Forecasting Model for Disease Propensity Using EHR Data

Students: Sue Sarafrazi, Jill Han, Matthew Domingo, Michael Chang, Omar Sharif, Anil Kemisetti, Omid Khazaee
USF Supervisor: Professor Patricia Francis-Lyon
Healthgrades Project Supervisors: Nathan Stott, Tom Brander
Healthgrades

- Offers a comprehensive rating and comparison database on the quality of physicians, hospitals, and providers in the U.S. sourced from 500M+ federal and private claims
- Healthgrades Hospital Solutions: CRM, marketing, patient acquisition, etc.
Healthgrades CRM Solutions

- Healthcare-specific predictive modeling: allowing providers to design effective, data-driven strategies through intervention and preventative care
- This is done by:
  - Identifying patients at risk for a medical event
  - Predicting future health needs
Introduction

- Certain conditions such as diabetes and heart failure are actionable
  - One to two years of early warning would represent a huge advance in preventing further complications

https://www.creative-diagnostics.com/Cardiac-Disease.htm

Objectives:

- Assessing risk of multiple actionable morbidities with EHR data on Amazon Web Service
- Ultimate goal is for project to be deployed into production after this semester at Healthgrades
Project Overview

Raw Data → Loading the raw data into S3 for Processing → S3 Bucket → Transforming Data Using → Data Preparation → Options

EMR → Spark

S3 Bucket for Processed Data → Visualizing Processed Data Using → Amazon Quick Sight

Local Installation → Model Training, Model Tuning, and Model Deployment

& SageMaker
Introduction of the Database (EHR data)

- All our practices and data are HIPAA compliant
- 170 Million records from 14 clients
- Does **not** contain information such as:
  - Blood pressure
  - Weight
  - Height
  - Body mass index (BMI)
  - Blood sugar level
  - A1C score
  - Medication
  - Smoking
Condition Forecasting Setup

- **Phenotyping:**
  - Identifying patients with targeted conditions using ICD codes: positive cases for training model
  - Adding boolean features for risk factors
- Aggregating each patient medical history into one record

Diabetes Phenotyping

- Find patients that were diagnosed with type II diabetes using ICD codes for diagnosis
  - ICD-10: 'e11', 'e13', ...
  - ICD-9: '250', '249', '249', ...
- Excluding patients with type I diabetes
  - ICD-10: 'e10', ...
  - ICD-9: '25001', '25003', ...
- Adding a boolean feature for prediabetes
  - ICD-10: 'e161', 'e162', ...
  - ICD-9: 'r73', 'r81', 'r824', ...
Industry needs large scale solutions:

- Common practices of distributed computing programming
- Strict adherence to SDLC with emphasis on:
  - Extensibility:
    - Data validation at the beginning of the pipeline
    - Generic codes as much as possible
  - Continuous Integration:
    - Automating unit and integration tests using Travis CI
  - User Experience (UX):
    - Restful API for UI

http://www.pinevalleyinstitute.co.za/introduction-software-development/sdlc/
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Area Under Curve (AUC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boosting Methods</td>
<td>78-79%</td>
</tr>
<tr>
<td>XGBoost, Catboost, Light GB</td>
<td></td>
</tr>
<tr>
<td>Wide and Deep</td>
<td>76%</td>
</tr>
<tr>
<td>Deep Neural Networks (DNN)</td>
<td>70%</td>
</tr>
<tr>
<td>Random Forest, Naive Bayes</td>
<td>69%</td>
</tr>
</tbody>
</table>
Feature Importance

We use XGBoost to find the most important predictors of diabetes:
Make a Business Decision: Precision vs. Recall

- Depending on marketing purposes we sometimes need high precision and sometimes high recall (sensitivity).
  - High recall: Email campaign
  - High precision: Mail campaign
Findings and Contributions

1. Developed extensible forecasting models to assess imminent risk of diabetes, CV events and CVD in general for Healthgrades outreaching campaigns.
2. Through feature importance analysis we obtained a good insight into which health conditions affect the risk of developing certain diseases.
Future Work

Cardiovascular Diseases

- Heart Failure
- Arrhythmia
- Cardiomyopathy
- Heart Valve Disease
- Aortic Diseases

Atherosclerosis

- Coronary Heart Disease
- Stroke
- Peripheral Arterial Disease

DOI: 10.1161/CIR.0000000000000659
Thank You for Your Attention!

We Welcome Your Questions, Comments & Suggestions!