Who Determines United States Healthcare Out-of-pocket Costs?

Factor Ranking and Selection Using Ensemble Learning

Chengcheng Zhang
Claremont Graduate University, Department of Economic Sciences

Yujia Ding, PhD
Claremont Graduate University, Institute of Mathematical Sciences

Qidi Peng, PhD
Claremont Graduate University, Institute of Mathematical Sciences

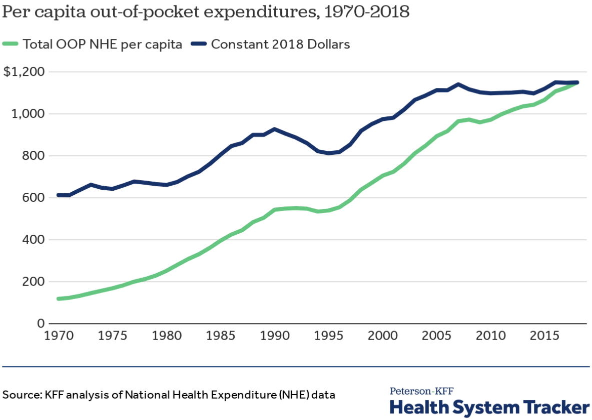
January 7-9, 2022

Motivation

- OOP costs: Medical care expenses that aren't reimbursed by insurance.

Motivation

- OOP costs: Medical care expenses that aren't reimbursed by insurance.
- Fast rise in OOP



.

Motivation

- High OOP increases financial burden.

Motivation

- High OOP increases financial burden.
- Various factors that cause high OOP costs have been discovered by research.
    - e.g. Females spend considerably more than males (Cylus et al 2010)

Motivation

» High OOP increases financial burden.

» Various factors that cause high OOP costs have been discovered by research.

- e.g. Females spend considerably more than males (Cylus et al 2010)

» However, up until now, no one has considered studying the above factors jointly or has ranked their importances.

Motivation

- High OOP increases financial burden.
- Various factors that cause high OOP costs have been discovered by research.
  - e.g. Females spend considerably more than males (Cylus et al 2010)
- However, up until now, no one has considered studying the above factors jointly or has ranked their importances.
- Our research goal is to fill this gap.

Overview

- Self-designed voting ensemble learning procedure to detect OOP costs level change.

Overview

- Self-designed voting ensemble learning procedure to detect OOP costs level change.
- Research Question:
  - Who determines OOP costs in the United States? Which factor is more important than the others when they are considered jointly?

Overview

» Self-designed voting ensemble learning procedure to detect OOP costs level change.

» Research Question:
  - Who determines OOP costs in the United States? Which factor is more important than the others when they are considered jointly?

» Findings:
  - The selected top-ranking factors, in order of importance, are: insurance type, age, asthma, family size, race, and number of physician office visits.
  - The predictive models using these factors perform better.

Structure of the research

» 2016-17 MEPS Data

Demographic		Age Sex Race Region Family size Primary language not English English proficiency Marital status Born in the U.S. Years in the U.S. Year
Socioeconomic		Family income Individual's wage income Hourly wage level Employment status Self-employment status Occupation groups Purchased food stamps
Health status	Chronic condition	High blood pressure Coronary heart disease Stroke Bronchitis High cholesterol Cancer Diabetes Asthma Arthritis Joint pain
	Functional limitation	Serious hearing difficulties Serious seeing difficulties Serious cognitive difficulties Cognitive limitation Physical functioning limitation Work/Housework/School limitation (Any limitation) Used assistive devices
	Self-reported health status	Perceived health status Perceived mental health status Number of physician office visits
Health insurance		Type of health insurance coverage

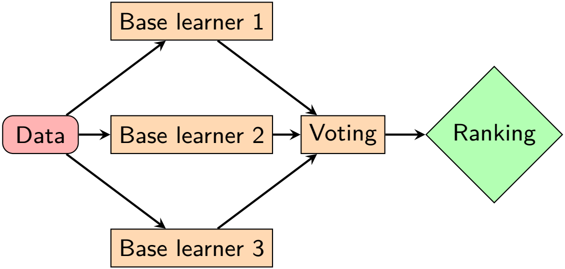
Structure of the research

- 2016-17 MEPS Data
  - 18 - 64
  - 2016 - 2017
  - 39 predictors
- Correlation detection
  - strong dependencies, yield inconsistent and misleading variable selection outputs
  - reduce collinearity
  - 13 variables are removed

Structure of the research

- 2016-17 MEPS Data
  - 18 - 64
  - 2016 - 2017
  - 39 predictors
- Correlation detection
  - strong dependencies, yield inconsistent and misleading variable selection outputs
  - reduce collinearity
  - 13 variables are removed
- Perform an ensemble learning for variable selection on the dataset obtained from correlation detection process.

Ensemble learning procedure



Ensemble learning procedure

Data

↓

Step 1: Response variable transformation ($p \to \log(p/(1-p))$)

↓

Step 2: Variable selection (best subset, random forests, LASSO)

↓

Step 3: Variable votes ranking (forward stepwise, backward stepwise, mean decrease in accuracy by random forests, mean decrease in node purity by random forests)

↓

Step 4: Model validation

Results

Table 5 Variable importance rankings.

Variable	Ranking				Score
	Forward	Backward	Mean decrease in accuracy	Mean decrease in node purity	
Type of insurance coverage	2 *	2 *	1 *	2 *	4
Age	3 *	3 *	6	3 *	3
Asthma	1 *	1 *	20	18	2
Family size	4 *	7	5 *	6	2
Race	6	5 *	2 *	8	2
Number of physician office visits	20	20	3 *	1 *	2
Family income	5 *	10	16	9	1
Primary language not English	7	4 *	12	10	1
Sex	10	9	4 *	12	1

$*$ denotes the variable ranks among top five under the corresponding criterion.

Results

Table 6 Variables recommended by literature and data-driven solutions.

Recommended by Literature	Recommended by Data-driven Solutions
Type of insurance coverage	Type of insurance coverage
Age	Age
Sex	Asthma
Family income	Family size
Race	Race
Number of physician office visits	Number of physician office visits

Table 8 Comparison of the training MSE and test MSE.

Method	Recommended by Literature		Recommended by Data-driven Solution	
	Training MSE	Test MSE	Training MSE	Test MSE
Linear regression	0.456971	0.457194	0.371514	0.371977
Random forests	0.339891	0.473549	0.311900	0.393755
Ridge	0.458609	0.458814	0.375946	0.376285
LASSO	0.459222	0.459375	0.384613	0.384935

MSEs of the four models are calculated using variables (in Table 6) recommended by literature and data-driven solutions.

» data-driven recommended variables all result in lower training MSE and test MSE

Future study

- examining how determinants influence OOP costs for individuals who have more healthcare needs.
  - different age groups
  - people have chronic diseases

Thank you!

Contact: chengcheng.zhang@cgu.edu

Publication version: https://doi.org/10.1007/s13755-021-00153-9

Source code: https://github.com/health-care-cost-data-analysis/factor-ranking-and-selection