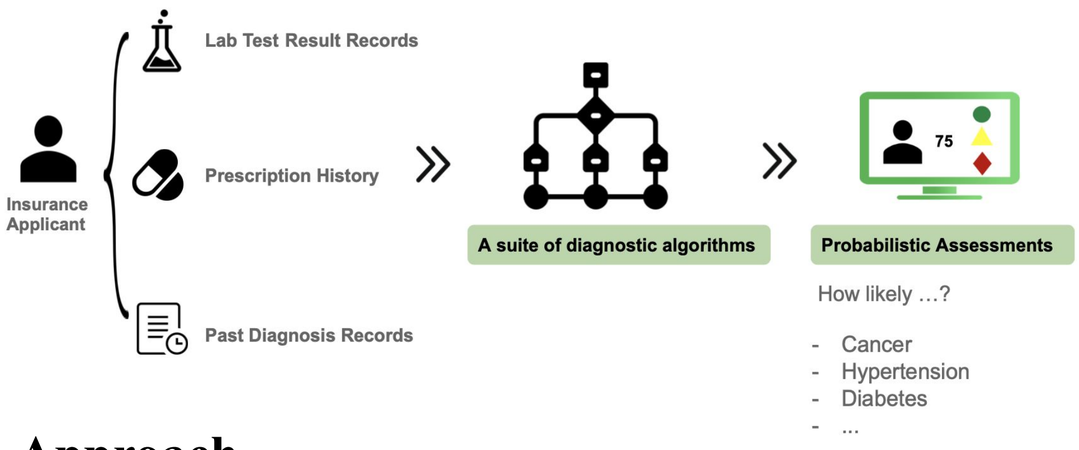
Disease Risk Evaluations in Life Insurance Underwriting via Laboratory and Prescription-Driven Diagnosis Models

Millie Mao, Jonathan Tukiman
Company Advisors: Brian J. Lanzrath, Yan Yan, Jennifer Delzeit, Benjamin Abbott
Faculty Advisor: Dimitris Bertsimas

1 Project Overview

Quest Diagnostics is the world's largest provider of clinical testing services. ExamOne is a Quest subsidiary that provides underwriting (risk assessment) solutions for life insurance industry. For a life insurance company, it is crucial to carefully evaluate applicants' health conditions to properly make decisions, such as determining suitable plans and determining prices for insurance products. Quest and ExamOne have access to multiple healthcare databases. Our project aims to leverage available historical health records and build a suite of diagnosis models to output risk assessments on insurance applicants for underwriters, providing valuable insights for them to make business decisions.

2 Project Flow



3 Data Description

Target — 29k Individuals
Whether a applicant ***has ever been diagnosed*** with a given medical condition

Lab Test Results — 4M Records
Name and result values of clinical lab tests taken by the patient in the past decade

Prescription Histories — 9M Records
Name and dosage of prescriptions that have ever been prescribed to a patient

4 Approach

Data Cleaning

Diagnosis Records - Re-encode medical conditions encoded in ICD10 codes to customize header levels utilized in underwriting assessments

Lab Test Results - Determine the nature of each lab test (uniquely identified by one *Loinc code*), uniform inconsistent result values, and convert to numerical

Prescription Records - Remove units and special characters in prescription dosage and keep numerical values

Feature Engineering and Data Aggregation

Lab test result records and prescription records are *encounter level*, indicating one particular clinical lab test taken or one ordering of prescription

» Aggregate heterogeneous historical lab test results and prescription records to individual level via summary statistics on each Loinc code, including min/max result values, count/frequency of tests taken in past years, count of tests taken in the past year, average dosage of prescription, etc.

Dimension of Resulting Feature Space: 3234

Imputation on Feature Matrix

Heterogeneous availability of records among individuals: for Loinc codes that have never been taken by individuals, it is necessary to fill in missing result values to obtain complete feature matrix for modeling.

» Zero Imputation, KNN Imputation, Adjusted Mode Imputation (impute by the mode of results values *given gender and age group* and count each individual once to avoid bias towards sick individuals)

What are the requirements for models qualified for implementation in underwriting purpose?

- **Good predictive power** - a model qualified for usage in risk assessment must have a test AUC above 0.8
- **Good interpretability** - the decision making process must be transparent to end users and match medical interpretations
- **Effective feature selection mechanism** - from feature space with dimension over 3000, effective feature selection for each disease is necessary

Modeling

Logistic Regression with LASSO	Sparse Regression	Tree-based Models	Combined Models
Embedded Feature Selection Method	Optimization-based regression seeking optimal combination of predictors	CART, Random Forest, XGBoost, Optimal Classification Tree	Sparse Regression for initial 100 features selection and feed into CART

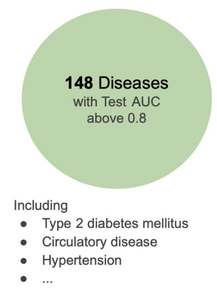
5 Model Results and Findings

- Logistic Regression with LASSO regularization fails to achieve desirable sparsity level in selected features
- Sparse Regression has overall better performances than CART
  MedianTest AUC among 300 experimented headers
  Sparse Regression 0.799
  CART 0.784
- Using Sparse Regression for initial 100 feature selection before feeding into tree-based models improves overfitting
- **Sparse Regression** is chosen as the main modeling approach by ExamOne for interpretability and ease of implementation

6 Model Example

Hypertension (I10)	
Feature	**Weights**
8480-6 min (systolic blood pressure)	0.00214
8480-6 max (systolic blood pressure)	-0.00229
2345-7 mean (tyriglyceride blood test)	0.00114
M4D average (Anatihyperlipidemic)	0.01116
A4D average (Antihypertensives)	2.28886
C4L average (Metformin)	0.00364
Past history of E11 (Type 2 Diabetes mellitus)	0.67554
Past history of I25 (Chronic ischemic heart disease)	0.95150

7 Business Impacts



$26.4 Million
Estimated Savings for End Users

8 Recommended Next Steps

- Collect more complete clinical lab results data - we observed that for some patients with over 10 diagnosed diseases in the past years have very few lab results records available, which might result from switching healthcare providers
- Evaluate and develop the potential of using more complex models for production usage