EMPLOYEE ATTRITION PROJECT

A Machine leaning Project by :

Paul Sentongo
Margret Peninah Nankya

- **According to the Uganda Bureau of Statistics' National Labour Force Survey 2021, there is a notable movement of employees between jobs, which impacts various sectors differently**

- **Human Resource Managers in Uganda have highlighted that this high turnover rate is challenging for companies, as it affects their ability to invest in and retain skilled employees.**

- **Hiring and retaining employees are extremely complex tasks that require capital, time and skills.**

- **“Small business owners spend 40% of their working hours on tasks that do not generate any income such as hiring”.**

- **“Companies spend 15%-20% of the employee's salary to recruit a new candidate”.**

Continued...

- **“An average company loses anywhere between 1% and 2.5% of their total revenue on the time it takes to bring a new hire up to speed”.**
- **Hiring a new employee costs an average of $7645 (0-500 corporation).**
- **It takes 52 days on average to fill a position.**

- **Our goal is to develop a predictive model to identify employees who are likely to quit. This will enable the HR teams to proactively address potential attrition issues.**
- **The dataset used includes various attributes of employees such as Job Involvement, Education, Job Satisfaction, Performance Rating, Relationship Satisfaction, and Work-Life Balance. This data is sourced from a public dataset available on Kaggle. We acknowledge that this dataset is from a different geographical location and may require contextual adaptation for our specific scenario in Uganda.**

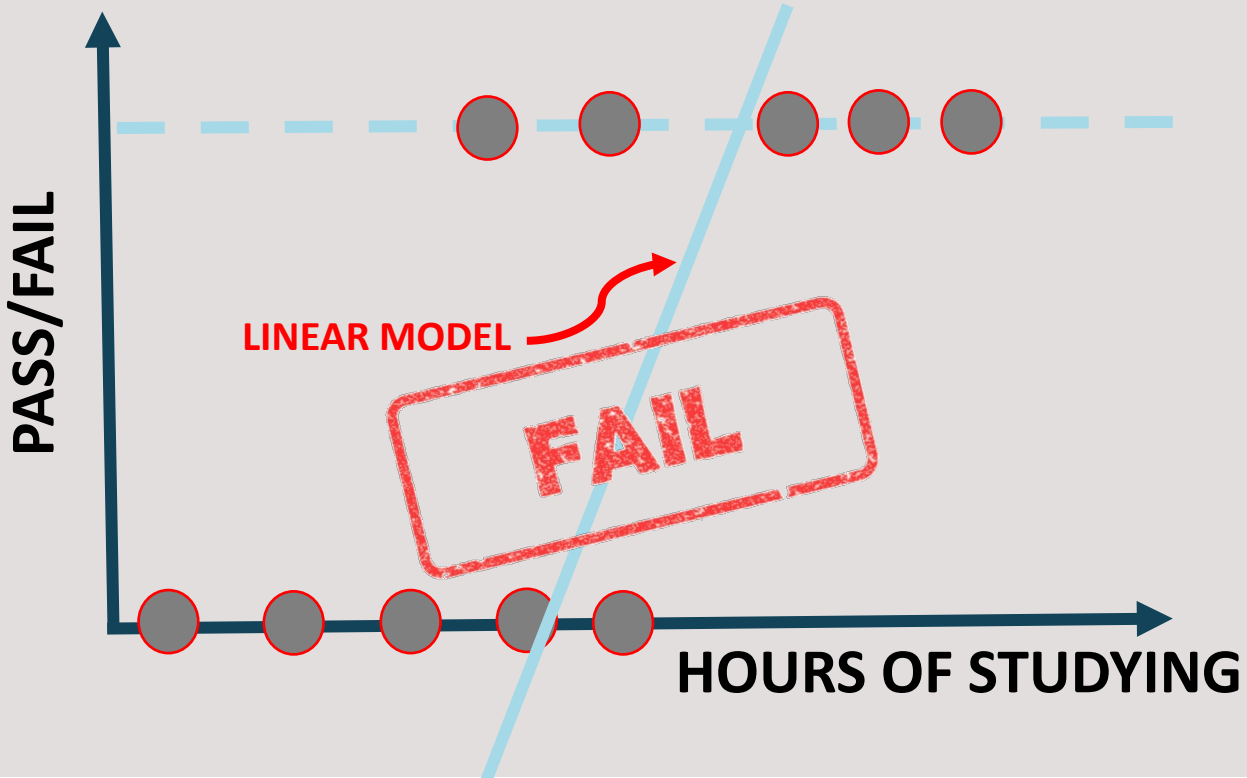
Data Source: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

MODEL SELECTION

We have selected the logistic regression classifier model as the best model with accuracy of 89% compared to the Artificial Neural Network and the random forest classifier

LOGISTIC REGRESSION CLASSIFIER

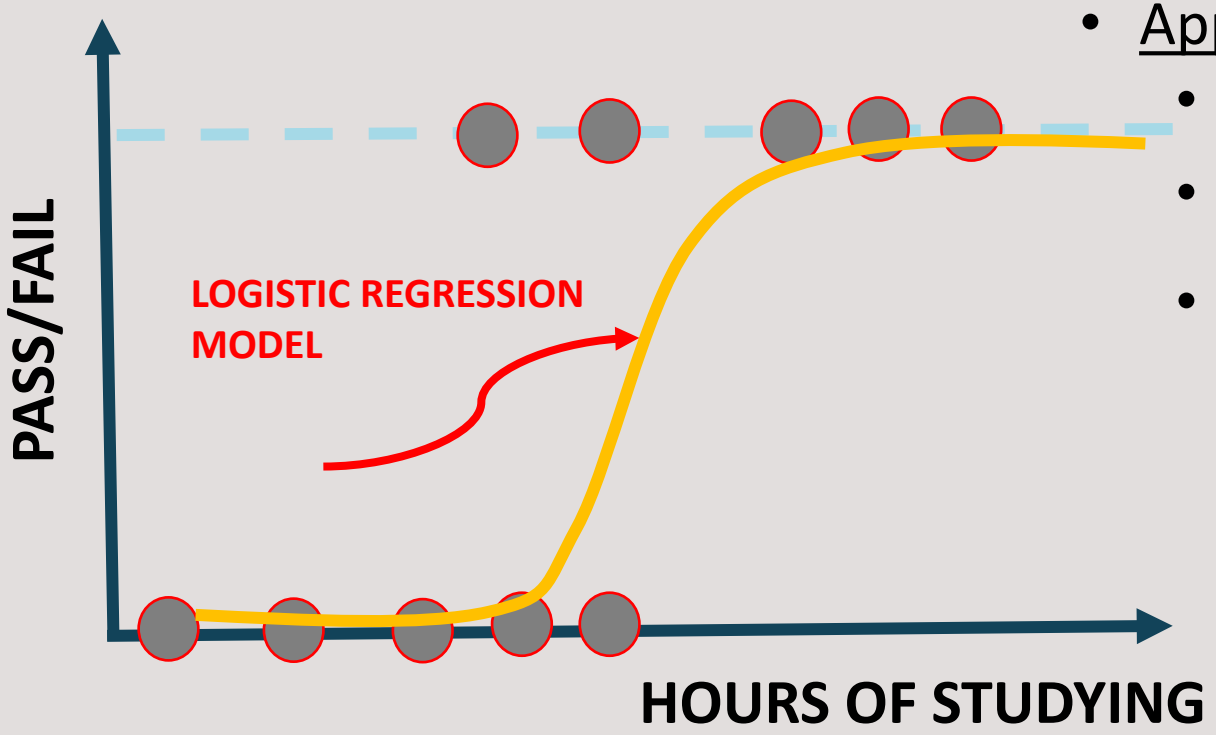
- **Linear regression** is used to predict outputs on a continuous spectrum.
- **Logistic regression is used to predict binary outputs** with two possible values labeled "0" or "1"
    - Logistic model output can be one of two classes: pass/fail, win/lose, healthy/sick etc.
    - Example as below.



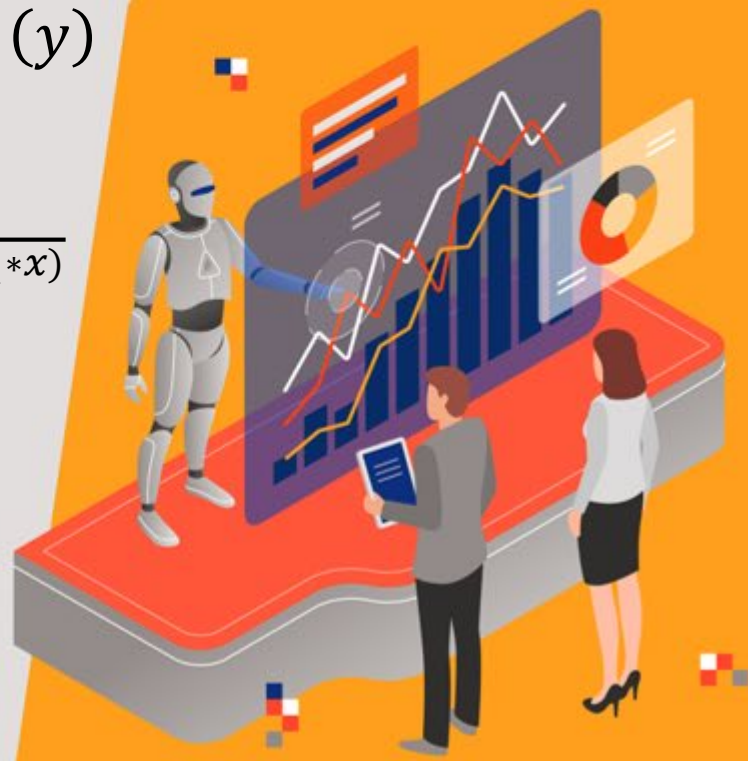
Hours Studying	Pass/Fail
1	0
1.5	0
2	0
3	1
3.25	0
4	1
5	1
6	1

LOGISTIC REGRESSION CLASSIFIER

- **Logistic regression algorithm works by implementing a linear equation first with independent predictors to predict a value.**
- **We then need to convert this value into a probability that could range from 0 to 1.**

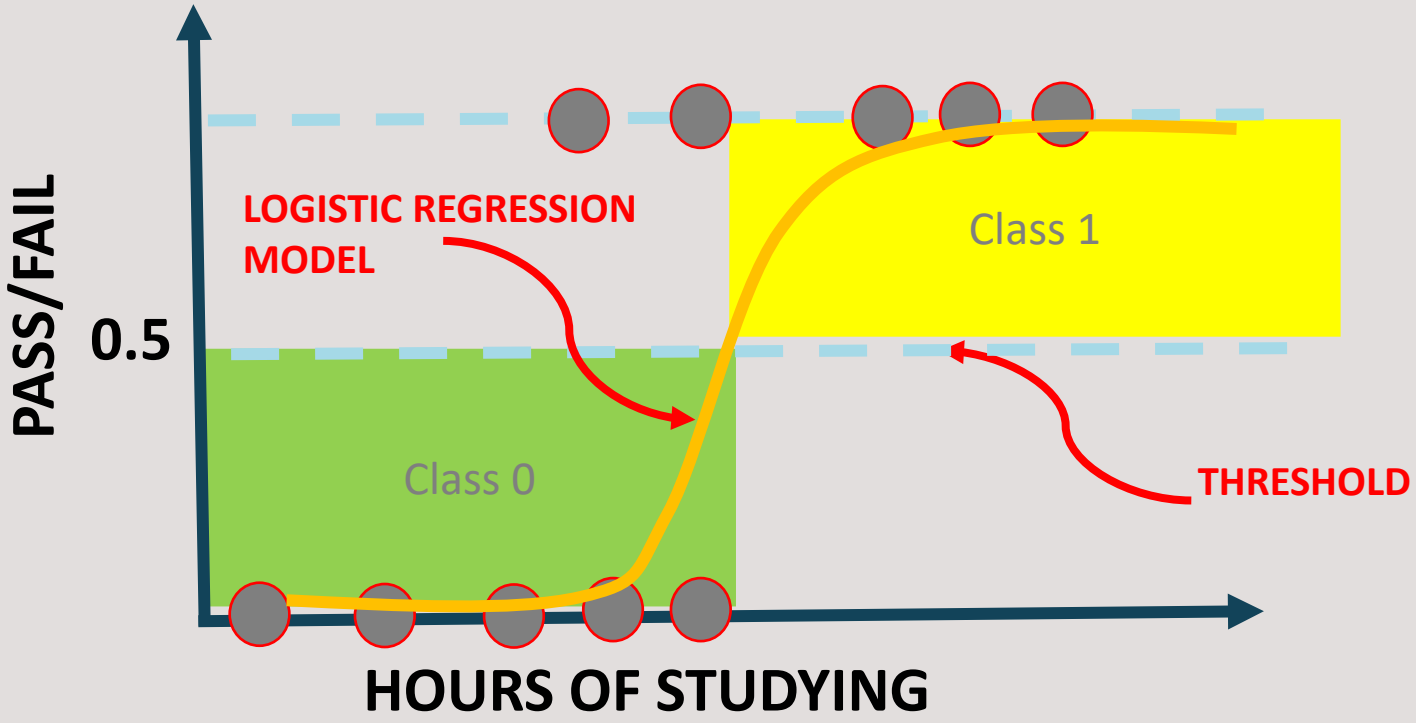


- Linear equation:
  - $y = b_0 + b_1 * x$
- Apply Sigmoid function:
  - $P(x) = sigmoid\ (y)$
  - $P(x) = \frac{1}{1+e^{-y}}$
  - $P(x) = \frac{1}{1+e^{-(b_0+b_1*x)}}$

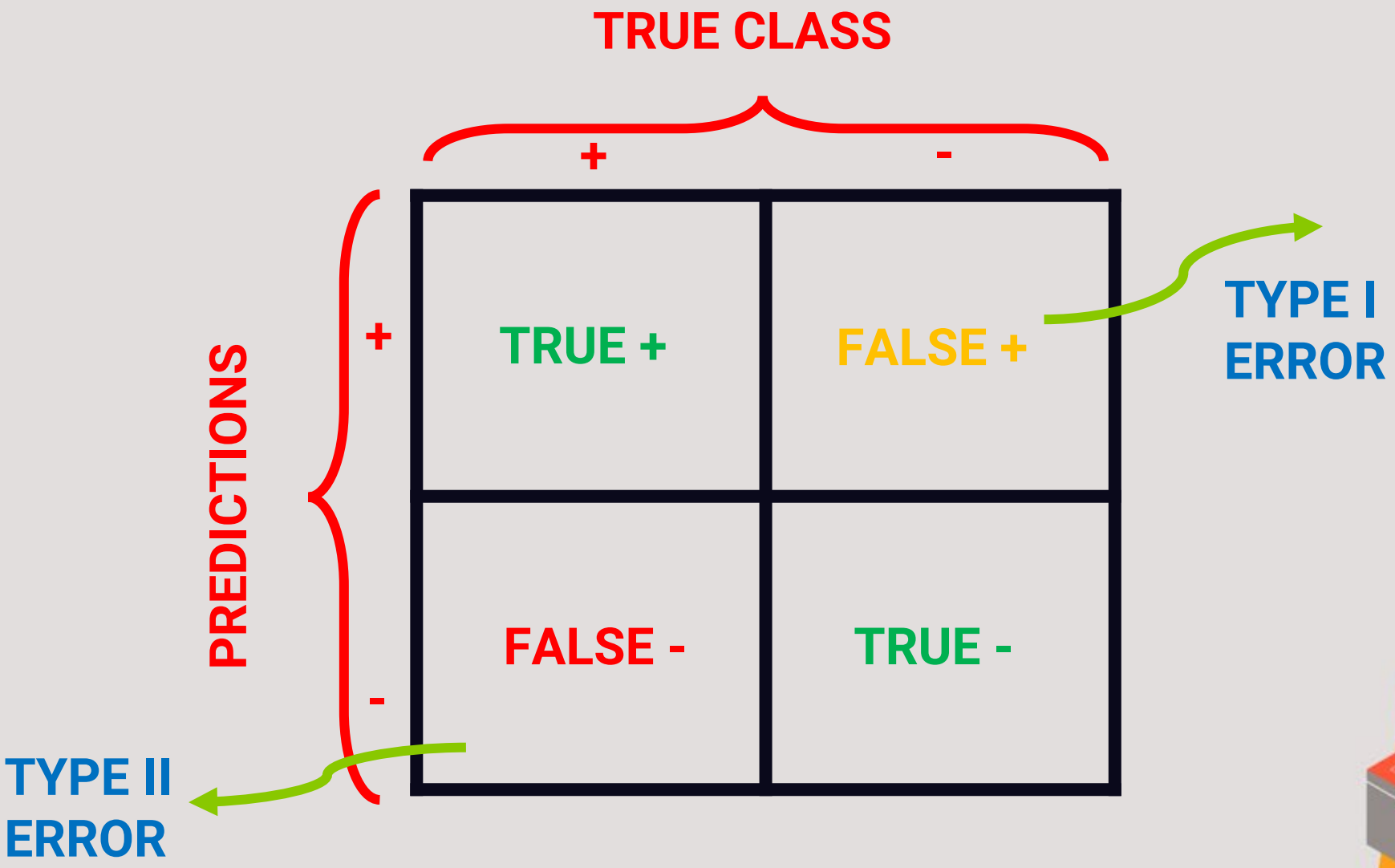


LOGISTIC REGRESSION CLASSIFIER

- **Now we need to convert from a probability to a class value which is “0” or “1”.**



CONFUSION MATRIX

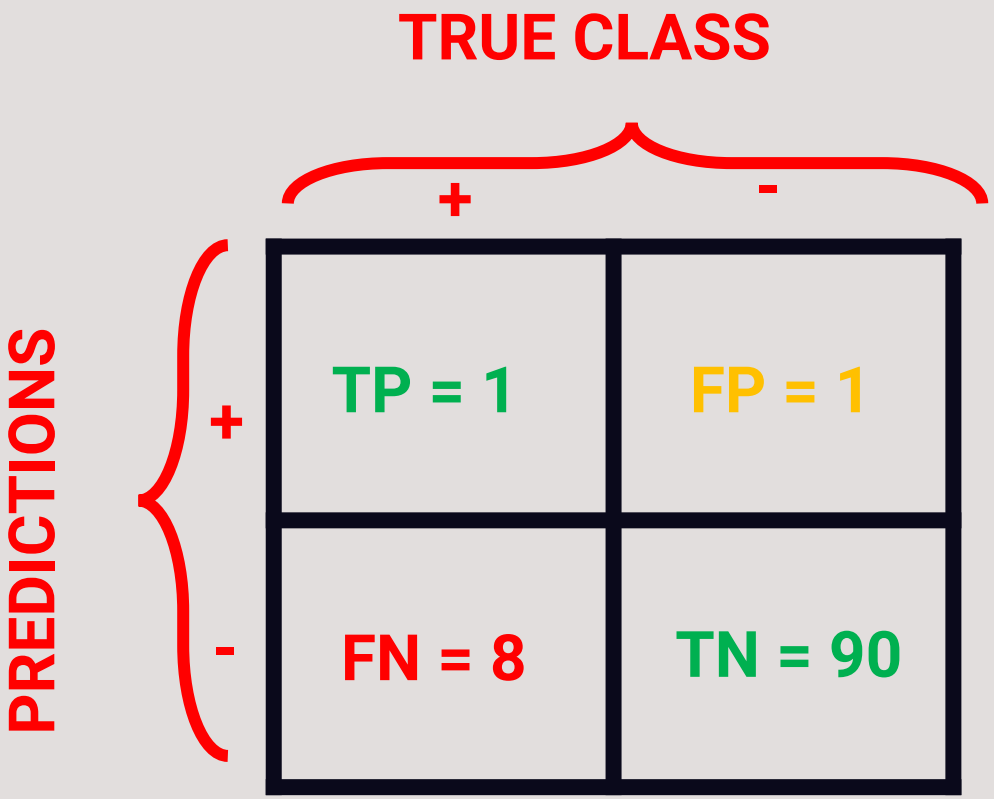


CLASSIFICATION MODEL KPIs

- A confusion matrix is used to describe the performance of a classification model:

    - **True positives (TP):** cases when classifier predicted TRUE and correct class was TRUE
    - **True negatives (TN):** cases when model predicted FALSE, and correct class was FALSE
    - **False positives (FP) (Type I error):** classifier predicted TRUE, but correct class was FALSE
    - **False negatives (FN) (Type II error):** classifier predicted FALSE, but they actually correct.
    - **Classification Accuracy** = (TP+TN) / (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?)

PRECISION VS. RECALL



- Accuracy is generally misleading and is not enough to assess the performance of a classifier.
- Recall is an important KPI

- 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%

F1-SCORE

$$F1\ Score\ = \frac{2\ *(PRECISION\ *RECALL)}{(PRECISION + RECALL)}$$

$$F1\ Score\ = \frac{2\ *TP}{2 * TP + FP + FN}$$

- F1 Score is an overall measure of a model's accuracy that combines precision and recall.
- F1 score is the harmonic mean of precision and recall.
- What is the difference between F1 Score and Accuracy?
- In unbalanced datasets, if we have large number of true negatives (healthy patients), accuracy could be misleading. Therefore, F1 score might be a better KPI to use since it provides a balance between recall and precision in the presence of unbalanced datasets.