CS 109A/Stat 121A/AC 209A: Final Project
Protopapas, Rader, Tanner

Mortality prediction and interpretation

Problem statement

What are the factors that increase the risk of someone dying? Many, one would say. We will not examine them all in this project, we will concentrate on a subset of the factors included in the NHANES I Epidemiologic Follow-up Study. NHEFS is a national longitudinal study that was jointly initiated by the National Center for Health Statistics and the National Institute on Aging in collaboration with other agencies of the Public Health Service. The NHEFS was designed to investigate the relationships between clinical, nutritional, and behavioral factors assessed in the first National Health and Nutrition Examination Survey NHANES I and subsequent morbidity, mortality, and hospital utilization, as well as changes in risk factors, functional limitation, and institutionalization.

For more details visit: `https://wwwn.cdc.gov/nchs/nhanes/nhefs/`.

Project goal

The goal for this project is to build and evaluate a data-driven model for predicting a person's risk of dying based on a set of clinical data and biochemical measurements. You will then try to **interpret** the model's decisions by looking at how each of the features contributed to the outcome of the model both individually and through pair-wise interactions with other features. For the latter part you may use one of the tools for interpreting models such as SHAP [**?**].

Data

The data is available from the National Center for Health Statistics and getting access to them can be a long process; one needs to register and apply for access. For the purposes of this project we will provide you with a pre-processed subset of NHANES I data. You may find them in files *NHANESI_X.csv* (for the features) and *NHANESI_y.csv* (for the return variable).

Because the data have been pre-processed for you, there is minimal data cleansing and imputation, so the focus of this project will be more on the building and interpretation of the model. You will also need to do some reading of relevant literature to understand the features. You may use any skills learned in class (or not learned in class for the 209 component) such as feature selection, feature engineering, trying more than one learning models, evaluating the models, and interpreting the importance of each feature.

Features in the dataset are:

1. Age,
2. Diastolic BP,
3. Poverty index
4. Race
5. Red blood cells
6. Sedimentation rate
7. Serum Albumin

8. Serum Cholesterol
9. Serum Iron
10. Serum Magnesium
11. Serum Protein'
12. Sex
13. Systolic BP
14. TIBC
15. TS
16. White blood cells
17. BMI
18. Pulse pressure

Our target variable is the log-odds of someone with those characteristics dying after some time.

High-level project goals

1. Perform thorough exploratory data analysis, using lots of plots.
2. Find any missing data and make a decision on how to handle them
3. Choose a performance metric to evaluate your model. Is accuracy the best one? What about AUC? Search the literature on what are common evaluation metrics in medical papers.
4. Construct a baseline model and then build other models to improve performance. **Hint:** Since we are interested in using interpretation techniques, such as SHAP values, and because we have structured tabular data, it is advisable to use a **tree model** such as Random Forests, or Gradient Boosted Trees.
5. Look at feature importance as well as interpretation values, such as SHAP values, that indicate how each feature affects the outcome of the model. Choose four features and construct interaction plots between them.
6. Feature selection: could you remove some of the features and achieve the same performance? Finding the smallest feature subset could be useful in clinical practice.
7. Discuss your findings and conclusions, in a way that a member of the medical community with little data science background could understand.

References

- Lundberg, Scott M, "Explainable AI for Trees: From Local Explanations to Global Understanding"