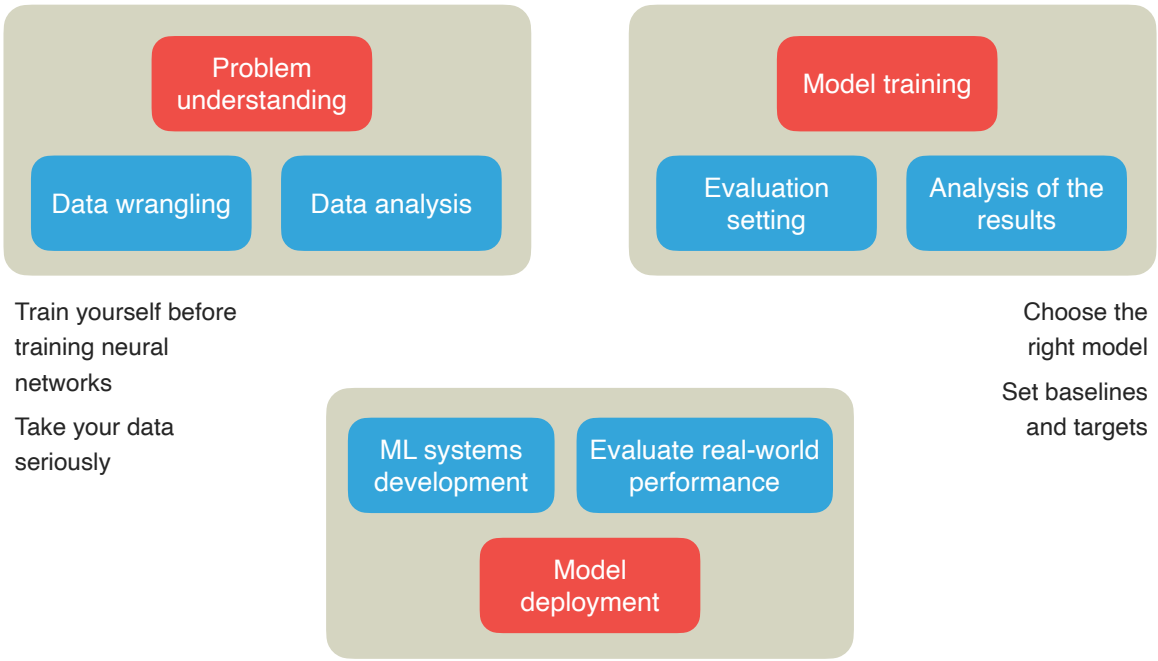
AI in practice - Part 1

Machine learning engineering: the logistic regression case

HELLENIC INSTITUTE of ADVANCED STUDIES

Machine learning engineering



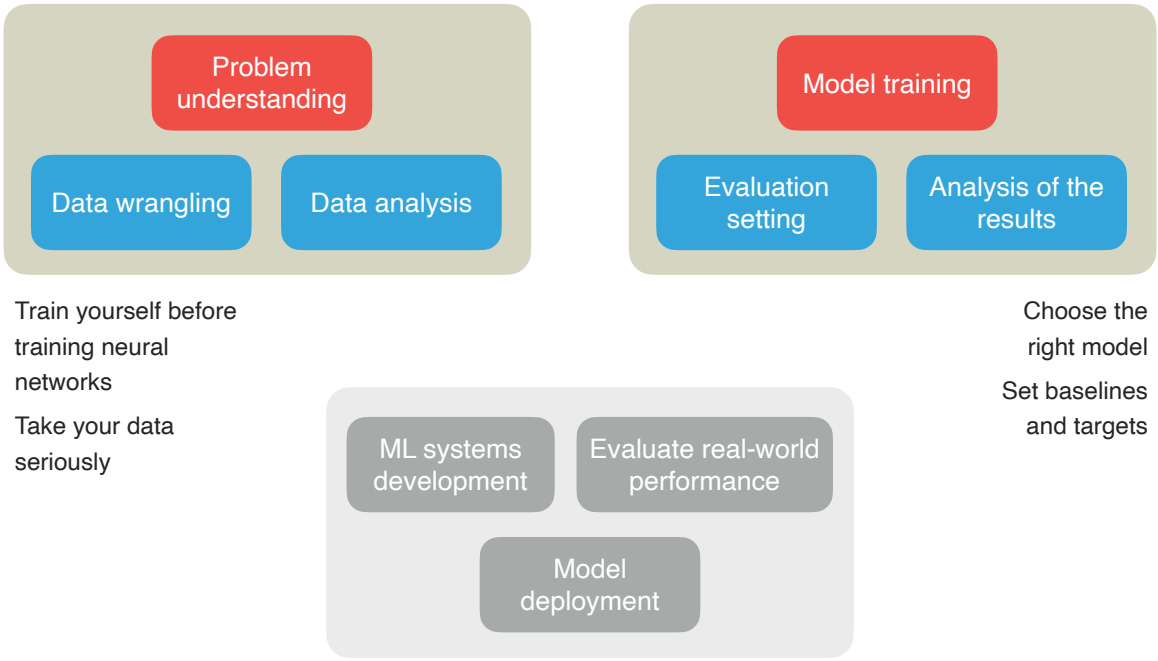
Train yourself before training neural networks

Take your data seriously

Choose the right model

Set baselines and targets

Machine learning engineering



Train yourself before training neural networks

Take your data seriously

Choose the right model

Set baselines and targets

Basic toolkit for problem understanding, data engineering and model training

Data analysis tools

- Data manipulation
- Data analysis
- Data visualisation

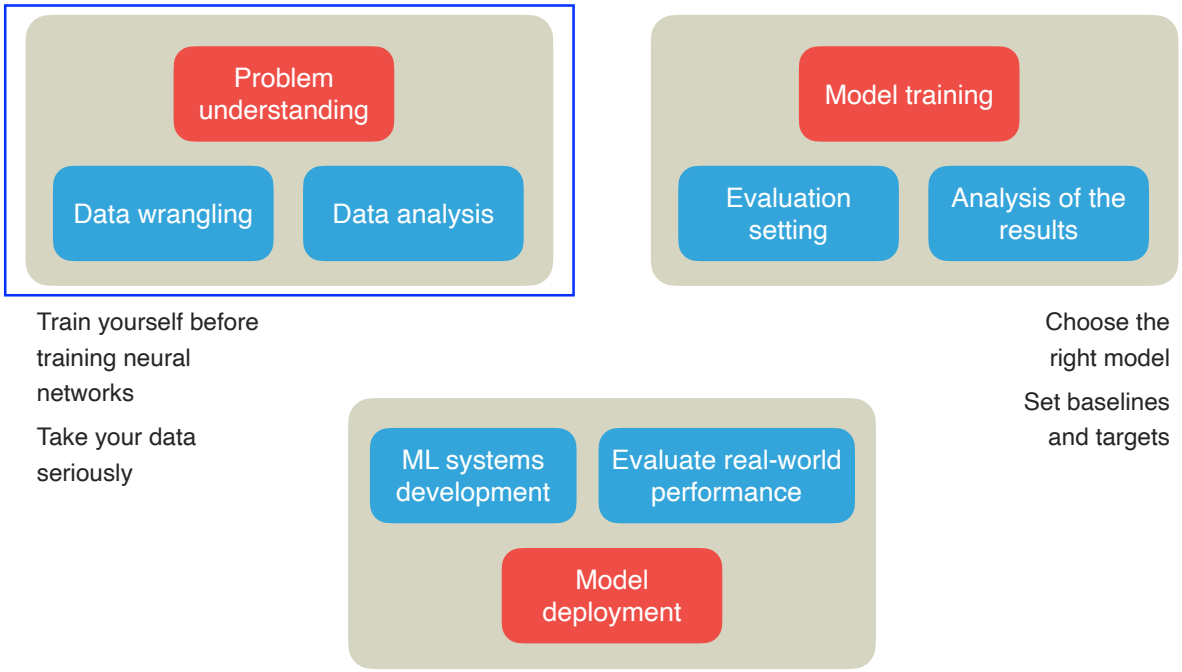
Programming tools

- Core programming languages
- Packages for mathematical functions
- Packages for mathematical functions

Development, evaluation, documentation and presentation tools

- Collaborative notebooks
- Interactive computing

Machine learning engineering



Train yourself before training neural networks

Take your data seriously

Choose the right model

Set baselines and targets

Problem: Heart Disease Prediction

Context
World Health Organisation has estimated 12 million deaths occur worldwide, every year due to Heart diseases. Half the deaths in developed countries are due to cardio vascular diseases.

Motivation
The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications.

Task
This research intends to pinpoint the most relevant/risk factors of heart disease as well as predict the overall risk using machine learning models (logistic regression).

Problem: Heart Disease Prediction

Dataset

Publicly available on the Kaggle website, it is from an ongoing ongoing cardiovascular study on residents of the town of Framingham, Massachusetts.
Provides the patients' information.

Dataset characteristics

Includes over 3.658 records and 16 attributes (possible risk factors).
The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD).

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
1	1	39	4	0	0	0	0	0	0	195	106	70	26.97	80	77	0
2	0	46	2	0	0	0	0	0	0	250	121	81	28.73	95	76	0
3	1	48	1	1	20	0	0	0	0	245	127.5	80	25.34	75	70	0
4	0	61	3	1	30	0	0	1	0	225	150	95	28.58	65	103	1
5	0	46	3	1	23	0	0	0	0	285	130	84	23.1	85	85	0
6	0	43	2	0	0	0	0	1	0	228	180	110	30.3	77	99	0
7	0	63	1	0	0	0	0	0	0	205	138	71	33.11	60	85	1
8	0	45	2	1	20	0	0	0	0	313	100	71	21.68	79	78	0

Data exploration

Preliminary and more complex tasks

Load the data

Use Pandas to load the CSV file
Perform basic operations on the data (access, visualise, filter etc)
Data handling (handle missing values, preprocessing etc)

Data exploration

Attributes

Demographic

Sex: male or female (Nominal)
Age: Age of the patient; (Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)
Education: no further information provided

Behavioural

Current Smoker: whether or not the patient is a current smoker (Nominal)
Cigs Per Day: the number of cigarettes that the person smoked on average in one day (can be considered continuous as one can have any number of cigarettes, even half a cigarette)

Information on medical history

BP Meds: whether or not the patient was on blood pressure medication (Nominal)
Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)
Prevalent Hyp: whether or not the patient was hypertensive (Nominal)
Diabetes: whether or not the patient had diabetes (Nominal)

Data exploration

Attributes

Information on current medical condition

Tot Chol: total cholesterol level (Continuous)
Sys BP: systolic blood pressure (Continuous)
Dia BP: diastolic blood pressure (Continuous)
BMI: Body Mass Index (Continuous)
Heart Rate: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)
Glucose: glucose level (Continuous)

Target variable to predict

TenYearCHD: 10 year risk of coronary heart disease (CHD) - (binary: “1”, means “Yes”, “0” means “No”)

Data exploration

Preliminary and more complex tasks

Load the data

Use Pandas to load the CSV file
Perform basic operations on the data (access, visualise, filter etc)
Data handling (handle missing values, preprocessing etc)

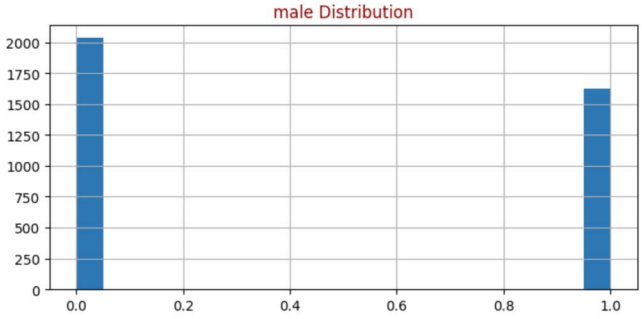
View the data

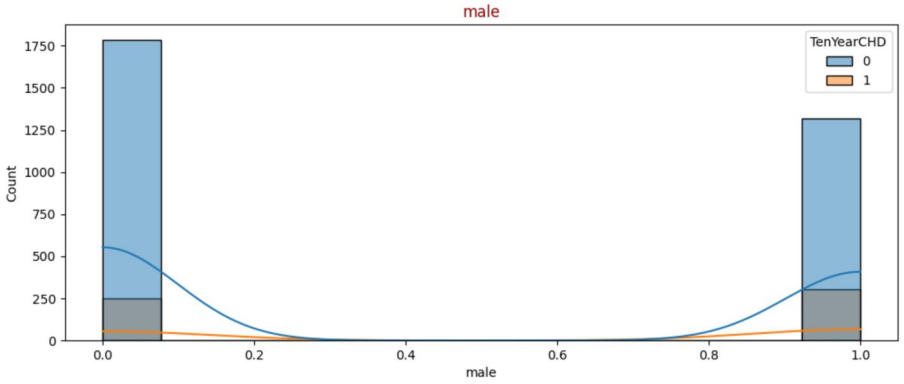
Get a glimpse of the data
Preview simple data statistics
Visualise complex data statistics

Understand the data

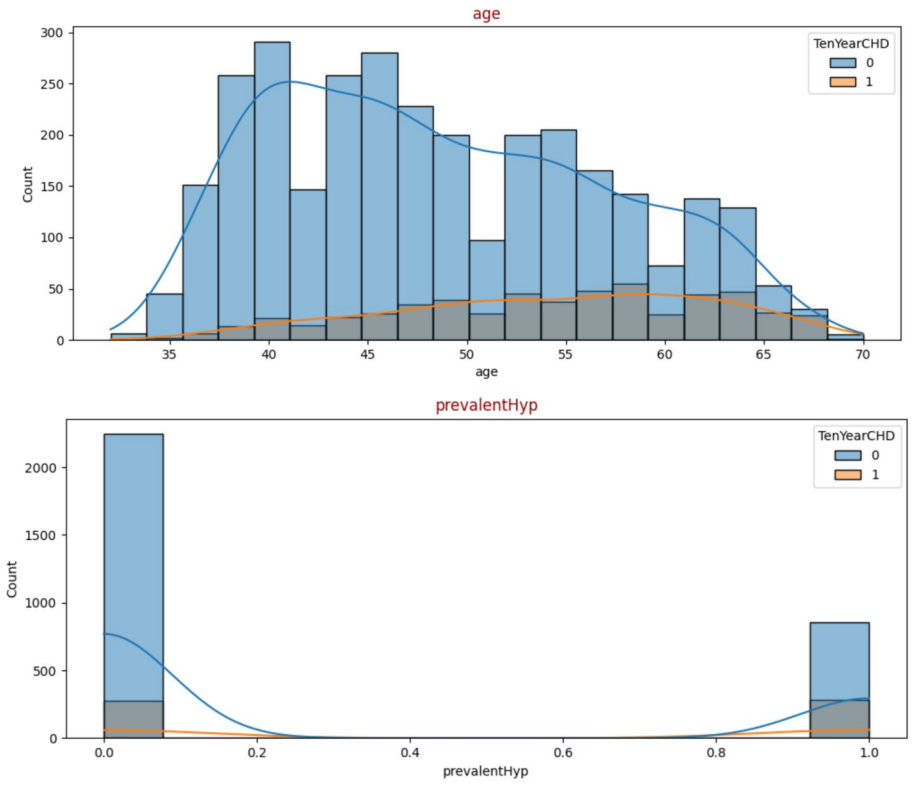
Understand the impact of the different attributes
Understand what are the attributes that are important
Understand possible biases

Data exploration

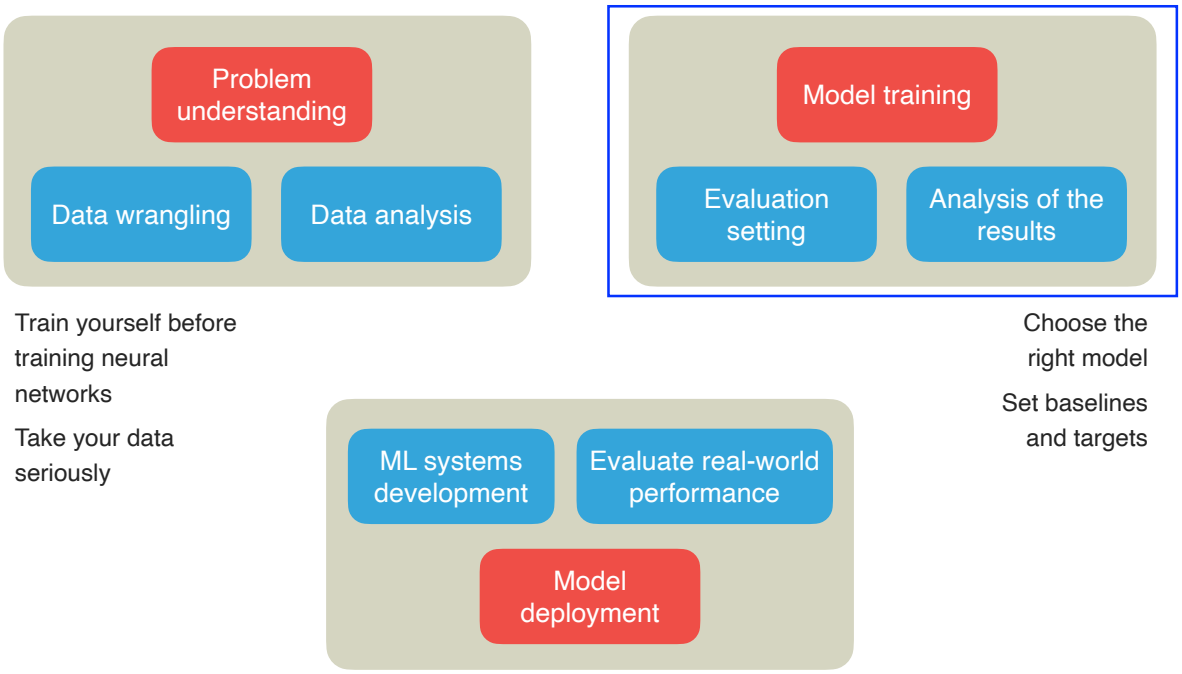




Data exploration



Machine learning engineering



Train yourself before training neural networks

Take your data seriously

Choose the right model

Set baselines and targets

Model training

Evaluation settings

Prepare the training and testing dataset
Set the baseline
Set the basic evaluation measures (model accuracy, confusion matrix etc)

The baseline

Can we use a very simple classifier?
Can we use specific attributes for classification?
What is the baseline accuracy?

Preparing the experimentation

Split the dataset: training and testing dataset
What is the training process info?
What are the measures that we are going to use for understanding the impact of the attributes?
Experimentation means (change the dataset splitting, reject some attributes, set the threshold etc)

Model training

Prepare the training and testing dataset

Split the dataset into training and testing datasets (training dataset between 50% and 80%)

Evaluation measures

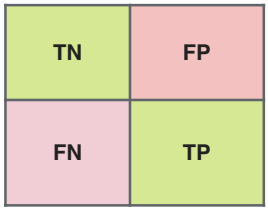
$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

Confusion matrix

True Positive (TP): a sample belonging to the positive class being classified correctly
True Negative (TN): a sample belonging to the negative class being classified correctly
False Positive (FP): a sample belonging to the negative class but being classified wrongly as belonging to the positive class
False Negative (FN): a sample belonging to the positive class but being classified wrongly as belonging to the negative class

Confusion matrix



Model training

The baseline

Always predict 0!

$\text{TenYearCHD} = \text{baseline}(x_1, \ldots, x_{16}) = 0$

Confusion matrix

TN	FP
FN	TP

=

3101	0
557	0

$$\text{Accuracy(baseline)} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{3101}{3101 + 557} = 0.8477310005467469$$

$$\text{Recall(baseline)} = \frac{\text{True Positive}}{\text{True Positive + False Negative}} = \frac{0}{0 + 557} = 0$$

Model training

Another baseline

Predict only based on PrevalentHyp (IF PrevalentHyp THEN TenYearCHD)

$\text{TenYearCHD} = \text{baseline2}(x_1, \ldots, x_{16}) = \text{PrevalentHyp}$

Confusion matrix

TN	FP
FN	TP

=

2245	856
273	284

$$\text{Accuracy(baseline2)} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{2245 + 284}{2245 + 856 + 273 + 284} = 0.6913613996719519$$

$$\text{Recall(Baseline2)} = \frac{\text{True Positive}}{\text{True Positive + False Negative}} = \frac{284}{273 + 284} = 0.5098743267504489$$

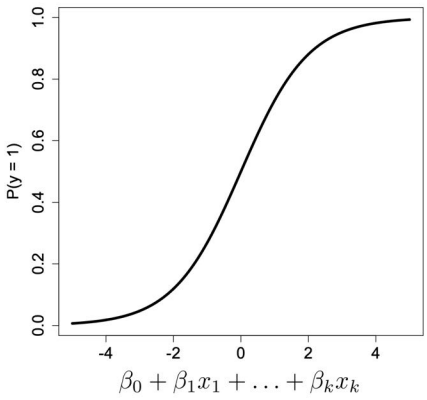
Training the model

The training process

What is the model that we are going to use? (logistic regression)
What are the important parameters of the model? (coefficients)

Understanding the trained model

What are the parameters of the model?
What is the influence of the attributes?



Logistic regression model $P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}}$

Training the model

The model coefficients

male	0.5545
age	0.0557
education	−0.0813
currentSmoker	0.0357
cigsPerDay	0.0142
BPMeds	0.1076
prevalentStroke	0.0820
prevalentHyp	0.4495
diabetes	0.4038
totChol	0.0007
sysBP	0.0150
diaBP	−0.0060
BMI	0.0116
heartRate	−0.0069
glucose	0.0047

Messages

- This fitted model shows that, holding all other features constant, the odds of getting diagnosed with heart disease for males (sex_male = 1) over that of females (sex_male = 0) is exp(0.5545) = 1.741. In terms of percent change, we can say that the odds for males are 74.1% higher than the odds for females.
- The coefficient for age says that, holding all others constant, we will see 6% increase in the odds of getting diagnosed with CDH for a one year increase in age since exp(0.0557) = 1.05728.
- Similarly , with every extra cigarette per day one smokes thers is a 1.43% increase in the odds of CDH.
- There is a 1.5% increase in odds for every unit increase in systolic Blood Pressure.

Testing the model

The model prediction

Compute the prediction of the model for the testing data

Evaluate the results using the basic measures

Understanding the predictive power of the model

Explore the intuitions behind the results

More experimentation (set different thresholds, reject the less important attributes etc)

Testing the model

The model prediction

Compute the prediction of the model for the testing data

Evaluate the results using the basic measures

Confusion matrix (testing data) (threshold=0.5)

TN	FP
FN	TP

=

1079	10
169	23

$$\text{Accuracy(model)} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{1079 + 23}{1079 + 10 + 69 + 23} = 0.8602654176424668$$

$$\text{Recall(model)} = \frac{\text{True Positive}}{\text{True Positive + False Negative}} = \frac{23}{23 + 169} = 0.11979166666666667$$

Testing the model

The model prediction

Compute the prediction of the model for the testing data
Evaluate the results using the basic measures

Confusion matrix (testing data) (threshold=0.3)

TN	FP
FN	TP

=

993	96
130	62

(threshold=0.5)

1079	10
169	23

$$\text{Accuracy(model2)} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{993 + 62}{993 + 96 + 130 + 62} = 0.8235753317720531$$

$$\text{Recall(model2)} = \frac{\text{True Positive}}{\text{True Positive + False Negative}} = \frac{62}{62 + 1130} = 0.3229166666666667$$

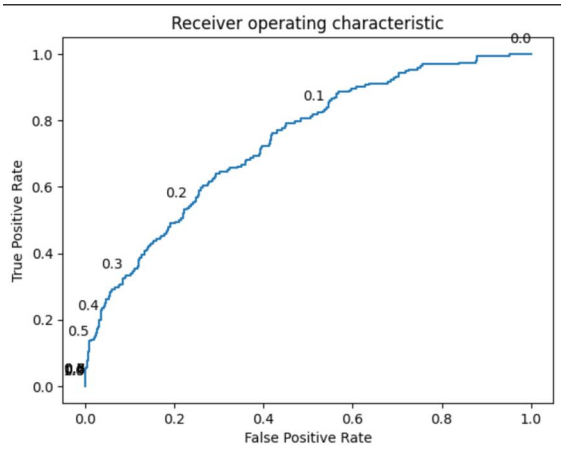
Testing the model

Understanding the predictive power of the model

Use more advanced measures (more informative measures)

ROC curve

A Receiver Operating Characteristic (ROC) curve is a graphical representation of the performance of a binary classifier system. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.



Area under the ROC curve

The area under the ROC curve (AUC) is a performance metric for binary classification problems. It represents the degree to which the predicted probabilities of a model are able to distinguish between the true positive and true negative instances.

Conclusions

- Men seem to be more susceptible to heart disease than women. Increase in Age, number of cigarettes smoked per day and systolic Blood Pressure also show increasing odds of having heart disease.
- We can eliminate attributes from the model that are not important for predicting the disease.
- Different values of thresholds may significantly change the accuracy of the model.
- The Area under the ROC curve is 0.74 which is satisfactory.
- Overall model could be improved with more data and experimentation.

Welcome to experimentation!