Cardiovascular Disease Risk Prediction using EHRs

Minh Nguyen
Department of Biomedical Informatics, School of Medicine

Predicting

- Cardiovascular disease (CVDs) is the number 1 cause of death globally. Identifying at-risk individuals is important for early prevention and timely treatments.
- Although there is an abundance of prediction models for CVDs, the use of EHRs is still at the beginning.
- We utilized the large Electronic Health Records from the Stanford Translational Research Integrated Database Environment (STRIDE 8) to derive a valid cohort and build a gradient boosting model for predicting CVD risks.
- The model performed very well compared to the standard risk prediction score and neural networks results from the literature, but raised questions regarding fairness when applying to subgroups with sensitive attributes such as gender, age, and race/ethnicity.

Data & Features

- **256,583 unique patients and 39,558 features**
- Features and outcomes are all binary values
- **Outcomes**: presence or absence of CVD events after prediction time randomly chosen during valid intervals
- Extremely imbalanced outcomes: 1.7% incident rate
- SQL codes to derive a valid cohort from multiple databases in STRIDE 8
- **Feature extractions**: Observational Medical Outcomes Partnership Common Data Model (OMOP CDM)
  - Convert a series of time-stamped clinical elements to a static presentation for each patient.
  - Diagnoses, conditions, procedures, med orders, lab tests, clinical encounter types, and others.
  - Sparse feature matrix as input

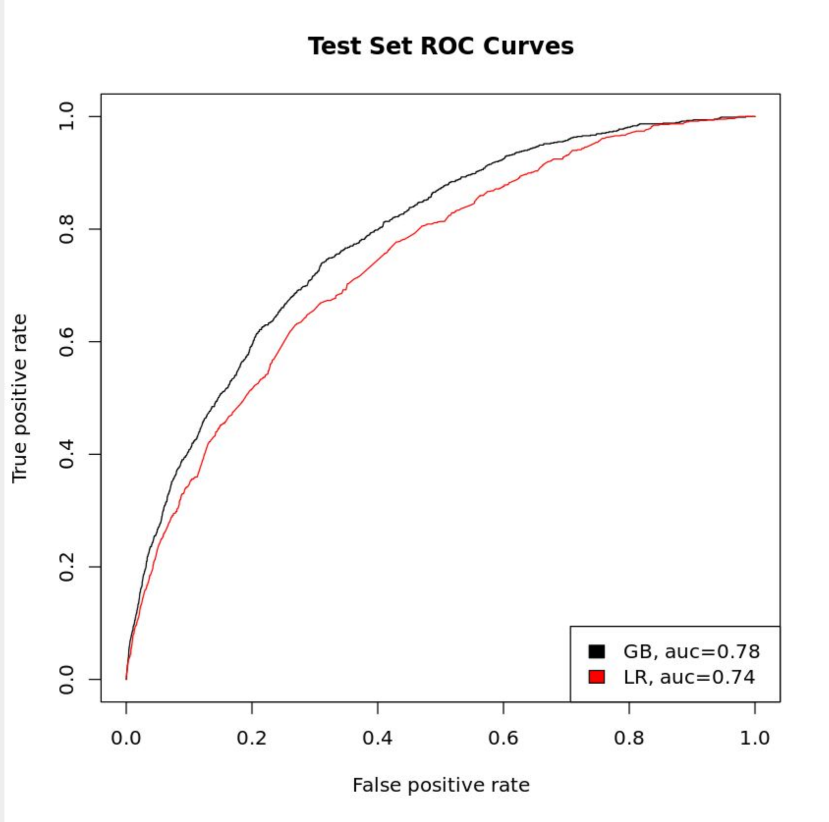
Models

- Splits: 80% train and 20% test
- Training used 10 fold cross validation

Baseline: Logistic Regression (glmnet)
With large number of features, regularization helps prevent overfitting by adding constraints to shrink the coefficient estimates toward 0, reducing model complexity using:
- L1 for Lasso regression, L2 for Ridge regression
- Elastic net uses a combination of both L1 and L2 reg

Gradient Boosting Trees: (xgboost is blazingly fast)
- Boosting refers to an ensemble technique of building a decision tree consisting of many sequentially built trees
- Each subsequent tree learns from the its precedents and aims to reduce the errors of the previous trees.
- Hyperparameter search first manually, then grid search
  - Max number of trees, max tree depth, learning rate,
  - L1 and L2 regularization
  - Max delta step allowed in each tree weight estimation to help with convergence for imbalanced data

Results (GBM)



Confusion Matrix (threshold is 3rd quantile of predicted probabilities)

```
           Reference
Prediction     0     1
         0 38193   294
         1 12275   554
```

Other metrics	
Sensitivity	0.6533
Specificity	0.7568
Accuracy	0.7511

Results (continued)

Model (built)	AUROC	Model (literature review)	AUROC
Logistic Reg (w/ Ridge)	0.728	Risk score eq (QRISK2, FRS, PCE)	0.60 - 0.75
Logistic Reg (w/ Lasso)	0.731	Logistic regression, Tree-based	0.765 - 0.782
Logistic Reg (w/ elastic net)	0.741	GBM with longitudinal data	~ 0. 790
Best tuned GBM	*0.782*	NN with longitudinal data	*~ 0.790*

AUROC evaluated on subgroups w/ sensitive attributes (*gender*, race/ethnicity)			
Male	*0.744*	Asians	0.792
Female	*0.794*	Black	0.781
White	0.767	Hispanic/Latino	0.743

Discussion & Conclusions

- Gradient boosting tree performed well compared to results from literature review, whereas logistic regression did worse, but not completely a disaster (best with elastic net)
- Heavy regularization prevented overfitting, with narrow AUROC gap between validation and test data
- Xgboost performance was extremely fast, helpful for large data set such EHR data. Glmnet was too slow, and not scalable.
- Limited support for sparse matrix limited use of more algorithms

Future Work

- Approaches to handle extreme class imbalance were explored, did not help much, but raised more questions.
- Restricting outcome time frame to handle right censored data, which will also help in handling imbalanced classes by using inverse probability of censoring weights.
- Fairness in machine learning needs work as performance varied a lot across different subgroups with sensitive attributes.

References

[1] Damen, Johanna AAG, et al. "Prediction models for cardiovascular disease risk in the general population: systematic review." bmj 353 (2016).
[2] Karmali, Kunal N., and Donald M. Lloyd-Jones. "Implementing cardiovascular risk prediction in clinical practice: the future is now." (2017).
[3] Pike, Mindy M., et al. "Improvement in cardiovascular risk prediction with electronic health records." Journal of cardiovascular translational research 9.3 (2016).
[4] Wolfson, Julian, et al. "Use and customization of risk scores for predicting cardiovascular events using electronic health record data." Journal of the American Heart Association 6.4 (2017).
[5] Goldstein, Benjamin A., et al. "Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review." Journal of the American Medical Informatics Association 24.1 (2017).
[6] Zhao, Juan, et al. "Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction." Scientific reports 9.1 (2019).
[7] Weng, Stephen F., et al. "Can machine-learning improve cardiovascular risk prediction using routine clinical data?." PloS one 12.4 (2017).
[8] Rajkomar, Alvin, et al. "Scalable and accurate deep learning with electronic health records." NPJ Digital Medicine 1.1 (2018): 18