China Aging Population Analysis:

A Machine Learning Method of Health Status Prediction

Grace Huang, Qingfeng Yu,
Qingya Yang, Yuyan He

Introduction

Research Question

– Which individual characteristics are most important for predicting self-reported health?

– How well can overall health status be predicted using variables that are commonly available in survey or administrative data?

Research Goals

- Develop predictive models of self-reported health using CHARLS data.
- Compare model performance using out-of-sample prediction metrics.
- Examine the relative importance of predictors across different age groups.

Research Importance

- An aging Chinese society: Need for timely and scalable tools to assess health risks among older populations.
- Limited social security resources: Approximate health risks without relying on costly large-scale surveys.
- Regional inequalities: Identifying key predictors of poor health across age groups to better target interventions

Data Source

China Health and Retirement Longitudinal Study (CHARLS), 2018

– Organizer: Peking University, Wuhan University

– Sample: Nationally representative survey of Chinese adults aged 45 and above. Multi-stage sampling, >10,000 households, >19,000 individuals.

– Variables: Provides individual-level information on demographics, family structure, health status, biomarkers, income, insurance…

National Bureau of Statistics (Provincial-level data), 2018

– Variables: GDP, registered population, health-related variables

Individual-level CHARLS data merged with provincial-level indicators based on province of residence

Allows incorporation of regional economic and healthcare context into health prediction models

Variable Selection

Demographic

- Age
- Gender
- Marital status
- Living with a partner
- Number of family members
- Religion

Socio-economic

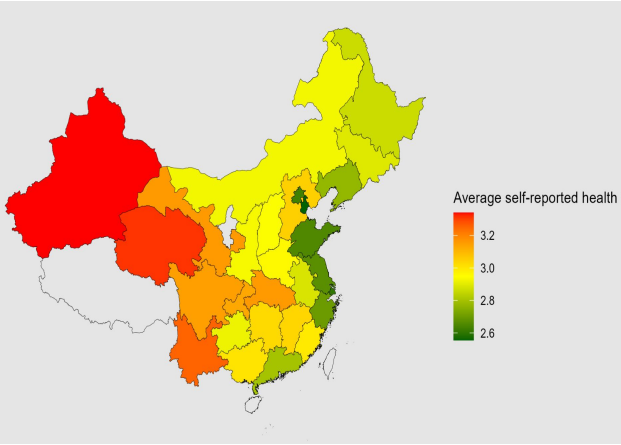
- Income
- Pension
- Farming status
- Living area (urban/rural)
- Real-estate ownership

Healthcare resources

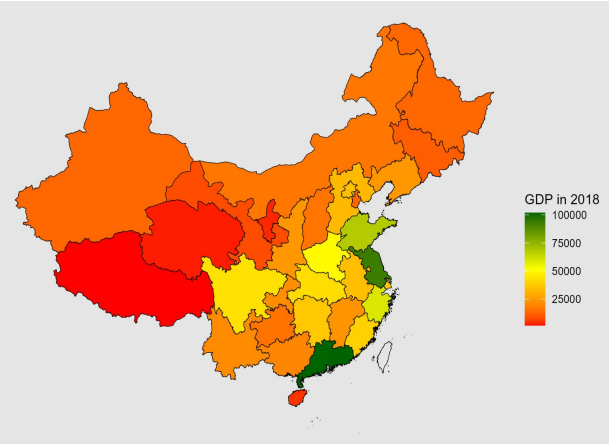
- Hospital density per 10k
- Healthcare professionals per 10k
- Medical insurance coverage

GIS Analysis

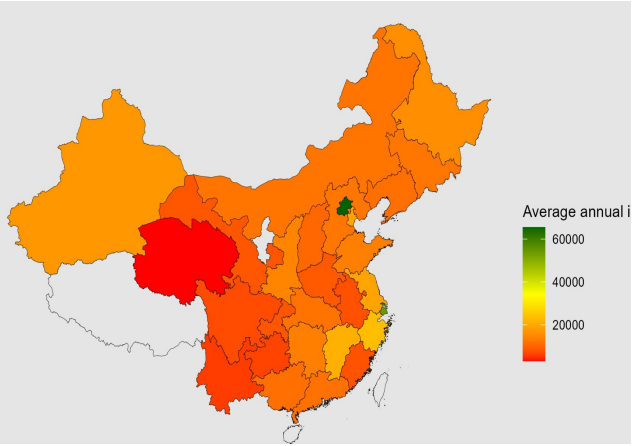
Average Self-reported Health



GDP by Province, 2018



Average Annual Individual Income by Province



Takeaway from GIS

– The spatial distribution of residents' health is related to their economic status.
– Residents in coastal areas generally enjoy better health and economic conditions.

Model Methodology

Model Selection
LASSO and Random Forest chosen

- Handle high-dimensional data

- Tracking variable importance.

Dual Approach:
Regression and Classification

- Regression preserved 5-point health scale
- Classification recoded health status into binary: Good (1-3) vs Bad (4-5).

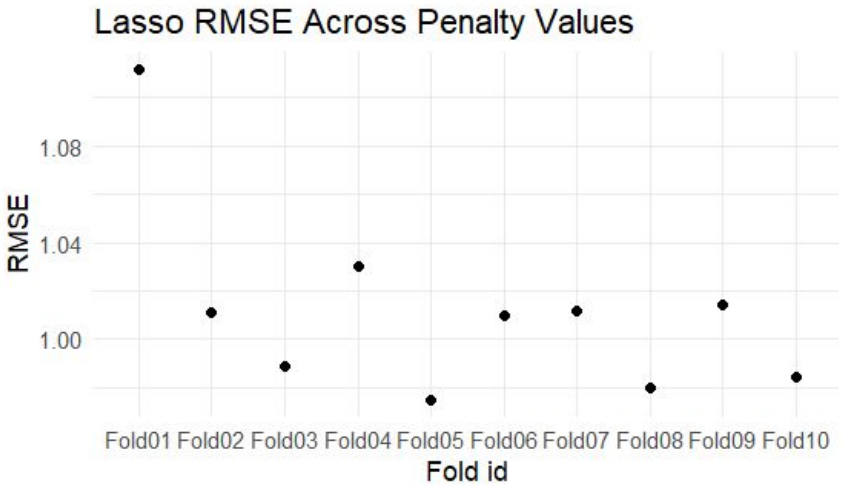
Standard Machine Validation Process

- 80/20 train-test split
- 10-fold cross-validation
- Hyperparameter tuning for robust results.

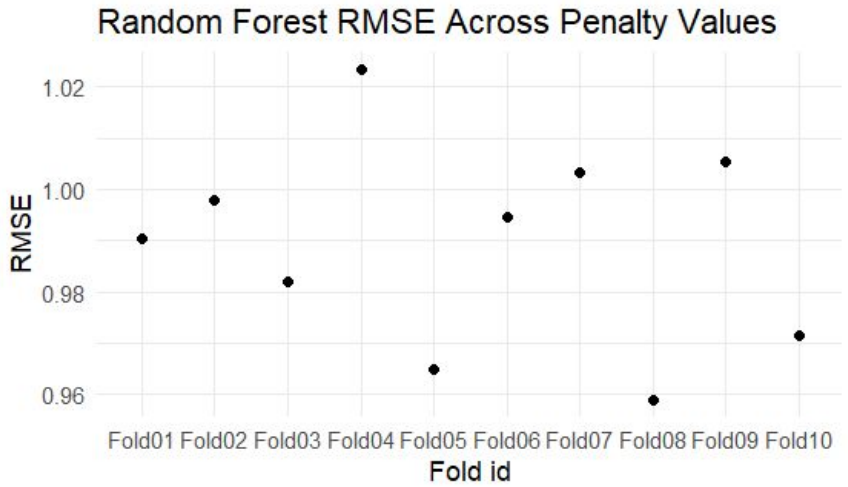
Model Performance & Selection

Regression Task

Random Forest achieved lower average RMSE on full 5-point scale



Avg RMSE: 1.008

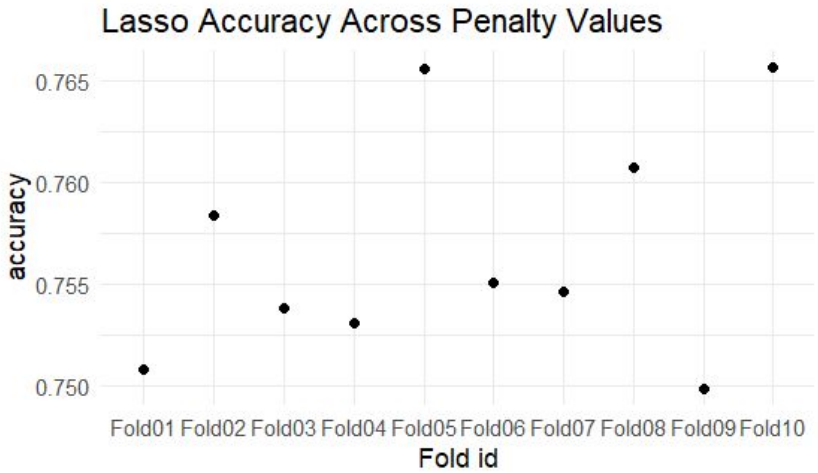


Avg RMSE: 0.989

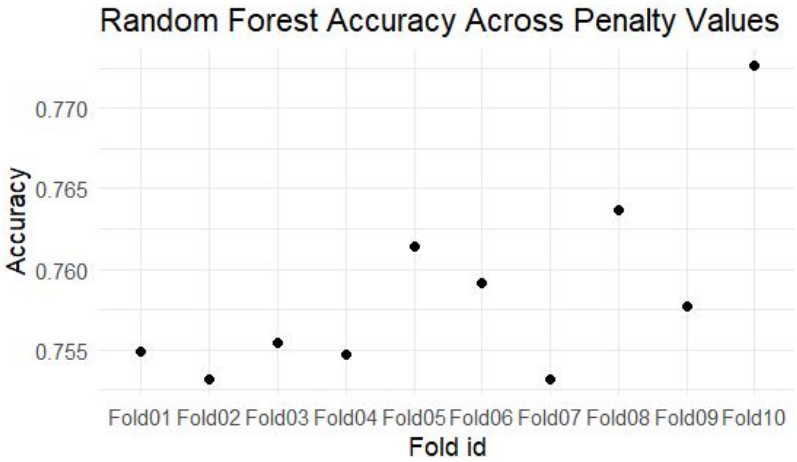
Model Performance & Selection (cont')

Classification Task

Both LASSO and RF yielded 75% accuracy after recoding; RF slightly outperforming Random Forest



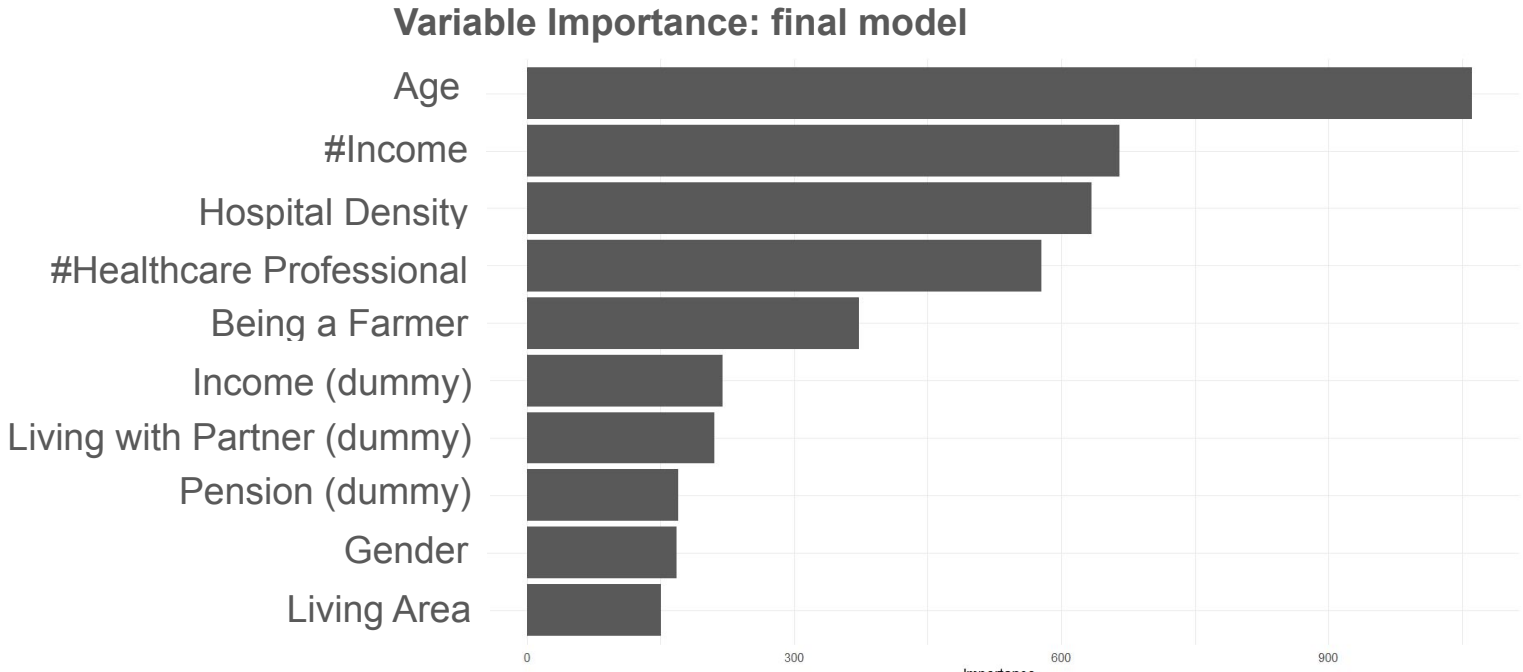
Avg Accuracy: 0.757



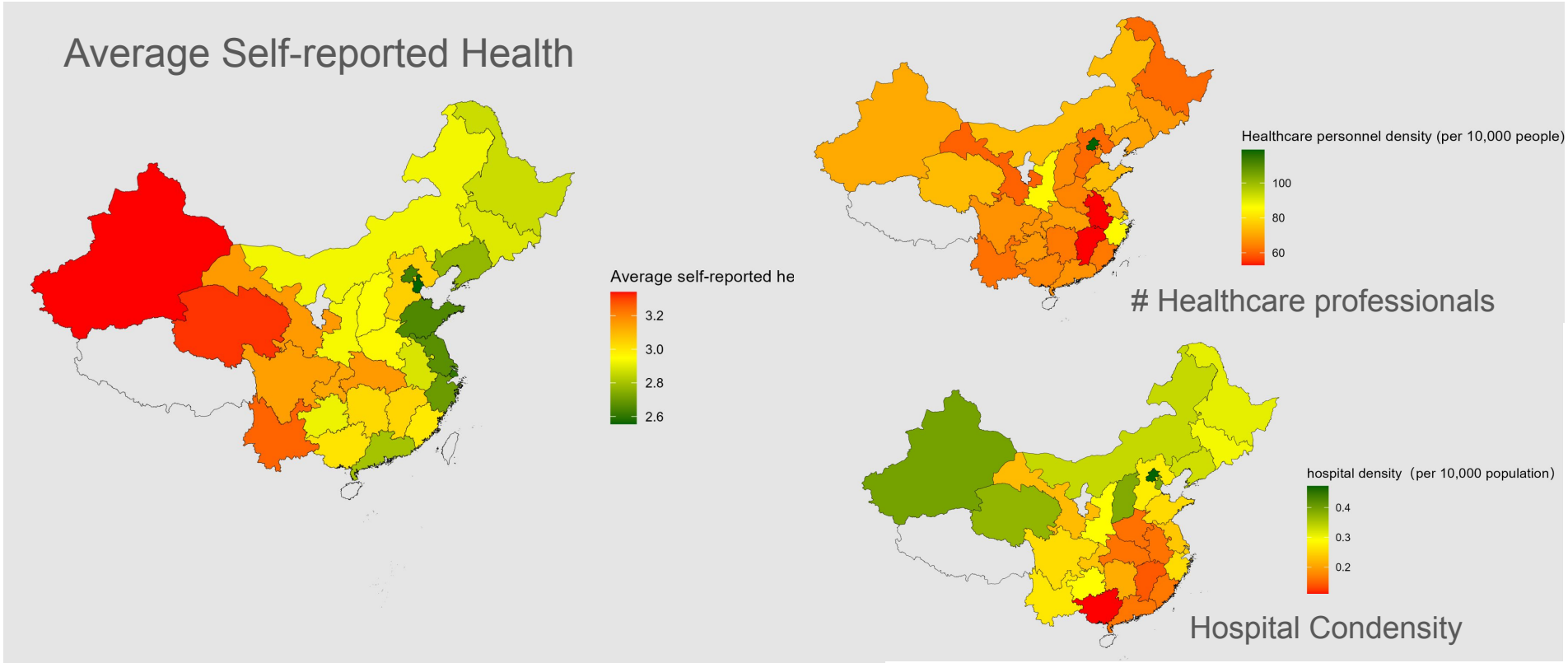
Avg Accuracy: 0.759

After Model Selection: Predictors Deep-dive Analysis

Age and **Healthcare Resources** emerged as most significant predictors. Two-part follow-up analysis revealed critical patterns



Follow-up #1: Spatial Alignment of Healthcare Resources



Follow-up #2: Variable Importance by Age

Age groups:

Younger: < 53 y.o.

Younger to Middle: 53 ~ 61 y.o.

Middle to Elder: 61 ~ 68 y.o.

Elderly: >68 y.o.

Variable categories:

Demographic: marital status, living with partner, # children, gender etc.

Socio-economic: income, pension, #property owned, living area etc.

Healthcare Resource: # hospital, # healthcare professionals, medical insurance coverage.

Follow-up #2: Variable Importance by Age (cont')