PREDICTION OF
HEART ATTACKS

Joan Akibode
Sanaa Mouzahir
Minh Nguyen
Gloria Hwoang
Andrea Lin

FALL 2021 • MECE 4520 • PROF. CHANGYAO CHEN

Introduction

Goal: With the given dataset, explore various models that can sufficiently predict if a person is prone to heart attack or not.

Data Sample:

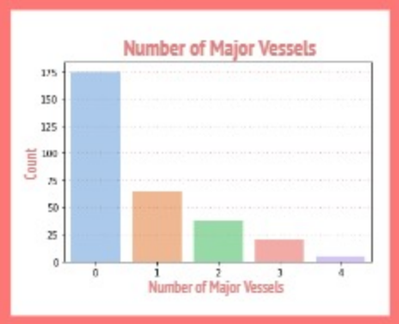
	age	sex	Chest Pain	Resting BP	chol	Fasting Blood Sug	Resting Elec Result	Max Heart Rate	Exercise	Previous Peak	slp	caa	thall	Heart Attack
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

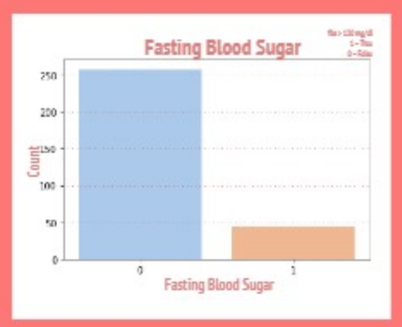
Data Dictionary

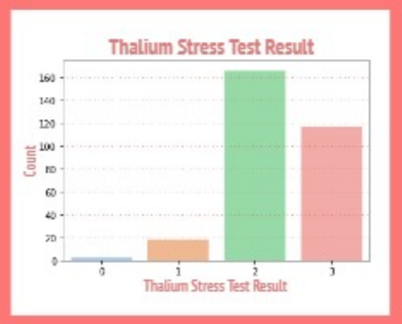
age	Age of the patient
sex	Sex of the patient
Chest Pain	Chest pain type ~ 0 = Typical Angina, 1 = Atypical Angina, 2 = Non-anginal Pain, 3 = Asymptomatic
Resting BP	Resting blood pressure (in mm Hg)
chol	Cholestoral in mg/dl fetched via BMI sensor
Fasting Blood Sug	(fasting blood sugar > 120 mg/dl) ~ 1 = True, 0 = False
Resting Elec Result	Resting electrocardiographic results ~ 0 = Normal, 1 = ST-T wave normality, 2 = Left ventricular hypertrophy
Max Heart Rate	Maximum heart rate achieved
Previous Peak	Previous peak
slp	Slope
caa	Number of major vessels
thall	Thalium Stress Test result ~ (0,3)
Exercise	Exercise induced angina ~ 1 = Yes, 0 = No
output	Target variable (Heart Attack Classification)

Exploratory Data Analysis

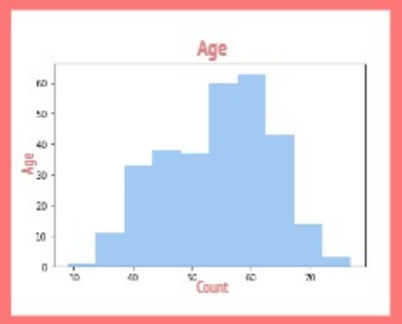
Categorical Variables :

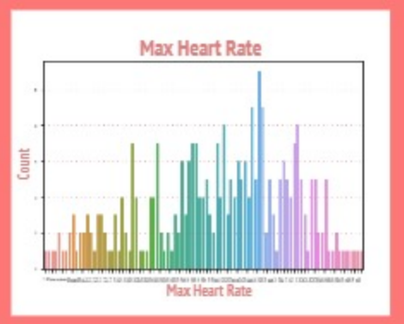


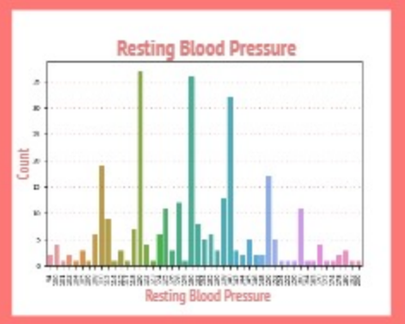


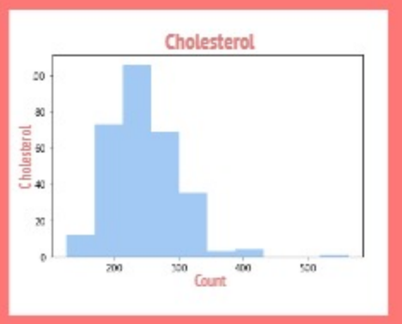


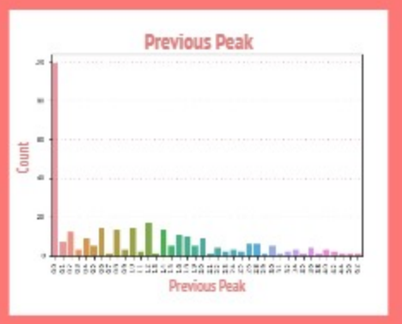
Continuous Variables :



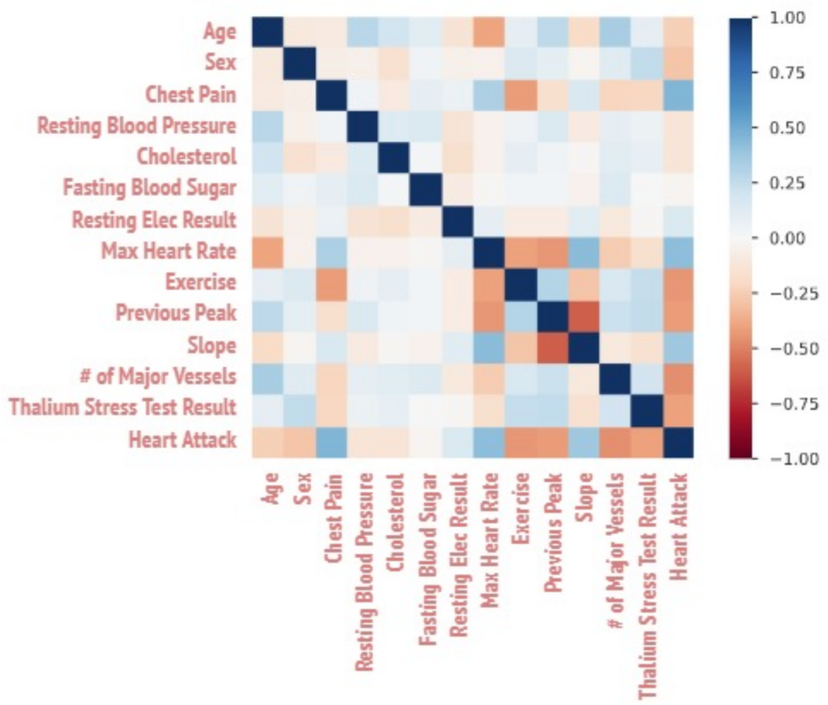








Correlations



Preprocessing

Removed Features

Exercise
Sex
Slope
Resting Electrocardiographic Results

Preprocessing

Selected Features

Age
Chest Pain
Resting Blood Pressure
Cholesterol
Fasting Blood Sugar
Max Heart Rate
Previous Peak
Number of Major Vessels
Thalium Stress Test Result

choosing features

Why did we choose the specific features?

When multiple features were correlated between each others, we eliminated the features that were the least correlated with "Heart Attack".
If they were equally correlated to "Heart Attack", we chose the the feature that was the most telling.

Example:

- sex and thall are highly correlated : we kept thall because it is more correlated to heart attack than sex
- slp and peak are highly correlated : we kept peak because it is the most telling feature

Conclusions from EDA

No features can singlehandly explain Heart attacks

The features the most correlated to heart attacks are Chest pain, Max heart rate, exercise (inversly correlated to Max heart rate) and Thalium Stress Test Result

The importance of some features were expected, such as Chest pain.

In contrary to our expectations, age is very slighly correlated to Heart Attack

Method of Analysis

1 Logistic Regression

Why?

- Logistic regression is used in classification problems.
- Binary output
- Clear threshold

Hyper Parameters

- Penalty
- Max-Number of Iterations
- # of Bootstrap Samples

2 SVM

Why?

- Binary output
- Data is separable.
- Works well on low dimensional data.

Hyper Parameters

- Linear Kernel
- Regularization Term "C"

3 Random Forest

Why?

- Decreases variance
- Increases robustness
- Easy to determine feature importance

Hyper Parameters

- Max Depth
- Number of Trees
- Number of Features to Split

Results

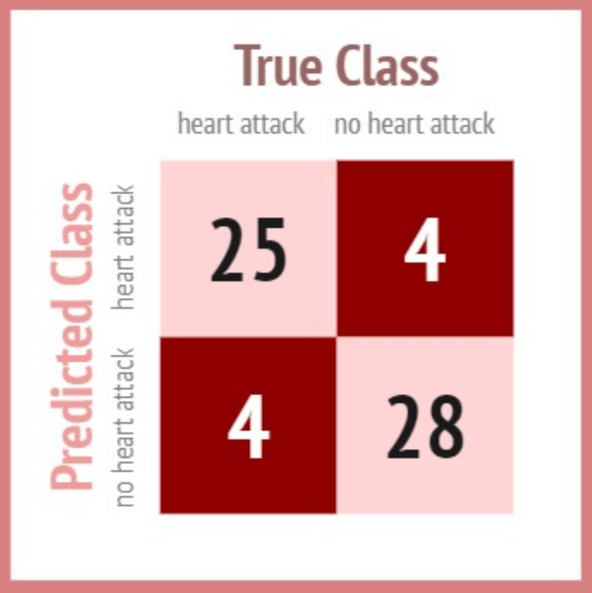
TP	FP
FN	TN

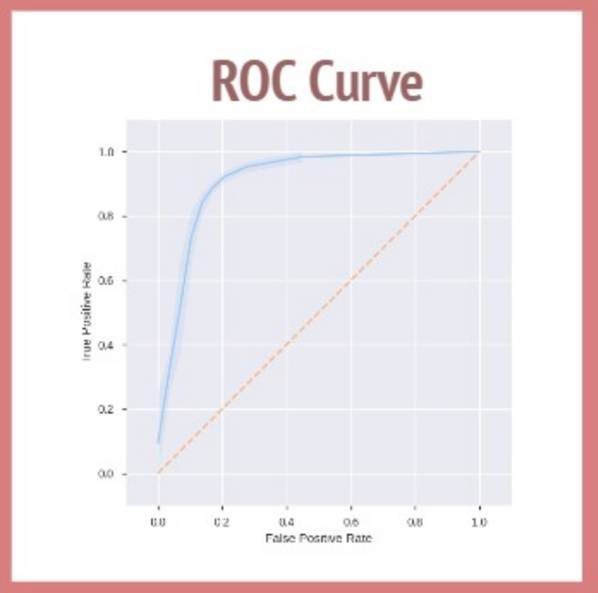
Logistic Regression Model

1

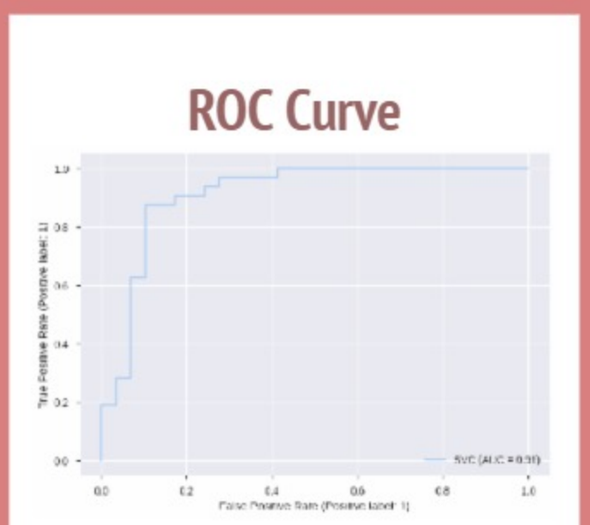
Accuracy Score of :

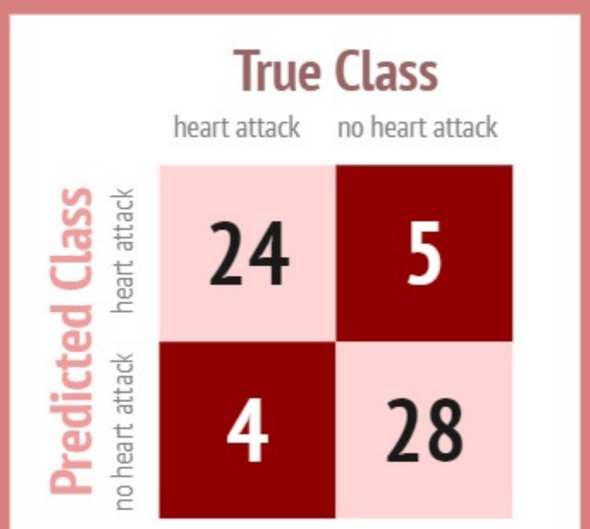
87%





SVM - Support Vector Machines





Accuracy Score of :

85%

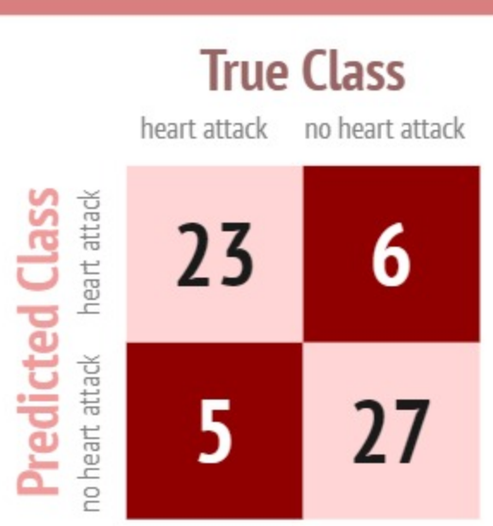
2

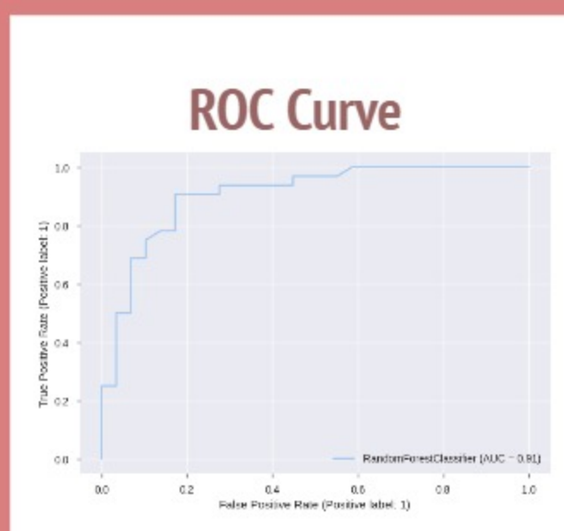
Random Forest

3

Accuracy Score of :

82%





Discussion

	The Outcomes	Possible Improvements
1 Logistic Regression	Larger dataset preferred Possible overfitting	Collect more data Tune parameter C to control complexity Use other classification methods...
2 SVM	SVM allows for a larger margin and greater versatility if kernels are taken advantage of, but we only produced a model using the linear kernel	Conduct a search for optimal hyperparameters and tune accordingly
3 Random Forest	Possible increased bias Incomplete tuning of parameters	Conduct a more thorough optimization of parameters

Conclusion

How can we take the project further?

We can use these models to predict cases of other health related issues.

Improve results of the SVM and Decision Tree models as they have many tunable parameters

Improve the model to have an accuracy prediction of over 90%?

Focus on life-style related features, such as diet, average sleep hours, etc.

Obtain more features to see other correlations that may exist, such as weights (of people), family health history, etc.

References

MECE 4520 Lecture Slides

Text book:

- Data-Driven Science and Engineering (Link)

Data science

- An Introduction to Statistical Learning ((Link)
- The Elements of Statistical Learning (Link)
- Python for Data Analysis (Link)
- Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (Link)

https://www.kaggle.com/namanmanchanda/heart-attack-eda-prediction-90-accuracy/data

questions?