Detecting Risk of Heart Disease

Precious Smith
DSIR-22221E
General Assembly
Presented 5.14.21

Agenda

1. Background
2. Data Collection
3. Exploratory Data Analysis
4. Modeling
5. App Demo
6. Conclusions & Future Direction

Background

Heart Disease Stats

- Heart disease is the #1 cause of death.
- There are several types of heart disease but **coronary artery disease** is the most common.
- Symptoms include:
  - Chest pain or discomfort (angina)
  - Weakness, light-headedness, nausea, cold sweat
  - Pain or discomfort in arms or shoulders
  - Shortness of breath

Sources: CDC

655,000 deaths

1 in 4 deaths each year

$219 billion

Cost of heart disease in the US, 2014-2015

47% of Americans

Key risk factors: hypertension, high cholesterol, smoking

Source: CDC-Million Hearts

Other Risk Factors

Source: CDC

Background

Prevention & Treatment

- Eat a healthy, balanced diet
- Live an active lifestyle
- Don’t smoke
- Take medications as prescribed

ABCS of Heart Health

- Take **aspirin** as prescribed
- Control your **Blood** pressure
- Manage your **Cholesterol**
- Don’t **Smoke**

Source: CDC-Million Hearts

How can we help?

Empower patients to initiate informed conversations with their physicians.

Educate patients on things they should be aware of and equip them with tools to make better decisions for themselves.

“

Problem Statement:

Can we predict whether someone is at risk of heart disease using a few key features, then create an app that informs users of their risk and provides them with test and discussion suggestions to take to their physicians?

Data Collection

National Health Interview Survey
CDC, 2019

2

CDC - National Health Interview Survey

- 530+ features
- Nearly 32,000 observations
- A lot of nulls
- Target: Coronary Heart Disease
  - 94% No, 6% Yes

Source: CDC

Data Cleaning

- Dropped columns with excessive nulls
- Reclassified target column as binary (0=no, 1=yes)
- Changed data types for categorical data
- Created dummies
- Used SelectKBest to find top 10 predictive features
- My own selection of features
- Used SMOTE to create balanced classes

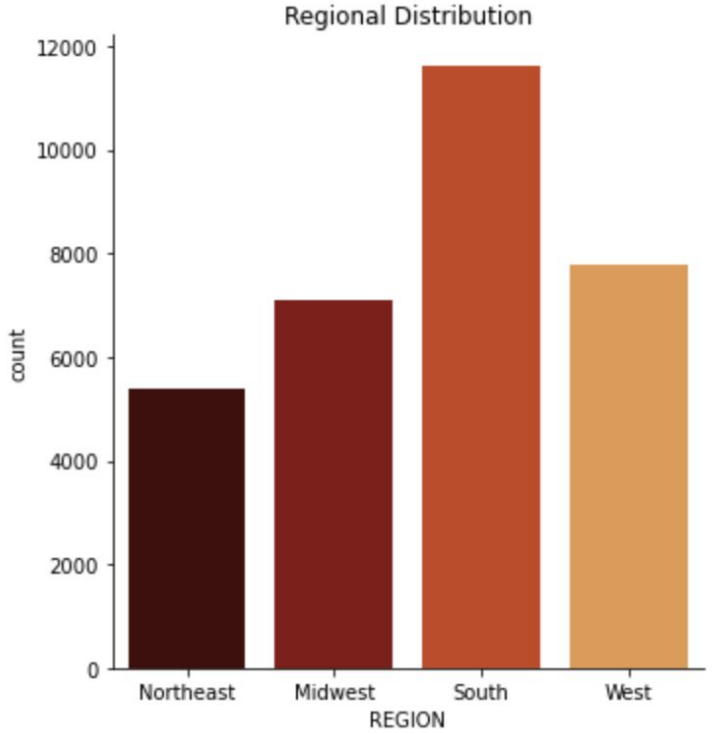
Source: CDC

Exploratory Data Analysis

What is the data telling us about America's health?

3

Where is the data from?



Heart Disease by Region			
Northeast	Midwest	South	West
6.2%	5.6%	6.5%	4.6%

What else is the data telling us?

	% of Respondents	% Who Also Have Heart Disease
Hypertension	36%	13%
High Cholesterol	29%	14%
Current Smoker	13%	9%
Diabetes	10%	18%

Modeling

4

Feature Selection & Balancing Classes

SelectKBest

Narrowed down features after dropping null columns and dummifying.

k=10

Hand-selected

Chose features based off research for what may have been good predictors of heart disease.

56 features

SMOTEN/NearMiss

Heart Disease: 94% no, 6% yes

SMOTEN: 50%, 45,072 observations

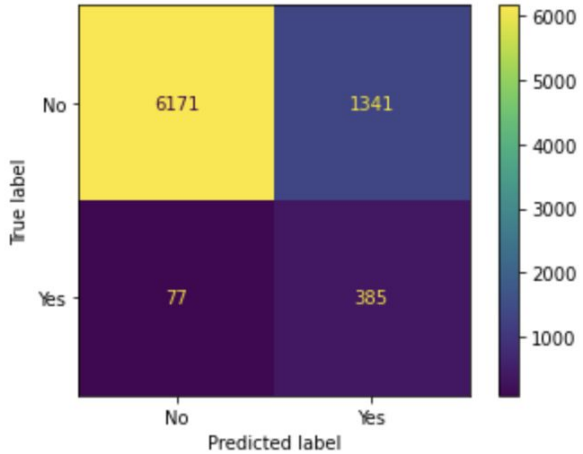
NearMiss: 50%, 2768 observations

Heart Disease Models & Scores

	Accuracy Score	Recall Score
SelectKBest, SMOTEN, Logistic Regression	.82	.83
SelectKBest, SMOTEN, Decision Tree	.84	.76
SelectKBest, SMOTEN, AdaBoost	.82	.83
SelectKBest, SMOTEN, Gradient Boost	.83	.81
SelectKBest, NearMiss, Logistic Regression	.83	.80
SelectKBest, NearMiss, Decision Tree	.78	.77
Hand-picked, SMOTEN, Logistic Regression	.95	.41
Hand-picked, SMOTEN, Decision Tree	.92	.35
Hand-picked, NearMiss, Logistic Regression	.39	.89
Hand-picked, NearMiss, Decision Tree	.51	.84

Final Heart Disease Model

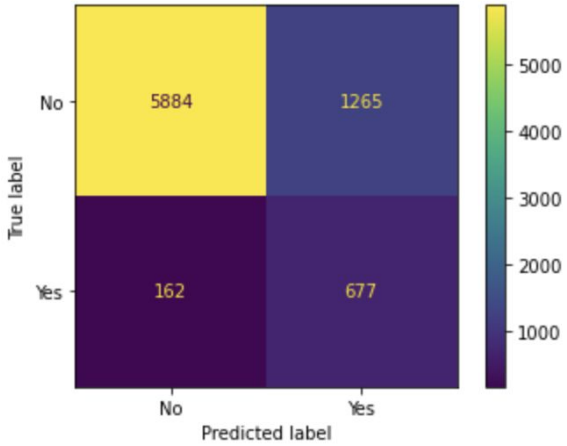
- SelectKBest, k=10
- SMOTEN
- Logistic Regression
- Accuracy: .82
- Recall: .83



	coefs	feature_names	questions
0	-0.171072	AFNOW_2.0	no one on active duty
1	0.724457	EMPWRKLSWK_A_2	did not work last week
2	-0.114547	MEDICARE_A_3	not on medicare
3	0.319874	OVER65FLG_A_1	one person in home over 65
4	-0.328528	INCSSRR_A_2.0	no income from railroad
5	-0.350059	HIKIND02_A_2	health insurance not mentioned
6	-3.383119	MIEV_A_2	no heart attack
7	-2.461329	ANGEV_A_2	no angina
8	-0.722358	CHLEV_A_2	no high cholesterol
9	-1.047147	HYPEV_A_2	no hypertension

Diabetes Model

- SelectKBest, k=10
- SMOTEN
- Logistic Regression
  - C=.01
- Accuracy: .82
- Recall: .81



	coefs	feature_names	questions
0	0.341789	EMPWRKLSWK_A_2	did not work last week
1	0.057768	MEDICARE_A_3	not on medicare
2	-0.493950	SHTPNUEV_A_2	no pneumonia shot
3	1.452884	DIABLAST_A_1	blood sugar test w/i last year
4	-1.207990	RX12M_A_2	no rx last 12 months
5	-0.139947	HIKIND02_A_2	health insurance not mentioned
6	-2.428958	PREDIB_A_2	no prediabetes
7	-0.331133	CHDEV_A_2	no heart disease
8	-0.580376	CHLEV_A_2	no high cholesterol
9	-0.802246	HYPEV_A_2	no hypertension

Heart Health Screening Tool

- 2 questionnaires
- 9 questions each
- At risk or not
- Suggestions for tests
- Conversation suggestions for physician

5

Conclusions & Future Direction

6

Conclusions

- We are able to build a model and an app that predicts whether someone is at risk of heart disease or diabetes with 82% accuracy.
- Minimized false negatives but false positives are still high.
- Medical data with test scores may improve model
- HIPAA laws make it difficult to access health data but data scientist with access may be able to create something better
- While app is informative, results may be too grim
- Don’t want to scare people away

Future Direction

- Doctor’s office waiting room
- Interactive version of a rack card
- Patient portal that can be submitted to nurse/doctor
- More information on the app instead of links leaving the app

Thank you!

- My instructors: Chuck, Varun & Grant
- My classmates, group mates, study partners

Questions?