CIS 9660 – FINAL PROJECT PRESENTATION

BINARY CLASSIFICATION ON CARDIOVASCULAR DATASET

AARIF JAHAN, JACOB BAYER, JOHN MAKHIJANI, SHAWN MENG

AUGUST 11TH , 2021

BACKGROUND

- Cardiovascular disease (CVD) includes
  - heart disease
  - stroke
  - heart failure
  - cardiomyopathy
  - other issues
- Cardiovascular disease accounts for ~30% of global deaths
- Up to 90% of CVD may be preventable through lifestyle improvement
- High blood pressure accounts for 13% of cardiovascular disease

PROBLEM STATEMENT

How can we use known risk factors to build a machine learning model that identifies patients who might be more likely to have a cardiovascular disease, or already have it?

PROBLEM STATEMENT (CONT.)

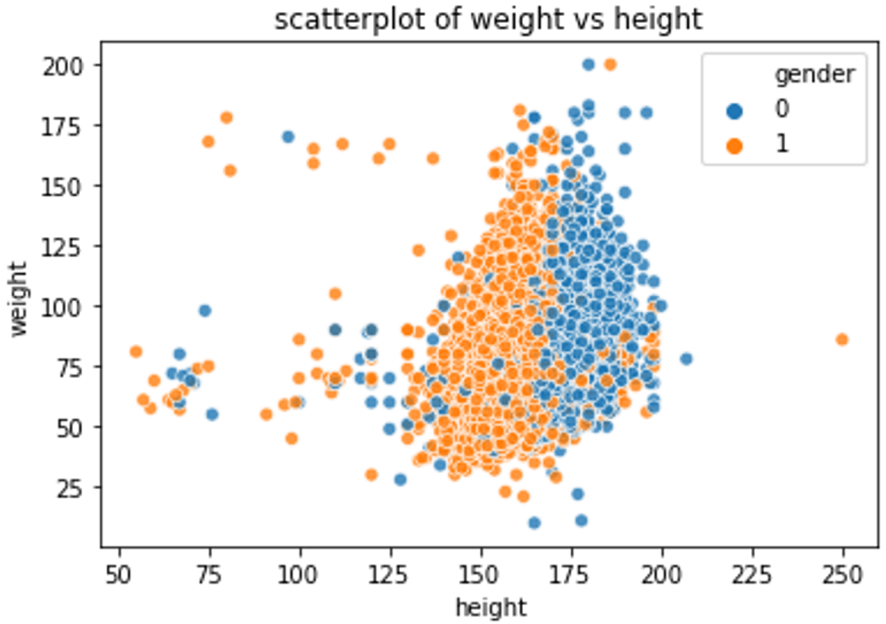
We attempt to answer this question by using several machine learning techniques to predict the presence or absence of cardiovascular disease in a labelled dataset of

- 70000 patients
- 11 predictive features, such as alcohol intake, glucose level, gender, and blood pressure
- ~50% positive labels (patients with CVD)
- ~50% negative labels (patients without CVD)
- No missing data

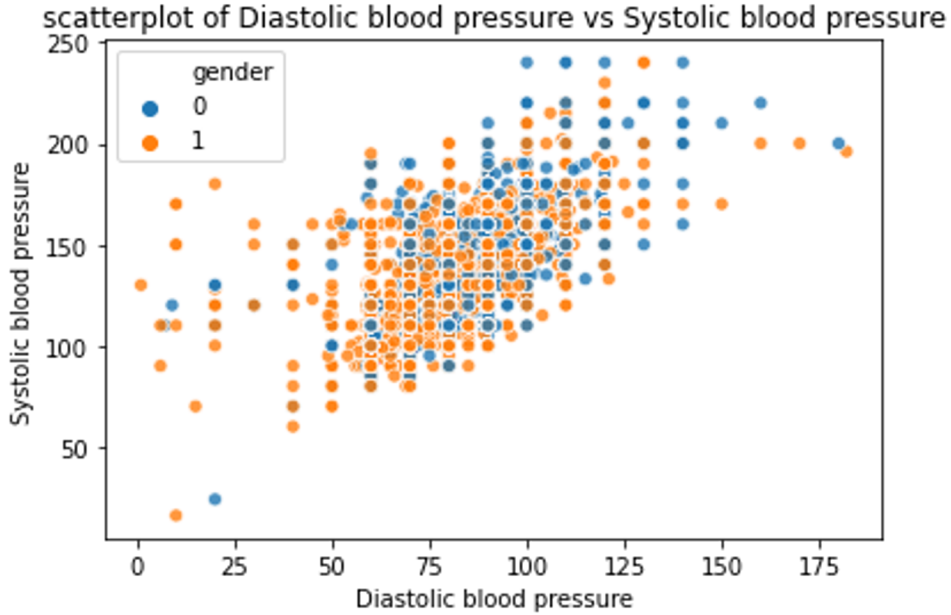
DATA CLEANING

- No missing values
- Abnormal data in height, weight, blood pressure.
- Abnormal data was prudently excluded based on subjective knowledge and the 1.5 IQR rule.
- For some predictor variables, the abnormal data might be caused by inconsistent units. But can't confirm with data source.
- Age distribution is skewed toward older population.

DATA EXPLORATION



- Several data clusters with shorter heights
- Male tends to be taller and heavier.



- Systolic blood pressure are correlated to diastolic blood pressure.
- No distinct difference between male and female.

Logistic Regression

Advantages

- Very fast predictions
- Simple to implement
- Interpretable

Disadvantages

- Requires feature engineering to eliminate feature interactions and correlated features

LOGISTIC REGRESSION PARAMETERS

```
logit_model = LogisticRegressionCV(random_state = 0,
                                   penalty = "l1",
                                   solver = "liblinear",
                                   max_iter = 250,
                                   Cs = range(90,210,10),
                                   cv = 10
                                  )
```

Using 10-fold cross validation and the L1 regularization parameter, the optimal regularization strength (C) was determined to be 120.

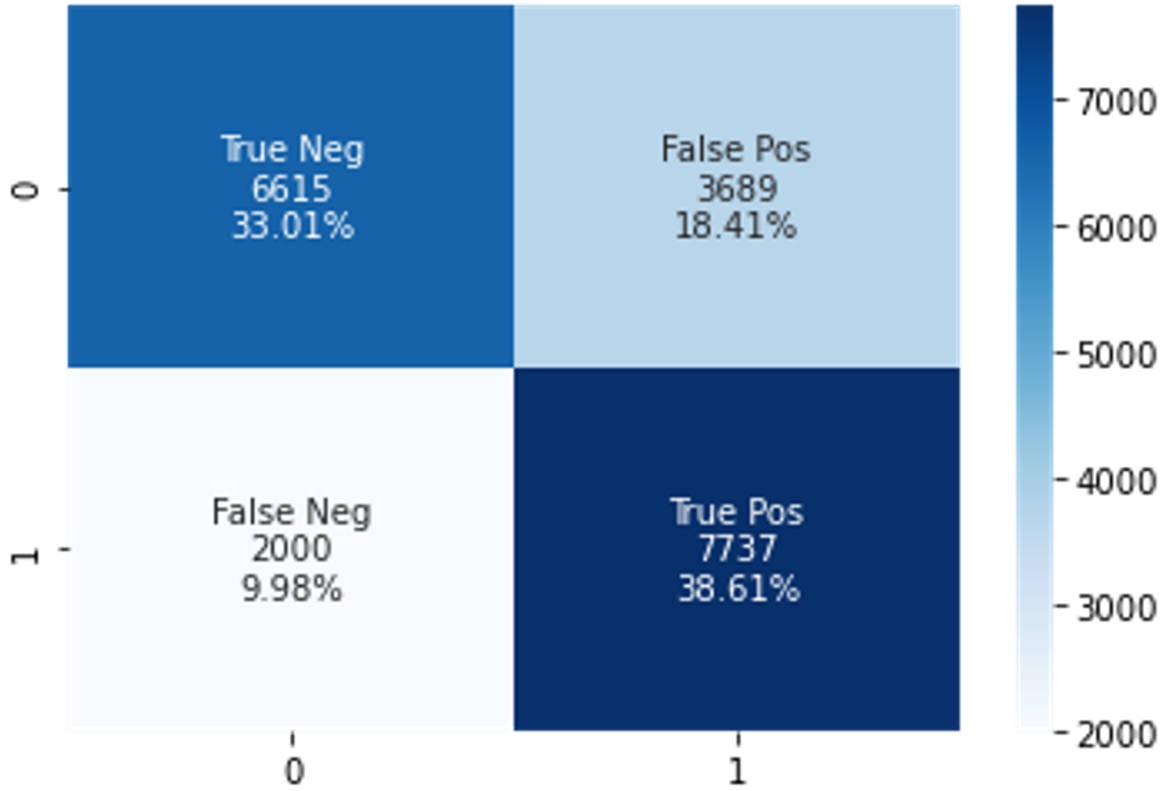
LOGISTIC REGRESSION PERFORMANCE

Evaluation Metrics (Without one-hot encoding)		
Metric	Training Data	Validation Data
Accuracy	0.726	0.729
Precision	0.751	0.753
Recall	0.661	0.660
F1	0.703	0.703
AUC	0.789	0.792
Training Time	154 seconds	

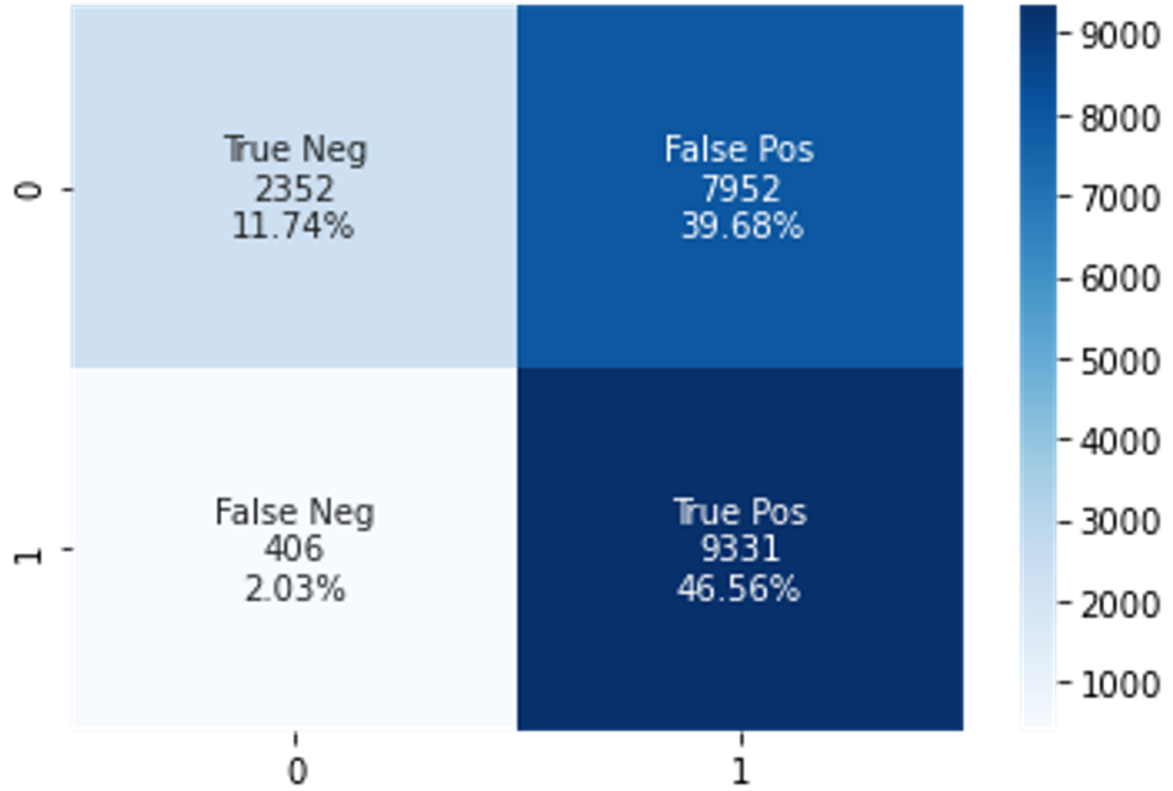
Evaluation Metrics (One-hot encoding)		
Metric	Training Data	Validation Data
Accuracy	0.727	0.730
Precision	0.754	0.755
Recall	0.659	0.658
F1	0.704	0.703
AUC	0.789	0.793
Training Time	121 seconds	

LOGISTIC REGRESSION RESULTS

Optimized probability threshold (40%)



20% probability threshold



Random Forests

Advantages

- Typically less variance, less overfitting and higher overall performance versus a single decision tree
- More stable feature importance metric

Disadvantages

- More computational time to train
- Each tree does not learn from other tree
- Hard to understand how model is making decisions

RANDOM FORESTS

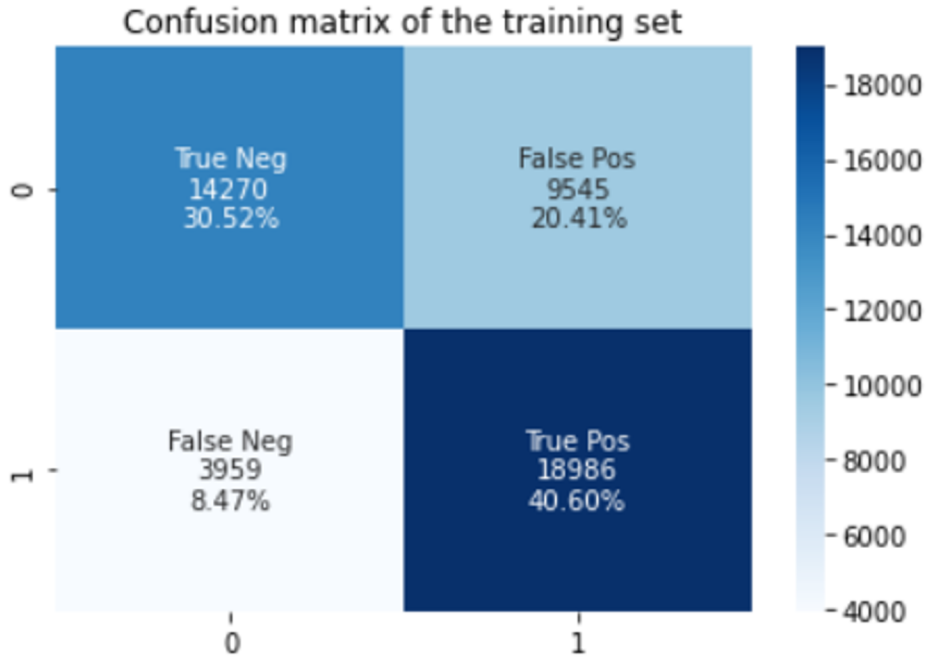
variable importance

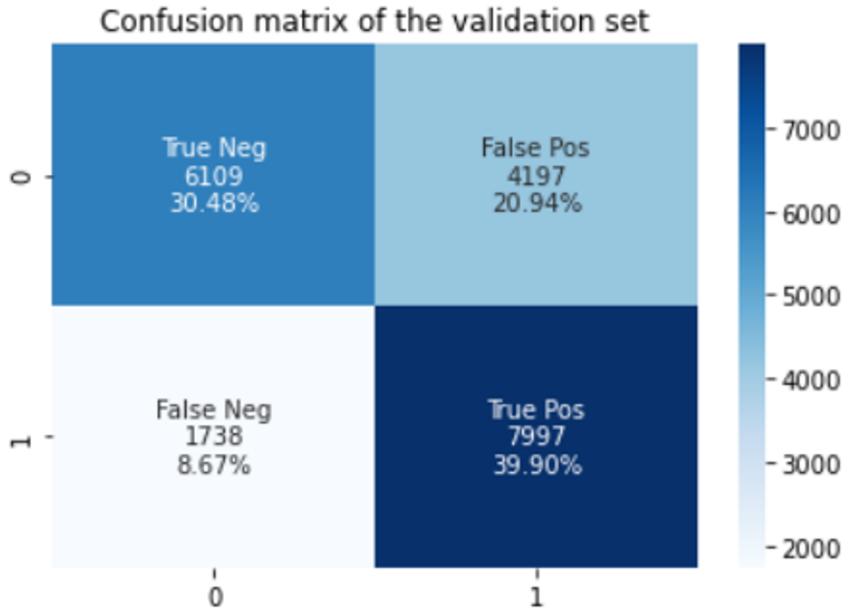
	var	imp
0	ap_hi	0.555829
1	ap_lo	0.163096
2	age_year	0.127071
3	cholesterol	0.083080
4	bmi	0.030142
5	weight	0.017646
6	height	0.009094
7	gluc	0.006009
8	active	0.004901
9	gender	0.001356
10	smoke	0.001070
11	alco	0.000706

	Training Set	Validation Set
Without hyperparameter tuning		
AUC	0.78908	0.78661
Training Time	1 second	
With hyperparameter tuning		
AUC	0.80683	0.79827
Training Time	7 minutes 47 seconds	
Best parameters	n=300, min sample split 70, min samples leaf 30, max samples 0.35, max features 6, max depth 7	

RANDOM FORESTS RESULTS

Based on best F1 score of the training set, the optimal threshold is 36%





	Precision	Recall	Accuracy	F1-Score
Results	0.66493	0.81893	0.71449	0.73393

XGBoost

Advantages

- Higher performance for structured data
- Built-in feature importance
- Suitable for parallel computing

Disadvantages

- Computational intensive
- Difficult to understand how the decision was made

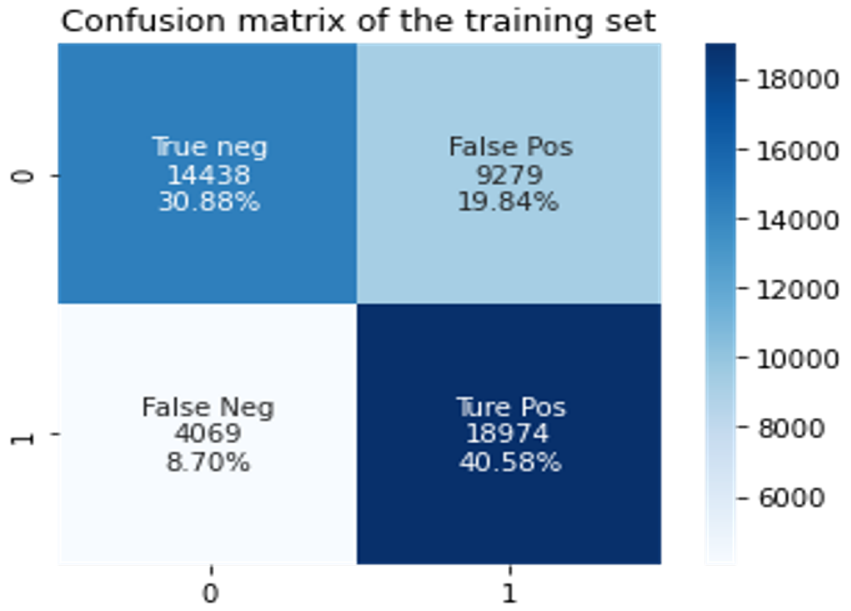
XGBOOST MODEL

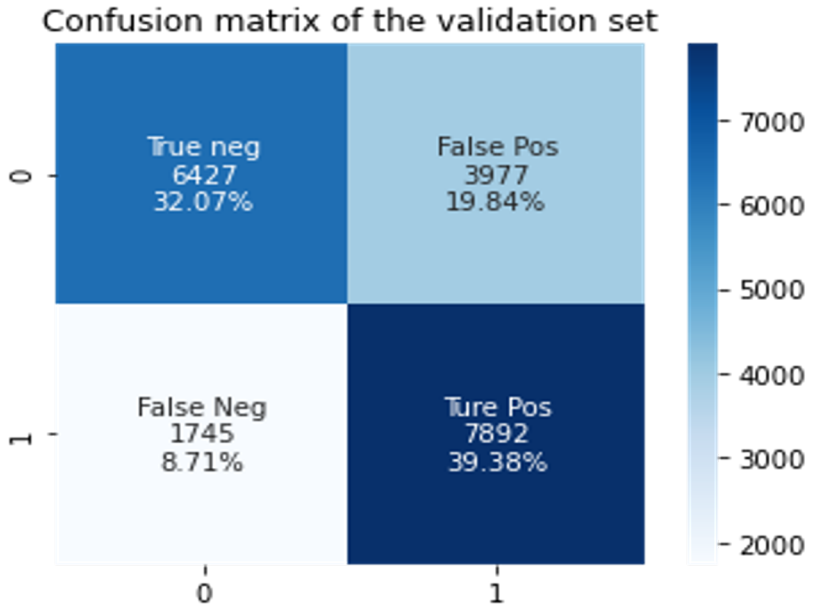
	feature	importance
0	ap_hi	0.34005
1	ap_lo	0.29973
2	cholesterol	0.12683
3	age	0.06896
4	active	0.03038
5	bmi	0.02563
6	gluc	0.02300
7	weight	0.02270
8	smoke	0.01976
9	alco	0.01699
10	gender	0.01455

XGBoost model evaluation		
Metric	Training Data	Validation Data
Accuracy	0.715	0.714
Precision	0.672	0.665
Recall	0.823	0.819
F1	0.740	0.734
AUC	0.808	0.805
Training Time	4148 seconds	

- Model was optimized for the F1-score.
- The optimal threshold of probability is 0.36
- Longer training time than other models.
- Slightly outperform other models.

XGBOOST MODEL RESULTS





	Precision	Recall	Accuracy	F1-Score
Results	0.65581	0.82147	0.703857	0.72935

K-MEANS CLUSTERING

Cannot Do

- Binary Classification Directly

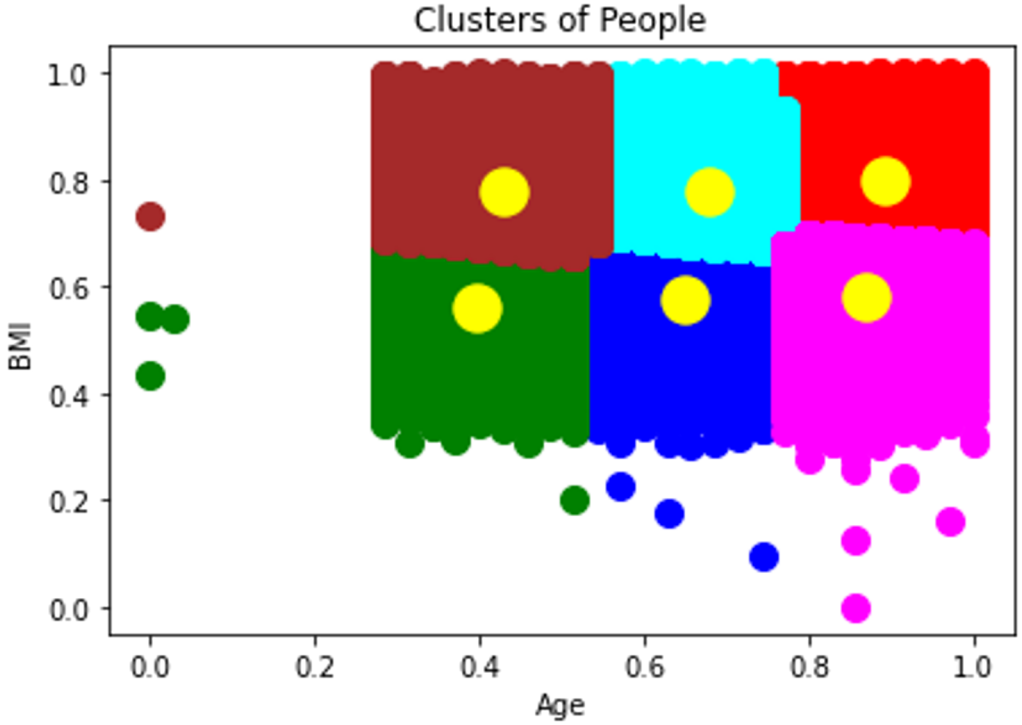
Can Do

- Data Segmentation
- Identify Feature Relationship
- Classify Weak Clusters
- Improve Data in Each Cluster

USING K-MEANS CLUSTERING

Step 1	Logistic regression model before Clustering
Step 2	Perform Data Segmentation on two important variables using K-Means Clustering
Step 3	Retrain Logistic Regression for data in each Cluster
Step 4	Evaluate individual Cluster performance and compare with performance before clustering

CLUSTERS – DATA SEGMENTATION ON AGE VS. BMI



- Cluster 0 - Red - Older People, High BMI
- Cluster 1 - Blue - Middle-Aged People, Low BMI
- Cluster 2 - Green - Younger People, Low BMI
- Cluster 3 - Cyan - Middle-Aged People, High BMI
- Cluster 4 - Magenta - Older People, Low BMI
- Cluster 5 - Brown - Younger People, High BMI

CLUSTERING RESULTS

Cluster	Has Cardio Disease?	
	No	Yes
0	2562	5775
1	9839	6689
2	7754	2971
3	4053	5316
4	7113	9370
5	2800	2559

Cluster	Cardio Disease Positivity
0	69%
1	40%
2	28%
3	57%
4	57%
5	48%

Cluster	AUC Train	AUC Val	Accuracy Train	Accuracy Val	Precision Train	Precision Val	Recall Train	Recall Val	F1 Train	F1 Val
Original	0.79	0.79	0.73	0.73	0.75	0.75	0.66	0.66	0.70	0.70
0	0.70	0.72	0.71	0.71	0.72	0.72	0.94	0.94	0.82	0.82
1	0.76	0.75	0.74	0.74	0.75	0.76	0.52	0.50	0.62	0.61
2	0.81	0.79	0.82	0.81	0.81	0.79	0.48	0.46	0.60	0.58
3	0.76	0.76	0.71	0.71	0.75	0.73	0.75	0.77	0.75	0.75
4	0.73	0.75	0.67	0.68	0.71	0.71	0.73	0.74	0.72	0.73
5	0.81	0.83	0.76	0.77	0.79	0.78	0.69	0.72	0.73	0.75

PERFORMANCE RESULTS

MODEL EVALUATION

Model	Validation AUC	Validation False Negatives (%)
Logistic Regression	0.793	9.98%
Random Forest	0.798	8.67%
XGBoost	0.805	8.71%
Logistic Regression (20% threshold)	0.793	2.03%

PERFORMANCE MONITORING

- In the first few years of model deployment, we should follow up with patients annually and label their observations with any potential test results
- This will allow us to continue to train and improve the model, as well as evaluate its performance
- If the model is misclassifying large numbers of people as false negatives, we should re-evaluate this model
- We can use K-means clustering for targeted data mining

IMPROVEMENT

- As discussed on the previous slide, more data will improve this model
- To collect more data, we will follow up with patients in order to label their observations corresponding to a test for cardiovascular disease that they took
- If the number of false negatives begins to rise significantly, the model may have to be taken out of production so that it can be re-evaluated

THANK YOU FOR WATCHING