Using Linked Data to Train and Validate Machine Learning Prediction Models

Orlando Davy

NCHS Data Linkage Program

FCSM Research & Policy Conference
October 26, 2023

Background

- Linked survey and administrative data can be used to facilitate richer analyses by augmenting the information collected from the surveys with vital records and other administrative data
- Data linkage requires survey participants to provide consent for linkage and sufficient personally identifiable information (PII)
- There has been a growing reluctance of survey participants to provide the PII needed for linkage

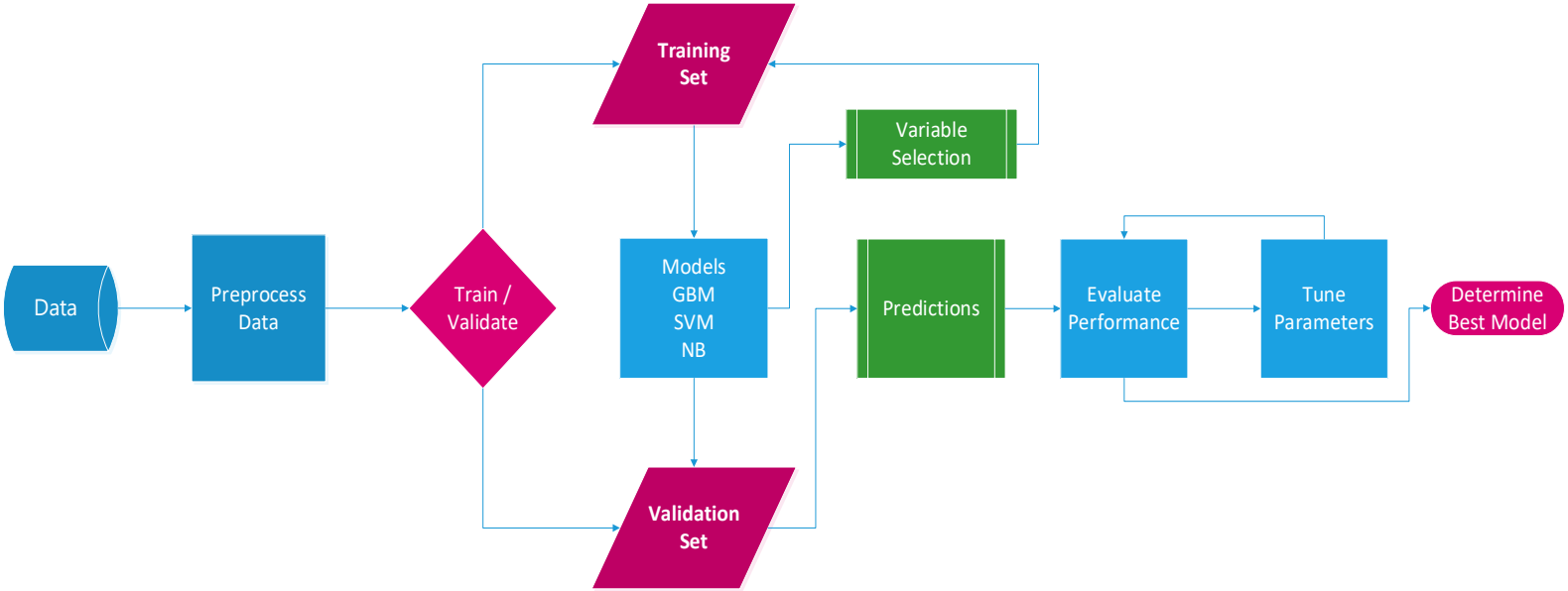
Machine Learning to Predict Outcomes

- When data linkage is not possible, machine learning (ML) prediction models can be used to predict outcomes, such as morbidity and mortality
- ML models require quality and accurate training data and a validation source
- NCHS Data Linkage Program has developed an extensive repository of high-quality linked data files that can be used to address a wide-range of health-related research topics and a variety data science applications

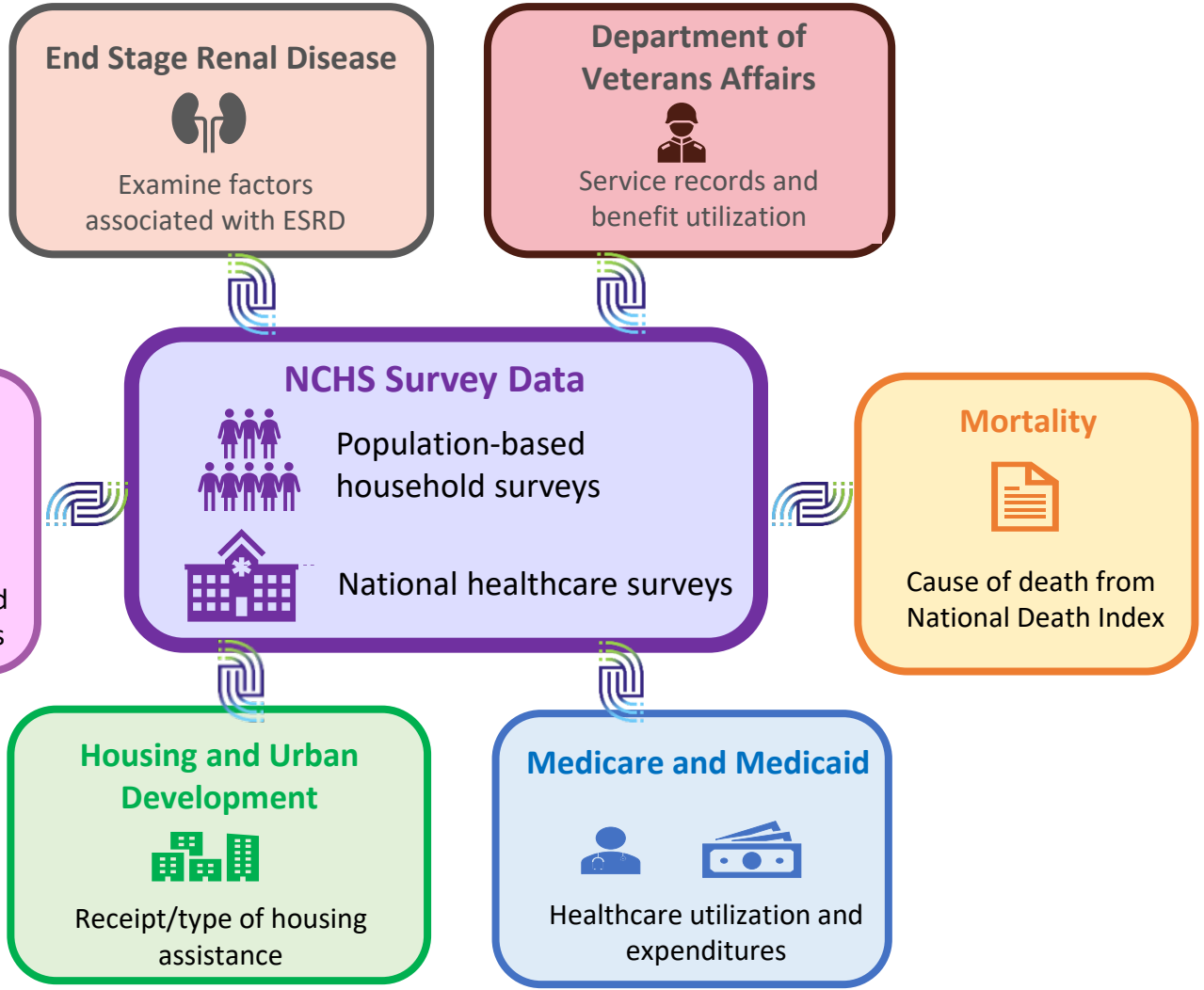
Project Goal

- To evaluate selected ML prediction models using linked data as the training data and validation source to assess model performance for predicting all-cause mortality

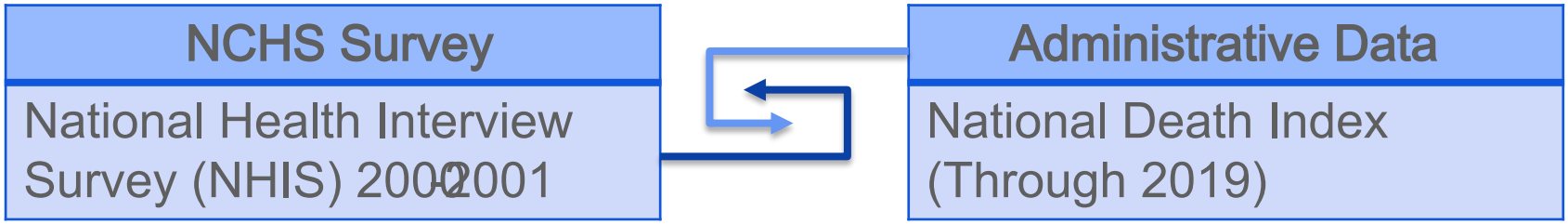
The ML Workflow



NCHS Data Linkage Program



Data Source: Linked NHIS-NDI Data



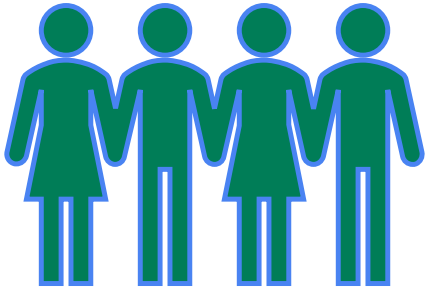
- Monitors Health of the Non-institutionalized US population
- Cross sectional design
- Geographically clustered
- Sampling weights

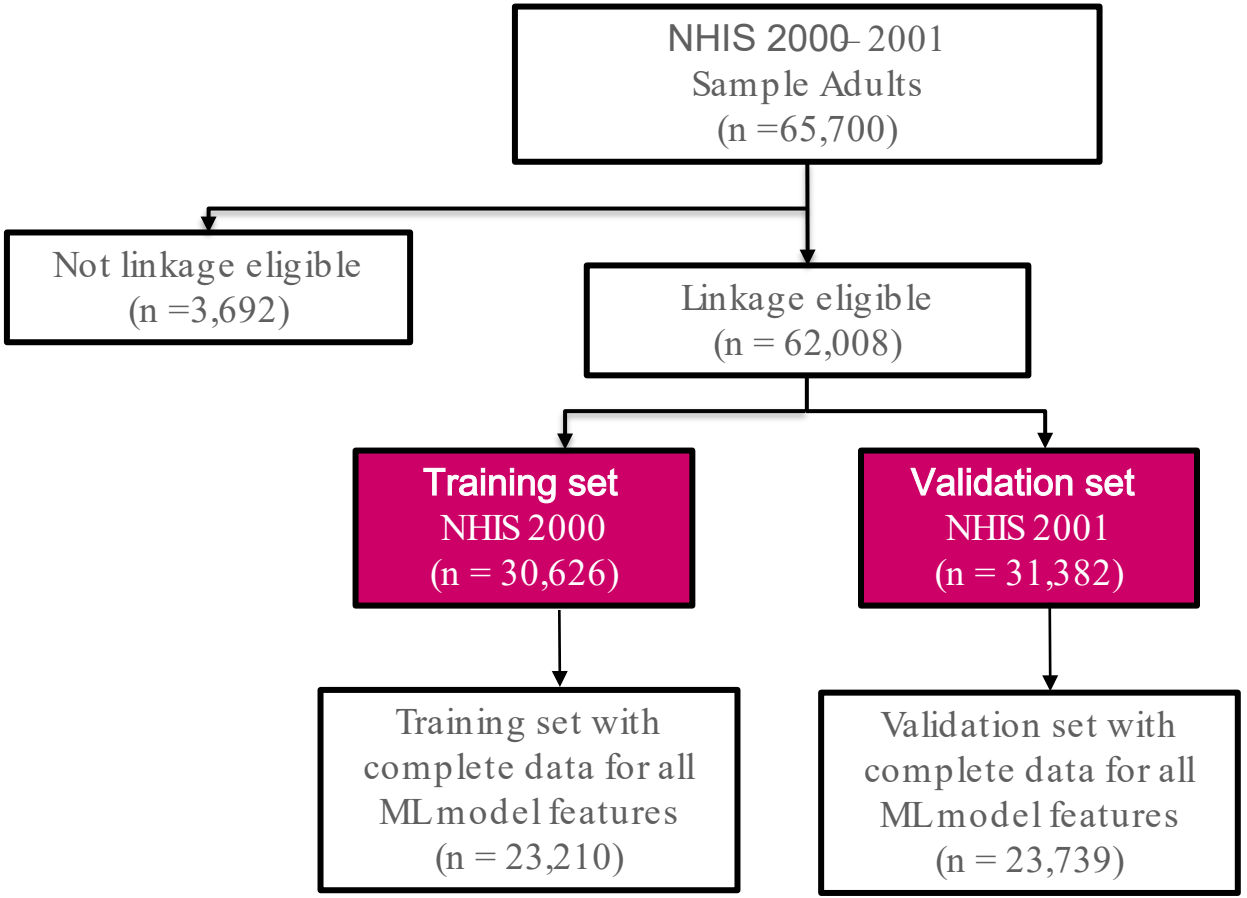
- A complete source of death information for the US
- Mortality status, date of death, and cause of death from death certificates

Selected Features

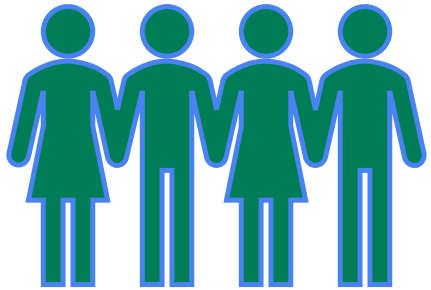
1. Age
2. Sex
3. Race and Ethnicity
4. Education
5. Marital Status
6. Poverty to Income Ratio (PIR)
7. Health Insurance
8. Inactivity
9. Smoking Status
10. Excessive Alcohol Consumption
11. Body Mass Index (BMI)
12. Hypertension
13. Diabetes
14. Coronary Heart Disease
15. Heart Condition
16. Heart Attack
17. Place for Care
18. Barrier to health Care: Cost
19. Psychological Distress

Sample Sizes





Distribution of Sample Size



Training set (NHIS 2000)
Alive: n = 17,969
Deceased: n = 5,241

Validation set (NHIS 2001)
Alive: n = 18,842
Deceased: n = 4,897

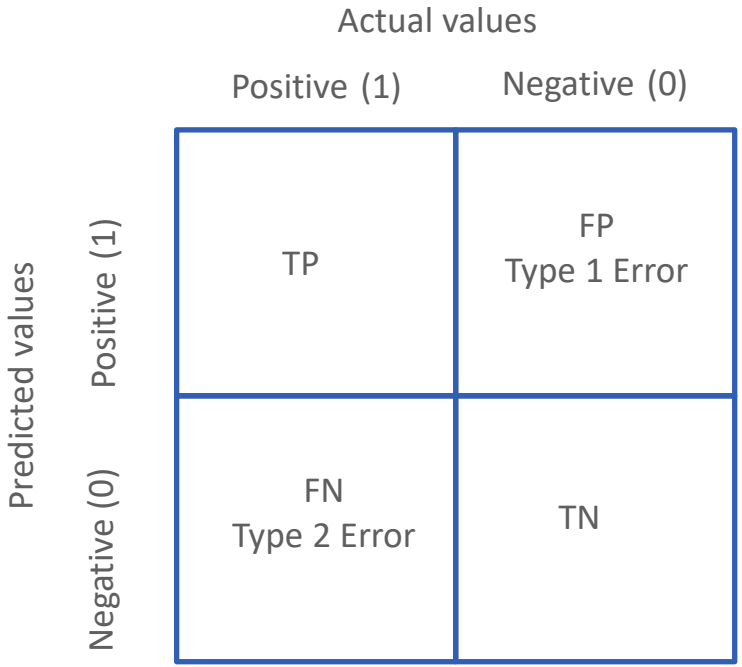
Selected ML Models

All analyses were conducted using R v4.2.2 and the *Caret* package v6.0-94

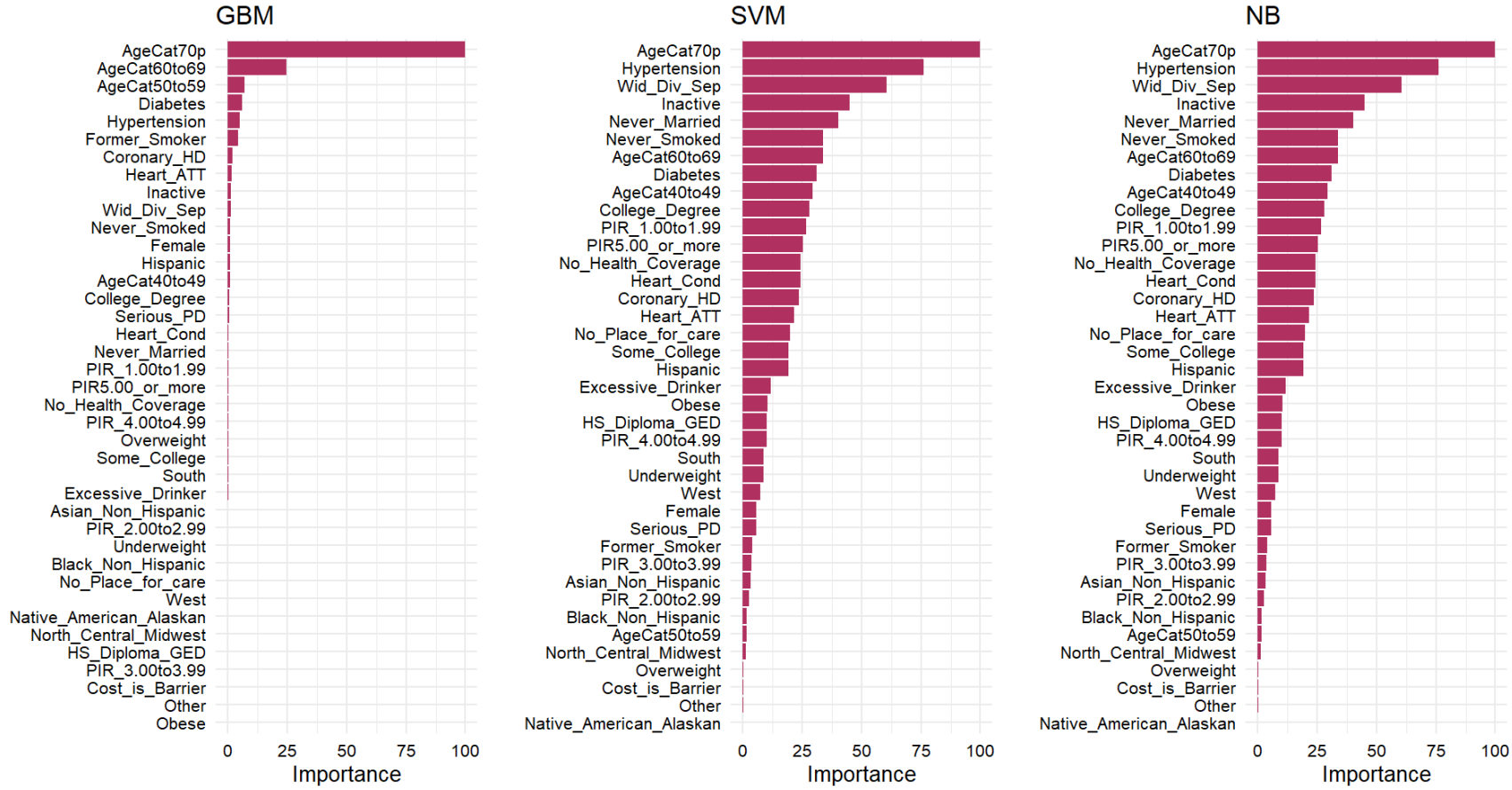
- **Generalized Boosting Model (GBM)**
  - No statistical assumptions
  - Sequence of independent trees that improves after each iteration
- **Support Vector Machines (SVM)**
  - No probabilistic explanation for classification
  - Slow to train and can be problematic for large datasets
- **Naive Bayes (NB)**
  - Assumes conditional independence and that all features contribute equally to the outcome
  - Easy to implement because only probabilities need to be calculated

Terminology and Evaluation Metrics

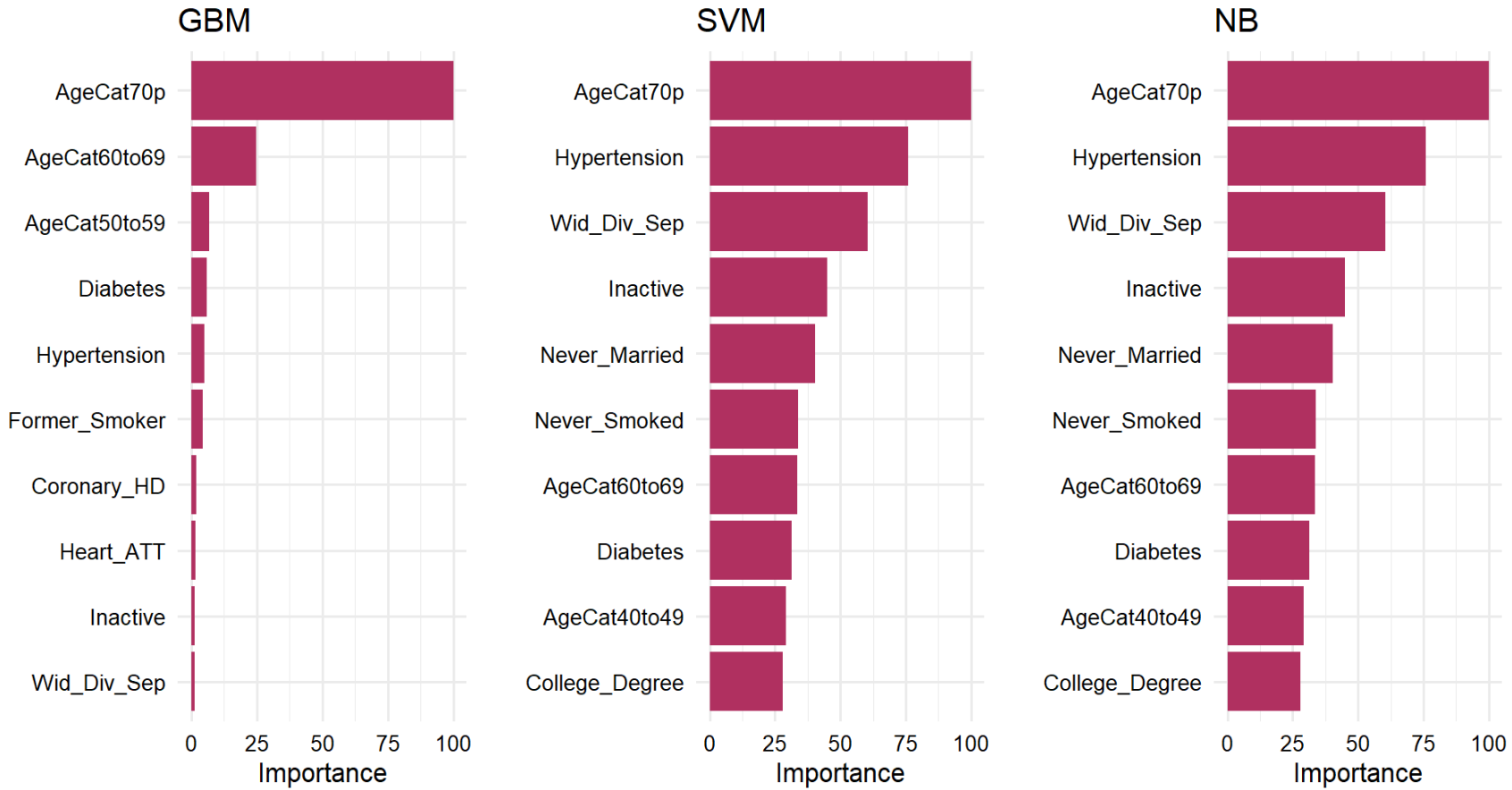
- Variable Importance
- Confusion Matrix
  - Misclassification Error
  - Sensitivity (Recall)
  - Precision
  - Balanced Accuracy
  - F1 Score
  - Area Under the Curve (ROC-AUC)



Variable Importance Plots

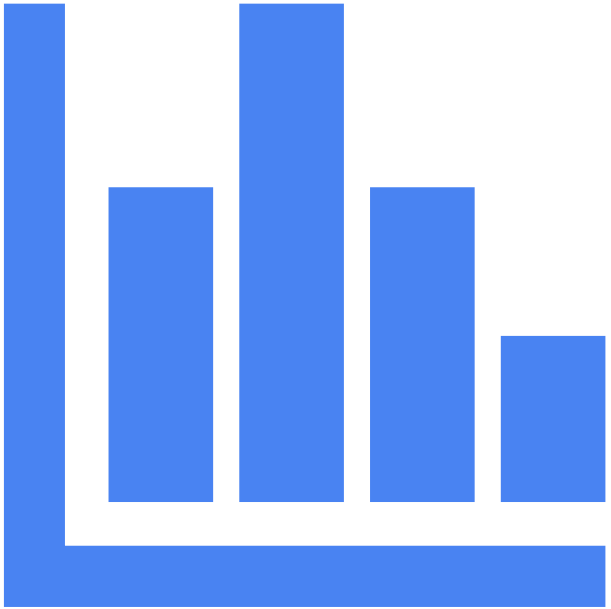


Variable Importance Plots: Top 10 Features

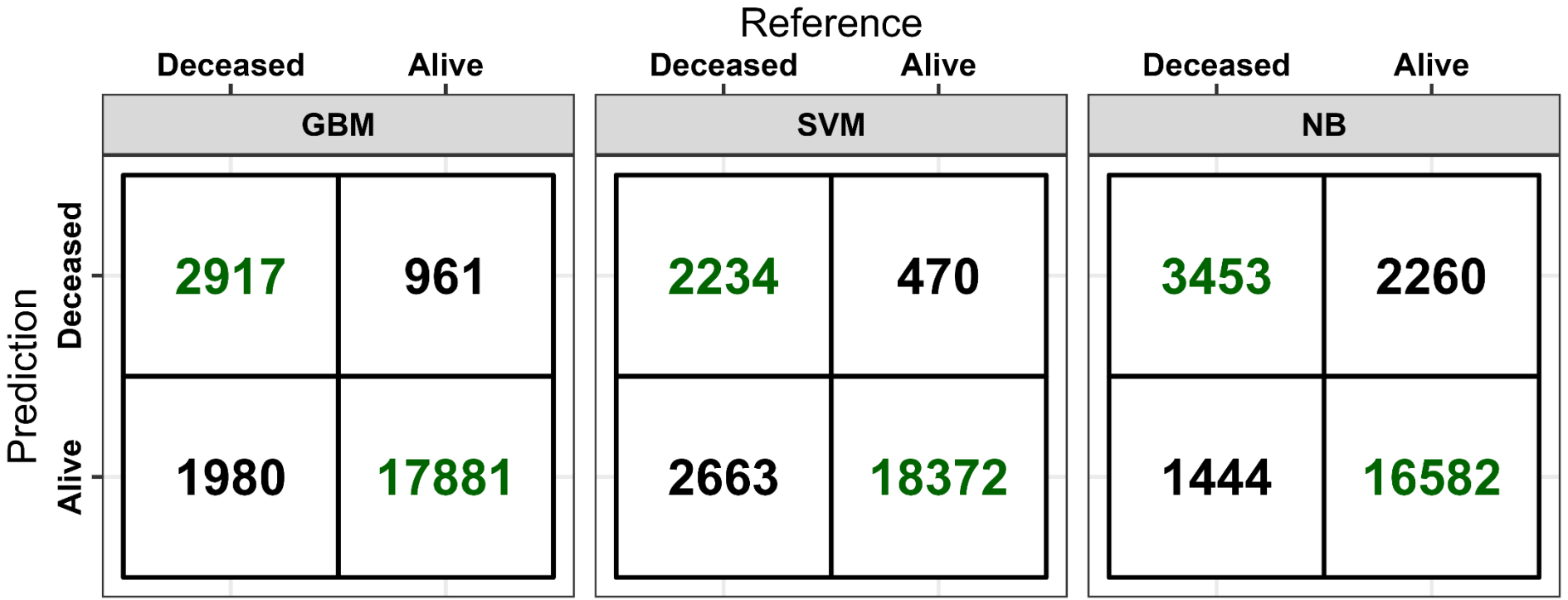


Results

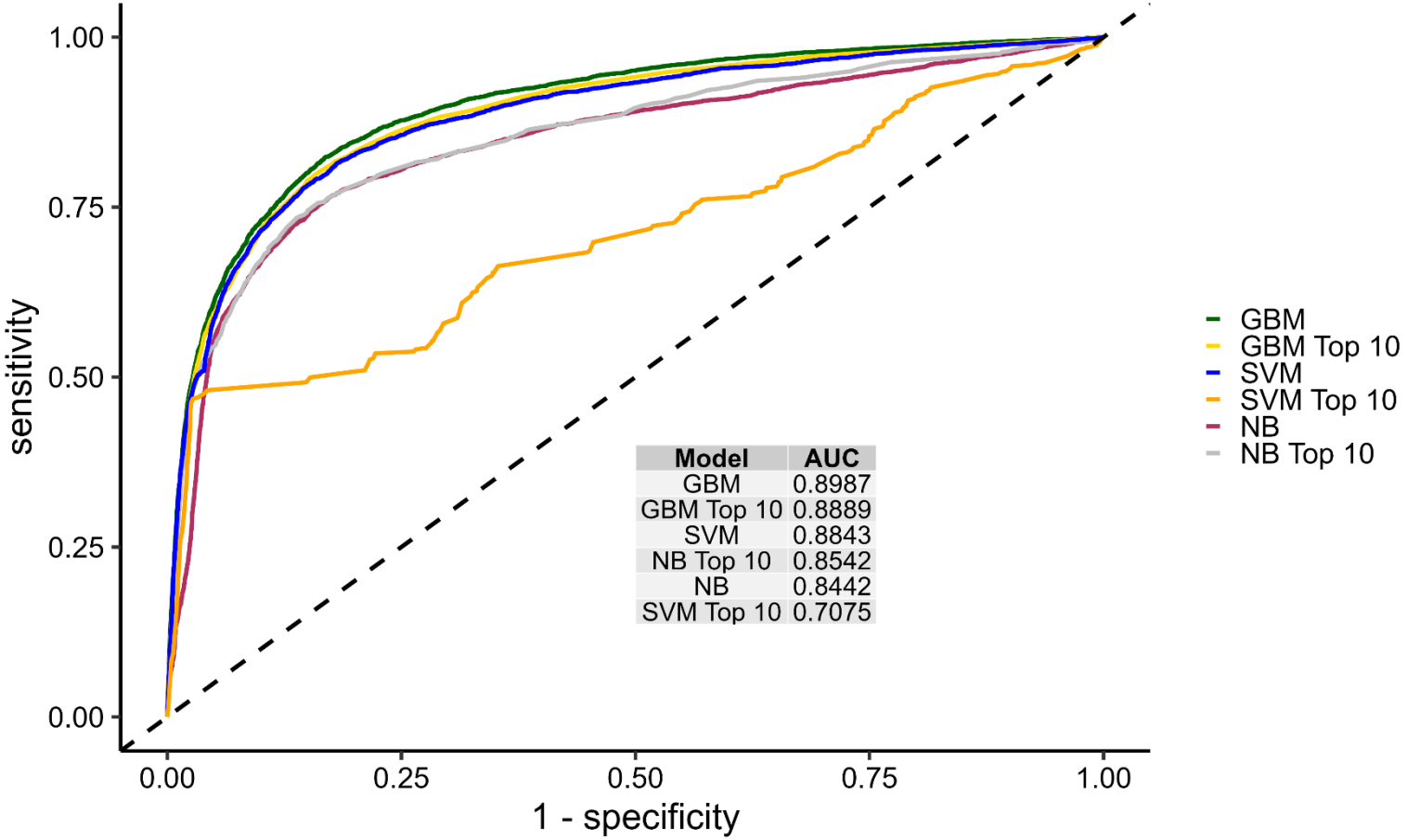
- Confusion Matrix
  - Top 10 features
- ROC-AUC
- Metrics
  - Error
  - Recall
  - Precision
  - Balanced Accuracy
  - F1-Score
  - *Run-time*



Confusion Matrix : Top 10 Features



Area Under the Curve (AUC-ROC)



Model	AUC
GBM	0.8987
GBM Top 10	0.8889
SVM	0.8843
NB Top 10	0.8542
NB	0.8442
SVM Top 10	0.7075

Evaluation Metrics

Model	Error	Recall	Precision	Balanced Accuracy	F1-Score	Run-time
GBM						
Complete	0.1212	0.6388	0.7383	0.7899	0.6849	2 min
Top 10	0.1239	0.5957	0.7752	0.7723	0.6648	35 sec
SVM						
Complete	0.1324	0.5089	0.7715	0.7349	0.6133	30 min
Top 10	0.1320	0.4562	0.8262	0.7156	0.5878	1 min
NB						
Complete	0.1618	0.7137	0.5891	0.7921	0.6454	2 sec
Top10	0.1560	0.7051	0.6044	0.7926	0.6509	1 sec

Summary

- GBM had lowest error rates and highest F1-scores
- SVM precision was greater than GBM, but was less efficient (longer run-time)
- NB was the most efficient and had higher recall and balanced accuracy measures
- Little change in overall performance across models when using limited features (only top 10 variables)
- When trained with high-quality data ML models perform reasonably well for predicting all-cause mortality

Thank You!

NCHS Data Linkage Program

Contact: Orlando Davy odavy@cdc.gov

Subscribe to the NCHS Data Linkage Program LISTSERV to receive updates! Email a message to list@cdc.gov Leave the subject line blank. In the body of the message, type or paste:

SUBSCRIBE NCHS-DATALINKAGE-PROGRAM last name, first name

where ‘last name, first name’ is your last and first name.

Definitions and Formulas

Definitions

- **Variable Importance (VI)**
  - A score indicating how much each variable contributes to the model prediction

- **Hyperparameters**
  - External configuration parameters used to manage ML model training

Definitions: Hyperparameters

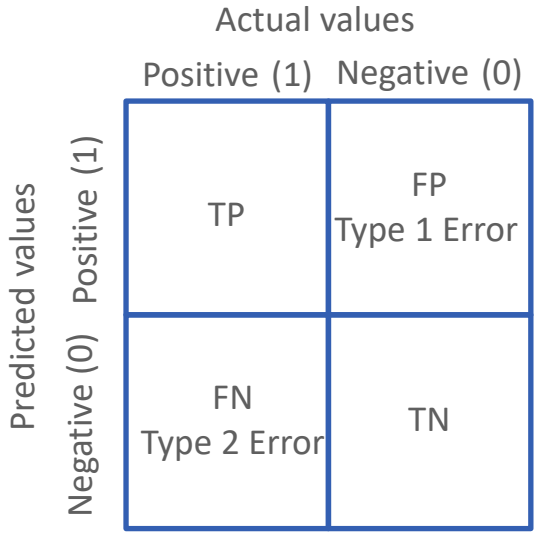
- GBM
  1. Interaction.depth = 3
  2. n.trees = 300
  3. n.minobsinnode = 25

- SVM
  1. C = 1.58

- NB
  1. Laplace = 1
  2. Kernel = False

Definitions

- **Confusion Matrix**
  - A type of contingency table used to summarize the performance of classification algorithms

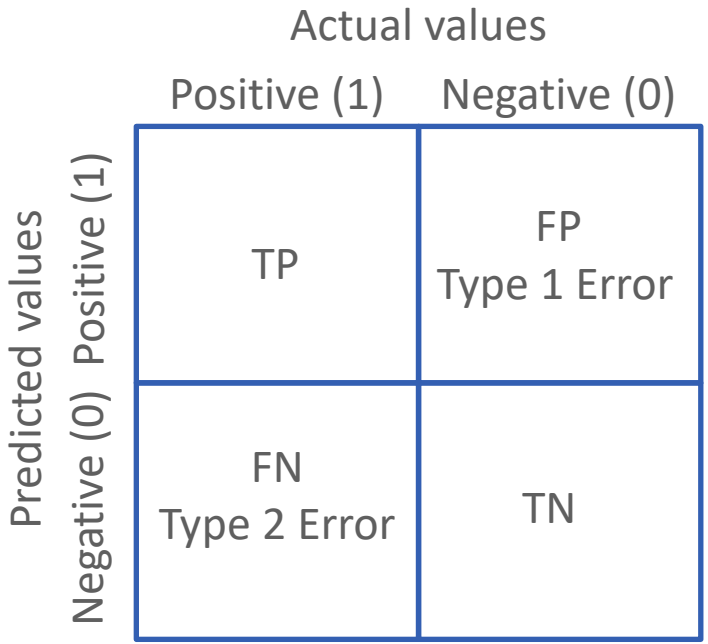


Performance measures
- Misclassification Error
- Precision
- Sensitivity (Recall)
- Specificity
- F1-Score
- Area Under the Curve (AUC)

Formulas

Misclassification Error

$$= \frac{FP+FN}{TP+TN+FN+FP}$$

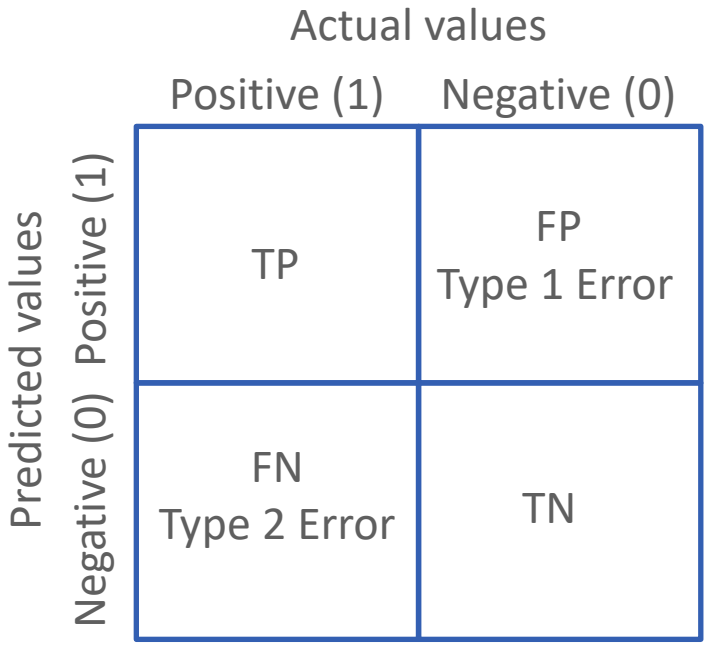


Formulas

$$\textbf{Precision} = \frac{TP}{TP+FP}$$

$$\textbf{Sensitivity (Recall)} = \frac{TP}{TP+FN}$$

$$\textbf{Specificity} = \frac{TN}{TN+FP}$$



Definitions

- The Area Under the Curve (AUC-ROC) or Receiver Operating Characteristic (ROC) curve is a summary measure of performance for classification problems at various thresholds
- The curve is plotted with sensitivity against 1-Specificity where sensitivity is on the y-axis and 1-Specificity is on the x-axis
- The 45° diagonal line serves as the reference line or random classification (AUC= .5)

Definitions

- Precision Recall Area Under the Curve (PR-AUC) gives a more informative picture of model performance when that data is highly imbalanced
- The curve is plotted with precision against recall where precision is on the y-axis and recall is on the x-axis
- Reference line is the fraction of positives in the data set

Definitions and Formulas

- F1-Score is the harmonic mean of Precision and Recall (Sensitivity)

$$= \frac{(1 + \beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}$$

$\beta = 1$

References

- Sidney-Gibbons, J.A., Sidney-Gibbons, C.J.. Machine Learning in medicine: a practical introduction. BMC Medical Research Methodology, (2019),19:64. https://doi.org/10.1186/s12874-019-0681-4
- Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM. A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. Front Bioinform. 2022 Jun 27;2:927312. doi: 10.3389/fbinf.2022.927312. PMID: 36304293; PMCID: PMC9580915.
- Borah, S., Mishra, S. K., Balas, V. E. and Polkowski, Z,. (2022) Advances in Data Science and Management: Proceedings of ICDSM 2021. (2002). Springer Nature.
- Data science websites:
  - Machine Learning | An Introduction | by Gavin Edwards | Towards Data Science
  - Understanding Confusion Matrix | by Sarang Narkhede | Towards Data Science
  - Data Science Stack Exchange

For more information, contact CDC
1-800-CDC-INFO (232-4636)
TTY: 1-888-232-6348 www.cdc.gov

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.