Cardiovascular Disease Risk Prediction using EHRs

Minh Nguyen
Department of Biomedical Informatics, School of Medicine

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)

<table>
<thead>
<tr>
<th>Model</th>
<th>AUROC</th>
<th>Model (literature review)</th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Reg (w/ Ridge)</td>
<td>0.728</td>
<td>Logistic score (EQ)</td>
<td>0.60 - 0.75</td>
</tr>
<tr>
<td>Logistic Reg (w/ Lasso)</td>
<td>0.731</td>
<td>Logistic regression, Tree-based</td>
<td>0.765 - 0.782</td>
</tr>
<tr>
<td>Logistic Reg (w/ elastic net)</td>
<td>0.741</td>
<td>GBM with longitudinal data</td>
<td>~ 0.790</td>
</tr>
<tr>
<td>Best tuned GBM</td>
<td>0.782</td>
<td>NN with longitudinal data</td>
<td>~ 0.790</td>
</tr>
</tbody>
</table>

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. Glimnet 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