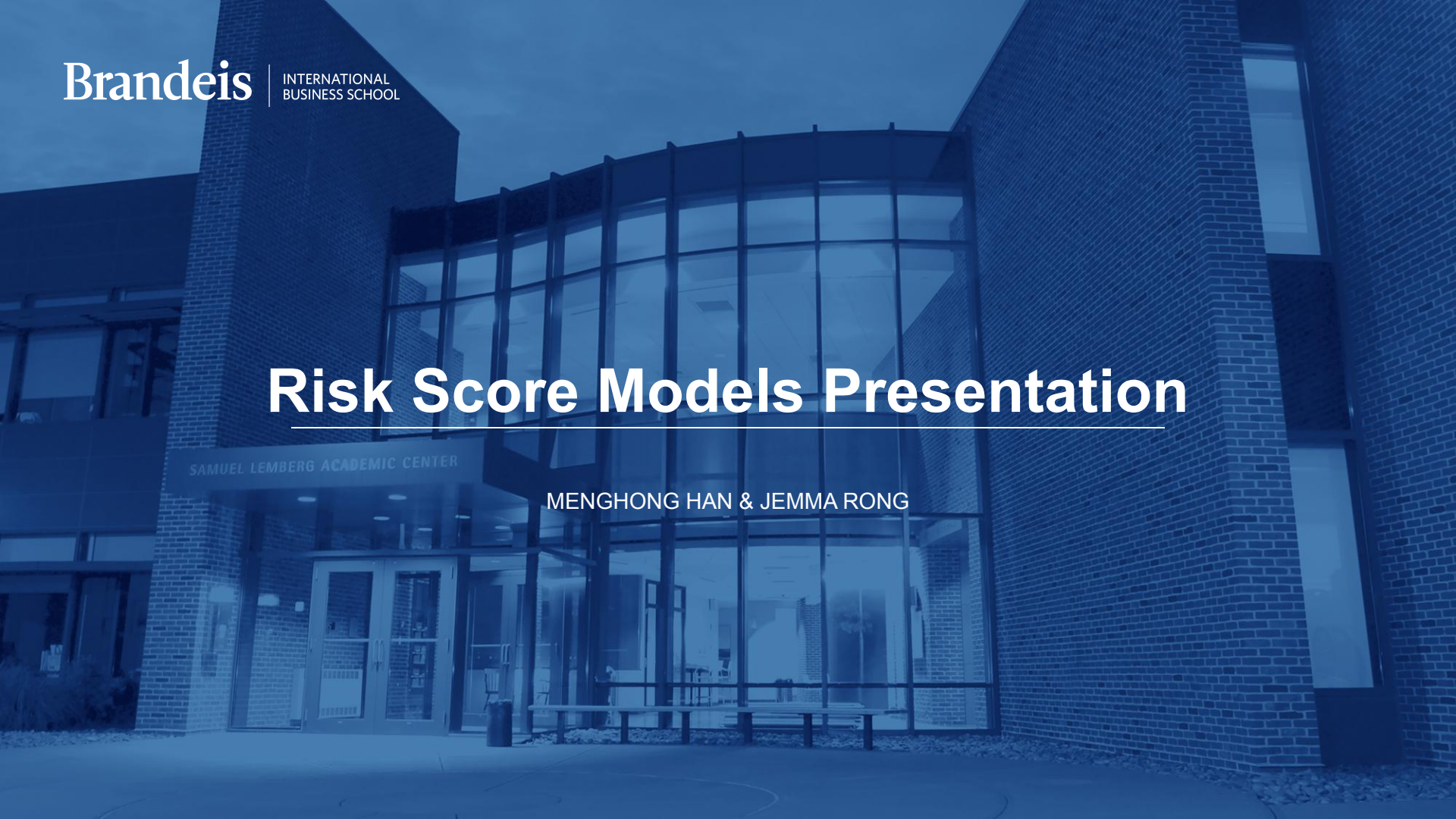


Brandeis | INTERNATIONAL BUSINESS SCHOOL

Risk Score Models Presentation

MENGHONG HAN & JEMMA RONG

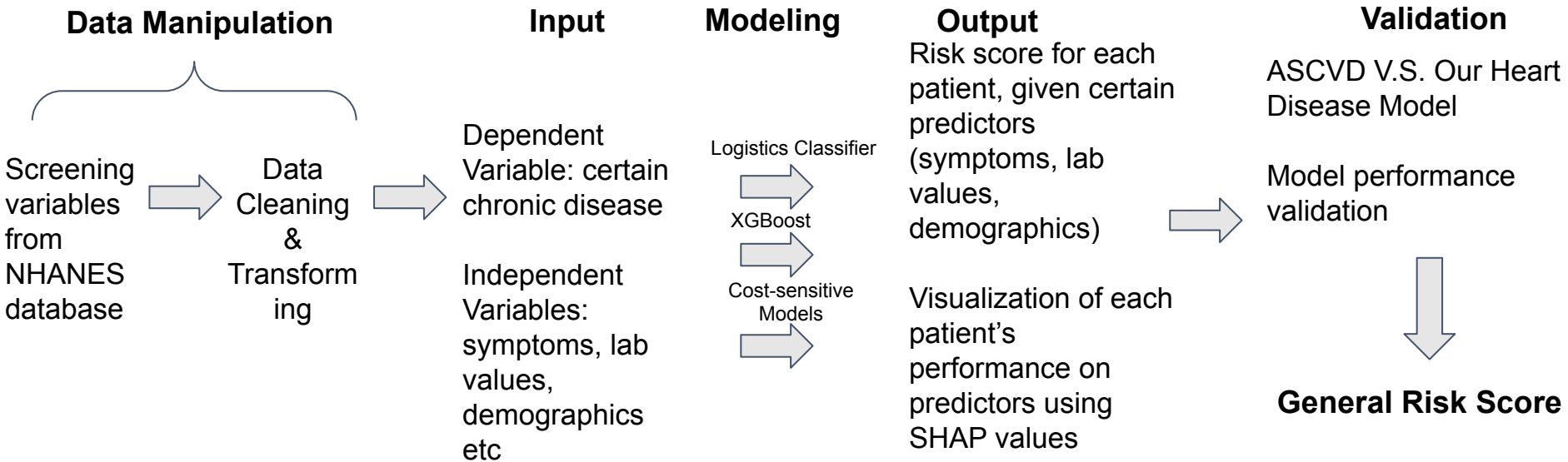
Current Achievements

1. Conducted exploratory analysis for Top 10 chronic diseases ranking by mortality in the US after screening variables in the CDC: National Health and Nutrition Examination Survey database
2. Trained machine learning models (Logistics/ XGBoost etc.) to predict risk scores of chronic diseases
3. Optimized the predictive models by adopting cost-sensitive machine learning methods

Project Overview

Jemma

Project Flow Chart



Chronic Disease Ranking By Mortality & Cost

By Mortality

Rank	Chronic Diseases	Deaths	Population	Crude Rate Per 100,000(95% Confidence Interval)
1	#Diseases of heart (I00-I09,I11,I13,I20-I51)	12,878,021	6,088,633,001	211.5 (211.4 - 211.6)
2	#Malignant neoplasms (C00-C97)	11,442,918	6,088,633,001	187.9 (187.8 - 188.0)
3	#Cerebrovascular diseases (I60-I69)	2,874,333	6,088,633,001	47.2 (47.2 - 47.3)
4	#Chronic lower respiratory diseases (J40-J47)	2,754,413	6,088,633,001	45.2 (45.2 - 45.3)
5	#Alzheimer disease (G30)	1,616,835	6,088,633,001	26.6 (26.5 - 26.6)
6	#Diabetes mellitus (E10-E14)	1,484,889	6,088,633,001	24.4 (24.3 - 24.4)
7	#Nephritis, nephrotic syndrome and nephrosis (N00-N07,N17-N19,N25-N27)	909,999	6,088,633,001	14.9 (14.9 - 15.0)
8	#Septicemia (A40-A41)	716,737	6,088,633,001	11.8 (11.7 - 11.8)
9	#Chronic liver disease and cirrhosis (K70,K73-K74)	646,963	6,088,633,001	10.6 (10.6 - 10.7)
10	#Essential hypertension and hypertensive renal disease (I10,I12,I15)	524,956	6,088,633,001	8.6 (8.6 - 8.6)
11	#Parkinson disease (G20-G21)	443,748	6,088,633,001	7.3 (7.3 - 7.3)

Not sure whether Septicemia belongs to chronic disease or not

data source: Underlying Cause of Death, 1999-2018 Results, CDC, https://wonder.cdc.gov/controller/datarequest/D76;jsessionid=52D79134D710A2677CEBD65268C42C90#Options

By Cost

Rank	Chronic Diseases	billion $ per year	Rank by morality
1	Cardiovascular diseases	$317	1
2	Smoking-related health issues	$300	4
3	Alcohol-related health issues	$249*	
4	Diabetes	$245	6
5	Alzheimer's disease	$236	5
6	Cancer (malignant neoplasms)	$171	2
7	Obesity	$147	
8	Arthritis	$128	
9	Asthma	$56	
10	Stroke (cerebrovascular diseases)	$33	3

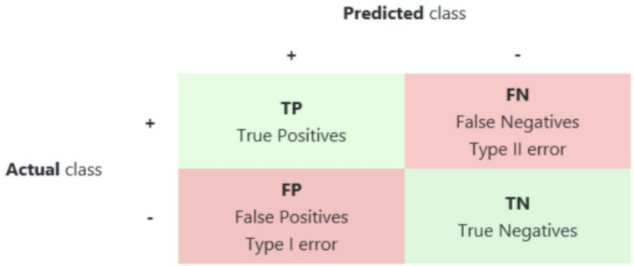
Data sourse: CDC, https://healthpayerintelligence.com/news/top-10-most-expensive-chronic-diseases-for-healthcare-payers

* data in 2010, others use 2016 data

Top 10 Chronic Disease:

1. Heart Disease
2. Cancer
3. Stroke
4. Chronic lower respiratory disease
5. Alzheimer's disease
6. Diabetes
7. Nephritis
8. Chronic liver disease
9. Hypertension and hypertensive renal disease
10. Parkinson Disease

Overview--Key concepts



1. **Confusion Matrix:**

In the project, for label==0 (healthy people):

True positives (TP): predicted positive, actual positive

False positives (FP): predicted positive, actual negative (healthy people who are diagnosed with certain disease) ⇒ mistakes that our models should avoid

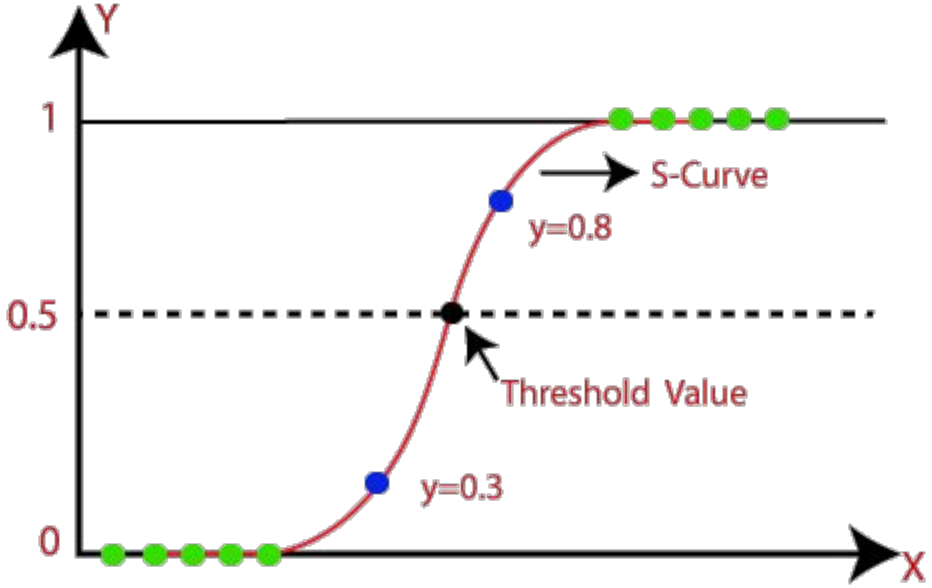
True negatives (TN): predicted negative, actual negative

False negatives (FN): predicted negative, actual positive

2. **Precision & Recall Score: ways to measure model performance**
   - Precision represents the proportion of the models' predictions of disease where disease is actually present.
   - **Recall represents the proportion of all cases of diseases that the model accurately predicted. (the metric that we pay the more attention to)**

Overview--Core Models

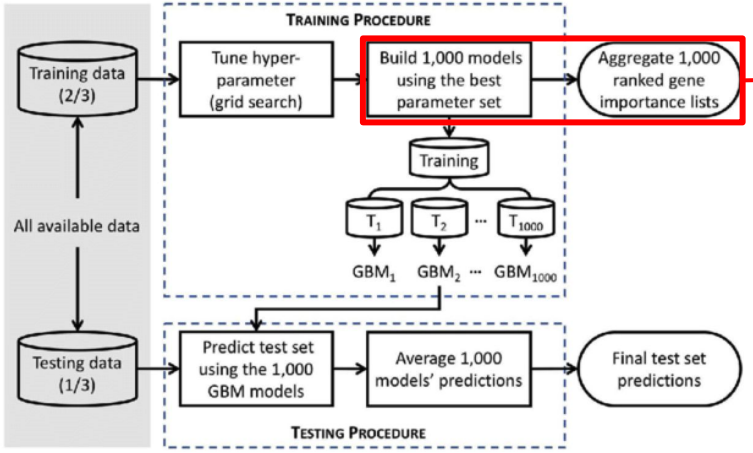
Logistic Regression Classifier: Predict the **probability** that a given data entry belongs to the category labeled as “1” ⇒ the probability represents the risk score



Overview--Core Models

XGBoost Classifier:

Rather than training all of the models in isolation of one another, boosting trains models in succession, with **each new model being trained to correct the errors made by the previous ones**. Models are added sequentially until no further improvements can be made.



Makes accuracy higher than regular models

Heart Disease

Menghong

Modeling Process -- Heart Disease

Target variable: the probability of answering "Yes" involve one or more following questions

MCQ160c - Ever told you had coronary heart disease
MCQ160b - Ever told had congestive heart failure
MCQ160e - Ever told you had heart attack

•Modeling Process flow:

Input data (predictors & target) ⇒ Split data (70% training & 30% testing) ⇒ Use training data to get the predictive model (**Logistic, Decision tree, XGBoost** ⇒ Use testing data to get the risk scores

Data Manipulation: Variables for predicting Heart Disease

No.	Factor	Variable Name	Question & Code	Type
	Target	Heart disease(general)	involve one or more follow Y	Categorical
		Coronary heart disease	MCQ160c - Ever told you had coronary heart disease	Categorical
		heart failure(chronic)	MCQ160b - Ever told had congestive heart failure	Categorical
		heart attact	MCQ160e - Ever told you had heart attack	Categorical
1	Predictor	Diabetes	DIQ010 - Doctor told you have diabetes	Categorical
2		BMI	BMXBMI - Body Mass Index (kg/m**2)	Continuous
3		Age	RIDAGEYR - Age in years at screening	Continuous
4		Gender	RIAGENDR - Gender	Categorical
5		Race	RIDRETH3 - Race/Hispanic origin w/ NH Asian	Categorical
6		Family history	MCQ300a - Close relative had heart attack?	Categorical
7			MCQ300b - Close relative had asthma?	Categorical
8			MCQ300c - Close relative had diabetes?	Categorical
9		Semi-health	MCQ365a - Doctor told you to lose weight	Categorical
10			MCQ365b - Doctor told you to exercise	Categorical
11			MCQ365c - Doctor told you to reduce salt in diet	Categorical
12			MCQ365d - Doctor told you to reduce fat/calories	Categorical
13		Angina	MCQ160d - Ever told you had angina/angina pectoris	Categorical
14		Pulse	BPXPULS - Pulse regular or irregular?	Categorical
15		Total Cholesterol	LBDTCSI - Total Cholesterol (mmol/L)	Continuous
16		Direct HDL-Cholestero	LBDHDDSI - Direct HDL-Cholesterol (mmol/L)	Continuous
17		Complete Blood Count with 5-part Differential	LBXMCVSI	Continuous
18		serum creatinine	LBXSCR	Continuous
19		UA	LBXSUA	Continuous
20		AI	TC/HDL-C	Continuous

Dataset: 5451 rows × 27 columns (Target == 0: 5062 rows, Target == 1: 389 rows)

Modeling Comparison -- Heart Disease

Model 1: Logistics Regression Classifier	Training Data			Testing Data		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Target == 0	0.98	0.78	0.87	0.97	0.77	0.86
Target == 1	0.22	0.80	0.35	0.17	0.69	0.27
Accuracy			0.78			0.77

Confusion Matrix for testing data		Predicted Target == 0		
		Negative	Positive	
Actual Target == 0	Negative	1182	351	Specificity: TN/ (FP+TN) = 0.771
	Positive	32	71	Sensitivity: TP/ (TP+FN) = 0.689

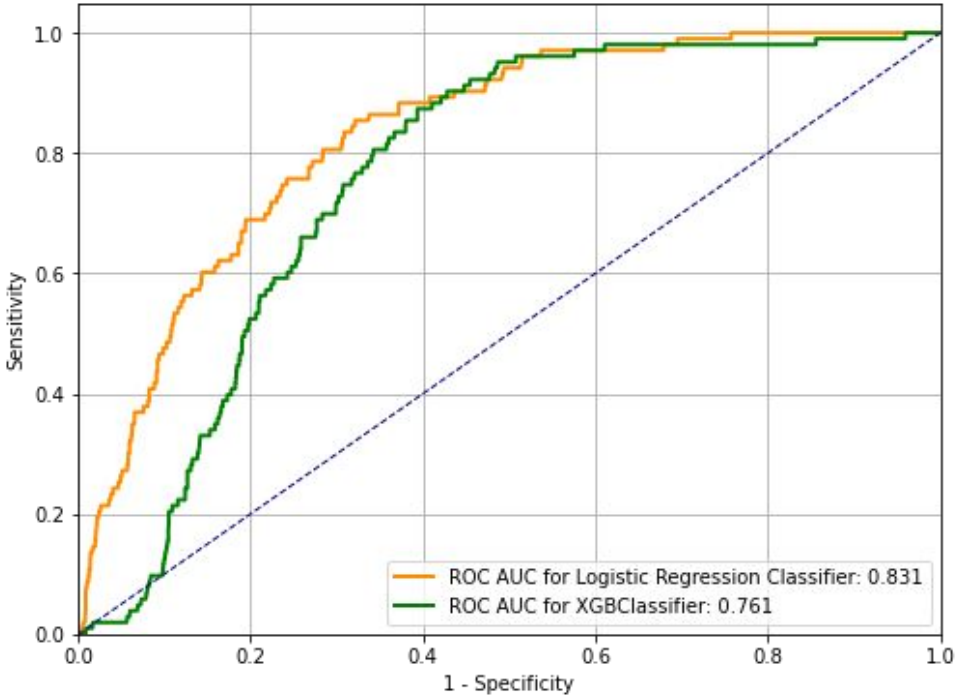
Modeling Comparison -- Heart Disease

Model 2: XGBoost Classifier	Training Data			Testing Data		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Target == 0	0.96	0.99	0.97	0.94	0.93	0.93
Target == 1	0.77	0.48	0.59	0.05	0.06	0.05
Accuracy			0.95			0.87

Confusion Matrix for testing data		Predicted Target == 0		
		Negative	**Positive**	
Actual Target == 0	**Negative**	1420	113	Specificity: TN/ (FP+TN) = 0.926
	Positive	97	6	Sensitivity: TP/ (TP+FN) =0.058

Modeling Comparison -- ROC-AUC Curve

- **A graph shows the performance of a classification model at all classification thresholds**
- **Could be used to determine which one performs better**
- **Logistic Regression performs better than XGB.**



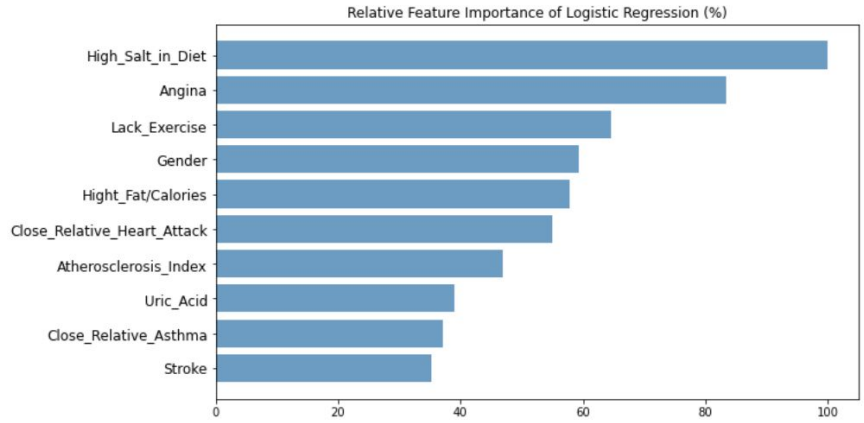
Choose the optimal threshold

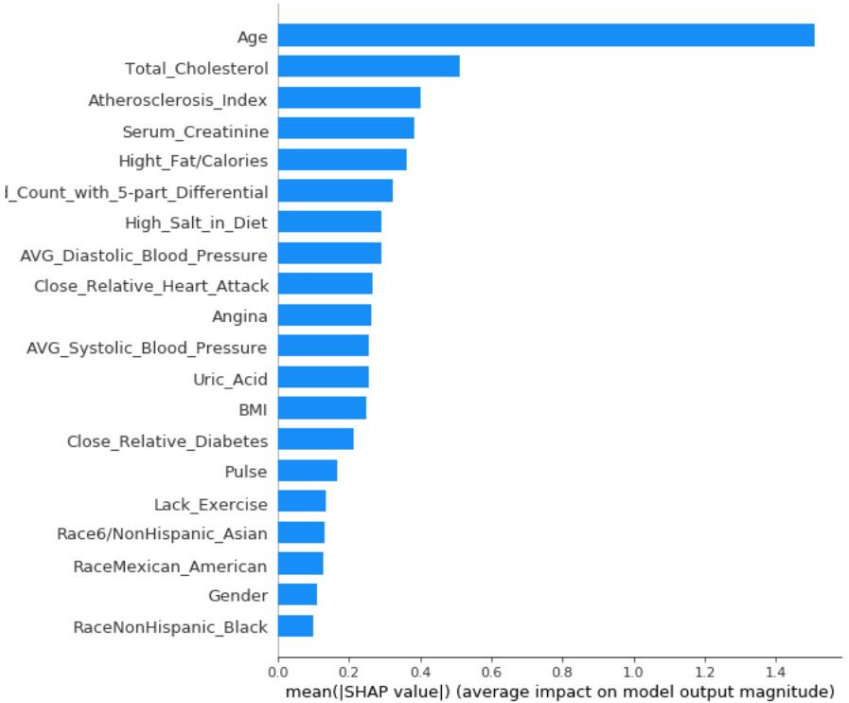
Best Threshold=0.391424, G-Mean=0.762

	precision	recall	f1-score	support
0	0.99	0.68	0.80	1533
1	0.15	0.85	0.26	103
accuracy			0.69	1636
macro avg	0.57	0.77	0.53	1636
weighted avg	0.93	0.69	0.77	1636

```
[[1041  492]
 [  15   88]]
```

Modeling Process -- Heart Disease





Problems need to be solved

Problem Statement:

- We can only get feature importance for the whole model, while each factor may have different impact on different individuals, we cannot see the performance of individuals in each specific factor.
- Feature importance could vary from models, it could even be different within a model with different calculation options. We need to find a measure can are both consistent and accurate.

Solution:

- Use “SHAP value” and show the visualization feature importance for **each individual** after showing the risk score.

SHAP: SHapley Additive exPlanation

SHAP (SHapley Additive exPlanation) is a game theoretic approach to explain the output of any machine learning model.

The goal of SHAP is to explain the prediction for any instance x_i as a sum of contributions from it's individual feature values. Individual feature values are assumed to be in a cooperative game whose payout is the prediction. In this setting, Shapley values provides a measure to fairly distribute the payout among the feature values.

$$\phi_i(v) = \frac{1}{|N|!} \sum_R \left[v(P_i^R \cup \{i\}) - v(P_i^R) \right]$$

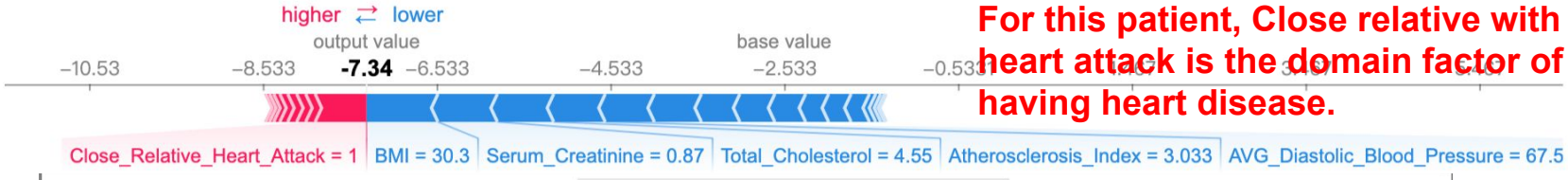
$$y_i = y_{\text{base}} + f(x_{i,1}) + f(x_{i,2}) + \cdots + f(x_{i,k})$$

Φ: Shapley value
N: Number of player (feature)
P_i^R: Set of player with order
$V(P_i^R)$: Contribution of set of player with order
$V(P_i^R \cup \{i\})$: Contribution of set of player with order and player i

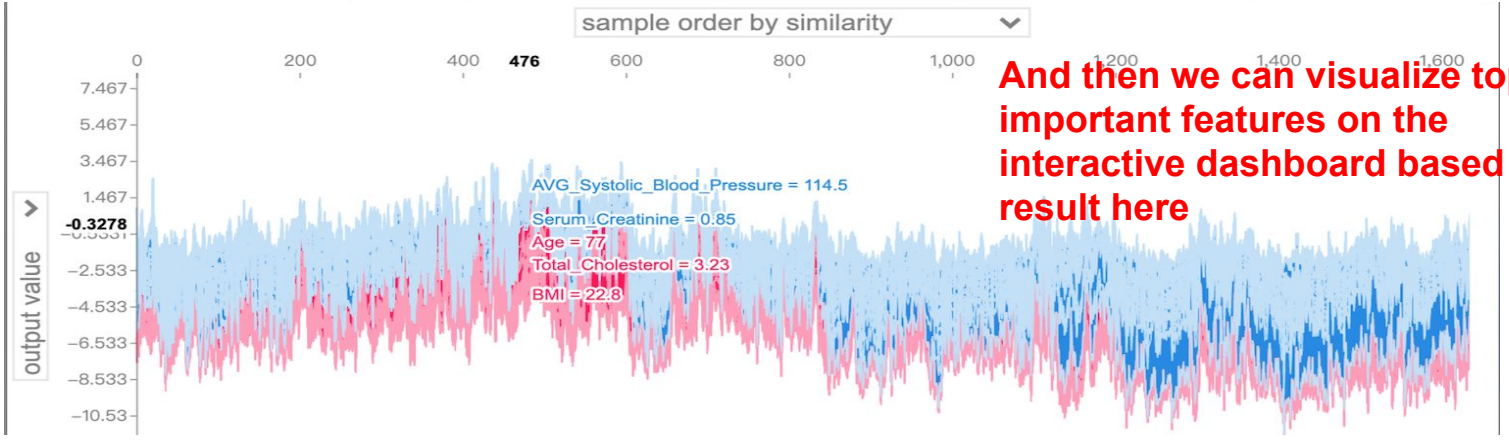
Visualization with heart disease dataset (cont.)

Risk Score: 98.382721

```
shap.force_plot(explainer.expected_value, shap_values[0,:], X_test.iloc[0,:])
```



For this patient, Close relative with heart attack is the domain factor of having heart disease.



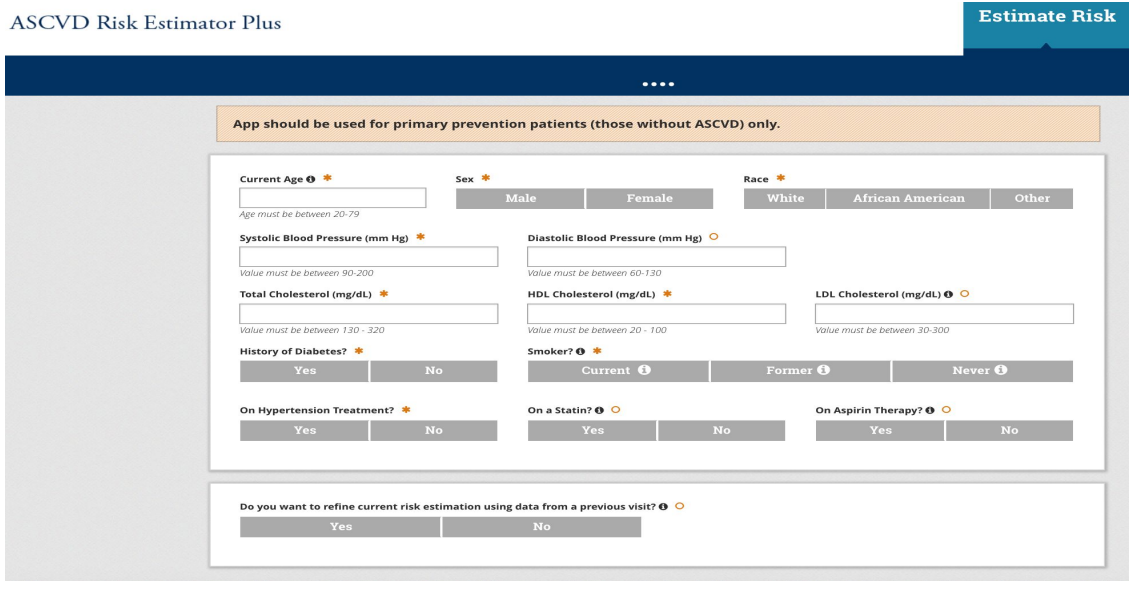
And then we can visualize top important features on the interactive dashboard based on the result here

Validation of Risk Score Model

Menghong

Validate Risk Score using ASCVD Risk Calculator

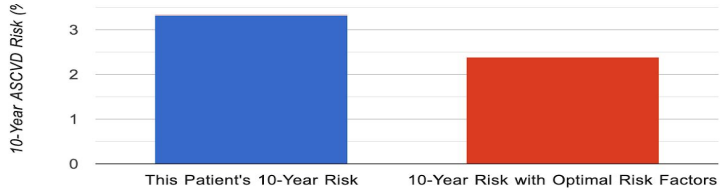
Key predictors: Age/Gender/Race/Systolic & Diastolic Blood Pressure/ Total & HDL & LDL Cholesterol (mg/dL)/ History of Diabetes/Smoker/Hypertension Treatment etc.



Validate Risk Score using ASCVD Risk Calculator

Our Model : 50.002934

ASCVD Lifetime risk: 50%



Lifetime risk of atherosclerotic cardiovascular disease :

50%
(95% CI 46% to 55%)

Lifetime risk for a 50-year-old with optimal risk factors :

5%
(95% CI 0% to 12%)

ASCVD Risk Interpretation

- This patient is at LOW 10-year risk (< 7.5%) for atherosclerotic cardiovascular disease (ASCVD)
- In individuals not receiving cholesterol-lowering drug therapy, recalculate the 10-year ASCVD risk every 4 to 6 years (assuming age 40-75 years, no clinical ASCVD or diabetes, and LDL 70-189 mg/dL)

Variables for ASCVD Assessment:

- Age: 41 year-old, African American, male
- Total cholesterol: 194 mg/dL
- HDL cholesterol: 45 mg/dL
- Systolic blood pressure: 94 mmHg
- Treatment for hypertension: No
- Diabetes: No
- Smoker: Yes

Result Table for All Chronic Diseases

Disease Name	"Best" Model	Recall Score of minority class before tuning	"Best" Model Recall Score of minority class after tuning
Heart Disease	**Logistics Regression Classifier**	0.69	0.85
Nephritis		0.58	0.68
Chronic Lower Respiratory Disease		0.74	0.74
Diabetes		0.75	0.79
Chronic liver disease		0.38	0.79
Hypertension and hypertensive renal disease		0.74	0.74
Stroke		0.7	0.76
Parkinson Disease		0	1
Cancer (Breast, Lung)		0.22/0	0.44/0

Cross-check Risk Score Model Performance through Papers

All chronic disease algorithms have been previously validated in Ontario (**sensitivity from 60.2–95.0%, specificity from 76.5–99.2%**)

Chronic Disease Population Risk Tool (CDPoRT): a study protocol for a prediction model that assesses population-based chronic disease incidence

https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-018-0042-5

Chronic disease	Algorithm	Hospital discharge codes		Physician claim codes (ICD-9)	Cancer registry codes (ICD-O-3)	Cause of death (ICD-9)	Sensitivity (95% CI)	Specificity (95% CI)
		ICD-9	ICD-10					
Congestive heart failure (CHF) [19]	2 hospitalization records in a 1-year period *or* 1 physician claim and 1 hospitalization record in a 1-year period *or* 2 physician claim records in a 1-year period *or* death certificate cause of death	428	I50.0, I50.1, I50.9	428	N/A	428	84.8 (77.7, 92.0)	97.0 (96.3, 97.9)
Chronic obstructive pulmonary disease (COPD) [20]	1 hospitalization record *or* 1 physician claim records *or* death certificate cause of death	491, 492, 496	J41, J42, J43, J44	491, 492, 496	N/A	491, 492, 496	85.0 (77.0, 91.0)	78.4 (73.6, 82.7)
Diabetes [21]	1 hospitalization record *or* 2 physician claim records in a 2-year period *or* death certificate cause of death	250	E10, E11, E13, E14	250	N/A	250	86	97
Lung cancer [22]	1 cancer registry record in the Ontario Cancer Registry *or* death certificate cause of death	N/A	N/A	N/A	C34.0, C34.1, C34.2, C34.3, C34.8, C34.9	162.2, 162.3, 162.4, 162.5, 162.8, 162.9	N/A	N/A
Myocardial infarction (MI) [23]	1 hospitalization record *or* death certificate cause of death	410	I21	N/A	N/A	410	95.0	88.0
Stroke including transient ischemic attack (TIA) [24]	1 hospitalization record *or* 2 physician claim records in a 1-year period *or* death certificate cause of death	362.3, 430, 431, 434, 435, 436	G45 (excluding G45.4), H34.0, H34.1, I60, I61, I63 (excluding 163.6), I64	432, 435, 436	N/A	362.3, 430, 431, 434, 435, 436	60.2 (50.7, 69.6)	99.2 (99.0, 99.5)

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

Cost-sensitive Machine Learning

Jemma

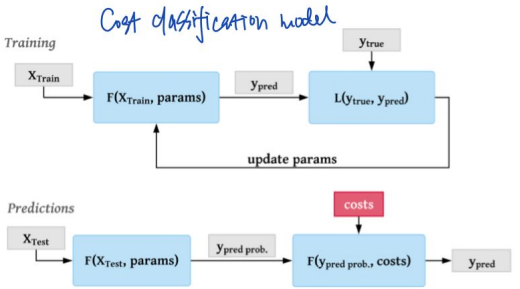
Cost Classification Model

- "As opposed to a cost-sensitive model that trains with a customized loss function, cost classification models calculate the expected costs based on **predicted probabilities**. The expected costs for a predicting a legitimate and a fraudulent transaction are calculated as follows:"

$$E_{classify\ fraud} = c_{TP} * y_{proba} + c_{FP} * (1 - y_{proba})$$

$$E_{classify\ legitimate} = c_{TN} * (1 - y_{proba}) + c_{FN} * y_{proba}$$

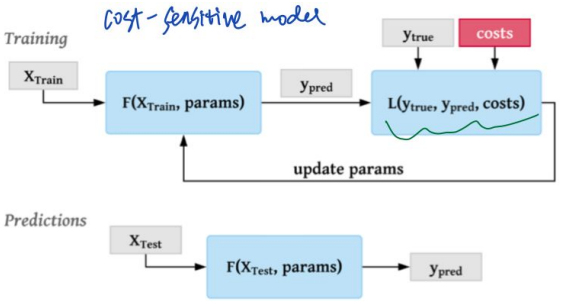
- The classifier chooses whichever prediction is expected to result in lower costs.



Cost-sensitive Machine Learning

- **Definition:**

Cost-sensitive learning is a subfield of machine learning that takes the **costs of prediction errors** (and potentially other costs) into account when training a machine learning model. It is a field of study that is closely related to the field of **imbalanced learning** that is concerned with classification on datasets with a skewed class distribution. **Trade-off between Accuracy and Cost**



- There are different perspectives towards cost-sensitive learning:

1) Model perspective
2) Hospital perspective
3) Patient perspective

- **Overall Hypothesis**:

H_0: After getting the positive prediction results, patients will immediately go to hospitals to get treated.

Perspective 1: **Model-based** Cost-sensitive Machine Learning

- Basic logic: Only generate **misclassification** cost (NOT REAL COST & NOT MEANINGFUL)
- TP = TN = 0:

FP = FN = C ∈ [10,1000]

Case: *Learning Cost-Sensitive Decision Trees to Support Medical Diagnosis*

Table 1. Attribute costs for Pima Indians Diabetes.

Test	Cost ($)	Group cost ($)
a. Number of times pregnant	1	
b. Glucose tolerance test	17.61	b + e = 38.29
c. Diastolic blood pressure	1	
d. Triceps skin fold thickness	1	
e. Serum insulin test	22.78	b + e = 38.29
f. Body mass index	1	
g. Diabetes pedigree function	1	
h. Age (years)	1	

Table 2. Cost matrix for Pima Indians Diabetes.

		Prediction	
		Healthy	**Diabetes**
Reality	**healthy**	$ 0	FP cost
	diabetes	FN cost	$ 0

Table 3. Average misclassification and attribute costs in the evaluation of 5 decision trees, for a range of misclassification costs (with FN cost equal to FP cost) and cost scale factors; 95% confidence intervals are included.

	Cost scale factor ω				
FN / FP cost	**0.0**	**0.1**	**0.2**	**0.5**	**1.0**
$10	23.4 ± 0.53	21.9 ± 0.57	19.5 ± 0.66	**6.4** ± 0.44	**5.7** ± 0.35
$20	26.0 ± 0.79	24.5 ± 0.81	22.2 ± 0.89	**9.5** ± 0.73	**8.8** ± 0.68
$50	33.8 ± 1.67	32.4 ± 1.67	30.6 ± 1.75	**19.0** ± 1.69	**18.4** ± 1.66
$100	46.9 ± 3.20	45.5 ± 3.21	44.4 ± 3.30	**34.7** ± 3.33	**34.2** ± 3.31
$200	73.1 ± 6.31	71.8 ± 6.31	72.2 ± 6.45	**66.2** ± 6.62	**66.0** ± 6.61
$500	**151.6** ± 15.7	**150.8** ± 15.7	**155.4** ± 16.0	**160.8** ± 16.5	**161.3** ± 16.5
$1,000	**282.5** ± 31.2	**282.3** ± 31.3	**294.0** ± 31.8	**318.3** ± 33.0	**320.2** ± 33.0
Accuracy (%):	73.8	73.7	72.3	68.5	68.2

Perspective 1: **Model-based** Cost-sensitive Machine Learning

- TP = TN = 0:

FP != FN but FP/FN = C ∈ [0,10] OR [10,200] ⇒ COST MIN & ACC MAX

Case: *Learning Cost-Sensitive Decision Trees to Support Medical Diagnosis*

Case: An Efficient Predictive Model for Myocardial Infarction Using Cost-sensitive J48 Model

Table 4. Average costs in the evaluation of 2 decision trees, for a range of misclassification costs (with FN cost different from FP cost) and cost scale factor (csf ω) 0.0 and 1.0.; 95% confidence intervals are included.

FN/FP ratio	FN ($)	FP ($)	csf ω = 0.0	csf ω = 1.0
1/10	10	100	34.3 ± 2.37	**17.3** ± 2.35
10/1	100	10	36.0 ± 2.52	**22.6** ± 2.75
3/1	150	50	47.9 ± 3.78	**37.2** ± 4.12
1/3	50	150	46.0 ± 3.58	**31.3** ± 3.59
½	500	1,000	**212.2** ± 24.58	**225.8** ± 25.20
2/1	1,000	500	**221.9** ± 25.67	**255.7** ± 28.09

Table 4:

The results of the proposed cost sensitive J48 model

		Cost ratio (cost of FN: cost of FP)					
		No costs	1: 10	1: 50	1: 100	1: 150	1: 200
Accuracy	No FS	58.67	62.67	65.33	68	68	68
	FS	60	64	80	85.33	85.33	82.67
Sensitivity	No FS	0	9.68	22.58	35.48	35.48	35.48
	FS	0	10	56.67	73.33	73.33	86.67
Specificity	No FS	100	100	95.45	90.91	90.91	90.91
	FS	100	100	95.56	93.33	93.33	80
F-measure	No FS	-	17.65	35	47.83	47.83	47.83
	FS	-	18.18	69.39	80	80	80

Perspective 2: **Hospital-based** Cost-sensitive Machine Learning

- Basic Logic:

TP = FP: Treatment Cost & FN = TN = 0

- Example: Diabetes

Treatment Cost = $9,601 per year

The largest components of medical expenditures are:

- hospital inpatient care (30% of the total medical cost),
- prescription medications to treat complications of diabetes (30%),
- anti-diabetic agents and diabetes supplies (15%), and
- physician office visits (13%)

		Predicted class	
		+	-
Actual class	+	**TP** True Positives	**FN** False Negatives Type II error
	-	**FP** False Positives Type I error	**TN** True Negatives

Perspective 3: Patient-based Cost-sensitive Machine Learning (Recommended)

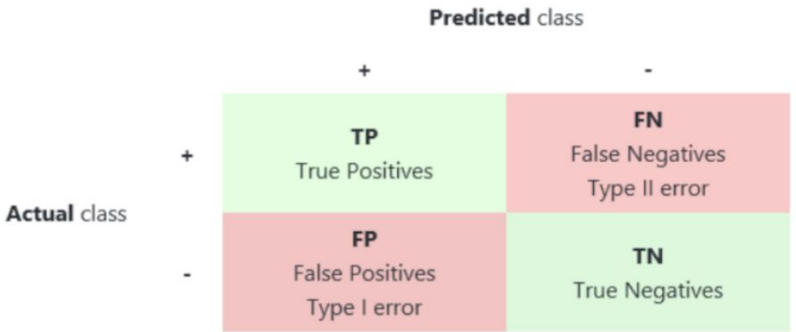
- H_1: For healthy people, their health would be affected negatively (-α) if treated
- H_2: For sick people, their health condition could be improved or remained the same after treatment (no immediate death), - β
- H_3: For sick people predicted healthy, the impact of health status - γ
- TN = 0, FP = - α (α>0), TP = - β (β>0), FN = - γ (γ >0 and γ is the largest among α, β, γ)

α = n * γ (0<n<1)

β = m * γ (m>0)

Therefore: β = (m/n) * α

1) Situation 1: β < α: (m/n)>1
2) Situation 2: β > α: (m/n)<1
3) Situation 3: β = α: (m/n)=1



Calculate Patient-based Cost Matrix Using DALY

Disability-Adjusted Life Year (DALY):

- Quantify the Burden of Disease from mortality and morbidity
- One DALY can be thought of as One lost year of "healthy" life
- DALYs across the population: a measurement of the gap between current health status and an ideal health situation where the entire population lives to an advanced age, free of disease and disability

DALY = YLL + YLD

YLL = N x L

where:

- N = number of deaths
- L = standard life expectancy at age of death in years

OR

YLD = P x DW

where:

- P = number of prevalent cases
- DW = disability weight

Calculate Patient-based Cost Matrix Using DALY

Reference of DALY values:

Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: A systematic analysis for the Global Burden of Disease Study 2010

GLOBAL DISABILITY-ADJUSTED LIFE YEARS, 291 CAUSES, ALL AGES, BOTH SEXES, 1990 TO 2010, REPORTED IN THOUSANDS, PER 100,000, AND PERCENT CHANGE

	All ages DALYs (thousands)			DALYs (per 100 000)		
	1990	2010	% change	1990	2010	% change
All causes	2,502,601 (2,389,053–2,639,606)	2,490,385 (2,349,250–2,637,538)	-0·5	47,205 (45,063–49,789)	36,145 (34,097–38,281)	-23·4
Communicable, maternal, neonatal, and nutritional disorders	1,181,610 (1,113,122–1,268,900)	868,024 (818,934–921,489)	-26·5	22,288 (20,996–23,934)	12,598 (11,886–13,374)	-43·5
HIV/AIDS and tuberculosis	79,368 (72,264–90,448)	130,944 (119,310–141,121)	65·0	1,497 (1,363–1,706)	1,900 (1,732–2,048)	26·9
Tuberculosis	61,250 (55,443–71,077)	49,396 (40,065–56,071)	-19·4	1,155 (1,046–1,341)	717 (581–814)	-37·9
HIV/AIDS	18,117 (15,012–22,260)	81,547 (75,003–88,367)	350·1	342 (283–420)	1,184 (1,089–1,283)	246·3

DALY for CKD: 21,151 (18,147–23,223)

	Predicted (sick)	**Predicted (healthy)**
Actual (sick)	TP = 18147 (treated)	FN = 23223 (untreated)
Actual (healthy)	FP < FN (Grid search for best FP)	TN = 0

Cost-sensitive Machine Learning Algorithm: Metacost

Procedure MetaCost (S, L, C, m, n, p, q)

For $i = 1$ to m
  Let S_i be a resample of S with n examples.
  Let M_i = Model produced by applying L to S_i.

For each example x in S
  For each class j
    Let $P(j|x) = \frac{1}{\sum_i 1} \sum_i P(j|x, M_i)$
    Where
      If p then $P(j|x, M_i)$ is produced by M_i
      Else $P(j|x, M_i) = 1$ for the class predicted by M_i for x, and 0 for all others.
      If q then i ranges over all M_i
      Else i ranges over all M_i such that $x \notin S_i$.
  Let x's class = argmin$_i \sum_j P(j|x)C(i, j)$.

Let M = Model produced by applying L to S.

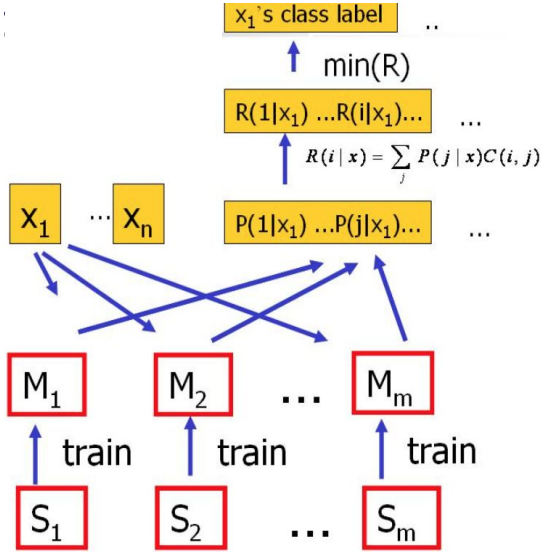
Return M.

lassifiers cost-sensitive, it
ox, requiring no
t and any kinds of
ro Domingos (1999).

itive classifier of the base
ation and interpretable
able).

del for Myocardial

32/#B37



Metacost - WEKA

WEKA: Open source, Java based software, a popular machine learning workbench with a development life of nearly two decades developed by University of Waikato, New Zealand.

Cost Matrix - CKD:

	Predicted (sick)	**Predicted (healthy)**
Actual (sick)	TP = 18147 (treated)	FN = 23223 (untreated)
Actual (healthy)	FP < FN (Grid search for best FP)	TN = 0

Metacost +

Underlying Classifier

J48 (1)
Multi-layer perceptron (2)
SMO (3)
Logistic (4)
Naive Bayes (5)
Random Forest (6)

Metacost - Model Comparison

Cross Validation: 10 folds, FP = 9000

Accuracy:

Dataset	(1) meta.Met	(2) meta.	(3) meta.	(4) meta.	(5) meta.	(6) meta.
'table_nephritis_age-weka(100)	77.36	74.86 *	78.93	77.18	75.29 *	72.58 *
	(v/ /*)	(0/0/1)	(0/1/0)	(0/1/0)	(0/0/1)	(0/0/1)

a	b	<-- class	a	b	a	b	a	b	a	b	a	b
1267	119	a = 0	1221	165	1386	0	1263	123	1209	177	1186	200
282	88	b = 1	286	84	370	0	279	91	252	118	284	86

Cost: J48 (avg cost = 16112.6344) has the lowest cost.

Chose J48 to find best value of FP (range 1-23223).

J48 (1)
Multi-layer perceptron (2)
SMO (3)
Logistic (4)
Naive Bayes (5)
Random Forest (6)

Metacost - FP value

Cross Validation: 10 folds, FP ∈ [1, 23223], let FP be integers to find a pattern

Accuracy:

Dataset	(1) meta.Met	(2) meta.	(3) meta.	(4) meta.	(5) meta.	(6) meta.
'table_nephritis_age-weka(100)	78.84	78.19	77.36	74.76 *	71.99 *	68.47 *
(v/ /*)		(0/1/0)	(0/1/0)	(0/0/1)	(0/0/1)	(0/0/1)

Findings:

FP value increase:

accuracy drop, cost increase

Solution:

choose smaller and sensible FP

```
Correctly Classified Instances        1355               77.164  %
Incorrectly Classified Instances       401               22.836  %
Kappa statistic                          0.1812
Total Cost                        28293786
Average Cost                         16112.6344
Mean absolute error                      0.2454
Root mean squared error                  0.4535
Relative absolute error                 73.723  %
Root relative squared error            111.2041 %
Total Number of Instances             1756

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
               0.914    0.762    0.818      0.914    0.863      0.192    0.605     0.820
               0.238    0.086    0.425      0.238    0.305      0.192    0.605     0.303
Weighted Avg.  0.772    0.620    0.735      0.772    0.746      0.192    0.605     0.711

=== Confusion Matrix ===

    a    b   <-- classified as
 1267  119 |    a = 0
  282   88 |    b = 1
```

Cost-sensitive Learning Challenges

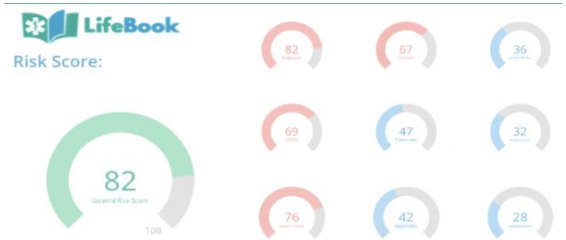
- **Insufficient information:** We assume the two ends of the DW (or DALY) ranges as the cost of treated and untreated, which is an approximate approach, note only in Global Burden of Disease(GBD) 1990 the cost of treated and untreated for certain diseases are given.
- **Subjective, still need manually predefine the cost of FP:** When costs of FN and TP are given, to minimize cost, the cost of FP should be set as a small but sensible and interpretable number.
- **Hard to consider comorbidities:** For instance, diabetes has lots of comorbidities and they all have different DW, we need more predictors includes those comorbidities and a larger database for filtering certain conditions.
- **Cannot visualize risk score for each patient:** SHAP doesn't have package for Metacost, also WEKA doesn't show feature importance.

Brandeis | INTERNATIONAL BUSINESS SCHOOL

General Risk Score

General Risk Score

- We want to show a **general risk score** on the dashboard for each individual instead of 10 for each chronic disease.



- Formula:

  General risk score =

  Risk score for each chronic disease * Corresponding weight

  Run model for each disease

  Check commonly used risk score system (CCI or EVCI)

Commonly used risk score system

Risk Score Systems			
For critically ill patients	Charlson Comorbidity Index (CCI)	CCI assigns a score to various chronic medical conditions and uses the sum to predict long-term mortality, e.g. predict 10-year survival in patients with multiple comorbidities.	In contrast to scoring modalities such as the Acute Physiology and Chronic Health Evaluation (APACHE) and Sequential Organ Failure Assessment (SOFA), the DCCI and EVCI can be calculated at the time of intensive care unit (ICU) admission and do not require the interpretation of laboratory and bedside clinical data. Thus, they can be easily derived from administrative databases and their use in the critical care literature has been increasing. Both the DCCI and EVCI have been widely used to predict survival in acute hospital settings.8–11
	Elixhauser Comorbidity Index (EVCI)	EVCI score is based on 30 acute and chronic comorbidities to predict in-hospital mortality.	
	Acute Physiology and Chronic Health Evaluation (APACHE)	The Acute Physiology and Chronic Health Evaluation (APACHE II) is a severity score and mortality estimation tool developed from a large sample of ICU patients in the United States	
	Sequential Organ Failure Assessment (SOFA)	The Sequential Organ Failure Assessment (SOFA) is a morbidity severity score and mortality estimation tool developed from a large sample of ICU patients throughout the world. Unlike other scoring systems, such as the SAPS II and APACHE II systems, the SOFA was designed to focus on organ dysfunction and morbidity, with less of an emphasis on mortality prediction. The authors designed the system with an emphasis on bedside applicability and simplicity using widely available variables.	
Insurance risk score	HHS-HCC	The Department of Health and Human Services Hierarchical Condition Category (HHS-HCC) diagnostic classification is the foundation of the HHS-operated risk adjustment program for the individual and small group markets under section 1343 of the Patient Protection and Affordable Care Act (PPACA). The HHS risk adjustment model uses patient diagnoses and demographic information, in addition to enrollment duration and a limited number of drugs for adults, to predict plan liability for medical and drug spending.	

We recommend **CCI or EVCI**

Charlson Comorbidity Index (CCI)

FORMULA

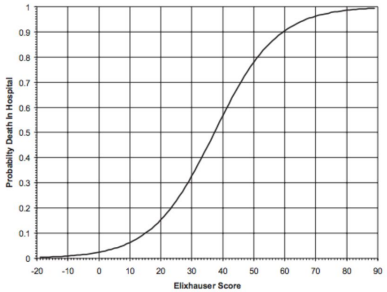
Addition of the selected points:

Variable	Definition	Points
Myocardial infarction	History of definite or probable MI (EKG changes and/or enzyme changes)	1
Congestive heart failure	Exertional or paroxysmal nocturnal dyspnea and has responded to digitalis, diuretics, or afterload reducing agents	1
Peripheral vascular disease	Intermittent claudication or past bypass for chronic arterial insufficiency, history of gangrene or acute arterial insufficiency, or untreated thoracic or abdominal aneurysm (≥6 cm)	1
Cerebrovascular accident or transient ischemic attack	History of a cerebrovascular accident with minor or no residua and transient ischemic attacks	1
Dementia	Chronic cognitive deficit	1
Chronic obstructive pulmonary disease	-	1
Connective tissue disease	-	1
Peptic ulcer disease	Any history of treatment for ulcer disease or history of ulcer bleeding	1
Mild liver disease	Mild = chronic hepatitis (or cirrhosis without portal hypertension)	1
Uncomplicated diabetes	-	1
Hemiplegia	-	2
Moderate to severe chronic kidney disease	Severe = on dialysis, status post kidney transplant, uremia, moderate = creatinine >3 mg/dL (0.27 mmol/L)	2
Diabetes with end-organ damage	-	2
Localized solid tumor	-	2
Leukemia	-	2
Lymphoma	-	2
Moderate to severe liver disease	Severe = cirrhosis and portal hypertension with variceal bleeding history, moderate = cirrhosis and portal hypertension but no variceal bleeding history	3
Metastatic solid tumor	-	6

https://www.mdcalc.com/charlson-comorbidity-index-cci#evidence

Elixhauser Comorbidity Index (EVCI)

Of the 30 original comorbidities, 21 were found to be associated with in-hospital death from a study population that included all admissions from The Ottawa Hospital, Canada between 1996 and 2008. Points were assigned to each of the 21 individual diseases (varying from -7 to +12) in order to create a weighted index. The newly created Elixhauser Comorbidity Index is obtained via the summation of points from each disease and the range of possible scores is from -19 (lesser disease burden) to +89 (greater disease burden). The association between the Elixhauser Comorbidity Index score and the probablity of inpatient death can be found using the below:



Source: van Walraven, Carl, et al.

The Elixhauser Comorbidity Measures and Elixhauser Comorbidity Index have also been associated with adverse outcomes and adverse hospital metrics after various orthopaedic conditions and procedures.

Elixhauser Comorbidity Index (EVCI)

Elixhauser Comorbidity Index Summary

Congestive heart failure Diagnosis of congestive heart failure	(7 points)
Cardiac arrhythmias Diagnosis of cardiac arrhythmias	(5 points)
Valvular disease Diagnosis of valvular disease	(-1 points)
Pulmonary circulation disorders Diagnosis of pulmonary circulation disorders	(4 points)
Peripheral vascular disorders Diagnosis of peripheral vascular disorders	(2 points)
Hypertension Diagnosis of hypertension	(0 points)
Paralysis Diagnosis of paralysis	(7 points)
Neurodegenerative disorders Diagnosis of neurodegenerative disorders	(6 points)
Chronic pulmonary disease Diagnosis of chronic pulmonary disease	(3 points)
Diabetes Diagnosis of diabetes with chronic complications	(0 points)
Hypothyroidism Diagnosis of hypothyroidism	(0 points)
Renal failure Diagnosis of renal failure	(5 points)
Liver disease Diagnosis of liver disease	(11 points)
Peptic ulcer disease, no bleeding Diagnosis of peptic ulcer disease, no bleeding	(0 points)
AIDS/HIV Diagnosis of AIDS/HIV	(0 points)
Lymphoma Diagnosis of lymphoma	(9 points)
Metastatic cancer Diagnosis of metastatic cancer	(12 points)
Solid tumor without metastasis Diagnosis of solid tumor without metastasis	(4 points)
Rheumatoid arthritis/collagen vascular diseases Diagnosis of rheumatoid arthritis/collagen vascular diseases	(0 points)
Coagulopathy Diagnosis of coagulopathy	(3 points)
Obesity Diagnosis of obesity	(-4 points)
Weight loss Diagnosis of weight loss	(6 points)
Fluid and electrolyte disorders Diagnosis of fluid and electrolyte disorders	(5 points)
Blood loss anemia Diagnosis of blood loss anemia	(-2 points)
Deficiency anemia Diagnosis of deficiency anemia	(-2 points)
Alcohol abuse Diagnosis of alcohol abuse	(0 points)
Drug abuse Diagnosis of drug abuse	(-7 points)
Psychosis Diagnosis of psychosis	(0 points)
Depressions Diagnosis of depression	(-3 points)

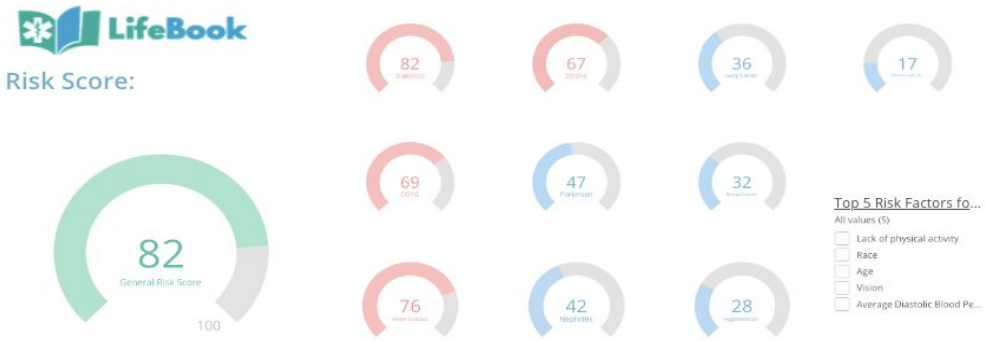
Pertinent Negative | Pertinent Positive | Pertinent Positive

van Walraven Elixhauser Comorbidity Index (range -19 to 89):
70 points or 17.6 percent.

Graphical van Walraven Elixhauser Comorbidity Index Score:

https://www.orthotoolkit.com/elixhauser-comorbidity-index/


Brandeis
INTERNATIONAL
BUSINESS SCHOOL
Dashboard Design



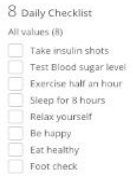
Diabetes :

Introduction:
Diabetes is a chronic (long-lasting) health condition that affects how your body turns food into energy.

Diagnosis:
There are several ways to diagnose diabetes. Each way usually needs to be repeated on a second day to diagnose diabetes.

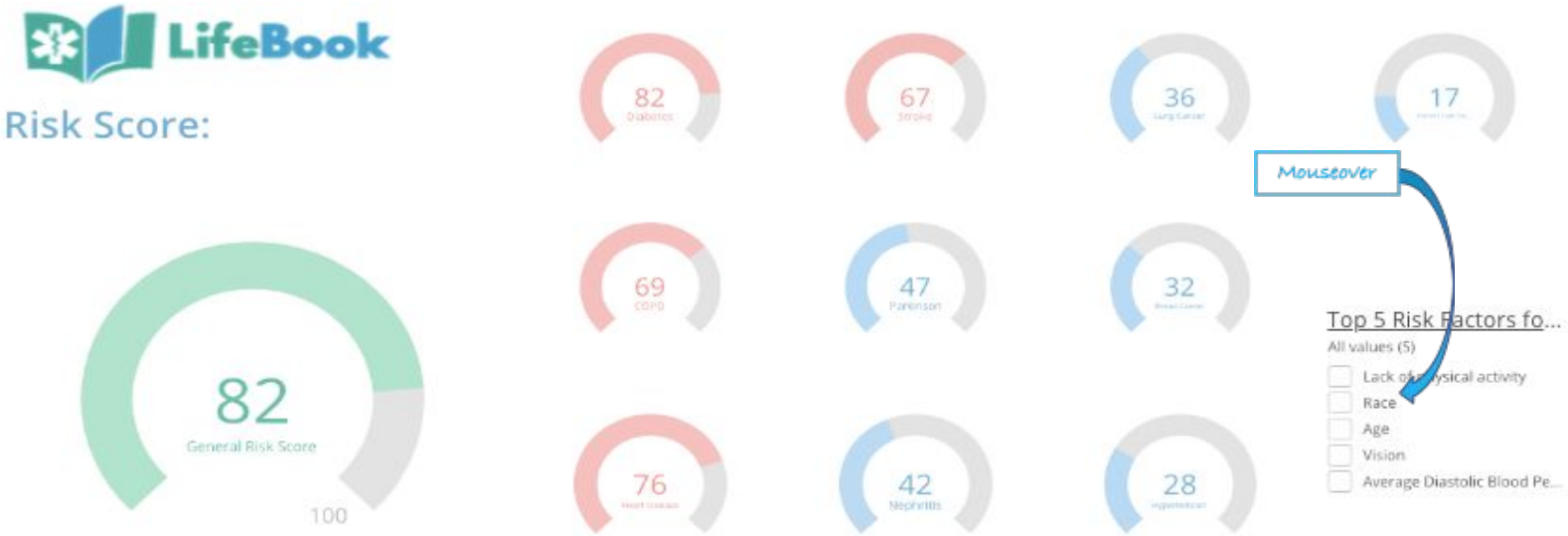
Types:
Prediabetes/Type 1 Diabetes/Type 2 Diabetes/Gestational Diabetes
Notes: You can prevent or delay type 2 diabetes with proven, achievable lifestyle changes even if you're at high risk.

Symptoms:
Urinating often/Feeling very thirsty



Type in Daily Stats:

- A1C Level: _____
- Fasting Plasma Glucose : _____
- Oral Glucose Tolerance Test: _____
- Random Plasma Glucose Test: _____


LifeBook
Risk Score:
82
General Risk Score
100
82
67
36
17
69
47
32
76
42
28
Mouseover
Top 5 Risk Factors fo...
All values (5)
Lack of physical activity
Race
Age
Vision
Average Diastolic Blood Pe...

Diabetes :

Introduction:

Diabetes is a chronic (long-lasting) health condition that affects how your body turns food into energy.

Diagnosis:

There are several ways to diagnose diabetes. Each way usually needs to be repeated on a second day to diagnose diabetes.

Types:

Prediabetes/Type 1 Diabetes/Type 2 Diabetes/Gestational Diabetes

Notes: You can prevent or delay type 2 diabetes with proven, achievable lifestyle changes even if you're at high risk.

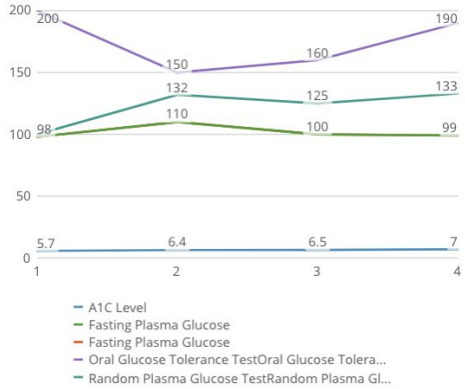
Symptoms:

Urinating often/Feeling very thirsty

3 Daily Checklist

All values (8) Show selected

- [x] Take insulin shots
- [x] Test Blood sugar level
- [x] Exercise half an hour
- [ ] Sleep for 8 hours
- [ ] Relax yourself
- [ ] Be happy
- [ ] Eat healthy
- [ ] Foot check



Type in Daily Stats:

- A1C Level: _____
- Fasting Plasma Glucose : _____
- Oral Glucose Tolerance Test: _____
- Random Plasma Glucose Test: _____

Brandeis | INTERNATIONAL BUSINESS SCHOOL

Prospect

Problems and Prospects

1. Larger real EHR dataset
2. Structured & unstructured data
3. Dataset including time-series

References

- NHANES

  https://www.cdc.gov/nchs/nhanes/index.htm

- DALY
  https://www.ncbi.nlm.nih.gov/books/NBK11802/
  https://www.who.int/healthinfo/global_burden_disease/metrics_daly/en/
  https://cevr.shinyapps.io/DALYcalculation/

- GBD

  https://www.who.int/quantifying_ehimpacts/publications/en/9241546204chap3.pdf

  https://www.who.int/healthinfo/global_burden_disease/GBD2004_DisabilityWeights.pdf

  https://docs.google.com/spreadsheets/d/1Pus5sEakO12ypVzZTEkI1UM6lvRP1j7N/edit#gid=1803487844

  https://www.who.int/healthinfo/global_burden_disease/estimates_regional_2004_2008/en/

- Cost-sensitive learning

  https://link.springer.com/content/pdf/10.1007/3-540-45164-1_42.pdf

  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5442282/#B37

  https://machinelearningmastery.com/cost-sensitive-decision-trees-for-imbalanced-classification/

  https://github.com/Albertsr/Class-Imbalance/tree/master/1.%20Cost%20Sensitive%20Learning

References

- SHAP

  https://towardsdatascience.com/interpretable-machine-learning-with-xgboost-9ec80d148d27

- CCI

  https://www.mdcalc.com/charlson-comorbidity-index-cci#evidence

- Elixhauser Comorbidity Index

  https://www.orthotoolkit.com/elixhauser-comorbidity-index/

- Accuracy validation

  https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-018-0042-5

  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6419763/

- ASCVD Risk Calculator

  http://www.cvriskcalculator.com/


Brandeis
INTERNATIONAL
BUSINESS SCHOOL
THANK YOU!