CARDIOVASCULAR DISEASE PREDICTION

Troy Zhongyi Zhang

AGENDA

- Study, Data, & Primary outcome
- Hypothesis
- Data Cleaning & Preprocessing: Predictors
- Table 1
- Exploratory Data Analysis: Visualization
- Unsupervised Learning
- Supervised Learning
- Ensemble Learning
- Extended Thinking

STUDY, DATA, & PRIMARY OUTCOME

- World Health Organization estimated 12 million deaths worldwide per year due to Heart diseases.
- **Half** of the deaths in the US and other developed countries are due to **Cardiovascular Diseases**.
- The early prognosis of cardiovascular diseases can aid in **making decisions** on lifestyle changes in high risk patients and in turn reduce the complications. This research intends to pinpoint the **most relevant/risk factors** of heart disease as well as predict the overall risk using Machine Learning models.
- Source The dataset is publicly available on the **Kaggle** website, and it is from an **ongoing cardiovascular study** on residents of the town of **Framingham, Massachusetts**. The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD)
- The dataset provides the patients' information. It includes over **4,000 records and 15 attributes**. Variables Each attribute is a potential risk factor. There are both demographic, behavioral and medical risk factors.
- TenYearCHD: 10-year coronary heart disease prediction – 0/1

```
df0.shape
```

```
(4238, 16)
```

HYPOTHESIS

- TenYearCHD is highly correlated with diabetes, glucose, and smoking.
- Education, age, and gender will not affect much to the TenYearCHD and the glucose level.
- Most of my prediction will be 0s, which means not risky for coronary disease in the next 10 years, because the glucose level for people is very low(mostly 55 - 100).
- (Normal blood sugar levels are less than 100 mg/dL after not eating for at least 8 hours; <140 mg/dL two hours after eating)
- Machine learning: Ensemble ML will give better prediction accuracy than Base ML models. – Accuracy comparison

Gender and risk

Although men tend to develop coronary artery disease earlier in life, after age 65 the risk of heart disease in women is almost the same as in men. Women have many of the same risk factors for heart disease as men, such as smoking, high blood pressure, and high cholesterol.

How Age and Gender Affect Your Heart | Kaiser Permanente ...

Over time, high blood glucose from diabetes can damage your blood vessels and the nerves that control your heart and blood vessels. The longer you have diabetes, the higher the chances that you will develop heart disease. ... In adults with diabetes, the most common causes of death are heart disease and stroke.

Diabetes, Heart Disease, and Stroke | NIDDK

https://www.niddk.nih.gov › diabetes › overview › preventing-problems › h...

Smoking raises your risk of getting CAD and dying early from CAD. Carbon monoxide, nicotine, and other substances in tobacco smoke can promote atherosclerosis and trigger symptoms of coronary artery disease. ... Clumping platelets can then block your coronary arteries and cause a heart attack. 2012年

Smoking and Coronary Artery Disease - CardioSmart

https://www.cardiosmart.org › Healthwise

CLEANING & PREPROCESSING

- number of predictors

```
df0.isnull().sum()
```

```
male                0
age                 0
education         105
currentSmoker       0
cigsPerDay         29
BPMeds             53
prevalentStroke     0
prevalentHyp        0
diabetes            0
totChol            50
sysBP               0
diaBP               0
BMI                19
heartRate           1
glucose           388
TenYearCHD          0
dtype: int64
```

- **Introduction:**
  - Systolic pressure; Diastolic blood pressure
  - Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)
  - Prevalent Hyp: whether or not the patient was hypertensive
- **Missing Values Imputation:**
  - Mean and medium to fill the null values in the cigsPerDay, BPMeds, totChol, BMI, and heartRate column.
  - KNN for education after scaling
  - Glucose: (compared RMSE without using the Education, training set = 92.5%)
    - Multiple Regression: 15.1074
    - Gradient Boosting Regressor: 17.7771
    - XGBoost Regressor: 18.3610
- Error: 15/100-150 =15/150 -> 15/100 = 10%-15%.

TABLE 1

- Outcome – TenYearCHD
- Age
- Education: 1, 2, 3, 4
- Male: 0, 1
- Blood Pressure Meds: 0, 1
- Diabetes: 0, 1
- H_0: $\mu_1=\mu_2$; no significance relationship
- H_a: some relationships
- P<0.05 -> Rejecting null hypothesis; there are relationships!

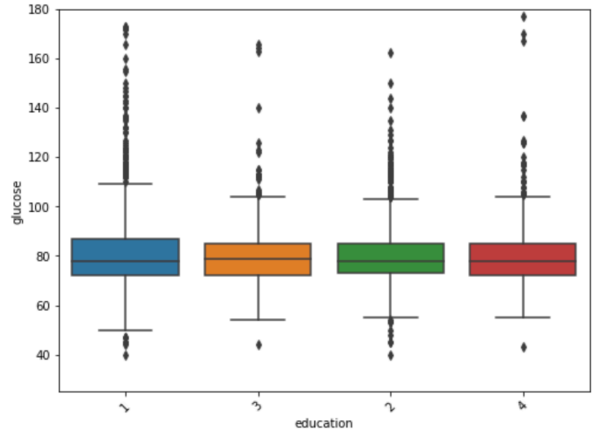
		Grouped by TenYearCHD				
		isnull	0	1	pval	ptest
variable	level					
n			3594	644		
age		0	48.8 (8.4)	54.1 (8.0)	<0.001	Two Sample T-test
education	1	0	1437 (40.0)	330 (51.2)	<0.001	Chi-squared
	2		1147 (31.9)	155 (24.1)		
	3		607 (16.9)	88 (13.7)		
	4		403 (11.2)	71 (11.0)		
male	0	0	2118 (58.9)	301 (46.7)	<0.001	Chi-squared
	1		1476 (41.1)	343 (53.3)		
BPMeds	0	0	3511 (97.7)	603 (93.6)	<0.001	Chi-squared
	1		83 (2.3)	41 (6.4)		
diabetes	0	0	3525 (98.1)	604 (93.8)	<0.001	Chi-squared
	1		69 (1.9)	40 (6.2)		

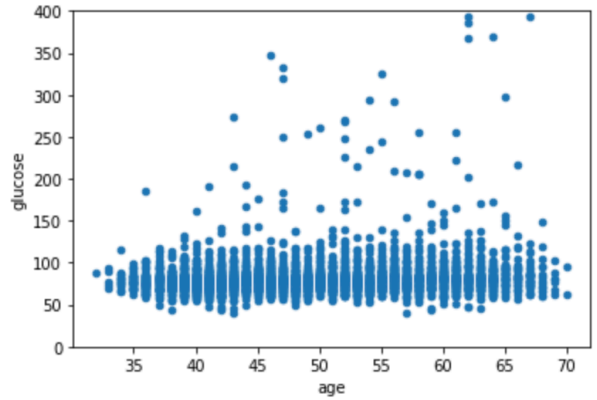
[1] Warning, Hartigan's Dip Test reports possible multimodal distributions for: age.
[2] Warning, test for normality reports non-normal distributions for: age.

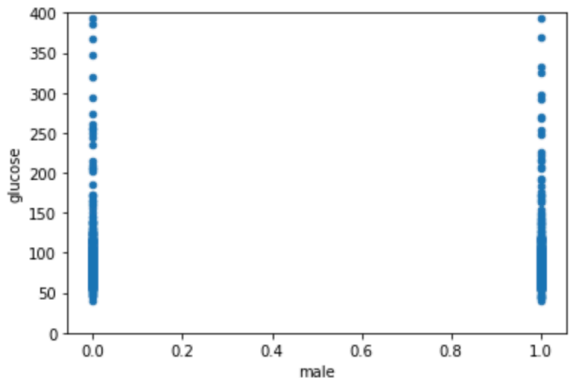
EXPLORATORY DATA ANALYSIS

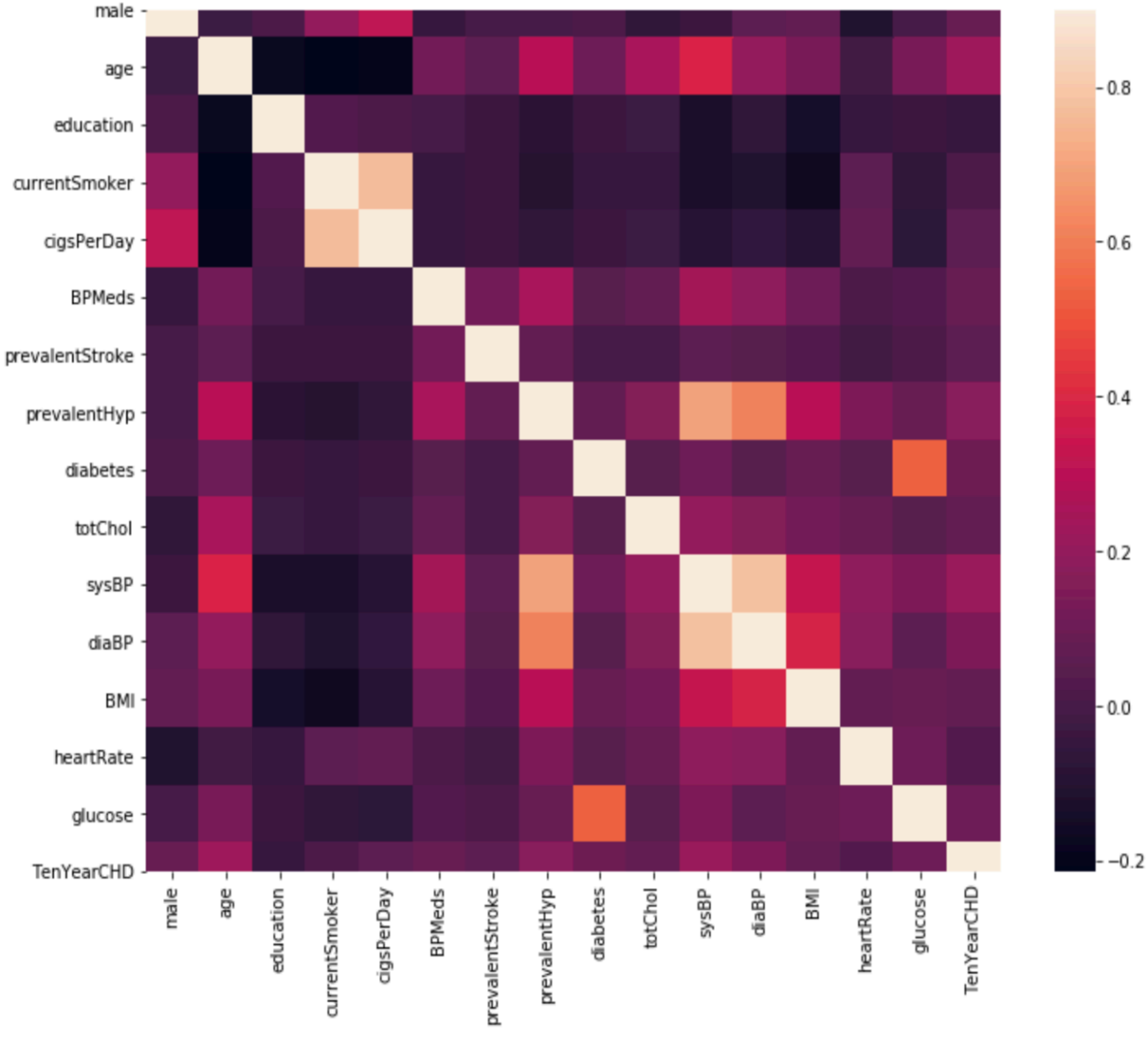
- Back to my hypothesis
- Age and education could indirectly affect **TenYearCHD** through affecting glucose.

- **Highly correlated with outcome:**
- Age
- prevalentHyp
- prevalentStroke
- Diabetes
- totChol
- sysBP
- diaBP



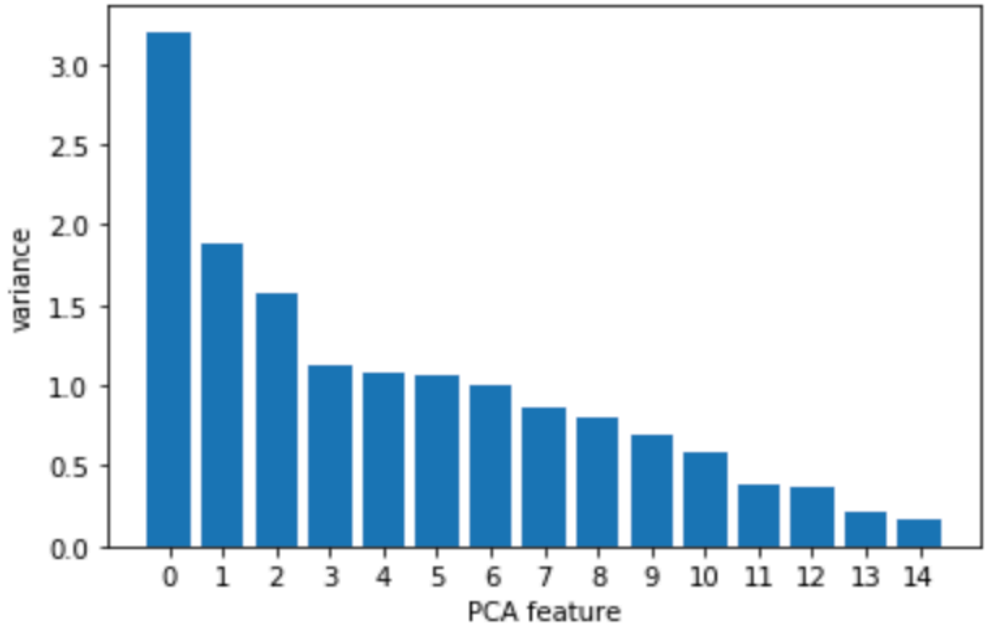


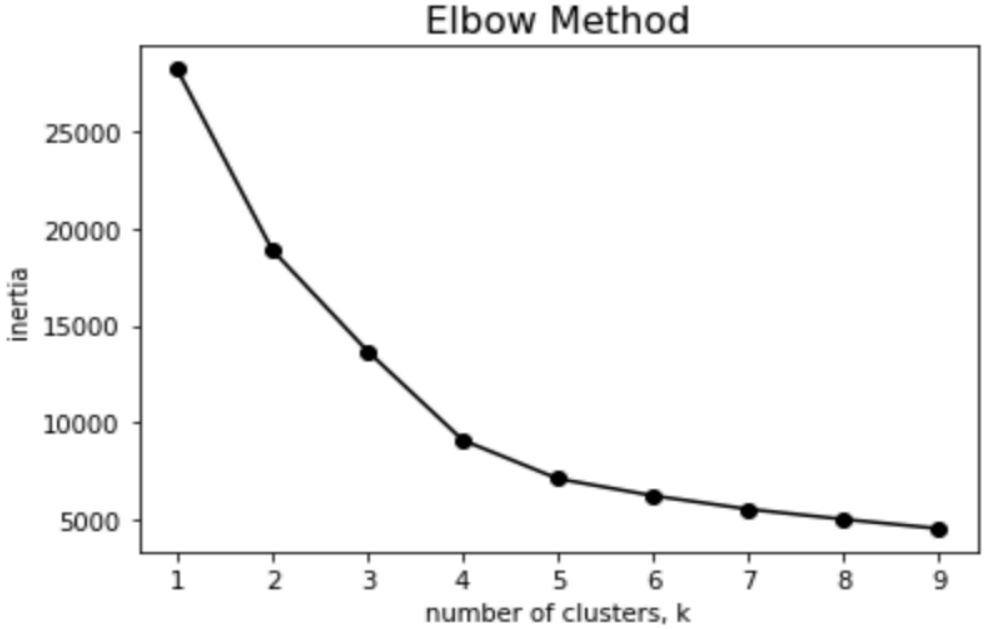




UNSUPERVISED

- Scaling dataset
- PCA
- 15 PCs in total
- PC1: 21.3%
- PC2: 12.5%
- PC3: 10.5%
- Total: 44.3%

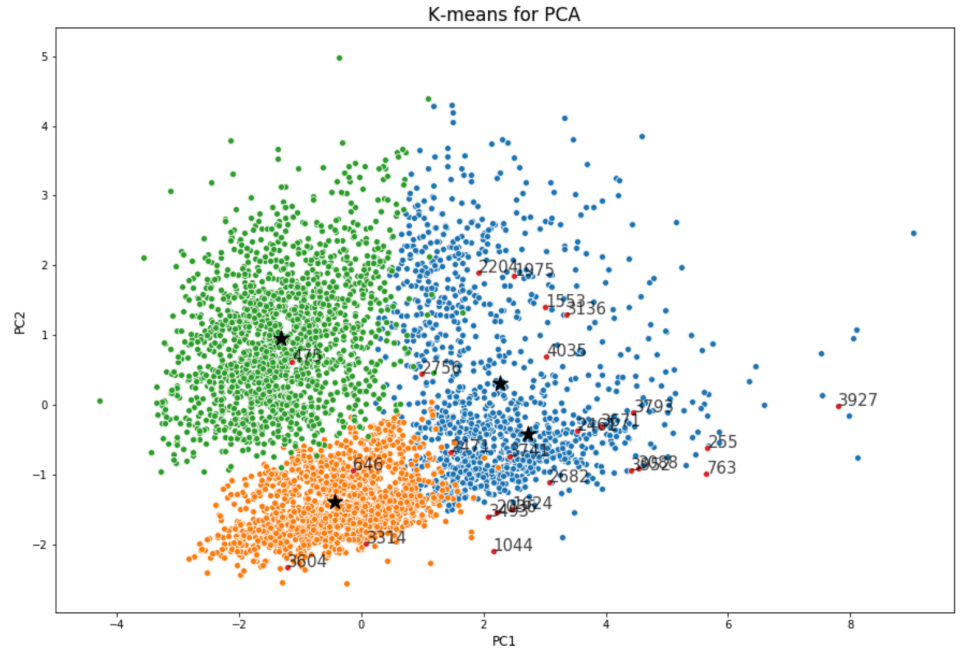




- Clusters to detect any anomalies or outliers

```
Variance:  Projected dimension
------------------------------
21.3%:    -0.05 * f1 +  0.30 * f2 + -0.11 * f3 + -0.20 * f4 + -0.17 * f5 +  0.20 * f6 +  0.07
* f7 +  0.43 * f8 +  0.14 * f9 +  0.19 * f10
12.5%:     0.35 * f1 + -0.11 * f2 + -0.02 * f3 +  0.59 * f4 +  0.63 * f5 +  0.04 * f6 + -0.02
* f7 +  0.16 * f8 + -0.02 * f9 +  0.02 * f10
10.5%:     0.06 * f1 +  0.02 * f2 + -0.03 * f3 +  0.06 * f4 +  0.05 * f5 + -0.05 * f6 + -0.02
* f7 + -0.12 * f8 +  0.69 * f9 + -0.02 * f10
 7.5%:     0.56 * f1 +  0.11 * f2 + -0.08 * f3 + -0.11 * f4 + -0.03 * f5 +  0.08 * f6 +  0.28
* f7 +  0.01 * f8 +  0.03 * f9 + -0.27 * f10
 7.1%:    -0.20 * f1 + -0.21 * f2 +  0.48 * f3 +  0.06 * f4 +  0.00 * f5 +  0.55 * f6 +  0.53
* f7 +  0.09 * f8 +  0.05 * f9 + -0.16 * f10
```

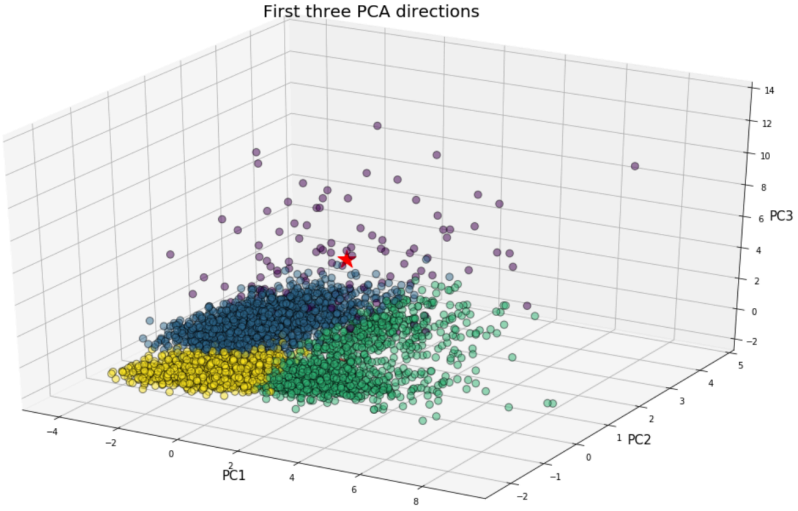
```
{0: array([   3,    5,    8, ..., 4231, 4232, 4233]), 1: array([   0,    1,    6, ..., 4221,
4226, 4237]), 2: array([   2,    4,    7, ..., 4234, 4235, 4236]), 3: array([ 255,  475,  64
6,  763, 1044, 1553, 1624, 1975, 2036, 2204, 2461,
       2471, 2682, 2756, 3088, 3136, 3314, 3493, 3604, 3671, 3741, 3793,
       3927, 3952, 4035])}

Text(0.5, 1.0, 'K-means for PCA')
```

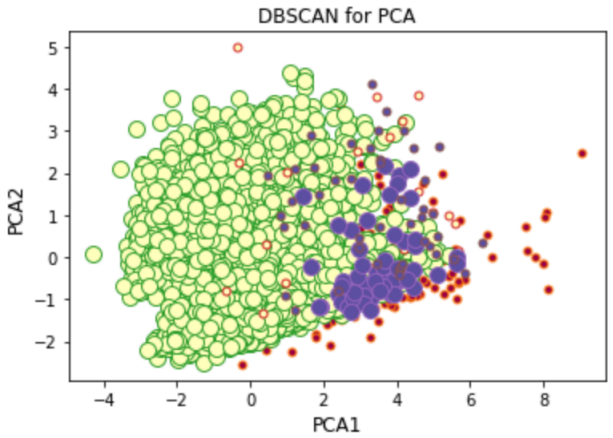



```
c = set(sus_l1) & set(sus_l2)
print(c)
```

```
{763}
```

```
{0: array([  37,   44,   56,   66,   96,  247,  249,  260,  284,  294,  311,
        357,  421,  443,  451,  471,  585,  763,  833,  903,  952, 1030,
       1068, 1111, 1123, 1165, 1238, 1268, 1303, 1333, 1340, 1363, 1485,
       1649, 1674, 1854, 1895, 1907, 1931, 1997, 2024, 2041, 2091, 2098,
       2180, 2217, 2234, 2378, 2393, 2406, 2498, 2503, 2528, 2570, 2600,
       2645, 2649, 2668, 2784, 2801, 2802, 2849, 2855, 2891, 2893, 2909,
       2926, 2961, 3002, 3051, 3112, 3140, 3203, 3242, 3256, 3300, 3321,
       3327, 3449, 3458, 3552, 3606, 3620, 3680, 3682, 3721, 3739, 3749,
       3763, 3778, 3797, 3809, 3817, 3839, 3844, 3849, 3868, 3895, 3971,
       3974, 4042, 4064, 4076, 4084, 4096, 4154, 4203, 4215, 4228]), 1: array([   2,    4,
7, ..., 4234, 4235, 4236]), 2: array([   3,    5,    8, ..., 4231, 4232, 4233]), 3: array([
0,    1,    6, ..., 4221, 4226, 4237])}
```





- The common outlier data is the patient at index number 763
- Very high glucose

totChol	sysBP	diaBP	BMI	heartRate	glucose
107–696	83.5–295	48–142.5	15.54–56.8	44–143	40–394

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose
763	0	58	1	0	0.0	0	1	1	1	267.0	157.0	94.0	33.32	92.0	205.0

Normal data! Keep it.

SUPERVISED LEARNING MODELS

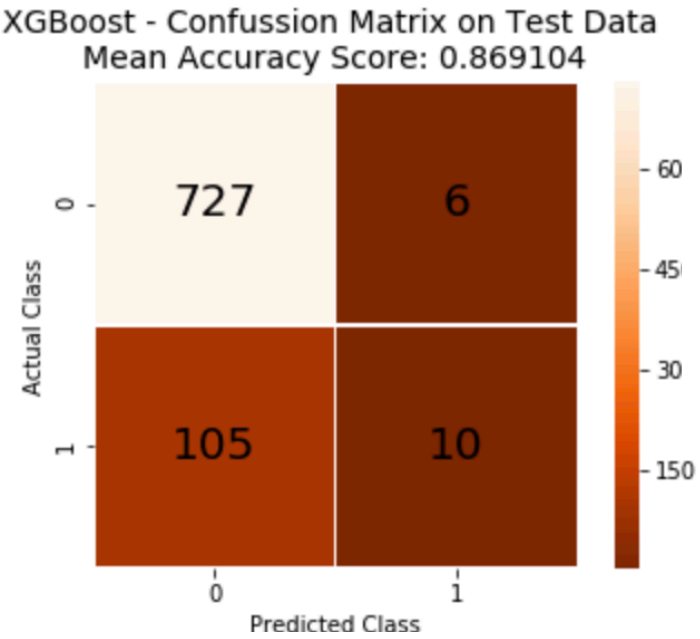
Training set = 80%; CV = 5

	Model Name	Accuracy on Testing	Accuracy on Training	AUC_ROC	DeLong 95% CI	1,000 Boostrapping 95% CI	Fitting condition
1	Logistic Regression	0.8656	0.8484	0.6629	(0.6074, 0.7184)	(0.6148, 0.7075)	Good
2	Decision Tree Classifier	0.7807	0.8484	0.5405	(0.4957, 0.5852)	(0.5016, 0.5790)	Overfitting
3	Random Forest Classifier	0.8620	0.9035	0.7125	(0.6599, 0.7651)	(0.6667, 0.7566)	A little overfitting
4	KNN with scaling	0.8514	0.8546	0.5952	(0.5387, 0.6518)	(0.5482, 0.6422)	Good
5	SVM with scaling	0.8585	0.8602	0.6050	(0.5426, 0.6674)	(0.5508, 0.6572)	Good
6	AdaBoost	0.8632	0.8584	0.7042	(0.6514, 0.7570)	(0.6594, 0.7474)	Good
7	GradientBoosting	0.8644	0.8584	0.7127	(0.6607, 0.7647)	(0.6689, 0.7554)	Good
8	**XGBoost**	**0.8691**	**0.8563**	**0.7155**	**(0.6644, 0.7667)**	**(0.6732, 0.7561)**	**Good**
9	ANN with SMOTE and scaling	0.7783	0.9834	0.5232	(0.4622, 0.5842)	(0.4737, 0.5754)	Overfitting
10	Bagging Classifier	0.8644	0.8440	0.5893	(0.5323, 0.6463)	(0.5410, 0.6354)	Good

XGBOOST

```
xgb.best_estimator_
```

```
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=1, gamma=0,
              learning_rate=0.1, max_delta_step=0, max_depth=1,
              min_child_weight=1, missing=None, n_estimators=700, n_jobs=1,
              nthread=None, objective='binary:logistic', random_state=1,
              reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
              silent=None, subsample=1, verbosity=1)
```



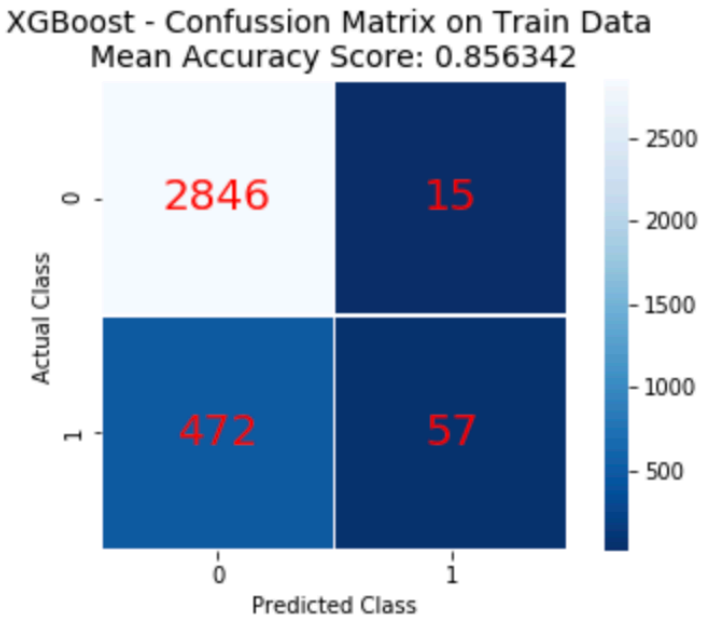
	precision	recall	f1-score	support
No risk in 10 years	0.87	0.99	0.93	733
Risky in 10 years	0.62	0.09	0.15	115
accuracy			0.87	848
macro avg	0.75	0.54	0.54	848
weighted avg	0.84	0.87	0.82	848

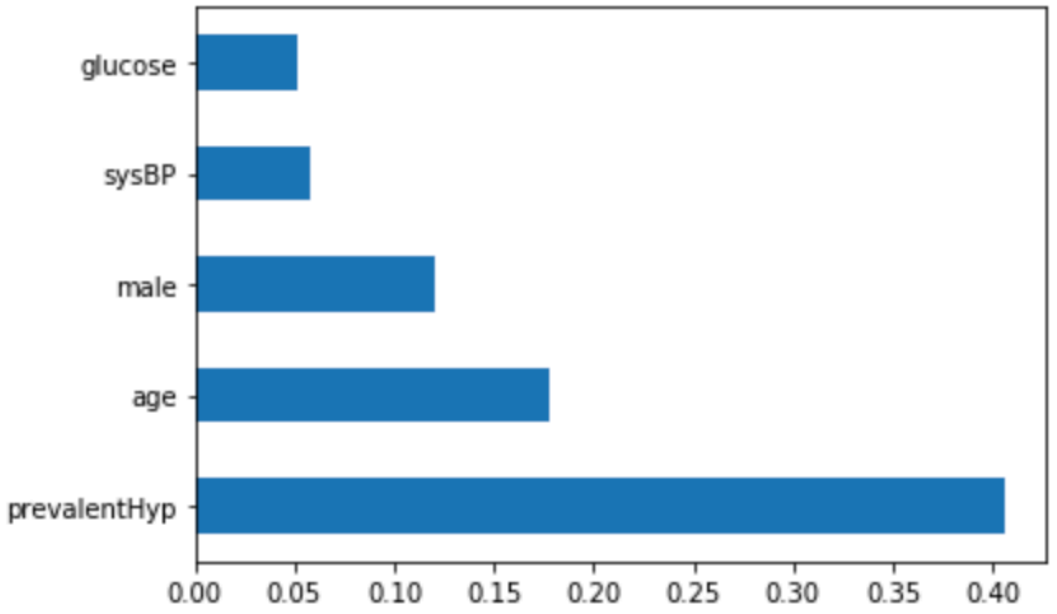
Sensitivity = TP/(TP+FN) = 10/(10+105) = 8.70%

Specificity = TN/(TN+FP) = 727/(727+6) = 99.18%

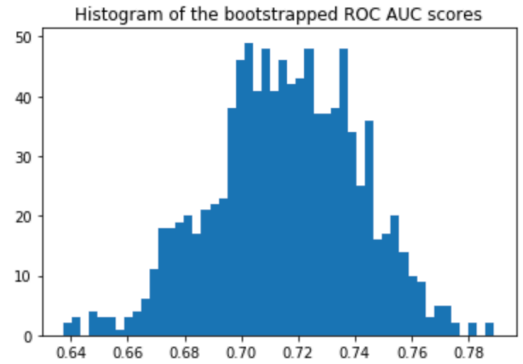
Sens: Ratio of ppl predicted to have CHD over ppl truly CHD

Spec: Ratio of ppl predicted not to have CHD over ppl truly no CHD

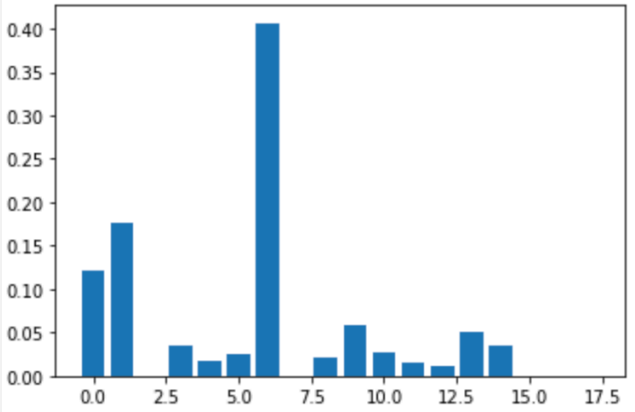





```
Original ROC area: 0.7155
```




```
Confidence interval for the score: [0.6732 - 0.7561]
```



EXTENDED THINKING

- Feature Engineering
- My dataset is a mixture of continuous and discrete. I couldn't apply Naïve Bayes Model directly on my dataset. GaussianNB – continuous; MultinomialNB – discrete. 2 ways to do:
  - I can transfer all the continuous variables into categorical variable, such as "Very Low – 0, Low – 1, Medium – 2, High – 3, Very High - 4" (20% each, (max – min)/5). Remove the previous continuous columns, and then I can fit a MultinomialNB.
  - I can break my dataset into 2 parts: 1 with only categorical, and 1 with only continuous. GaussianNB – continuous; MultinomialNB – discrete. Then transform all the dataset by taking the class assignment probabilities (with predict_proba method) as new features: np.hstack((multinomial_probas, gaussian_probas)).
    - Finally, refit a new model (e.g. a new gaussian NB) on the new features.
  - Found a package called "mixed_naive_bayes".

- **Stacking & Soft/Hard voting classifiers**
- Transform my dataset into a spectrum to apply **CNN** according to the color distribution to predict the 10-year coronary heart disease.

QUESTIONS

Thank you!