

EMPLOYEE ATTRITION PROJECT

A Machine learning 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 90% compared to the Artificial Neural Network and the random forest classifier and metrics summarized.

	precision	recall	f1-score	support
0	0.89	1.00	0.94	305
1	0.96	0.43	0.59	63
accuracy			0.90	368
macro avg	0.93	0.71	0.77	368
weighted avg	0.91	0.90	0.88	368

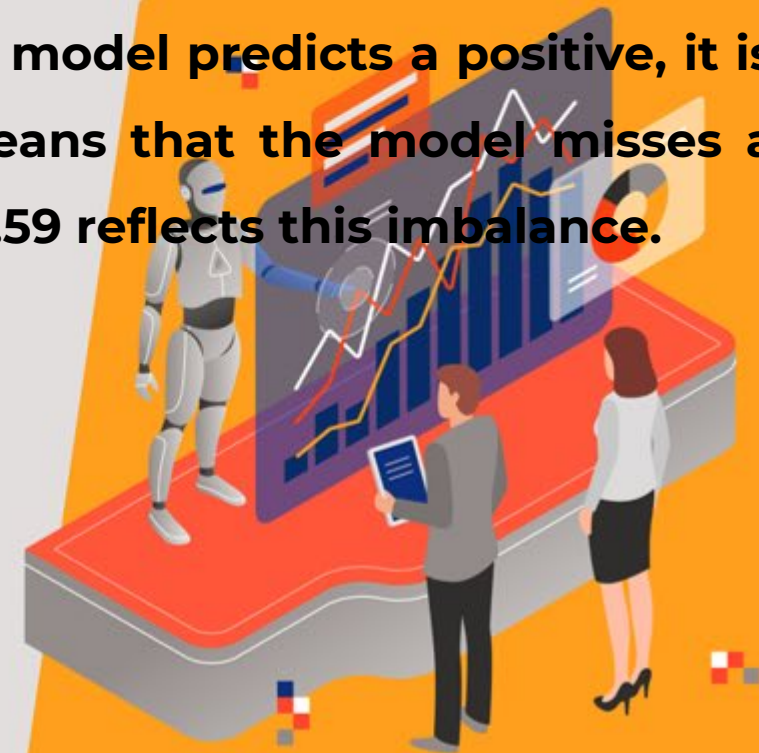


OBSERVATION

The model performs very well for class 0 but has difficulties correctly identifying all instances of class 1.

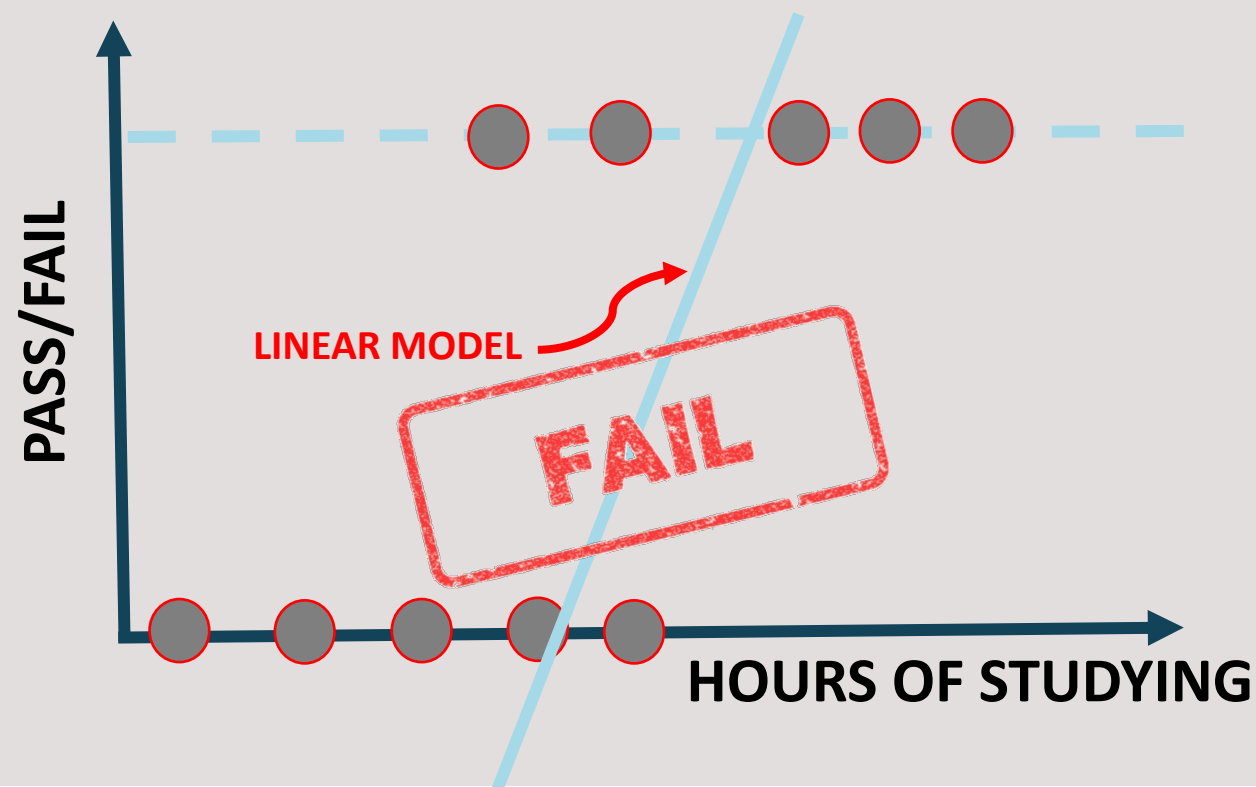
For class 0, the model is very good at identifying negative instances (high recall of 1.00). Precision is also quite high (0.89), meaning most predicted negatives are correct. The F1-score of 0.94 indicates a strong balance between precision and recall for this class.

For class 1, while the precision is high (0.96), meaning when the model predicts a positive, it is very likely to be correct, the recall is quite low (0.43). This means that the model misses a significant number of actual positive instances. The F1-score of 0.59 reflects this imbalance.



LOGISTIC REGRESSION CLASSIFIER EXPLAINED

- **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.

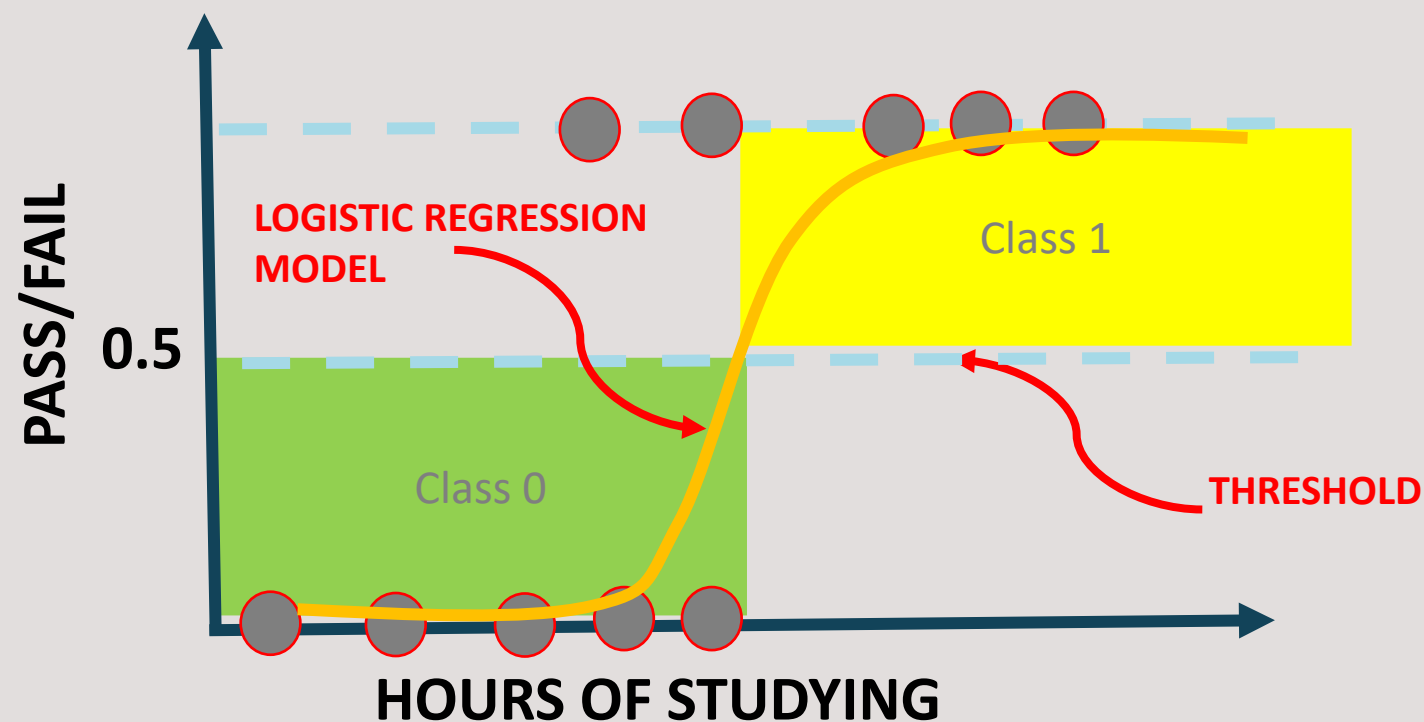


Hours Studying	Pass/Fail
1	0
1.5	0
2	0
3	1
3.25	0
4	1
5	1
6	1

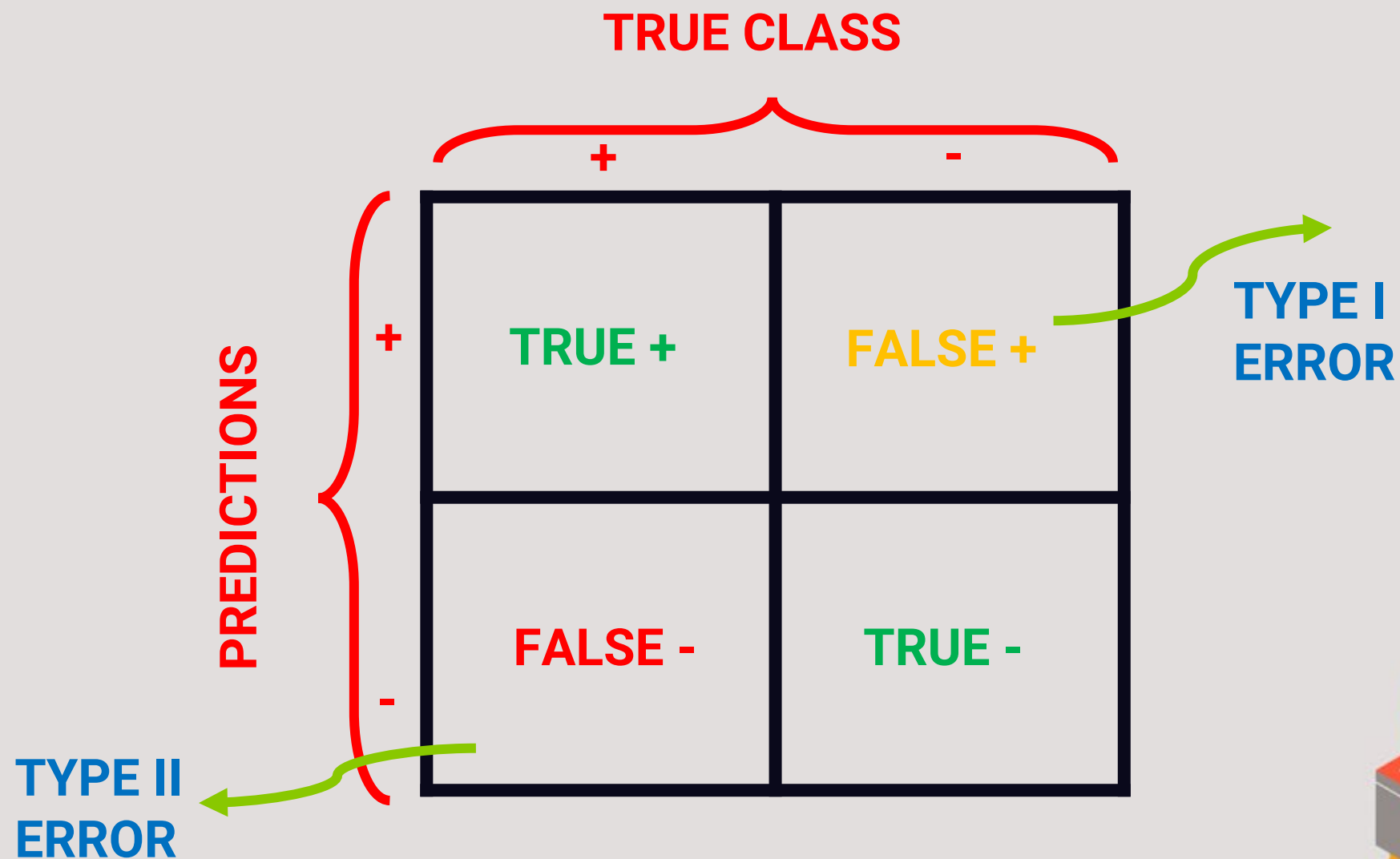


LOGISTIC REGRESSION CLASSIFIER

- 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** = $TP / \text{Total TRUE Predictions} = TP / (TP+FP)$ (When model predicted TRUE class, how often was it right?)
 - **Recall** = $TP / \text{Actual TRUE} = TP / (TP+FN)$ (when the class was actually TRUE, how often did the classifier get it right?)



PRECISION VS. RECALL

		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

- Classification Accuracy = $(TP+TN) / (TP + TN + FP + FN) = 91\%$
- Precision = $TP / \text{Total TRUE Predictions} = TP / (TP+FP) = \frac{1}{2} = 50\%$
- Recall = $TP / \text{Actual TRUE} = TP / (TP+FN) = \frac{1}{9} = 11\%$

