PREDICTING AGE AT DEATH

Wesley Hughes, Elma Kajtaz, Pranav Ram, Paul Schlueter, Rohan Somni

Georgia Institute of Technology

CDC Centers for Disease Control and Prevention

MOTIVATION

- Working with the CDC to analyze one of their datasets
- Identify possible patterns in diseases & causes of death
- Focus on creating a “Death Predictor”
- Questions motivating the project:
  - Given various predictive variables such as gender, education, race, place of residency, or marital status, can we predict the person’s age at death?
  - Does knowing their disease/disorder help in that prediction?
  - If successful, our prediction model could serve as a benchmark for future CDC research

DATA & APPROACH

DATA:

- Mortality data from National Vital Statistics System
- Demographic, geographic & cause-of-death information on US citizens
- 2.4 million records with 132 features/columns
- Total data > 10GB

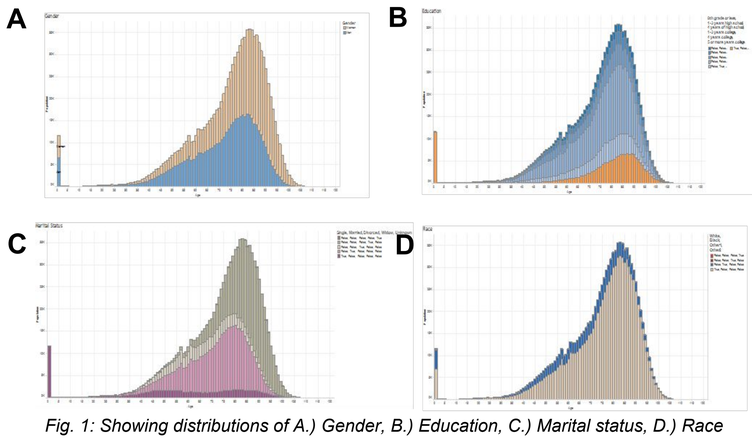


Fig. 1: Showing distributions of A.) Gender, B.) Education, C.) Marital status, D.) Race

MACHINE LEARNING:

- Categorical features to binary
- Scikit-learn
- Pickle + gzip for storage and retrieval

VISUALIZATION:

- Comparator view - “Cards” representing data slices for each state
- Additional details on demand option
- Predictor view - Visualization of probability function: Age at death

RESULTS

A. MACHINE LEARNING

Random Forest Classifier

PREDICTS

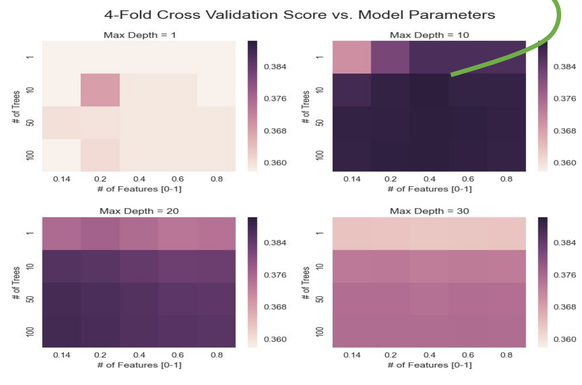
age at death

(1-59, 60-73, 74-81, 82-87, 88+)

using **education level, race, marital status, medical conditions, gender, location, county population**

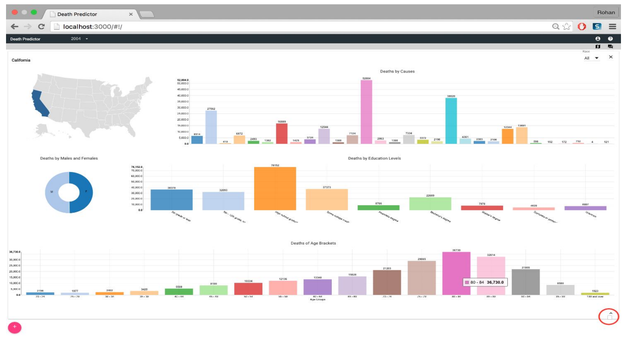
Model Optimization

1. Vary **tree depth**, **# of features**, **# of trees**
2. Score with **cross-validation**
3. Pick best score

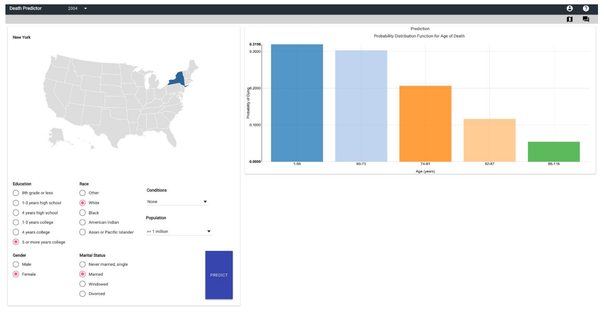


41.3% Cross-Validation **Accuracy**

B. VISUALIZATION



- Overview & Details on Demand View - California
- Highest death cause is heart disease, followed by neoplasms
- Most prevalent age death bin 80 - 84 years



- Predictor view - New York
- Insights - White female, college educated
- On average, married women tend to live a LOT longer

CONCLUSIONS

- Presented the final product to Paula Braun of the CDC and Dr. Mark Braunstein (Georgia Tech). Overall feedback was extremely positive.
- In discussion about how our project can be hosted online on the official NCHS visualization page. (http://blogs.cdc.gov/nchs-data-visualization/).

ACKNOWLEDGMENTS

Authors would like to thank Ms. Paula Braun of the CDC and Dr. Mark Braunstein (Georgia Tech) for providing the dataset, insights and feedback.