Causal Structure and Causal Inquiries for Health Features in the United States Population

Hiram G. Menendez

Springboard – Data Science

Capstone Project 3

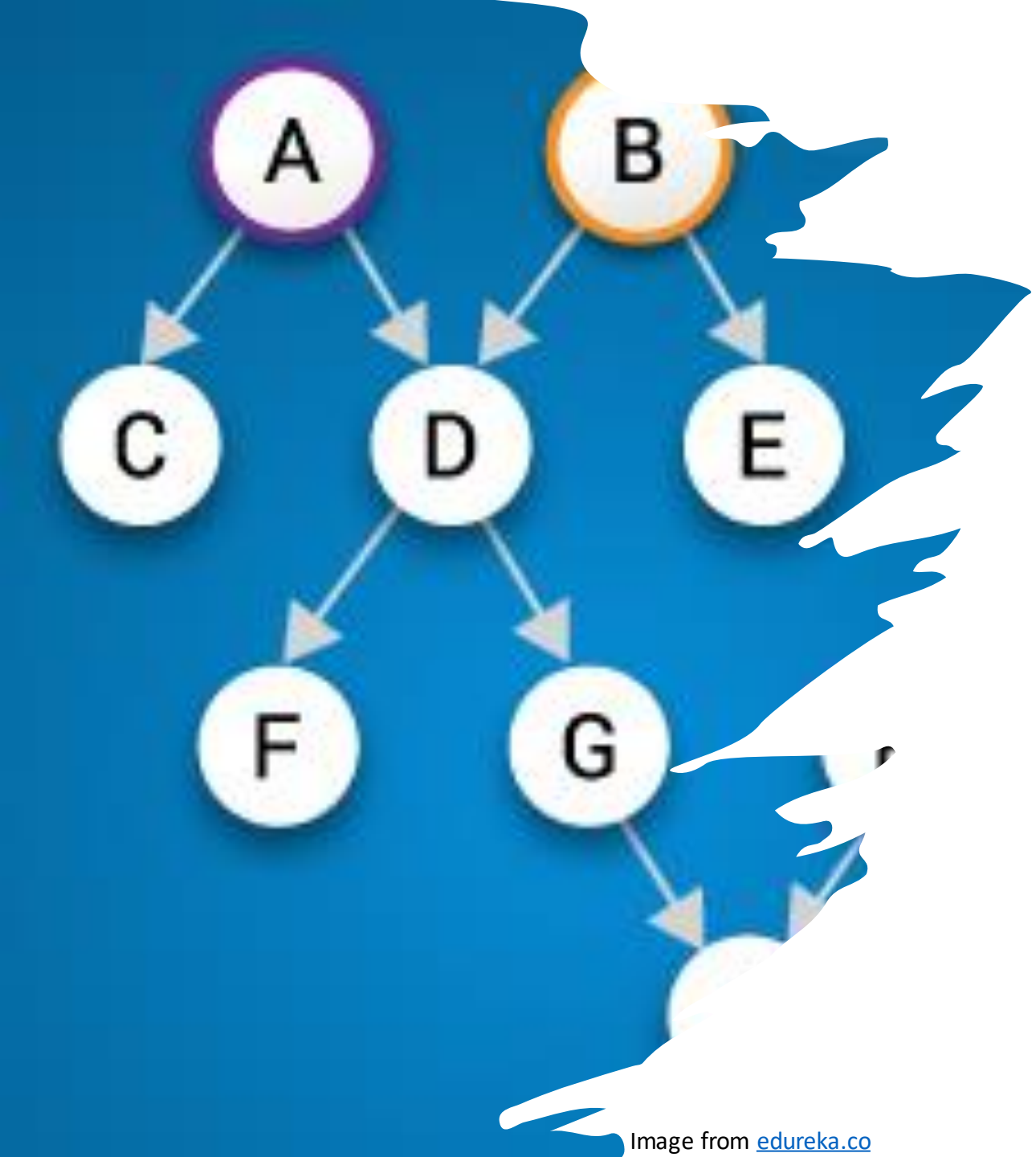


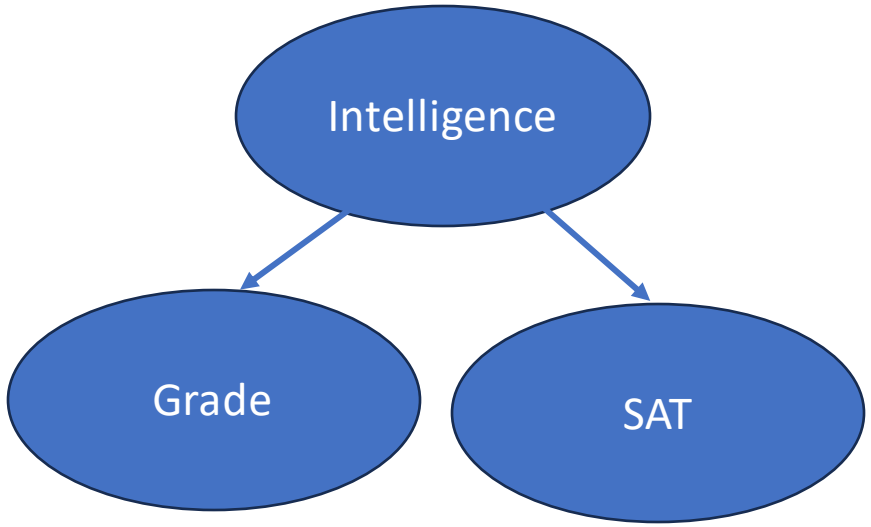
Image from edureka.co

Causality in Machine Learning

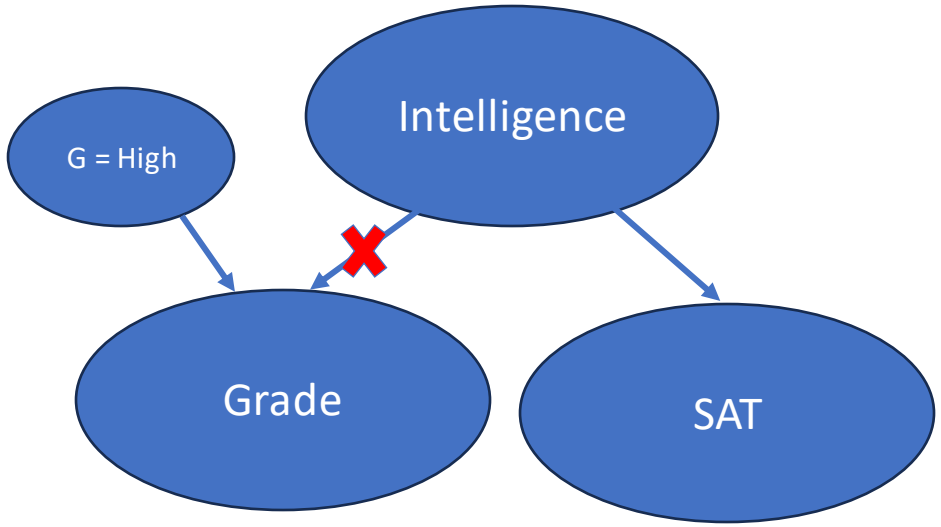
- Predicting feature values of data instances is an important goal in data science.
  - "Individuals who are smokers are more likely to have heart disease."
- In non-causal models, conclusions about the causal structure of the data cannot be made.
  - "Does smoking cause heart disease?"
- In some applications, knowing the causal structure is essential.
  - "If the patient quits smoking, will his risk of developing heart disease decrease?

DAGs and Interventions

- Causal structure is described with Directed Acyclic Graphs (DAGs).
- DAG changes when making an intervention in the random outcome-generating mechanism.



Correlation between "Grade" and "SAT" not causal



DAG after intervening on "Grade". Correlation between "Grade" and "SAT" no longer observed.

Warning: Correlation is still observed if we condition on Grade = High

Intervening ≠ Conditioning

Example source: "Probabilistic Graphic Models: Principles and Techniques", D. Koller and N. Friedman

Project Objectives

1. Develop models that describe the causal structure of health features of individuals in the United States Population.
2. Evaluate their performance in classification tasks.
3. Perform causal queries to evaluate to effectiveness of health interventions.

The Data

- Data used was the Behavioral Risk Factor Surveillance System for the year 2015.
- Contains responses to hundreds of health-related questions about:
  - General health
  - Hypertension
  - Arthritis
  - Depression
  - Demographics
  - Seatbelt use

Images from https://www.cdc.gov/

The Models

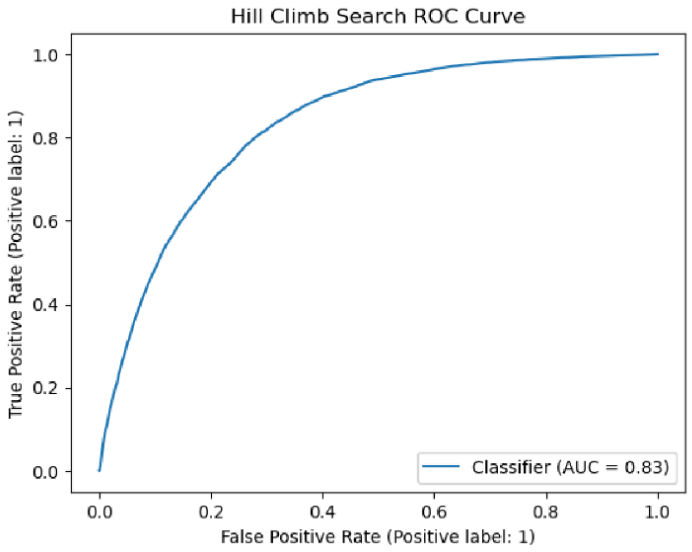
- 10 categorical features were used:
  - Good Health
  - Hypertension
  - High Cholesterol
  - Smoker Status
  - Age Category
  - Diabetes
  - High Sodium
  - Heavy Drinker
  - Heart Disease
  - Sex

- Model Components:
  1. DAG - Describes causal structure
  2. Joint Probability Distribution - Gives answers to queries

- Two algorithms used:
  1. Hill Climb Search
  2. Mmhc (Max-Min Hill Climb)

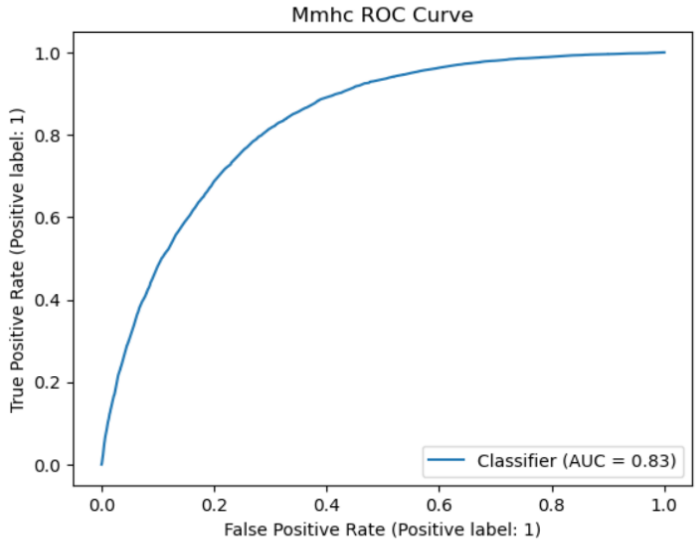
Classification Results

Causal Models



Hill Climb Search Classification Report

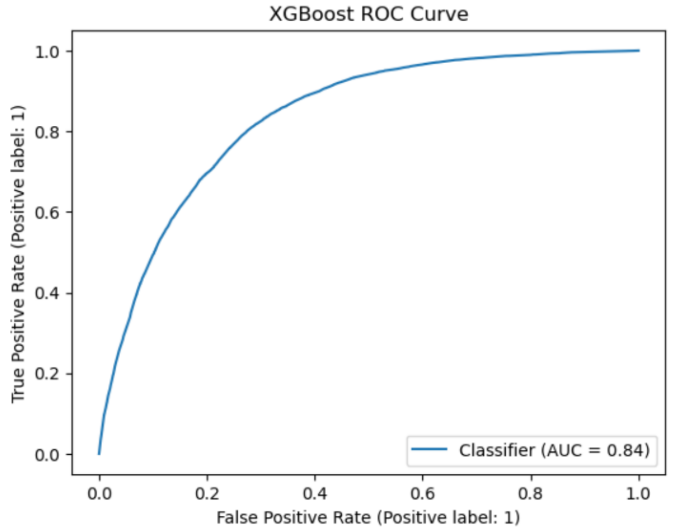
	precision	recall	f1-score	support
0	0.97	0.71	0.82	79776
1	0.21	0.81	0.34	7727
accuracy			0.72	87503
macro avg	0.59	0.76	0.58	87503
weighted avg	0.91	0.72	0.78	87503



Mmhc Classification Report

	precision	recall	f1-score	support
0	0.97	0.72	0.83	79776
1	0.21	0.80	0.34	7727
accuracy			0.72	87503
macro avg	0.59	0.76	0.58	87503
weighted avg	0.91	0.72	0.78	87503

Non-Causal Model



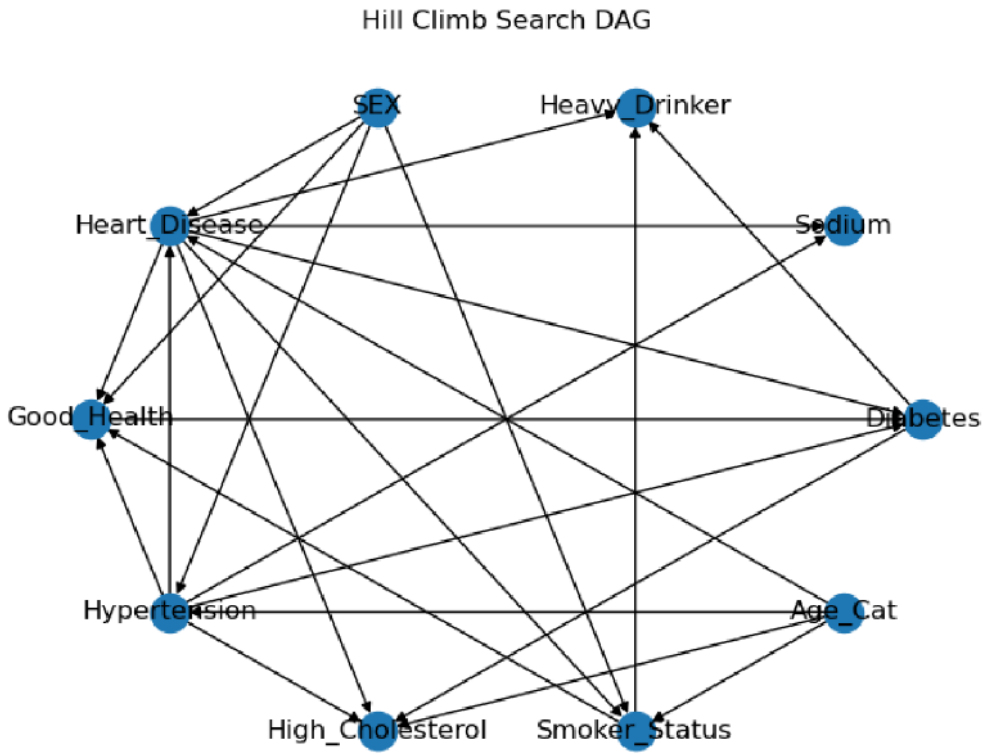
XGBoost Classification Report

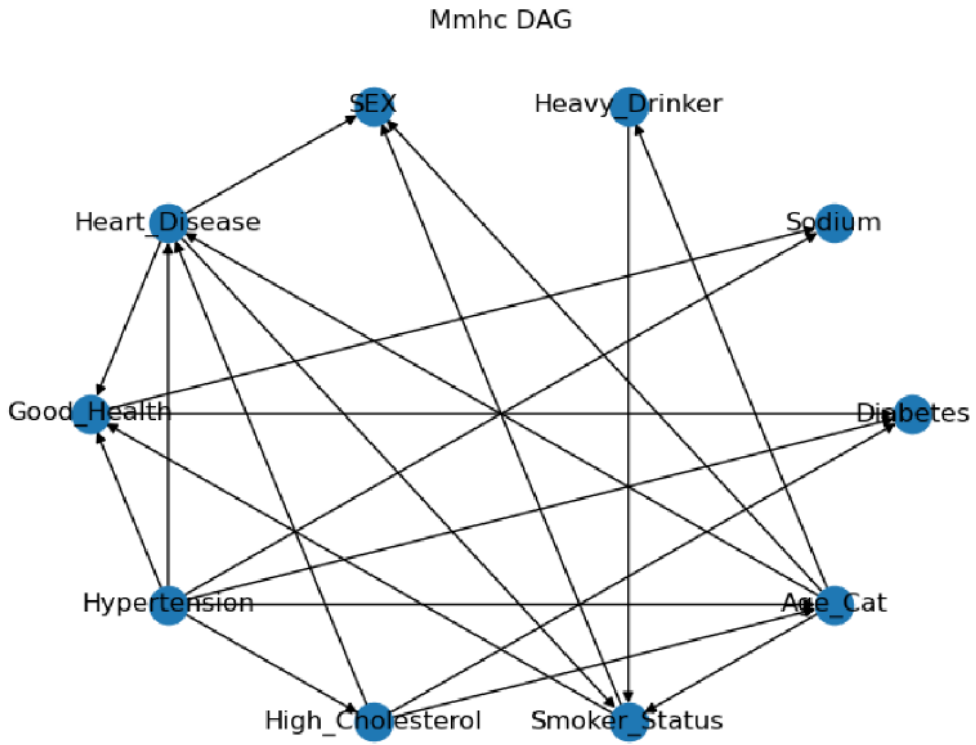
	precision	recall	f1-score	support
0	0.98	0.71	0.82	79776
1	0.22	0.81	0.34	7727
accuracy			0.72	87503
macro avg	0.60	0.76	0.58	87503
weighted avg	0.91	0.72	0.78	87503

Classification Performance is surprisingly
similar between Causal and Non-Causal Models

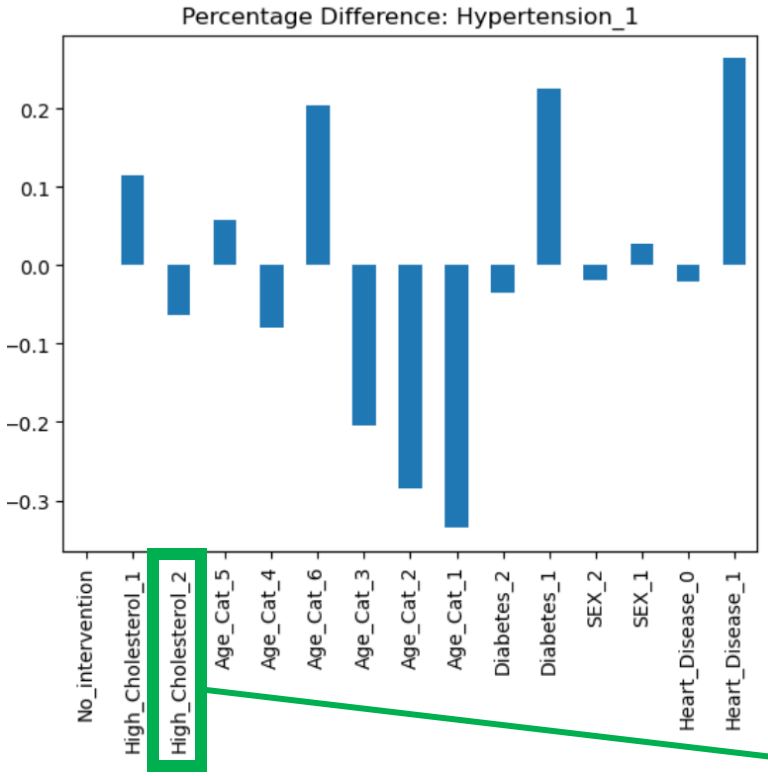
(Possible) Causal Structures

- Best DAGs found by algorithms based on how well the associated joint probability distribution can reconstruct the data.
- Algorithm cannot unambiguously determine correct causal structure (only observational data available).
- Expert domain knowledge often essential for arriving at the correct causal structure.





Causal Inference



- Each bar shows the difference in probability when doing an intervention and not doing one.
- The difference is due to the causal influence the intervened-on feature has on the feature of interest.
- In other words, the bar heights tell us how the probability of having a health feature changes if we intervene in the lifestyle of a patient.
- These conclusions depend on the DAG being a correct representation of the causal structure.

Example: If we "force" the patient to have low cholesterol, the probability of developing hypertension decreases by about 5%

Photo by Alexander Gray on Unsplash

Using Models in a Medical Setting

- The models can be used by physicians to inquire about how the risk of a health feature changes if the patient is forced to change lifestyle.
- The known features of the patient can be used in the inquiry.
- Example: "What is the probability that an individual develops heart disease if the patient is forced to not smoke given that we know the individual has diabetes".

Conclusions

- Causal models seem to be equally effective at classification tasks than non-causal models.
- We discovered possible causal structures for the health features of the model. However, it is hard to tell if they are correct without domain knowledge.
- The models are able to inquire about how the risk of a health feature changes if a patient is "forced" to change his or her lifestyle. However, this depends on the correctness of the causal structure.

Photo by Milad Fakurian on Unsplash

Thank You!