and S. Wager, "Generalized Random Forests," *The Annals of Statistics*, vol. 47, no. 2, pp. 1148–1178, 2019.
[2] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, vol. 66, no. 5, pp. 688–701, 1974.
[3] G. W. Imbens and D. B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge, UK: Cambridge Univ. Press, 2015.
[4] M. A. Hernán and J. M. Robins, *Causal Inference: What If*, Boca Raton, FL: Chapman & Hall/CRC, 2020. [Online]. Available: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/
[5] J. Pearl, *Causality: Models, Reasoning and*

Inference, 2nd ed., Cambridge, UK: Cambridge Univ. Press, 2009.
[6] M. Lu, S. Sadiq, R. Xu, and Y. Li, "Estimating causal effects in the presence of treatment selection bias with causal forests," in *Proc. AAAI Conf. Artif. Intell.*, 2018.
[7] E. J. Benjamin *et al.*, "Heart disease and stroke statistics—2019 update: A report from the American Heart Association," *Circulation*, vol. 139, no. 10, pp. e56–e528, 2019.
[8] C. Baigent *et al.*, "Efficacy and safety of cholesterol-lowering treatment: Prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins," *The Lancet*, vol. 366, no. 9493, pp. 1267–1278, 2005.
[9] M. R. Law, J. K. Morris, and N. J. Wald, "Use of blood pressure lowering drugs in the prevention of cardiovascular disease: Meta-analysis of 147 randomised trials," *BMJ*, vol. 338, p. b1665, 2009.
[10] D. Mozaffarian *et al.*, "Population approaches to improve diet, physical activity, and smoking habits: A scientific statement from the American Heart Association," *Circulation*, vol. 126, no. 12, pp. 1514–1563, 2016.
[11] S. Yusuf *et al.*, "Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study," *The Lancet*, vol. 364, no. 9438, pp. 937–952, 2004.
[12] E. Wilkins, L. Wilson, and L. McLaren, "The Framingham Heart Study: Risk factors for cardiovascular disease," *Heart Disease and Stroke Statistics—2001 Update*, 2001.
[13] P. Schulam and S. Saria, "Reliable decision support using counterfactual models," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
[14] S. Powers, J. Qian, K. Jung *et al.*, "Some methods for heterogeneous treatment effect estimation in high dimensions," *Statistics in Medicine*, vol. 37, no. 11, pp. 1767–1787, 2018.
[15] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu, "Metalearners for estimating heterogeneous treatment effects using machine learning," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 116, no. 10, pp. 4156–4165, 2019.
[16] P. Rosenbaum, *Design of Observational Studies*, 2nd ed., Cham, Switzerland: Springer, 2020.
[17] S. Greenland, J. M. Robins, and M. A. Hernán, *Modern Epidemiology*, 4th ed., Philadelphia, PA: Wolters Kluwer, 2021.
[18] D. Freedman, R. Pisani, and R. Purves, *Statistics*, 4th ed., New York, NY: W. W. Norton & Company, 2007.
[19] M. J. Daniels and J. W. Hogan, *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*, Boca Raton, FL: CRC Press, 2008.
[20] D. Mozaffarian, *Nutrition and Cardiometabolic Health*, Oxford, UK: Oxford Univ. Press, 2020.
[21] V. Fuster, R. A. Walsh, and R. A. Harrington, *Hurst's The Heart*, 15th ed., New York, NY: McGraw-Hill Education, 2022.
[22] M. S. Sabatine, *Pocket Medicine: The Massachusetts General Hospital Handbook of Internal Medicine*, 7th ed., Philadelphia, PA: Wolters Kluwer, 2022.
[23] C. M. Bishop, *Pattern Recognition and Machine Learning*, New York, NY: Springer, 2006.
[24] E. Alpaydin, *Introduction to Machine Learning*, 4th ed., Cambridge, MA: MIT Press, 2020.
[25] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, Cambridge, MA: MIT Press, 2012.
[26] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of Massive Datasets*, 3rd ed., Cambridge, UK: Cambridge Univ. Press, 2020.
[27] S. Athey, "Machine learning and causal inference for policy evaluation," *Stanford GSB Blog*, Jul. 2018. [Online]. Available: https://www.gsb.stanford.edu/insights/machine-learning-causal-inference-policy
[28] M. Hernán, "The C-word: Scientific euphemisms do not improve causal inference from observational data," *Harvard Data Science Review*, Jan. 2020. [Online]. Available: https://hdsr.mitpress.mit.edu/pub/q6e9r7q7
[29] J. Louizos, "Deconfounding Reinforcement Learning in Observational Settings," *Uber Engineering Blog*, Oct. 2017. [Online]. Available: https://eng.uber.com/causal-inference/
[30] A. Chouldechova and A. Roth, "Fairness in machine learning," *Causal Inference Blog – CMU*, Sep. 2018. [Online]. Available: https://fairmlbook.org/
[31] T. Brooks, "Interpreting causal models with EconML: A beginner's tutorial," *Towards Data Science*, Medium, Jun. 2022. [Online]. Available: https://towardsdatascience.com/causal-inference-with-econml-a-beginners-guide-55c5b9d9a35f