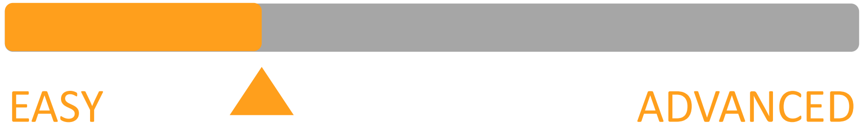
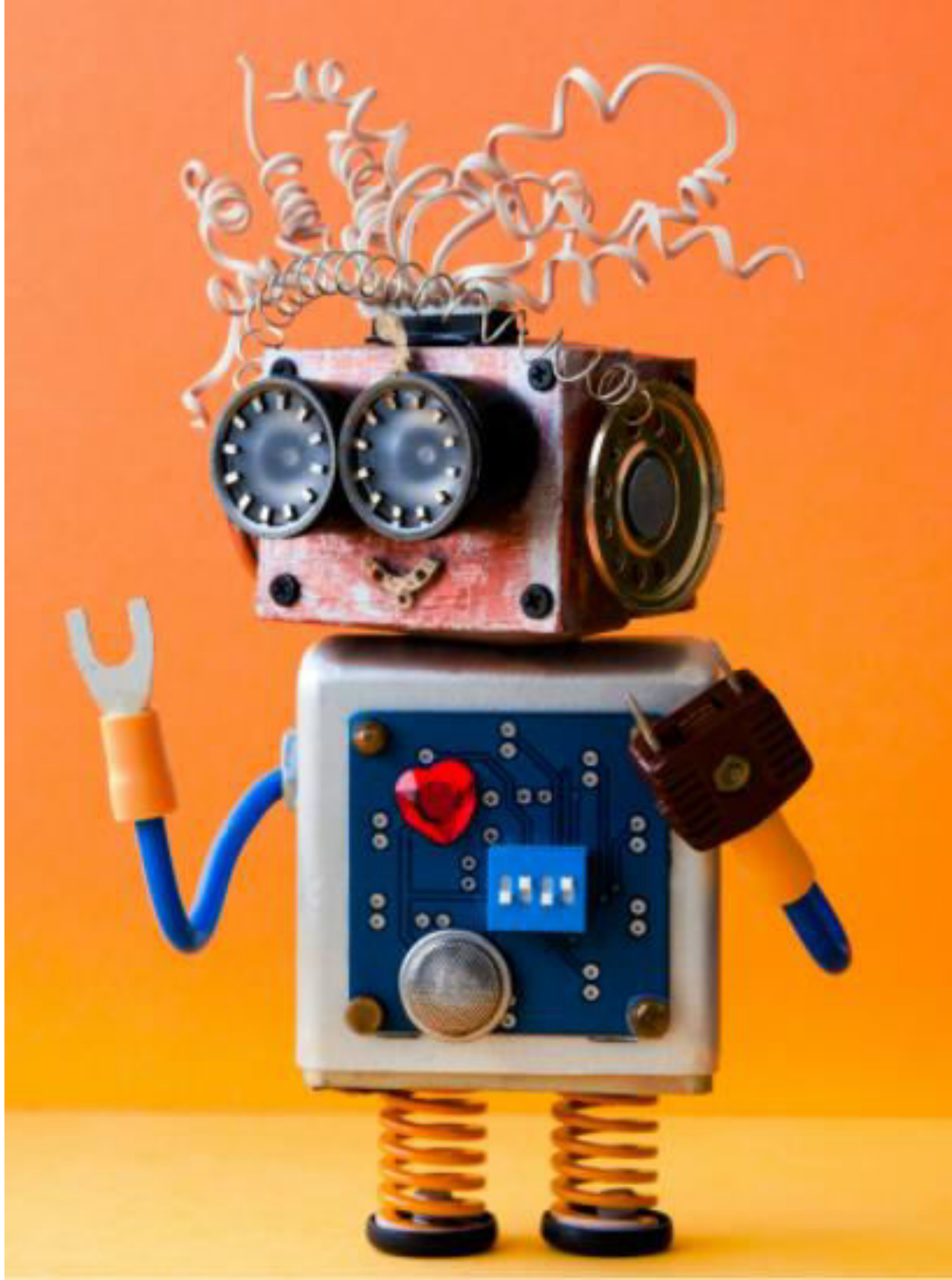
PROJECT OVERVIEW



EASY

ADVANCED



PROJECT OVERVIEW

- In this case study, we will assume that you work as a data scientist at a bank in Taiwan.
- The bank has collected extensive data about its customers such as demographics, historical payments record, amount of bill dollar values.
- Data has been collected between April 2005 to September 2005.
- The data consists of 25 variables. Let’s explore these variables in the next slide!
- Data Source: https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset

INPUTS/OUTPUTS

OUTPUT:

- default.payment.next.month: Default payment (1=yes, 0=no)

INPUTS:

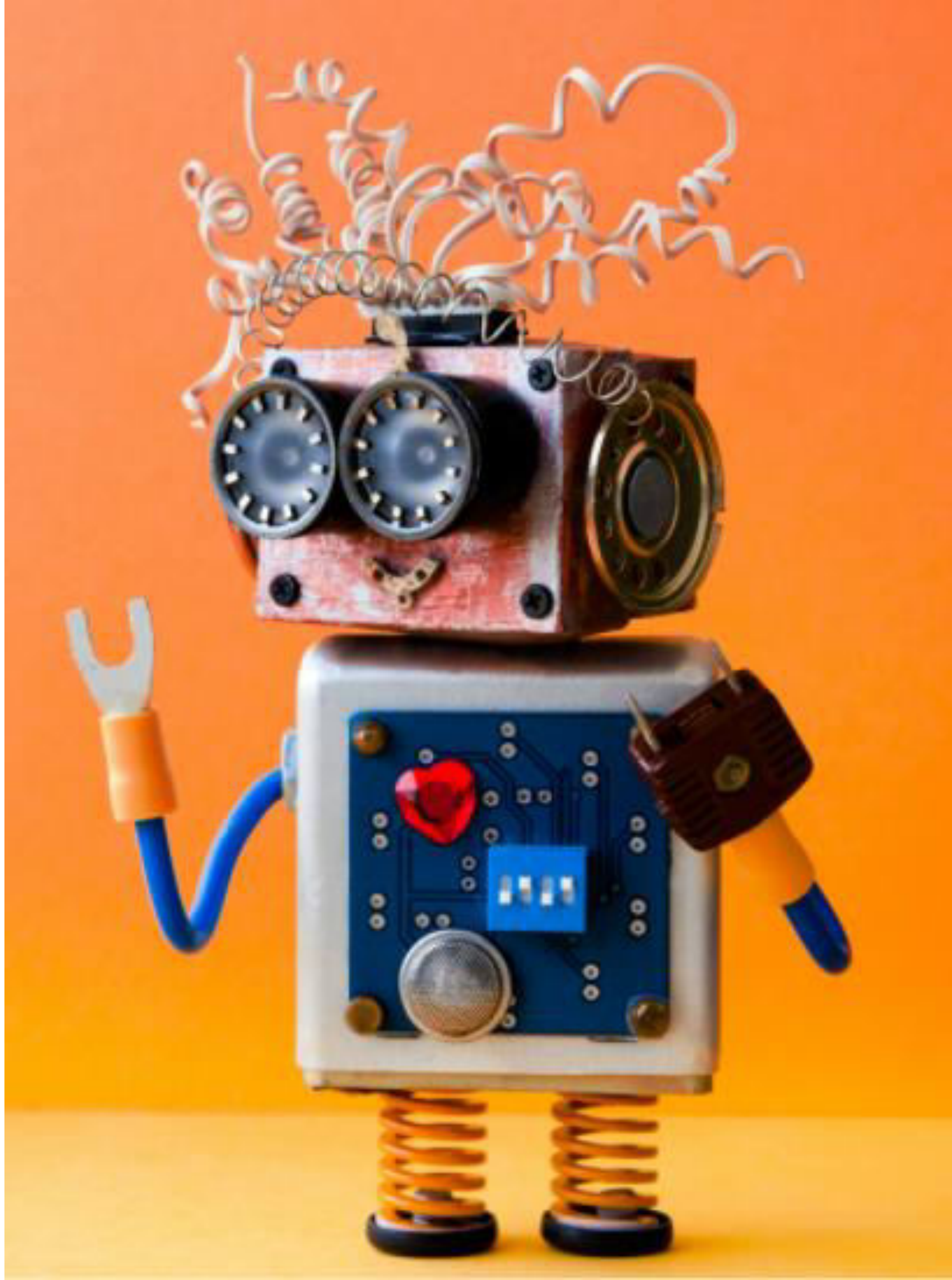
- ID: ID of each client
- LIMIT_BAL: Amount of given credit in NT (New Taiwan) dollars
- SEX: Gender (1=male, 2=female)
- EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
- MARRIAGE: Marital status (1=married, 2=single, 3=others)
- AGE: Age in years
- PAY_0: Repayment status in Sep, 2005 (-1=pay duly,
- 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight mont
  hs, 9=payment delay for nine months and above)
- PAY_2: Repayment status in August, 2005 (scale same as above)
- PAY_3: Repayment status in July, 2005 (scale same as above)
- PAY_4: Repayment status in June, 2005 (scale same as above)
- PAY_5: Repayment status in May, 2005 (scale same as above)
- PAY_6: Repayment status in April, 2005 (scale same as above)

INPUTS/OUTPUTS

- **INPUTS (CONTINUED):**
  - BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar)
  - BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)
  - BILL_AMT3: Amount of bill statement in July, 2005 (NT dollar)
  - BILL_AMT4: Amount of bill statement in June, 2005 (NT dollar)
  - BILL_AMT5: Amount of bill statement in May, 2005 (NT dollar)
  - BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar)
  - PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar)
  - PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar)
  - PAY_AMT3: Amount of previous payment in July, 2005 (NT dollar)
  - PAY_AMT4: Amount of previous payment in June, 2005 (NT dollar)
  - PAY_AMT5: Amount of previous payment in May, 2005 (NT dollar)
  - PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar)

XG-BOOST ALGORITHM REVIEW

EASY ADVANCED



XGBOOST: RECAP

- XGBoost or Extreme gradient boosting is the algorithm of choice for many data scientists and could be used for regression and classification tasks.
- XGBoost is a supervised learning algorithm and implements gradient boosted trees algorithm.
- The algorithm work by combining an ensemble of predictions from several weak models.
- It is robust to many data distributions and relationships and offers many hyperparameters to tune model performance.
- Xgboost offers increased speed and enhanced memory utilization.
- Xgboost is analogous to the idea of “discovering truth by building on previous discoveries”.

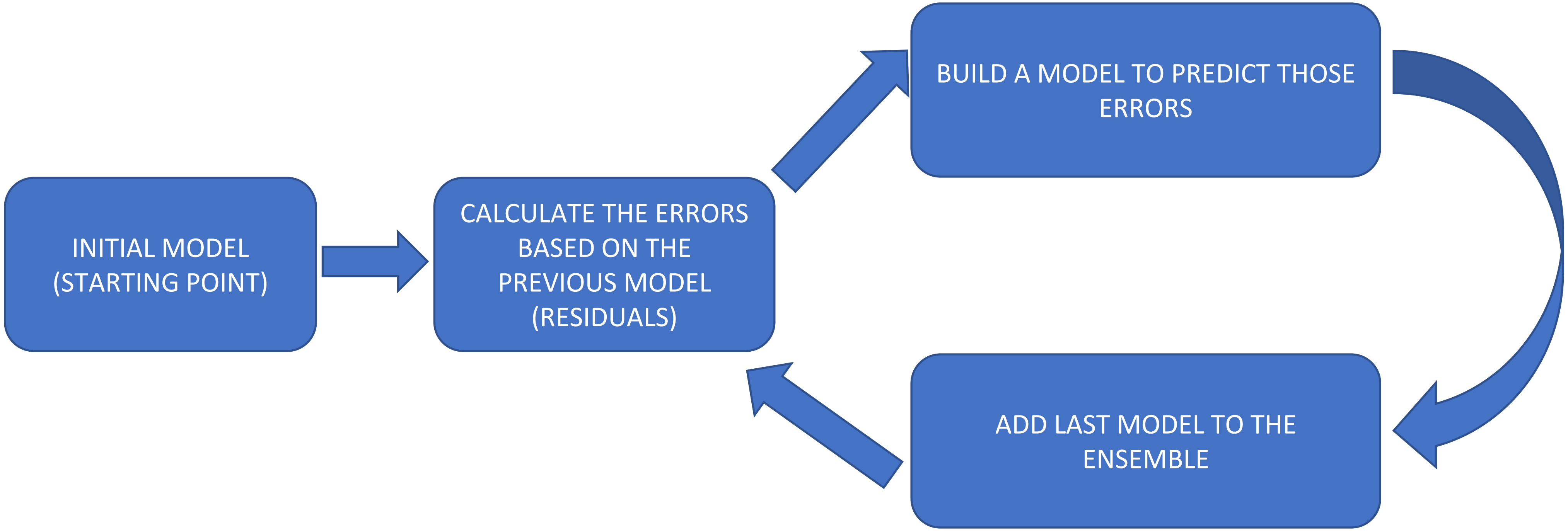
"If I have seen further it is by standing on the shoulders of Giants”, Isaac Newton

This picture is derived from Greek mythology: the giant Orion carried his servant Cedalion on his shoulders to act as the giant's eyes.

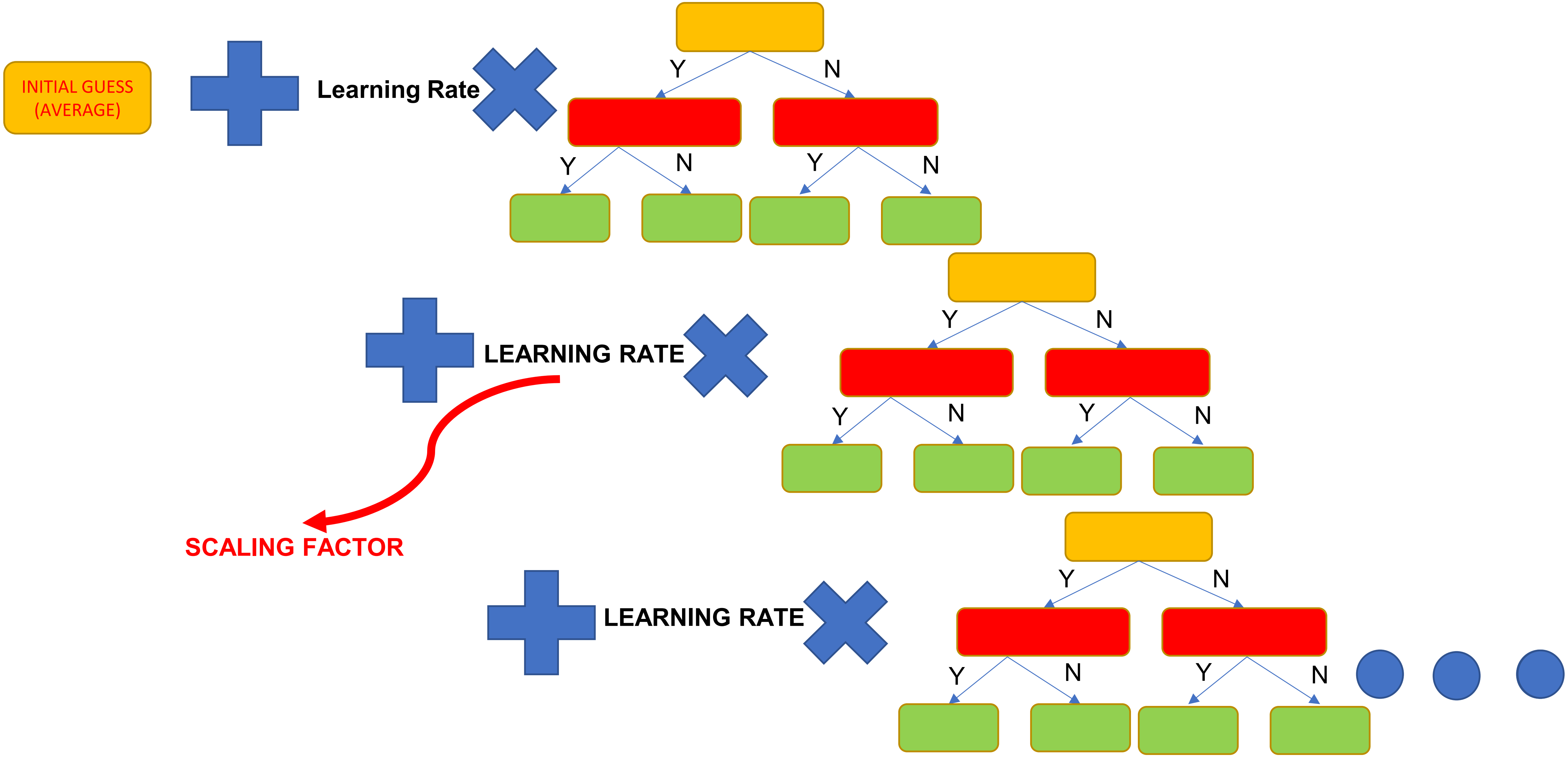
Source: https://commons.wikimedia.org/wiki/File:Library_of_Congress,_Rosenwald_4,_Bl._5r.jpg

XGBOOST: RECAP

- XGBoost repeatedly builds new models and combine them into an ensemble model
- Initially build the first model and calculate the error for each observation in the dataset
- Then you build a new model to predict those residuals (errors)
- Then you add prediction from this model to the ensemble of models
- XGboost is superior compared to gradient boosting algorithm since it offers a good balance between bias and variance (Gradient boosting only optimized for the variance so tend to overfit training data while XGboost offers regularization terms that can improve model generalization).


INITIAL MODEL (STARTING POINT)
CALCULATE THE ERRORS BASED ON THE PREVIOUS MODEL (RESIDUALS)
BUILD A MODEL TO PREDICT THOSE ERRORS
ADD LAST MODEL TO THE ENSEMBLE

XGBOOST: GRADIENT BOOSTING ALGORITHM



INITIAL GUESS (AVERAGE)

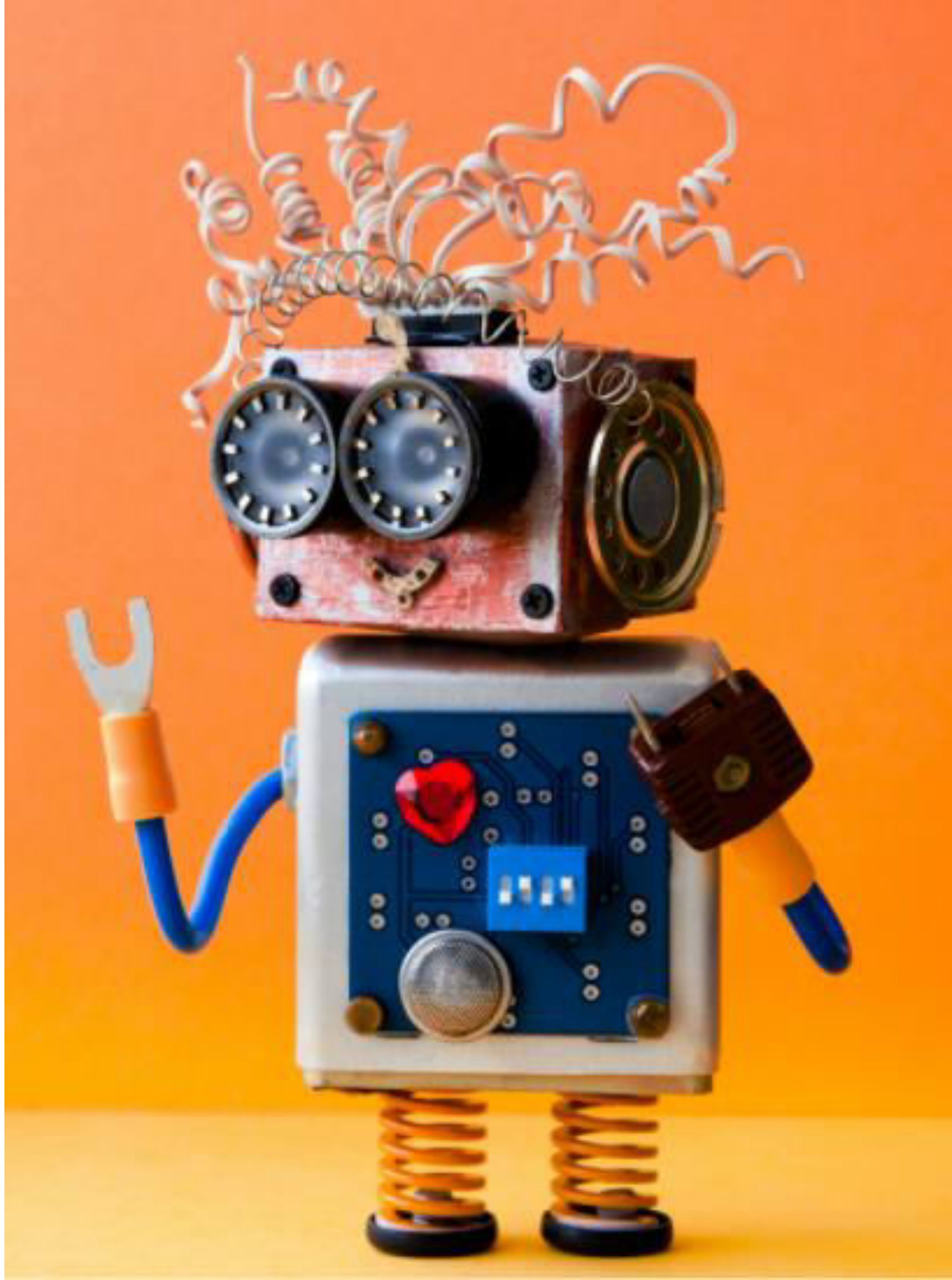
Learning Rate

LEARNING RATE

SCALING FACTOR

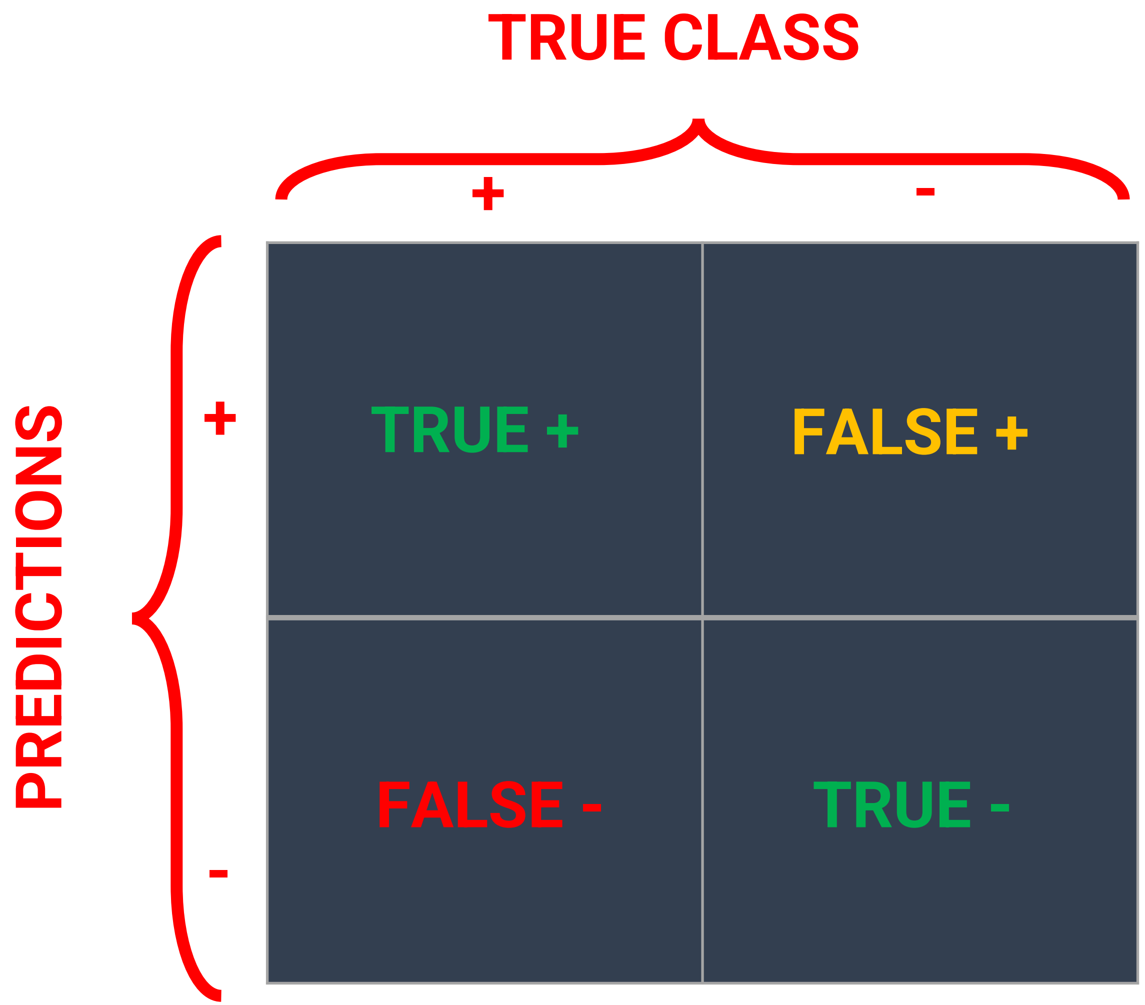
LEARNING RATE

CLASSIFICATION MODELS KPIs RECAP [SKIP IF FAMILIAR]



CLASSIFICATION MODEL KPIs

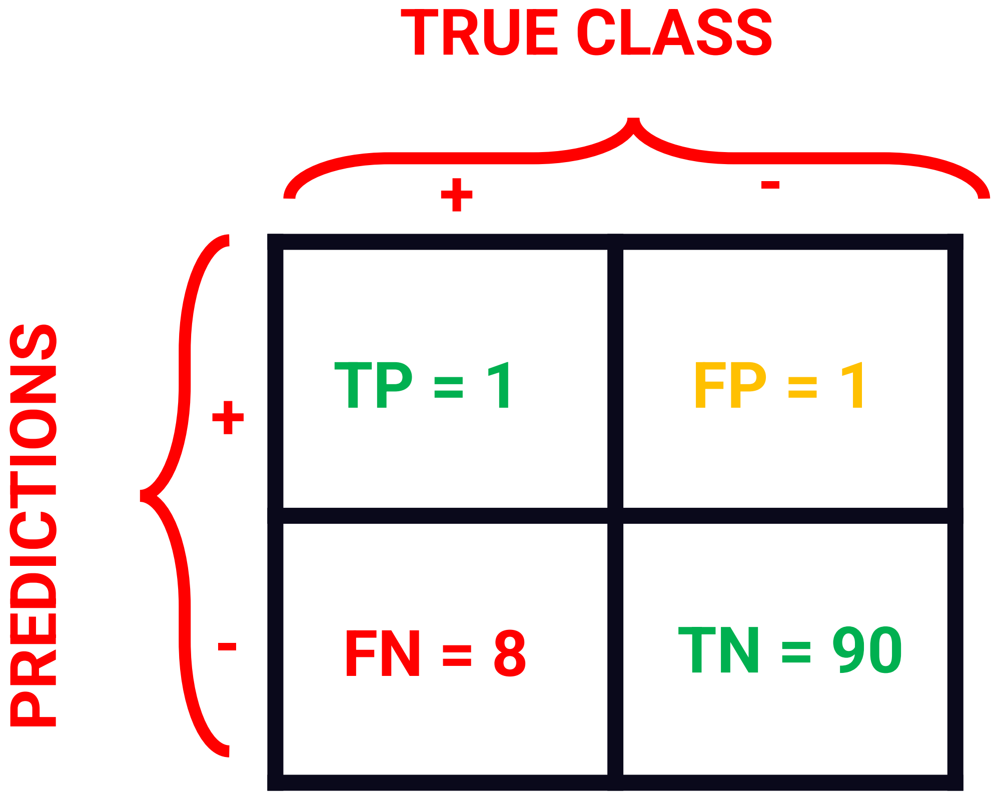
- Classification Accuracy = (TP+TN) / (TP + TN + FP + FN)
- Misclassification rate (Error Rate) = (FP + FN) / (TP + TN + FP + FN)
- Precision = TP/Total TRUE Predictions = TP/ (TP+FP) (When model predicted TRUE class, how often was it right?)
- Recall = TP/ Actual TRUE = TP/ (TP+FN) (when the class was actually TRUE, how often did the classifier get it right?)



		TRUE CLASS	
		+	-
PREDICTIONS	+	TRUE +	FALSE +
	-	FALSE -	TRUE -

PRECISION Vs. RECALL EXAMPLE

FACTS:
100 PATIENTS TOTAL
91 PATIENTS ARE HEALTHY
9 PATIENTS HAVE CANCER

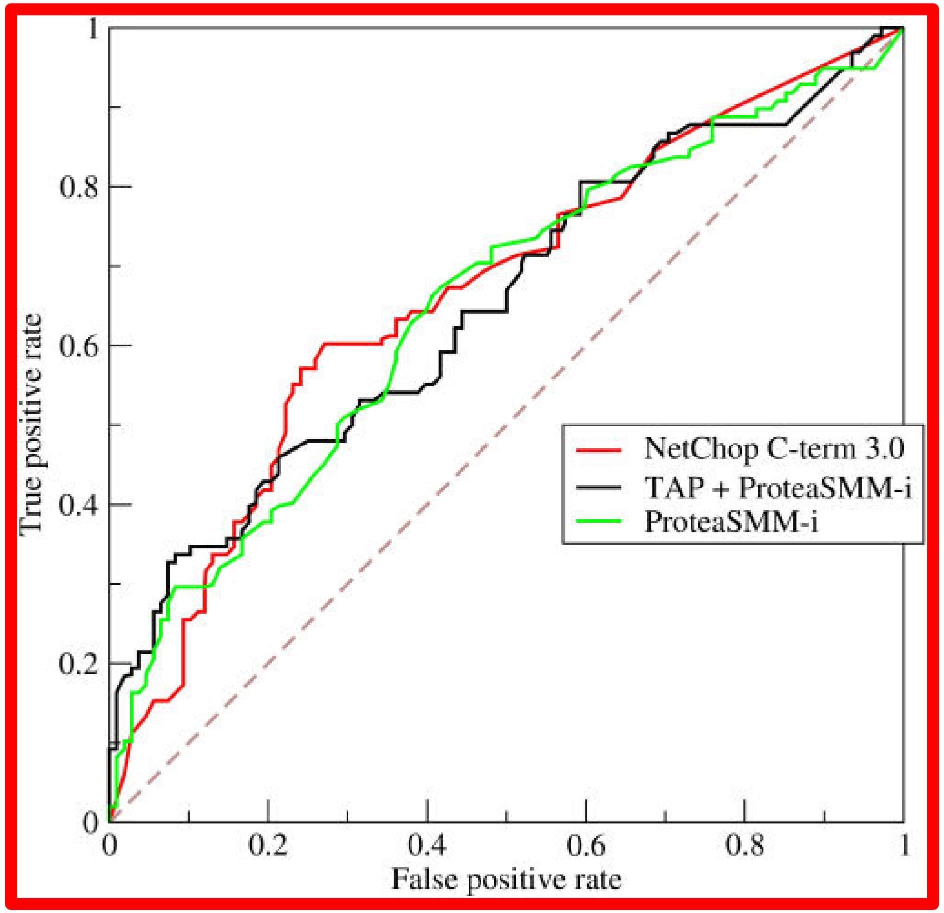


		TRUE CLASS	
		+	-
PREDICTIONS	+	TP = 1	FP = 1
	-	FN = 8	TN = 90

- Accuracy is generally misleading and is not enough to assess the performance of a classifier.
- Recall is an important KPI in situations where:
  - Dataset is highly unbalanced; cases when you have small cancer patients compared to healthy ones.

- Classification Accuracy = (TP+TN) / (TP + TN + FP + FN) = 91%
- Precision = TP/Total TRUE Predictions = TP/ (TP+FP) = ½=50%
- Recall = TP/ Actual TRUE = TP/ (TP+FN) = 1/9 = 11%

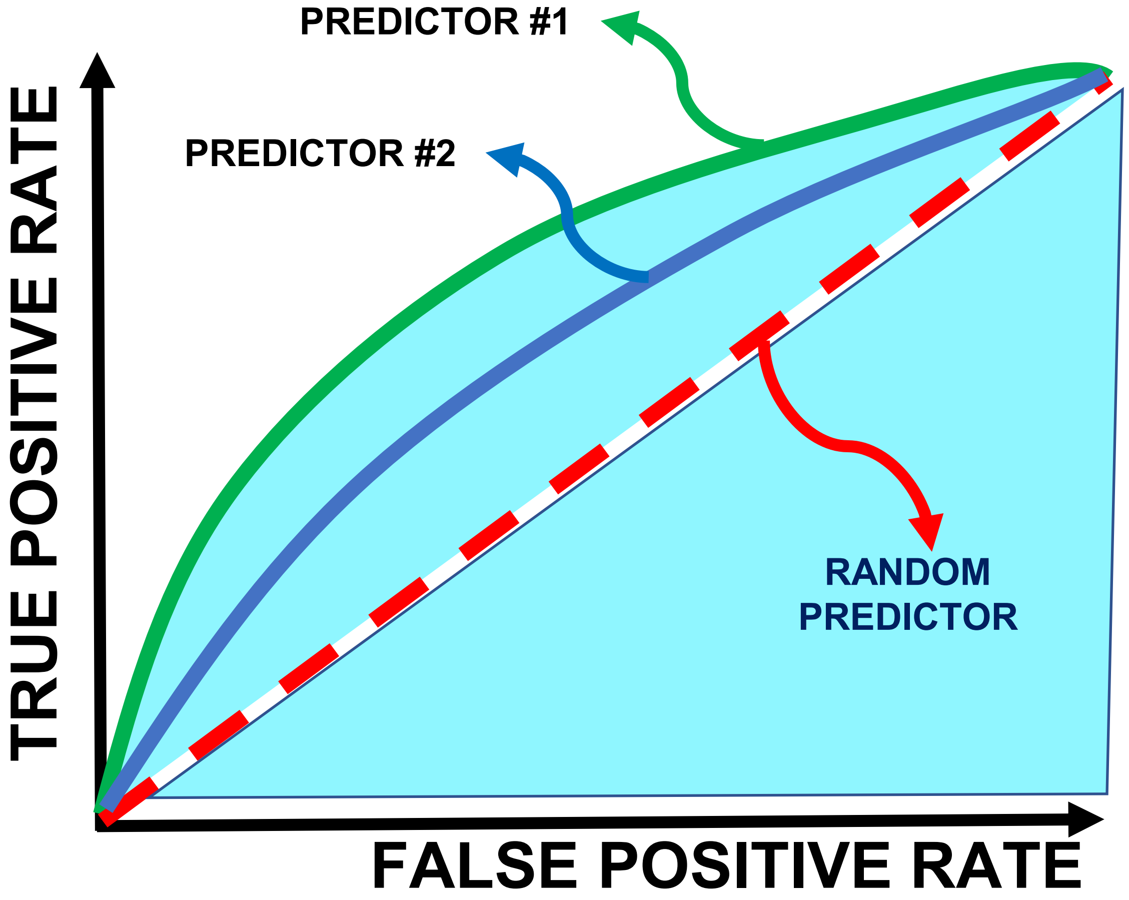
ROC (RECEIVER OPERATING CHARACTERISTIC CURVE)



- ROC Curve is a metric that assesses the model ability to distinguish between binary (0 or 1) classes.
- The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.
- The true-positive rate is also known as sensitivity, recall or probability of detection in machine learning.
- The false-positive rate is also known as the probability of false alarm and can be calculated as (1 − specificity).
- Points above the diagonal line represent good classification (better than random)
- The model performance improves if it becomes skewed towards the upper left corner.

Photo Credit: https://commons.wikimedia.org/wiki/File:Roccurves.png

AUC (AREA UNDER CURVE)

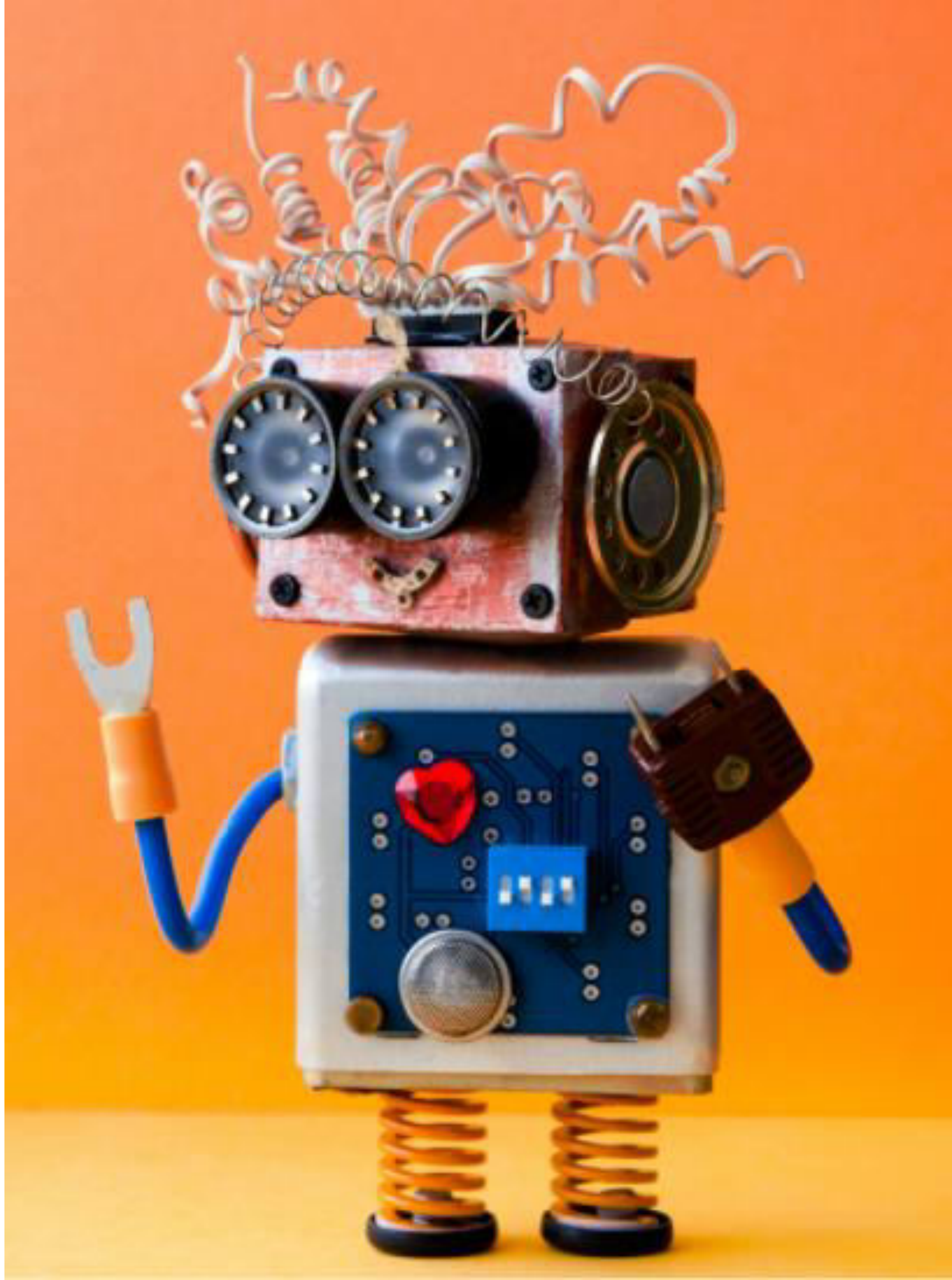


- The light blue area represents the area Under the Curve of the Receiver Operating Characteristic (AUROC).
- The diagonal dashed red line represents the ROC curve of a random predictor with AUROC of 0.5.
- If ROC AUC = 1, perfect classifier
- Predictor #1 is better than predictor #2
- Higher the AUC, the better the model is at predicting 0s as 0s and 1s as 1s.

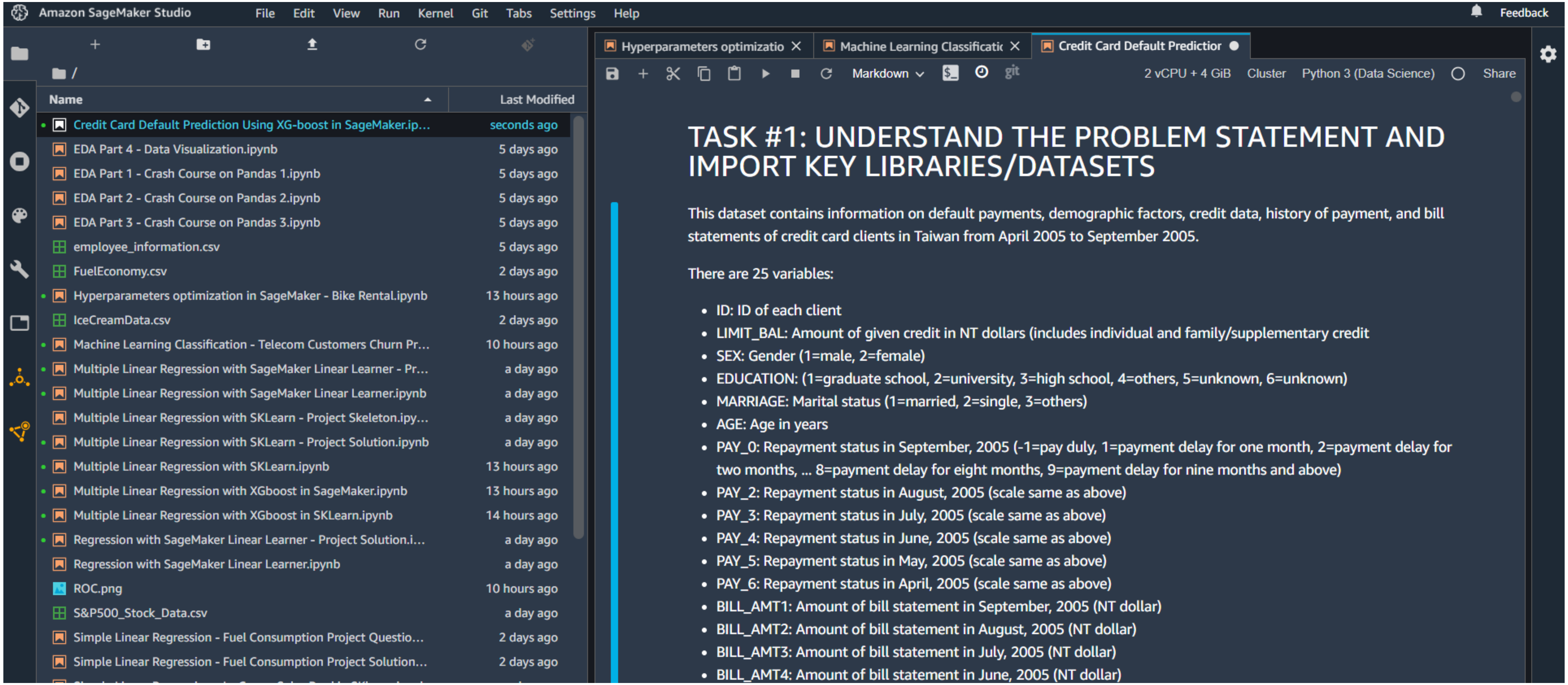
CODE DEMO

EASY

ADVANCED



CODE DEMO



Name	Last Modified
Credit Card Default Prediction Using XG-boost in SageMaker.ip...	seconds ago
EDA Part 4 - Data Visualization.ipynb	5 days ago
EDA Part 1 - Crash Course on Pandas 1.ipynb	5 days ago
EDA Part 2 - Crash Course on Pandas 2.ipynb	5 days ago
EDA Part 3 - Crash Course on Pandas 3.ipynb	5 days ago
employee_information.csv	5 days ago
FuelEconomy.csv	2 days ago
Hyperparameters optimization in SageMaker - Bike Rental.ipynb	13 hours ago
IceCreamData.csv	2 days ago
Machine Learning Classification - Telecom Customers Churn Pr...	10 hours ago
Multiple Linear Regression with SageMaker Linear Learner - Pr...	a day ago
Multiple Linear Regression with SageMaker Linear Learner.ipynb	a day ago
Multiple Linear Regression with SKLearn - Project Skeleton.ipy...	a day ago
Multiple Linear Regression with SKLearn - Project Solution.ipynb	a day ago
Multiple Linear Regression with SKLearn.ipynb	13 hours ago
Multiple Linear Regression with XGboost in SageMaker.ipynb	13 hours ago
Multiple Linear Regression with XGboost in SKLearn.ipynb	14 hours ago
Regression with SageMaker Linear Learner - Project Solution.i...	a day ago
Regression with SageMaker Linear Learner.ipynb	a day ago
ROC.png	10 hours ago
S&P500_Stock_Data.csv	a day ago
Simple Linear Regression - Fuel Consumption Project Questio...	2 days ago
Simple Linear Regression - Fuel Consumption Project Solution...	2 days ago

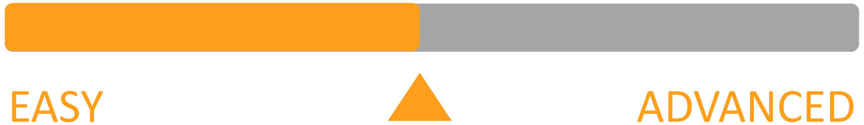
TASK #1: UNDERSTAND THE PROBLEM STATEMENT AND IMPORT KEY LIBRARIES/DATASETS

This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

There are 25 variables:

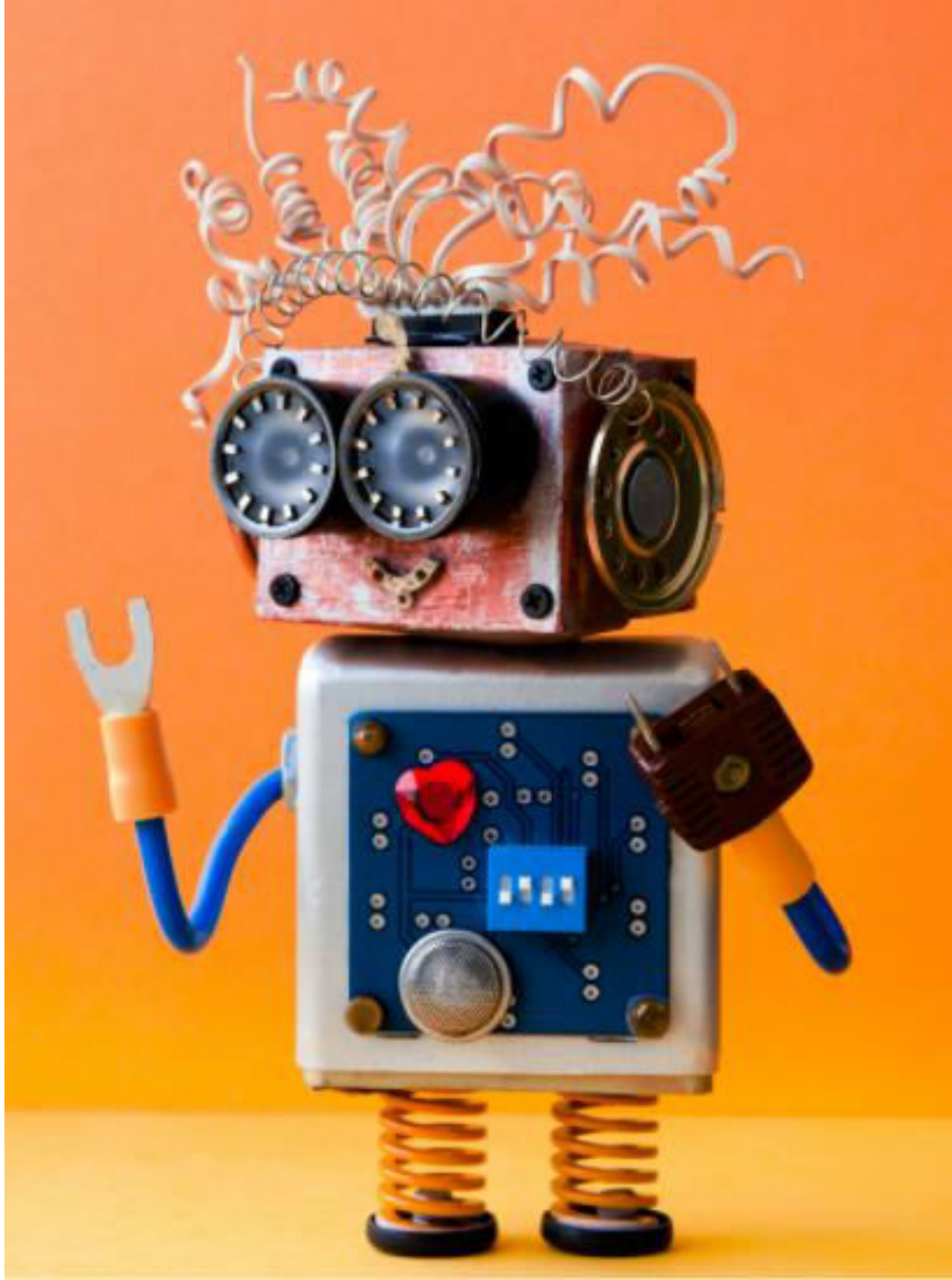
- ID: ID of each client
- LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit
- SEX: Gender (1=male, 2=female)
- EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
- MARRIAGE: Marital status (1=married, 2=single, 3=others)
- AGE: Age in years
- PAY_0: Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, ... 8=payment delay for eight months, 9=payment delay for nine months and above)
- PAY_2: Repayment status in August, 2005 (scale same as above)
- PAY_3: Repayment status in July, 2005 (scale same as above)
- PAY_4: Repayment status in June, 2005 (scale same as above)
- PAY_5: Repayment status in May, 2005 (scale same as above)
- PAY_6: Repayment status in April, 2005 (scale same as above)
- BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar)
- BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)
- BILL_AMT3: Amount of bill statement in July, 2005 (NT dollar)
- BILL_AMT4: Amount of bill statement in June, 2005 (NT dollar)

FINAL END-OF-DAY CAPSTONE PROJECT



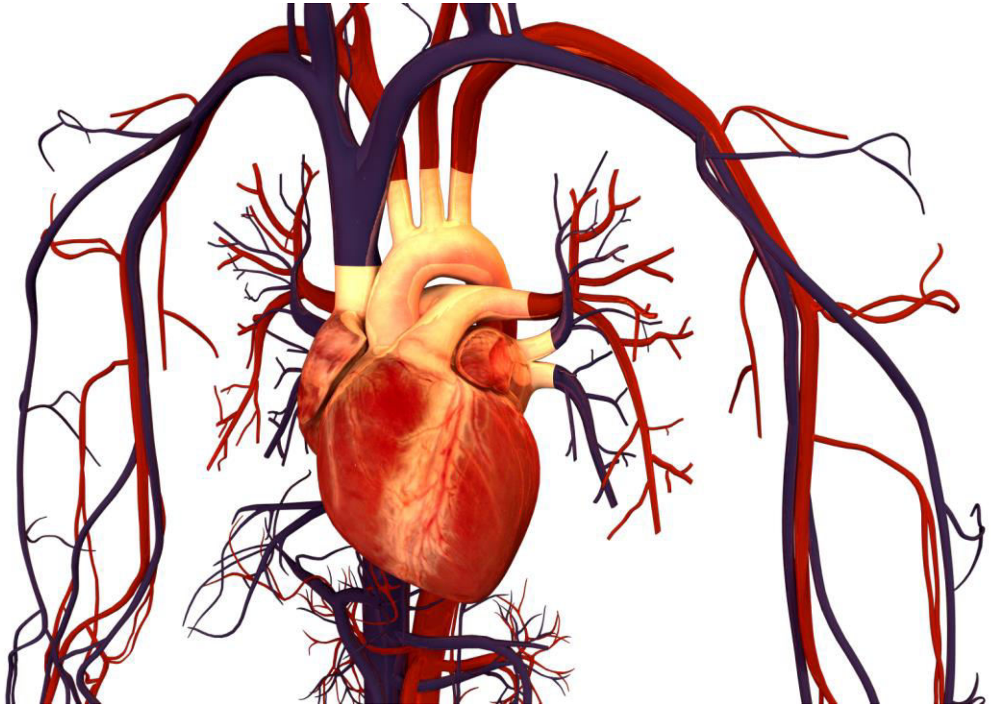
EASY

ADVANCED



PROJECT OVERVIEW:

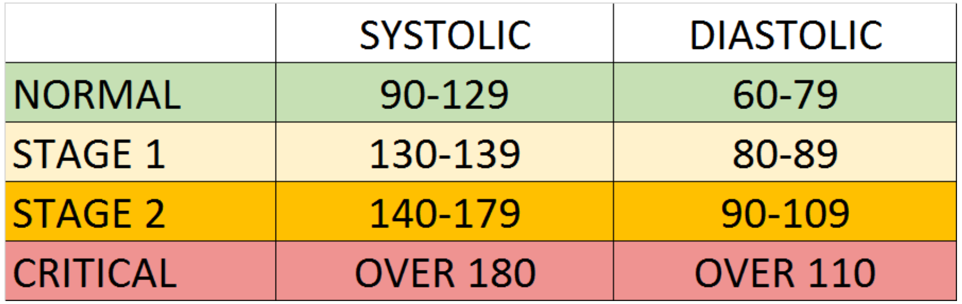
- Aim of the problem is to detect the presence or absence of cardiovascular disease in person based on the given features.
- Features available are:
  - Age
  - Height
  - Weight
  - Gender
  - Smoking
  - Alcohol intake
  - Physical activity
  - Systolic blood pressure
  - Diastolic blood pressure
  - Cholesterol
  - Glucose



- **Data Source:** https://www.kaggle.com/sulianova/cardiovascular-disease-dataset
- **Image Source:** https://commons.wikimedia.org/wiki/File:Human_Heart_and_Circulatory_System.png

PROJECT OVERVIEW: NOTES ON BLOOD PRESSURE

- **Blood Pressure notes:**
  - Blood pressure is represented by 2 numbers systolic and diastolic (ideally 120/80 mm Hg).
  - These two number are critical in assessing the heart health.
  - The top number represents **systolic** and the bottom number representing the **diastolic**.
  - Systolic pressure indicates the blood pressure in the arteries when the blood is pumped out of the heart.
  - The diastolic pressure indicates the blood pressure between beats (at rest, filling up and ready to pump again).
  - If these numbers are high, that means that the heart is exerting more effort to pump blood in the arteries to the body.



	SYSTOLIC	DIASTOLIC
NORMAL	90-129	60-79
STAGE 1	130-139	80-89
STAGE 2	140-179	90-109
CRITICAL	OVER 180	OVER 110

Photo Source: https://commons.wikimedia.org/wiki/File:Hypertension_ranges_chart.png

PROJECT OVERVIEW: NOTES ON CHOLESTEROL

- **Cholesterol notes:**
  - Cholesterol is a waxy material found in humans blood.
  - Normal level of cholesterol is necessary to ensure healthy body cells but as these levels increase, heart disease risk is elevated.
  - This waxy material can block the arteries and could result in strokes and heart attacks.
  - Healthy lifestyle and regular exercises can reduce the risk of having high cholesterol levels.
  - More information: https://www.mayoclinic.org/diseases-conditions/high-blood-cholesterol/symptoms-causes/syc-20350800

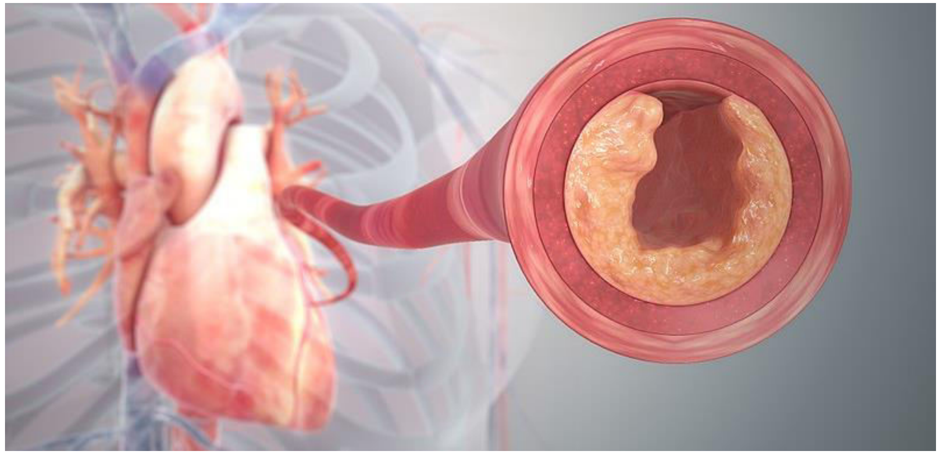


Photo Credit: https://commons.wikimedia.org/wiki/File:Clogged_Heart_Artery.jpg

PROJECT OVERVIEW: NOTES ON GLUCOSE

- **Glucose notes:**
  - Glucose represents the sugar that the human body receive when they consume food.
  - Glucose means “sweet” in Greek.
  - Insulin hormone plays a key role in moving glucose from the blood to the body cells for energy.
  - Diabetic patients have high glucose in their blood stream which could be due to two reasons:
    - They don’t have enough insulin
    - Body cells do not react to insulin the proper way
  - Read more: https://www.webmd.com/diabetes/glucose-diabetes

Photo Credit: https://commons.wikimedia.org/wiki/File:Clogged_Heart_Artery.jpg

PROJECT TASKS

Using SageMaker XG-Boost, perform the following:

- 1. Load the "*cardio_train.csv*" dataset to S3
- 2. Split the data into 80% for training and 20% for testing
- 3. Train an XG-Boost classifier model using SK-Learn Library
- 4. Perform GridSearch to optimize model hyperparameters
- 5. Train an XG-Boost classifier model using Amazon SageMaker
- 6. Deploy trained model as an endpoint
- 7. Assess trained model performance
- 8. Plot the confusion matrix
- 9. Delete the endpoint