Penalized Models and the California Teachers Study

Eugene Nguyen

PM 606

Summer 2022

Goals

- Develop a machine learning model capable of predicting mortality
- Compare 3 different classification models
- Utilize demographics, physical activity measures, diet measures, and primary diagnosis codes as inputs

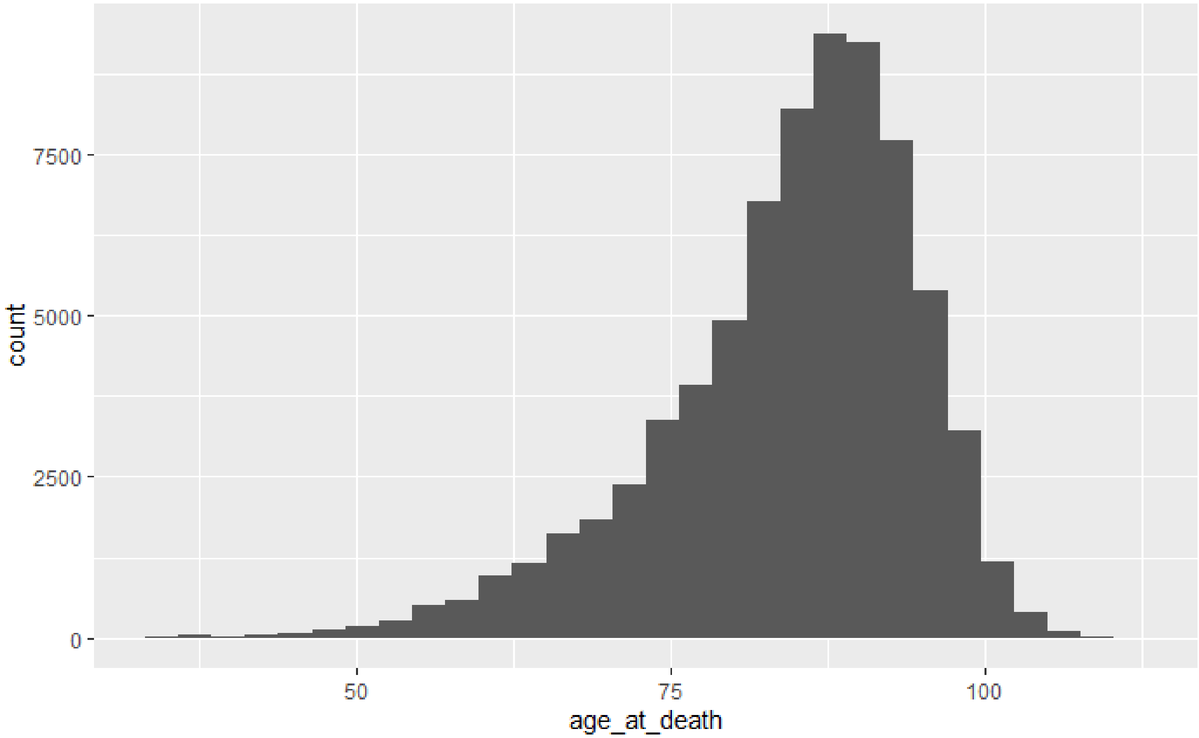
Contents

Overall Descriptive Statistics

- Total Observations = 154,315
- Total Unique Individuals = 48,324
- 18,474 (38.2%) of the total unique individuals have passed away

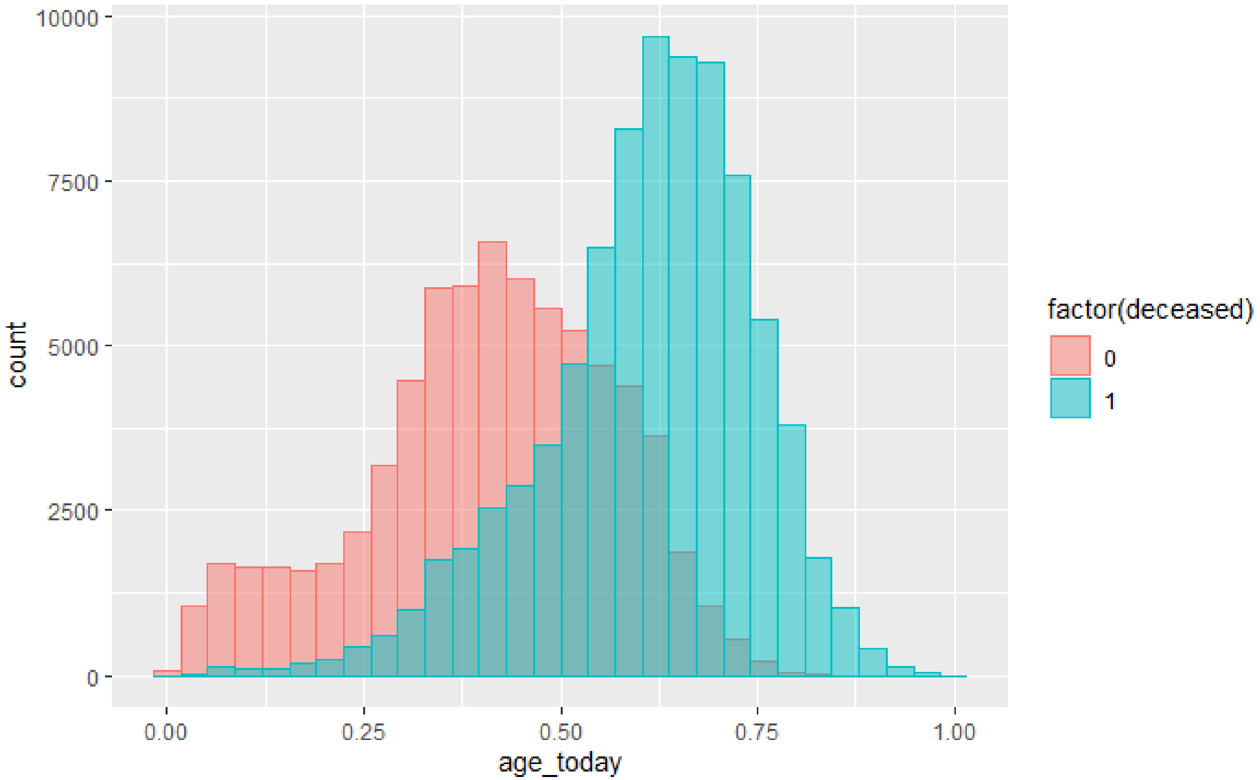
Deceased	N (%)
Yes	18474 (38.2%)
No	29850 (61.8%)

Age at Death



Minimum	25%	50%	Mean	75%	Max
33.78	79.13	86.29	84.42	91.57	110.83

Age Today vs. Mortality



Minimum	25%	50%	Mean	75%	Max
49.06	74.40	84.26	83.95	94.80	123.50

Contents

Methodology

- Compare 3 regularized classification models
  - Elastic Net
  - Lasso
  - Ridge

  - Note: Neural Networks and Random Forests were also explored, but the training time took too long (24-48 hours), therefore penalized models were chosen for balance of robustness and timeliness.

Inputs

- Demographics
  - Age, urbanization, residence status, adopted, twin, birthplace, race/ethnicity
- Clinical Factors
  - Number of admissions, total charges, primary ICD9 codes, primary ICD10 codes
- Physical Activity
  - Hours of exercise per week, hours standing/walking per day at work, hours sitting, hours sleeping
- Diet
  - Plant based, high protein/fat, high carb, ethnic diet, salad/wine, multivitamin, frequency of fat/oil in cooking

Data Handling

- Data Clean
  - Converted all factors to dummy variables
  - Min/Max normalization of all continuous variables
  - Clean ICD codes to bucket into parent categories
  - Missing integer columns were filled with 0
  - Missing numeric columns were filled with the mean
  - Dummy variables with sparse positive classes were dropped (anything below 0.5%)
- Data Split
  - Data split into a 70%/30% training and testing sets
  - Cross validation applied to the training set to tune lambda
  - The best parameters for each model will be tested against the final 30% set

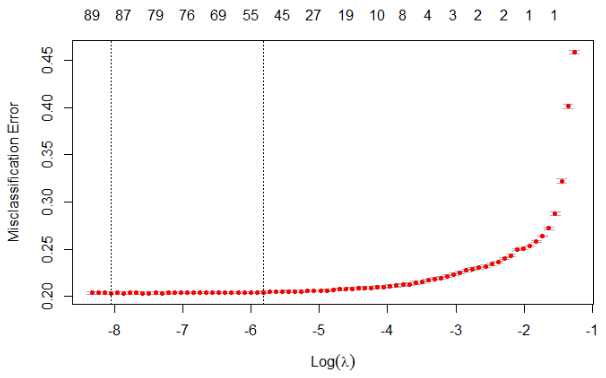
Contents

Models

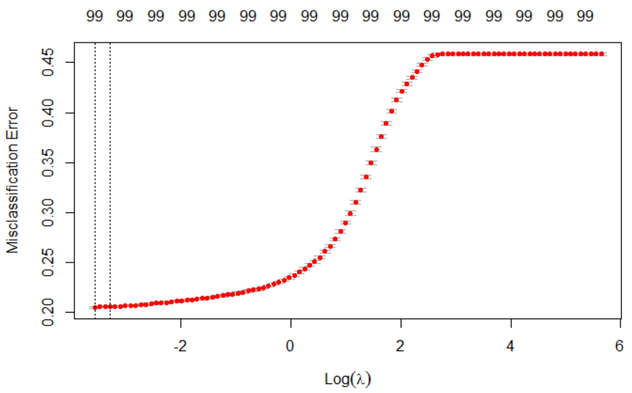
- Lasso
  - Can penalize coefficients to 0
  - Great when models contain a lot of **useless** variables
- Ridge
  - Will shrink coefficients, but not remove
  - Great when models contain a lot of **useful** variables
- Elastic Net
  - Combination of lasso and ridge

Cross Validation

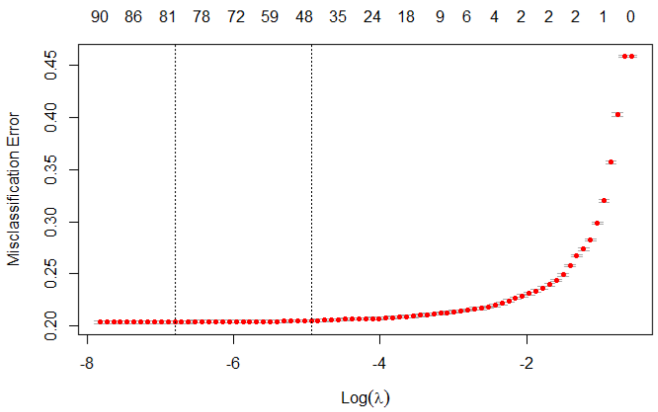
- 10-fold cross validation was used for each model (lasso, ridge, elastic net)
- The minimum lambda was extracted from each cross validation model and used on the testing set



Lasso



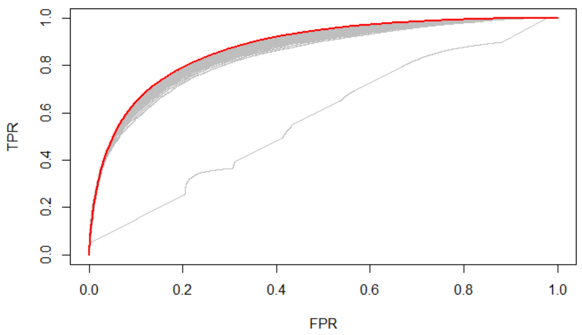
Ridge



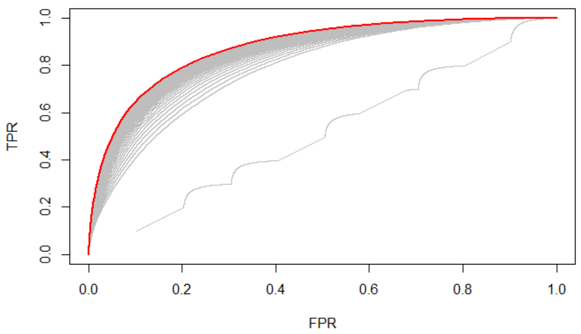
Elastic Net

ROC Curves

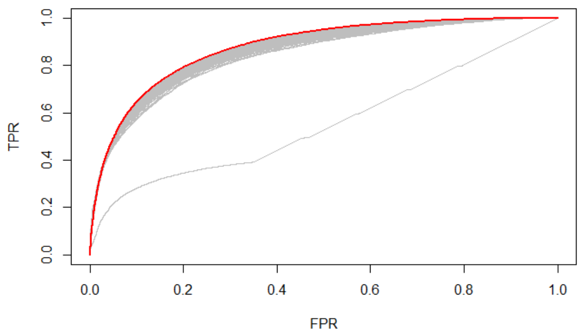
- ROC Curves were generated to visualize the cross validated models
- The red curve depicts the best model with the highest AUC values



Lasso



Ridge



Elastic Net

Variable Importance – Top 10

Variable	Overall
Age Today	8.1
Number of Admissions	5.6
Musculoskeletal System Disease (ICD10)	1.8
Total Charges	1.1
Neoplasms (ICD9)	0.87
Age at Baseline	0.81
Musculoskeletal System Disease (ICD9)	0.78
Injury/Poisoning (ICD10)	0.77
Respiratory System Diseases (ICD9)	0.77
Infectious Parasitic Diseases (ICD9)	0.75

Contents

- Descriptives
- Methodology
- Modeling
- Results

Cross Validation Results – Training Data

Model	Minimum Lambda	AUC	Misclassification Error	Accuracy
Lasso	0.00186	0.877195	0.2022681	0.7939
Ridge	0.0282	0.877015	0.2036937	0.7934
Elastic Net	0.00339	0.877243	0.2024193	0.7934

- Lambda – Regularization parameter
- AUC (Area Under the Curve) – Measure of the ability of a classifier to distinguish between classes; used as a summary of the ROC curve
- Misclassification Error – Percentage of observations that were incorrectly predicted
- Accuracy – Percentage of observations that were correctly predicted

Validation Results

Model	Accuracy	95% CI	Sensitivity	Specificity	PPV	NPV
Lasso	0.7977	0.794 – 0.801	0.752	0.837	0.798	0.798
Ridge	0.7963	0.793 – 0.800	0.745	0.840	0.800	0.794
Elastic Net	0.7976	0.794 – 0.801	0.751	0.837	0.798	0.797

- Sensitivity –Percentage of true positives. Proportion of observations that tested positive and are positive of all the labels that are actually positive.
- Specificity – Percentage of true negatives. Proportion of observations that tested negative and are negative of all the labels that actually are negative.
- Positive Predictive Value (PPV) – Also known as precision. If the test result is positive, how well does that predict an actual presence of disease?
- Negative Predictive Value (NPV) – Probability that observations with a negative predicted result truly should be negative.