DNAnexus
@Chai_Arkarachai chai@dnanexus.com

Prioritizing Risk Factors of Heart Failure from UK Biobank Using Explainable Artificial Intelligence

Arkarachai Fungtammasan, Chiao-Feng Lin, Daniel Quang, Yih-Chii Hwang , Peter Nguyen, Jason Chin

502K samples from biobank uk

Heart Failure label from ICD9, 10

Commonly studied traits for heart failure

Baseline
- Sex
- Age
- Ethnic
- BMI
- Townsend Deprivation

Blood Measure
- Total cholesterol
- LDL
- HDL
- C-reactive protein
- HbA1c

Self-reported Lifestyle Survey
- Alcohol consumption
- Smoking
- Eating fresh fruit
- Eating cooked vegie
- Eating raw vegie
- Eating processed meat
- Eating
- Sleeping
- Walking
- Sitting
- Moderate physical activity
- Vigorous physical activity

Data processing
- Combining multiple measurements
- Marking missing values
- Encode categorical variables
- Balance case vs control using undersampling
- Remove or impute missing values

Analysis

Glassbox (model that could natively provide insight into how features are used to make prediction)
e.g. Logistic Regression, Explainable Boosting Machine

Blackbox (model that decision making is convoluted and need inference method to elucidate)
e.g. Extreme Gradient Boosting (XGBoost), Light Gradient Boosted Machine (LightGBM), Random Forest.

Figure 1: Relative importance and trend of impact for each feature

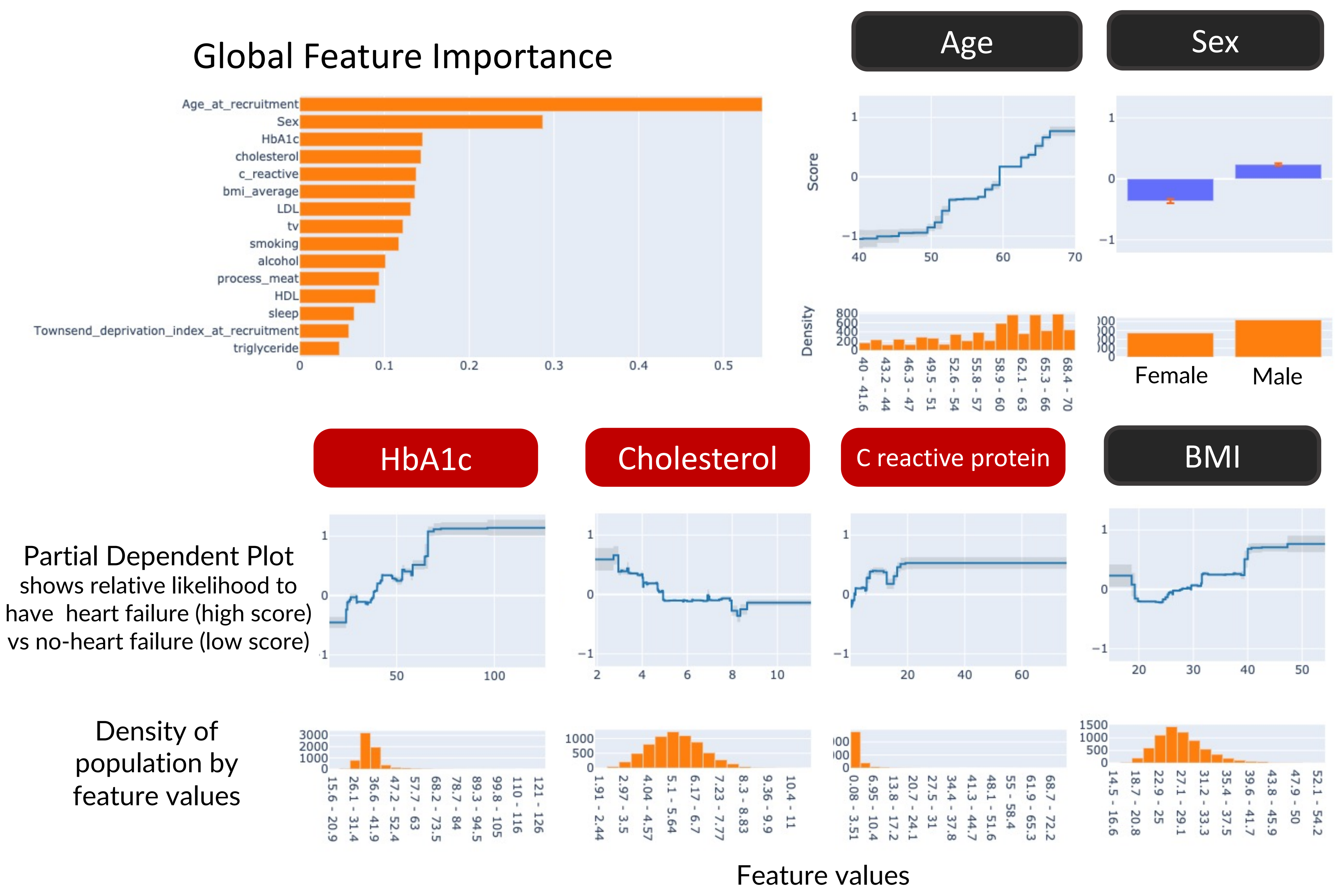

Table 1: Performance of various heart failure prediction model

	XGBoost 1	LightGBM 1	XGBoost 2	LightGBM 2	Random Forest	Logistic Regression	Explainable Boosting Machine
Feature importance inference	Blackbox: infer feature importance using SHAP					Glassbox: Natively assign feature importance	
Missing value handling	Sparsity-aware split finding (loss minimizing imputation)		Complete Case Analysis (remove any rows with missing values)				
Training F1	0.7677	0.7503	0.7918	0.7819	1	0.7190	0.7419
Test F1	0.7307	0.7380	0.7280	0.7340	0.7293	0.7357	0.7300
Top 5 most important features (Ranked)	Age Sex BMI Cholesterol HbA1c		Age Sex Cholesterol BMI HbA1c		Age Cholesterol Sex HbA1c BMI	Age LDL Sex Cholesterol BMI	Age Sex HbA1c Cholesterol C-reactive

Key Findings

1. Complex blackbox and glassbox machine learning methods yield similar predictive power (Table 1)
2. Age and sex are the most contributed features for boosted tree models regardless of algorithms, missing value handling strategies, or glassbox vs blackbox feature importance analysis. However, no good concordance is found with other types of models. (Table1 & Figure1)
3. Except age and sex, other most important features have strong impact in minority of population (see second row of Figure 1: high score impact locate in low density of population's feature values).

Notebook: https://github.com/Arkarachai/UKB46926_XAI_ASHG2021

Citations: https://github.com/scikit-learn/scikit-learn; https://github.com/scikit-learn-contrib/imbalanced-learn; https://github.com/interpretml/interpret; https://github.com/slundberg/shap

This research has been conducted using the UK Biobank Resource under Application Number '46926' using UKB Research Analysis Platform. We thank all the participants in the UK Biobank study.