Heart Attack App and Heart Attack Prevention

Group 3: Kyu Sung Cho, Liang Gong, Yezi Liu, Yijing Sun

March 2024

Business Problem & Objectives

- Heart disease is a leading cause of death for people of most races in US.
- Roughly half of all Americans face risks of having heart attack.
- Aim to build heart attack predicting models & investigate factors that cause heart attack to help educate the public & increase awareness
- Ultimate implementation: a self-testing App that takes in user information, predicts heart attack probability, and generates personalized health analysis

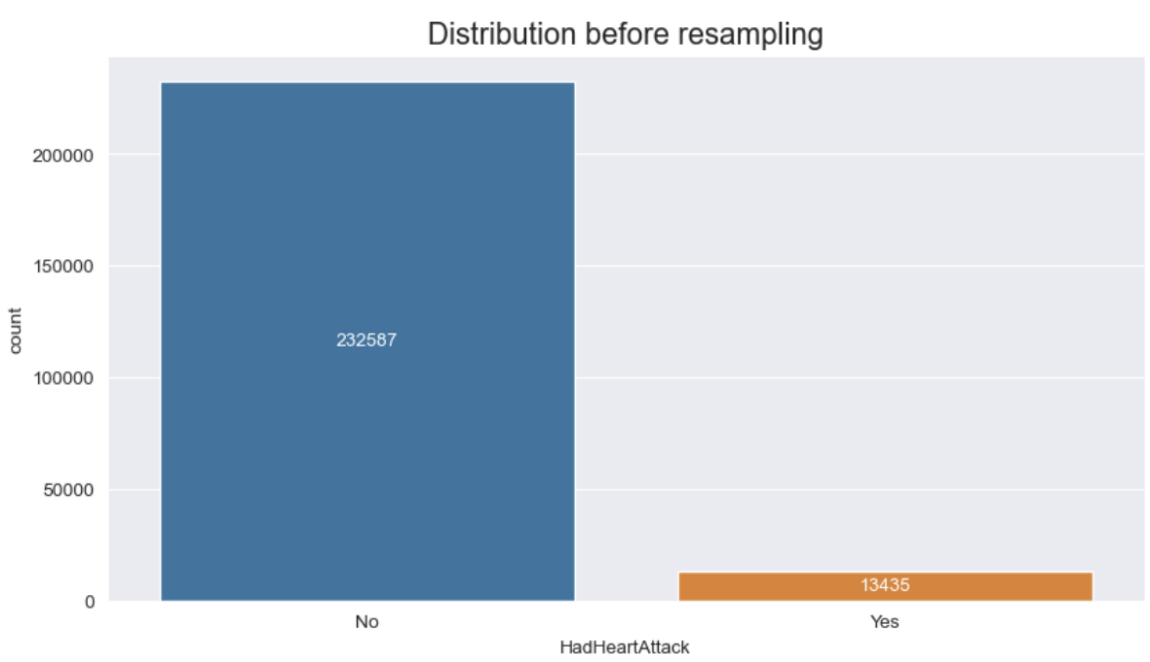
Data & Feature Selection

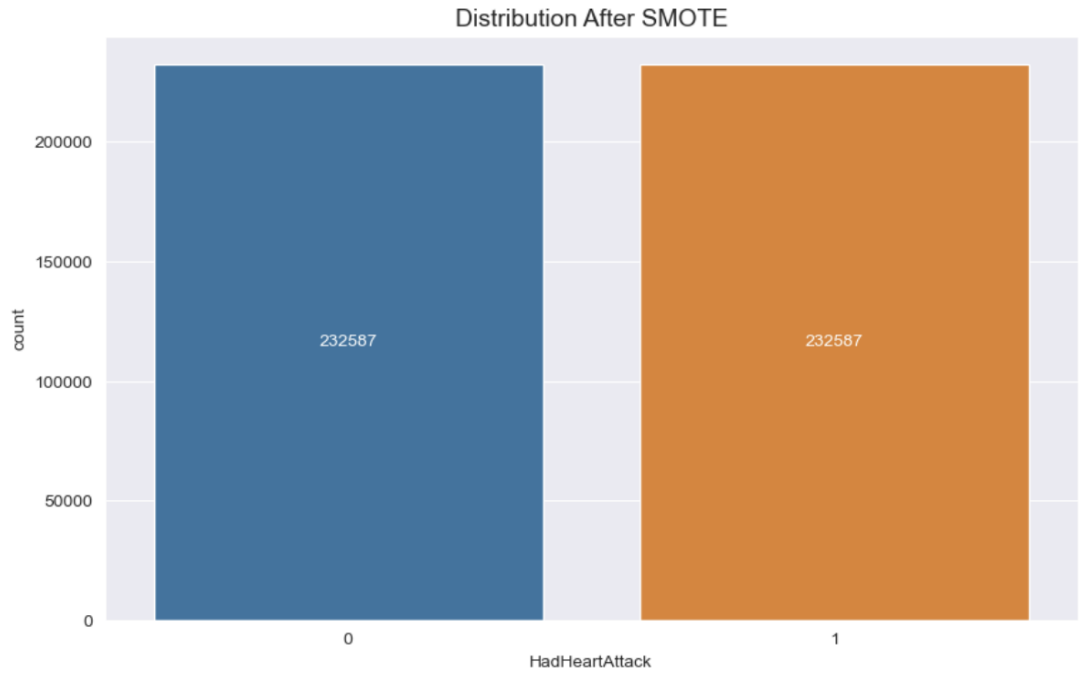
- Dataset Introduction
  - From CDC, major part of Behavioral Risk Factor Surveillance System that conducts annual telephone surveys to collect data on health status of US residents.
  - 40 columns with 246,022 records
  - Target binary variable is "*HadHeartAttack*"
  - Raw features include people's demographics, medical history, and important heath indexes.
    - Sex, BMI, SmokeStatus, SleepHours, HadAngina, HadStroke, HadKidneyDisease, DifficultyWalking, etc.

- Feature Selection & Transformation
  - Used SMOTE (oversampling technique) to balance target variable
  - Dropped '*WeightInKilograms*' due to multicollinearity with BMI
  - Transformed *'state'* to *'region'* with fewer categories
  - Applied Logistic Regression Backward Selection after one-hot encoding

Synthetic Minority Over-sampling Technique (SMOTE)

- A huge imbalance in *"HadHeartAttack"* classes (Target Variable)
- Oversampling balances the number of samples between classes
  - synthetic samples are generated for the minority class





Model 1 Logistic Regression

- Model Performance on Test Set: 0.945 AUC score; 0.872 F1 score
- Model Summary & Coefficients Interpretation

RemovedTeeth_All		-0.9856	0.023	-42.520	0.000
-1.031	-0.940				
RemovedTeeth_None of them		-1.2603	0.014	-92.827	0.000
-1.287	-1.234				
HadAngina_Yes		2.5544	0.019	137.029	0.000
2.518	2.591				
HadStroke_Yes		0.4309	0.026	16.372	0.000
0.379	0.483				
HadAsthma_Yes		-0.6426	0.022	-29.758	0.000
-0.685	-0.600				
HadSkinCancer_Yes		-0.6174	0.021	-29.043	0.000
-0.659	-0.576				

'*HadStroke_Yes*' Coefficient Interpretation

One-unit increase in "HadStroke" (from 0 to 1) is associated with an increase in the log-odds of having a heart attack by 0.4309.

e^0.4309 ≈1.54(sigmoid link function)

Individuals who have had a stroke are estimated to have approximately 1.54 times higher odds of experiencing a heart attack compared to individuals who have not had a stroke.

Model 1 Logistic Regression Cont.

- Model Summary & Coefficients Interpretation

	coef	std err	z	P>|z|	[0.025	0.975]
const	9.6304	0.222	43.434	0.000	9.196	10.065
PhysicalHealthDays	0.0138	0.001	16.760	0.000	0.012	0.015
MentalHealthDays	-0.0048	0.001	-5.019	0.000	-0.007	-0.003
SleepHours	-0.1309	0.004	-32.147	0.000	-0.139	-0.123
HeightInMeters	-1.1759	0.214	-5.503	0.000		

"*SleepHours*" Coefficient

One additional hour of sleep, the log odds of having a heart attack decrease by 0.1309

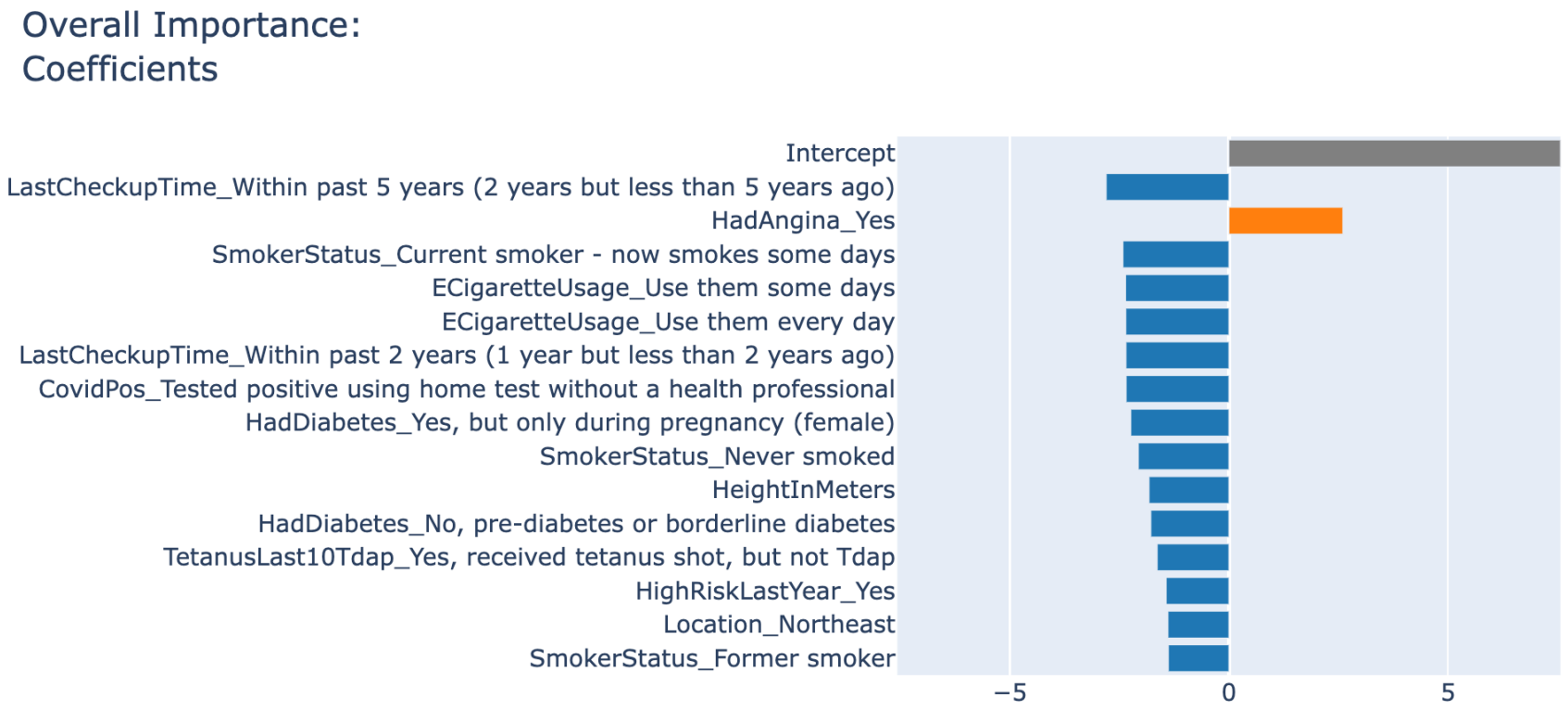
e^(-0.1309) ≈ 0.88(sigmoid link function)

For each additional hour of sleep, the odds of having a heart attack decrease by a factor of 0.88. So each additional hour of sleep reduces the odds of having a heart attack by about 12%.

1 − 0.88 = 0.12 or 12%

Model 1 Logistic Regression

- Feature Importance from '*interpret*' package



Model 2 Explainable Boosting Machine (Classifier)

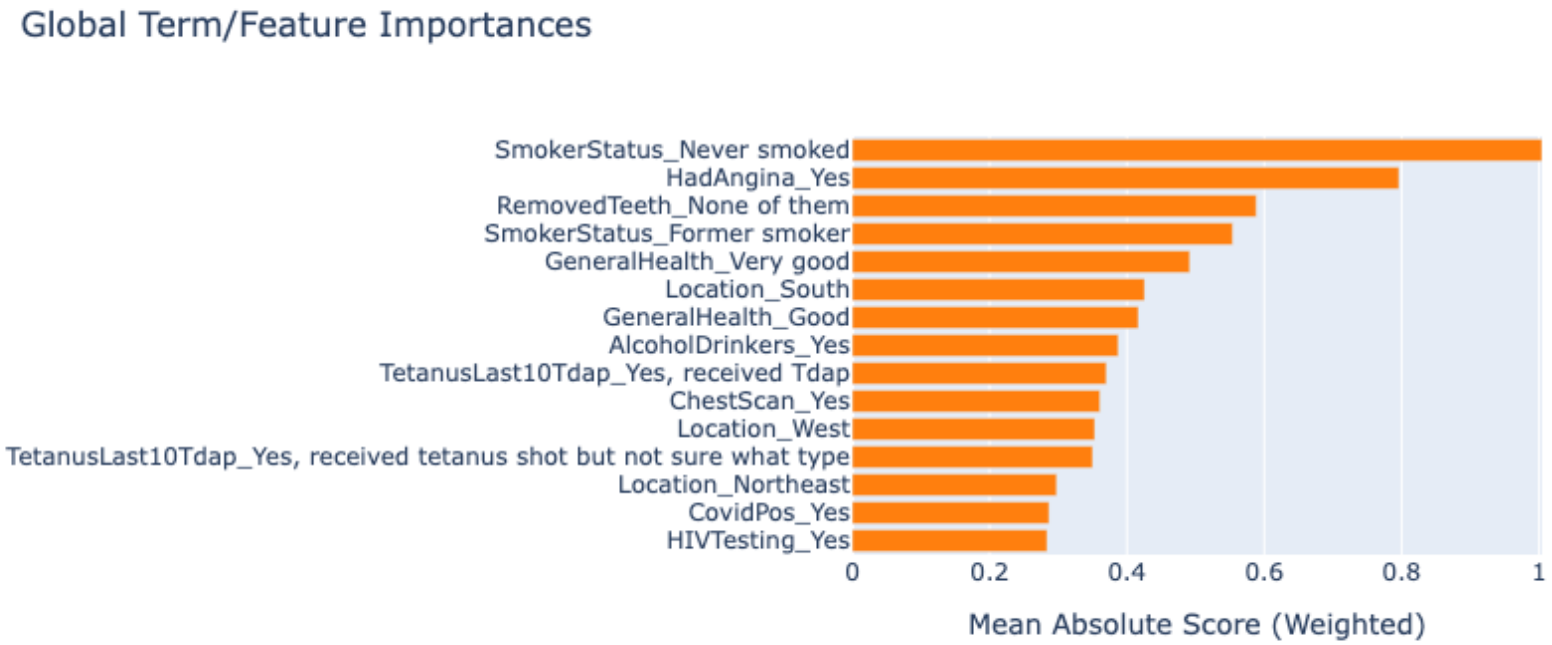
- What is Explainable Boosting Machine:
  - A tree-based, cyclic gradient boosting Generalized Additive Model
  - Can detect interactions between features
  - Can achieve accuracy comparable to black box models
- The form of the Generalized Additive Model that EBM is:

$$g(E[y]) = \beta_0 + \sum f_j(x_j)$$

- Key improvements over traditional GAMs:
  - learns each feature function (f_j) using techniques like bagging and gradient boosting
  - automatically detects and includes pairwise interaction terms
    - Enhance the accuracy while maintaining interpretability

Model 2 Explainable Boosting Machine (Classifier)

- Model Performance on Test Set: 0.951 AUC score; 0.880 F1 score
- Model Summary Graph (Feature Importances)



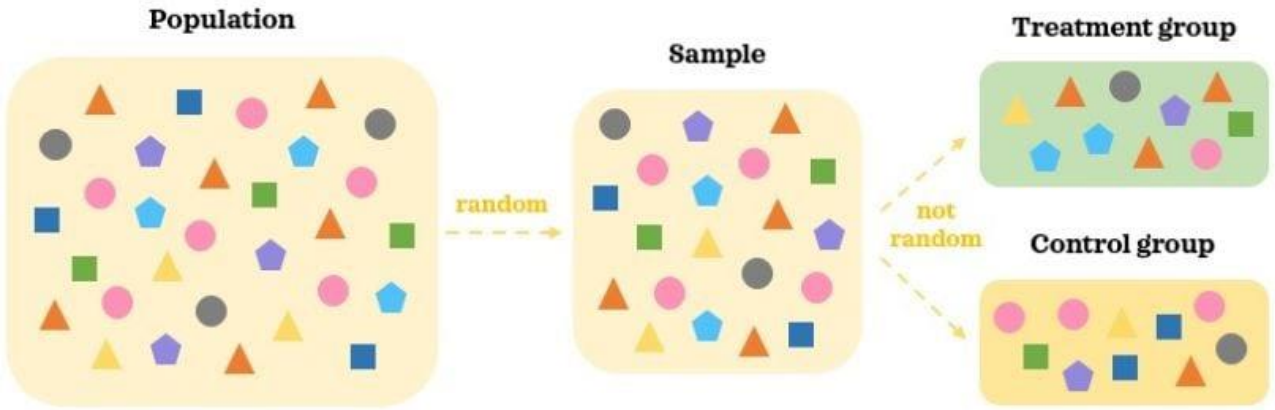
Why Causal Inference Analysis

- Based on the previous models, stroke and diabetes are correlated with heart attack.
- However, it is hard to conclude that stroke and diabetes cause the difference in heart attack by 53.7% (0.431 log odds) and –9.8% (-0.103 log odds).
- For example, if individuals with diabetes are more likely to consume fewer carbohydrates in their diet, the difference in heart attack may be caused by the diet rather than diabetes.
- To investigate whether there is a causal relationship between diabetes and heart attack, we need to use causal inference to estimate the effect of diabetes on heart attack.
- In real world, there are always some unobserved factors that may affect diabetes and heart attack at the same time, we call the bias associated with these factors as omitted variable bias.

Propensity Score Matching Approach

- Propensity scores matching approach balances differences between treatment and control groups for a more accurate assessment of the causal effect of the treatment.

- We used *causalinference* package to employ logistic regression for estimating the propensity score, then weighted the observations by propensity scores estimated using all covariates.



PSM Result Interpretation

Diabetes - Treatment Effect Estimates

	Estimate	P-value
ATE	0.016	0.002
ATC	0.012	0.035
ATT	0.037	0.000

PSM Result Interpretation

Kidney Disease - Treatment Effect Estimates

	Estimate	P-value
ATE	0.005	0.591
ATC	0.004	0.686
ATT	0.030	0.000

PSM Result Interpretation

Stroke Condition - Treatment Effect Estimates

	Estimate	P-value
ATE	0.054	0.000
ATC	0.052	0.000
ATT	0.107	0.000

Limitations regarding PSM

Limitations to proving causality with PSM

- **Unmeasured Confounders**: PSM can only account for observed and measured covariates. Any unmeasured or unknown confounders can still bias the results, potentially obscuring true causal relationships. Eg, cholestoral level or high blood pressure can influence both the treatment and the outcome.
- **Quality of the Model**: The accuracy of propensity score estimation depends on the selection of covariates and the model used. Mis-specification of the model or omission of relevant covariates can lead to biased estimates. Eg, omission of relevant covariates that we are not able to collect.
- **Assumption of No Unmeasured Confounders**: PSM relies on the strong assumption that there are no unmeasured confounders affecting both the treatment and outcome. This assumption is untestable with the data at hand and can be a significant source of bias.
- **Conduct Sensitivity Analyses**: These analyses assess how sensitive the results are to potential unmeasured confounding, providing insight into the robustness of the causal inference.

While PSM can reduce bias and support stronger causal inferences than simple observational analyses, it cannot fully prove causality, especially in the presence of unmeasured confounders.

Frameworks designed for causal analysis – Future work

Randomized Controlled Trials (RCTs)

- **Applicability**: it involves manipulating individuals' sleep hours in a controlled environment. researchers would follow a structured protocol to randomly assign participants to different sleep duration groups and then assess the incidence of heart attacks over a specified follow-up period. Randomization helps ensure that any differences in heart attack risk between groups are due to sleep duration and not other factors.

Instrumental Variable (IV) Analysis

- **Applicability**: IV analysis requires finding a variable that influences "SleepHours" but does not directly affect the risk of heart attacks except through its impact on sleep. This could be something like daylight saving time changes, work schedule policies, or other external factors that impact sleep patterns.

Difference-in-Differences (DiD)

- **Applicability**: DiD requires data over time and ideally a natural experiment or policy change that affects "SleepHours" for part of the population but not others. This method is powerful for longitudinal data where you can compare changes over time between groups.

Regression Discontinuity (RD)

- **Applicability**: RD relies on a cutoff-based assignment to treatment. For sleep hours, sleep less than 8 hours, or more than 8 hours.

The best approach involves using multiple methods where possible to **triangulate evidence** and strengthen the case for causal inference, acknowledging the limitations and assumptions inherent in each method.

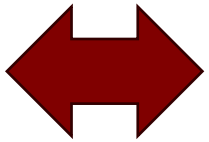
Thank You

Question?

Appendix 1: Logistic Regression vs. EBM

Logistic Regression

Accuracy: 0.873
Recall : 0.869
Precision: 0.875
F1-score: 0.872
AUC: 0.944



EBM

Accuracy: 0.880 (+0.007)
Recall : 0.879 (+0.01)
Precision: 0.881 (+0.006)
F1-score: 0.880 (+0.008)
AUC: 0.951 (+0.007)

Appendix 2: Model 2 Explainable Boosting Machine (Classifier)