Ghana Space Science & Technology Institute, GAEC

Classification in Deep Learning: A Beginner’s Guide

Theophilus Ansah-Narh, Ph.D.
theophilus.ansah-narh@gaec.gov.gh

February 19, 2025

Outline

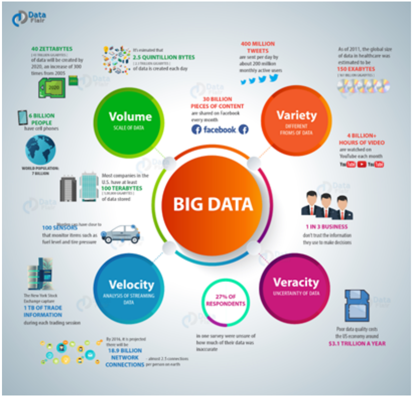
1. Big Data
2. Machine Learning
3. Deep Learning
4. Performance Metrics
5. DEMO
6. Remarks

Introduction

- **Importance of Bankruptcy Prediction:**
  - Accurately evaluating a company's financial health is crucial for stakeholders to mitigate risks and prevent bankruptcy.
  - Effective bankruptcy prediction helps stakeholders make informed decisions and proactively manage financial stability in a rapidly changing business landscape.
- **Overview of the Study's Methodology:**
  - The study presents an innovative approach by integrating Domain Adaptation Learning (DAL) and Genetic Algorithm (GA) techniques.
  - DAL addresses distributional changes in real-world scenarios, while GA excels in feature selection.
  - Six machine learning models are rigorously evaluated against the hybrid model to enhance corporate bankruptcy prediction.

Big Data

The Erra of Big Data I


Data Flair
40 ZETTABYTES
of data will be created by 2020, an increase of 300 times from 2005
2020
It's estimated that
2.5 QUINTILLION BYTES
of data is created each day
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
Volume
SCALE OF DATA
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
facebook
Variety
DIFFERENT FROMS OF DATA
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
YouTube
Most companies in the U.S. have at least
100 TERABYTES
of data stored
BIG DATA
1 IN 3 BUSINESS
don't trust the information they use to make decisions
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
Velocity
ANALYSIS OF STREAMING DATA
The New York Stock Exchange capture
1 TB OF TRADE INFORMATION
during each trading session
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
- almost 2.5 connections per person on earth

Facets and Elements of Big Data. *Image credit*: Dzone website

- **Big data** is an amount of data that is enormous in volume and is constantly expanding rapidly.
- No typical data management systems can effectively store or analyze this data because of its magnitude and complexity.
- fundamental characteristics of big data are listed below

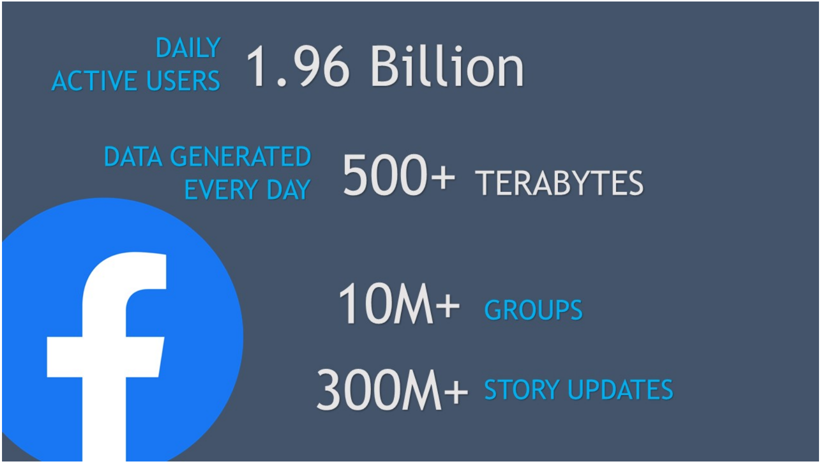
The Erra of Big Data II

Volume

The Erra of Big Data III

- Big Data is a vast *volume* of data generated from many sources daily, such as business processes, machines, social media platforms, networks, human interactions, and many more.
- Industry trends predict a significant increase in data volume over the next few years.
- Usually measured in gigabytes (GB), terabytes (TB), zettabytes (ZB), and yottabytes (YB)
- Nonetheless, Big data generally refers to datasets with a high volume of the order of magnitude of exabytes (10^{18}B $=10^{9}$GB $=10^{6}$TB $=1$EB) and greater (Jelic *et* al. 2019).

The Erra of Big Data IV



DAILY ACTIVE USERS 1.96 Billion

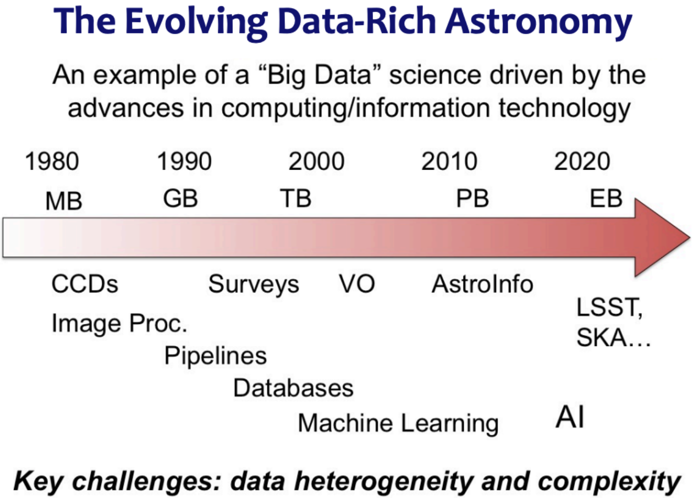
DATA GENERATED EVERY DAY 500+ TERABYTES

10M+ GROUPS

300M+ STORY UPDATES

Image credit: Analytics Vidhya

The Erra of Big Data V



The Evolving Data-Rich Astronomy

An example of a “Big Data” science driven by the advances in computing/information technology

1980 MB

1990 GB

2000 TB

2010 PB

2020 EB

CCDs

Image Proc.

Pipelines

Surveys

Databases

VO

Machine Learning

AstroInfo

AI

LSST, SKA…

Key challenges: data heterogeneity and complexity

Image credit: Djorgovski, 2019

The Erra of Big Data VI

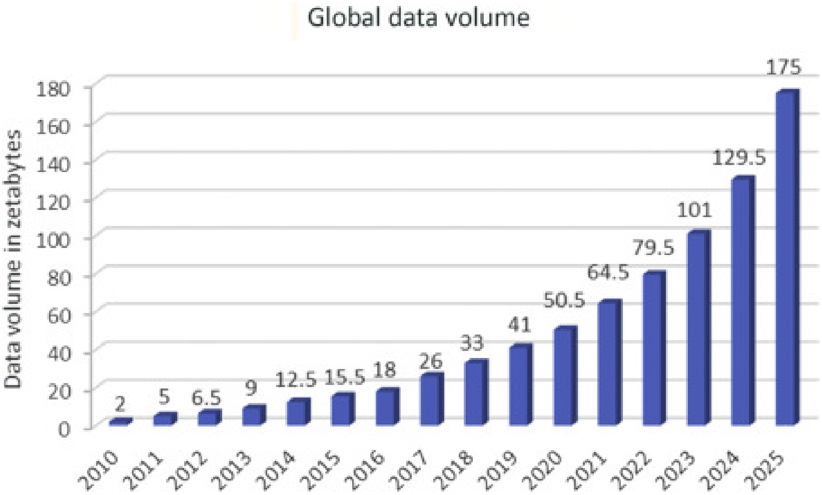


Image credit: Lei & Kong, 2020

The Erra of Big Data VII

Variety

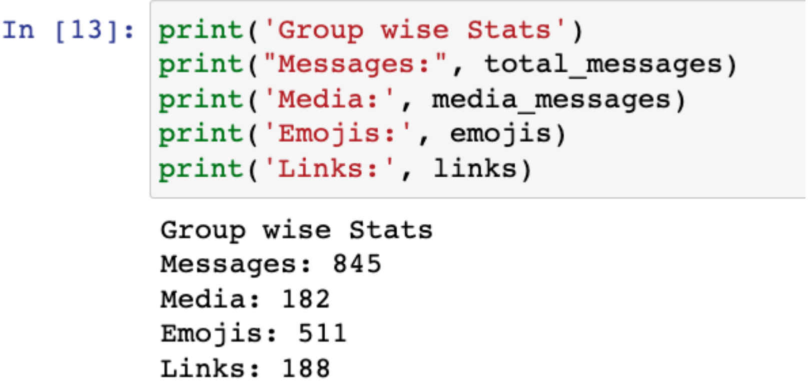
- In the past, data is only collected from **databases** and **sheets**.
- These days the data will come in array forms, that are **PDFs**, **Emails**, **audios**, **SM posts**, **photos**, **videos**, etc.
- Big Data can be **structured**, **unstructured**, and **semi-structured** that are being collected from different sources.
  - **Structured data:** In Structured schema, along with all the required columns. It is in a tabular form. Structured Data is stored in the relational database management system.
  - **Semi-structured:** In Semi-structured, the schema is not appropriately defined, e.g., JSON, XML, CSV, TSV, and email. OLTP (Online Transaction Processing) systems are built to work with semi-structured data. It is stored in relations, i.e., tables.
  - **Unstructured Data:** All the unstructured files, log files, audio files, video files, e-mails, word processing, and image files are included in the unstructured data.

The Erra of Big Data VIII

Veracity

- The accuracy of your findings can be severely harmed by poor data reliability.
- Making it one of the most crucial big data qualities
- There's a need to calibrate your data since most of the data you encounter is unstructured.

The Erra of Big Data IX

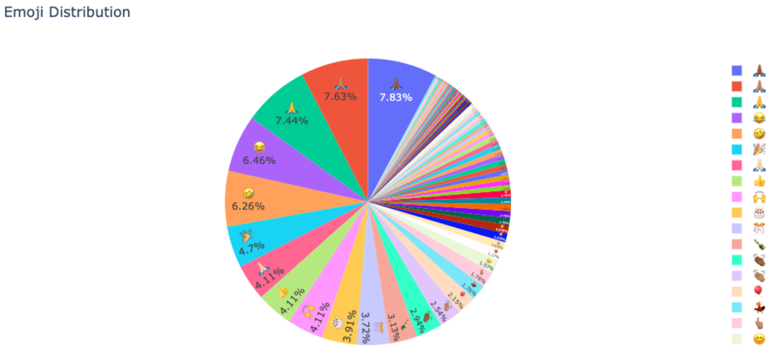


In [13]:

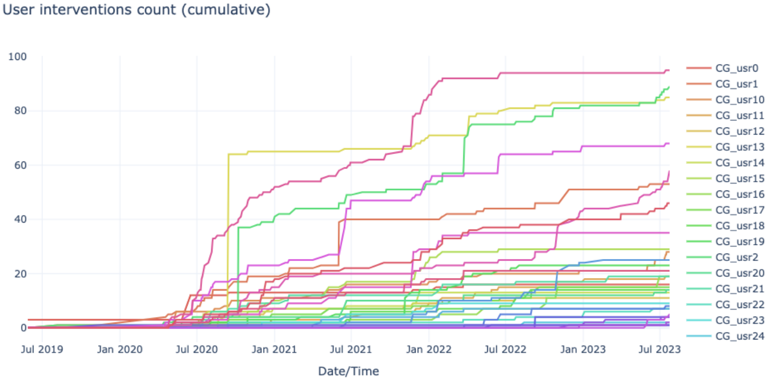
```python
print('Group wise Stats')
print("Messages:", total_messages)
print('Media:', media_messages)
print('Emojis:', emojis)
print('Links:', links)
```

```
Group wise Stats
Messages: 845
Media: 182
Emojis: 511
Links: 188
```

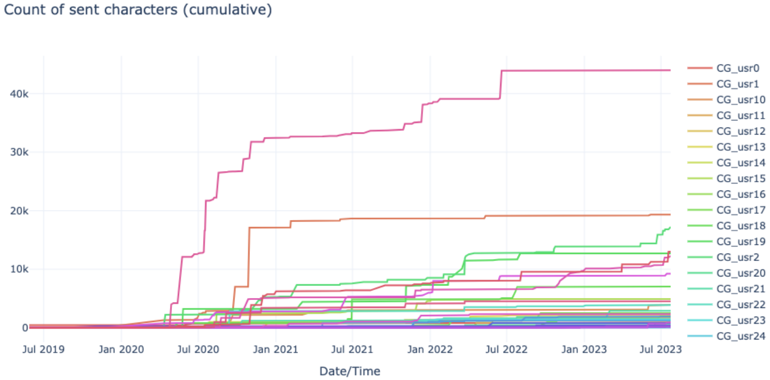
The Erra of Big Data X



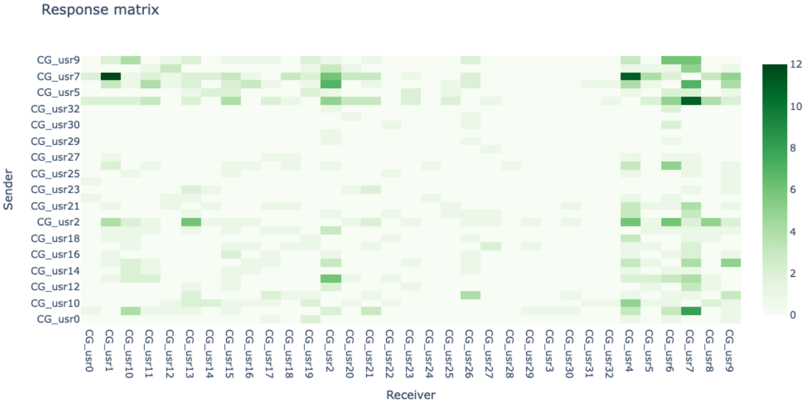
The Erra of Big Data XI



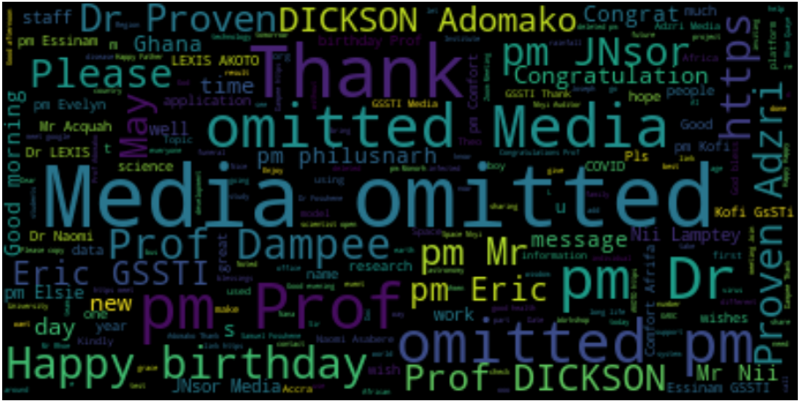
The Erra of Big Data XII



The Erra of Big Data XIII



The Erra of Big Data XIV

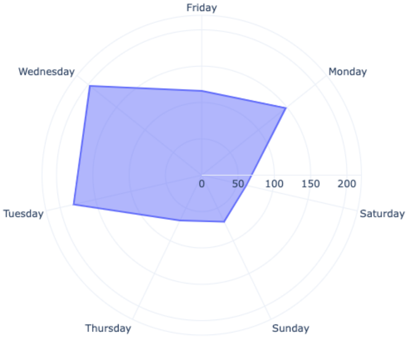


The Erra of Big Data XV

The Erra of Big Data XVI

Thank
GSSTI
https
Happy birthday
Congratulation
Please
Good morning
research
people
message
Noted
link
application
University
name
team
hope
let
test
GAEC
staff
May
deleted
Kindly
time
Dear
wish
using
make
around
Sir
platform
information
funeral
Institute
event
inviting
org
good health
virus
technology
everyone
long life
boy
Good
day
new
work
science
well
Ghana
great
Space
add
COVID
GSSTI Thank
Happy Happy
Enjoy
know
Accra
Africa
God bless
Amen
Nsor
Congrat
year
Topic
God
project
AI
data
one
first
disease
much
model
Pls
wishes
used
way
rainfall
individual
grace
family
system
today
go
help
Workshop
u
part
office
meet
google
countries
blessings
html
best
Space Nkyi
infected
study
Region
sharing
students
astronomy
link https
scientist

The Erra of Big Data XVII



The Erra of Big Data XVIII

Value

- On this data set, analysis and pattern recognition are performed.
- The results of the method may be used to determine the value of the data.
- Making it one of the most crucial big data qualities.

Machine Learning

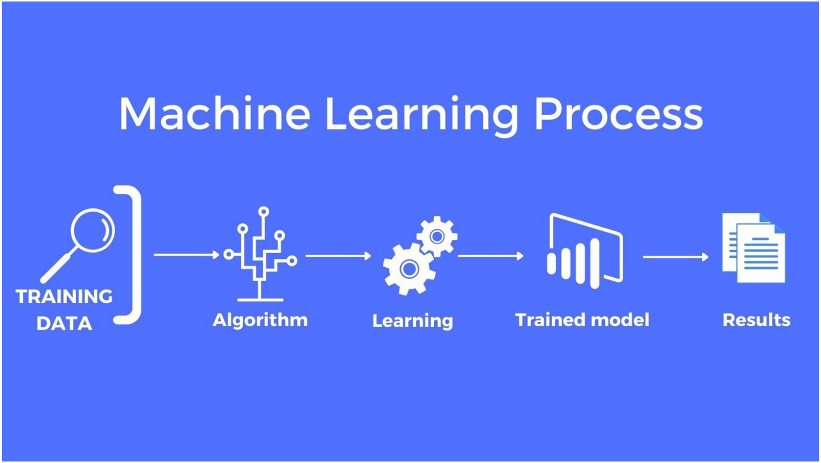


Image credit: mapendo site

Machine Learning II

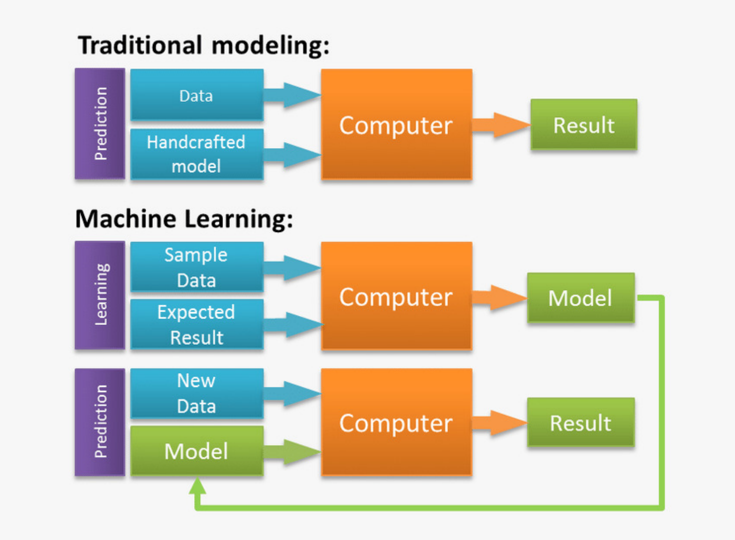


Image credit: Mehra & Hasanuzzaman, (2020)

Deep Learning

Deep Learning I

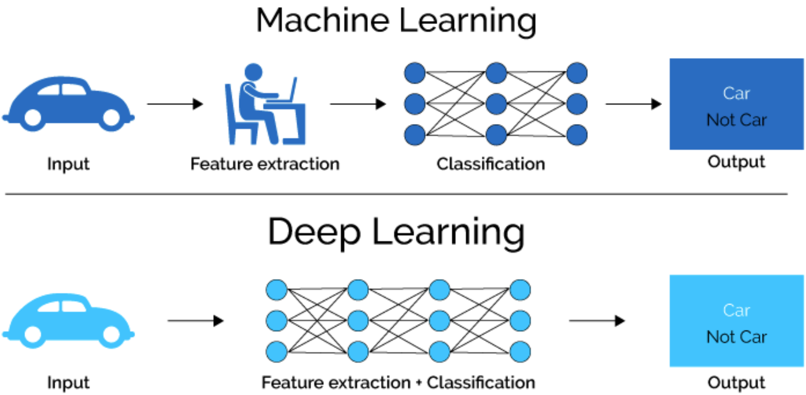
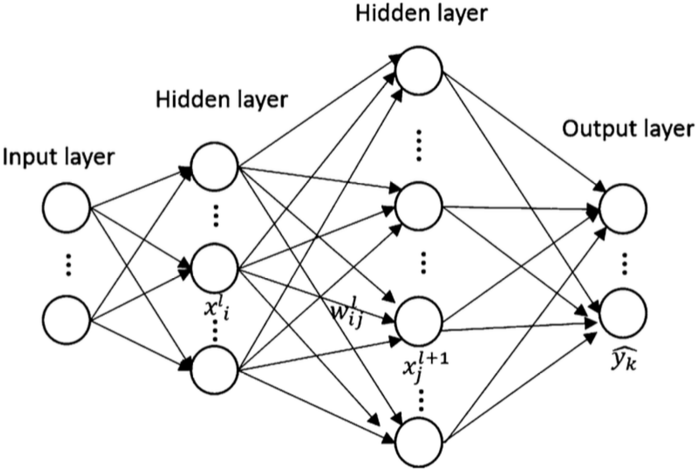


Image credit: Odi & Nguyen, (2018)

Deep Learning II



Deep Learning III

Schematic of a feed-forward neural network

Deep Learning IV

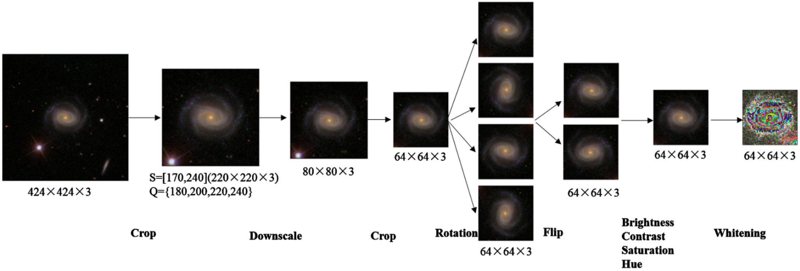
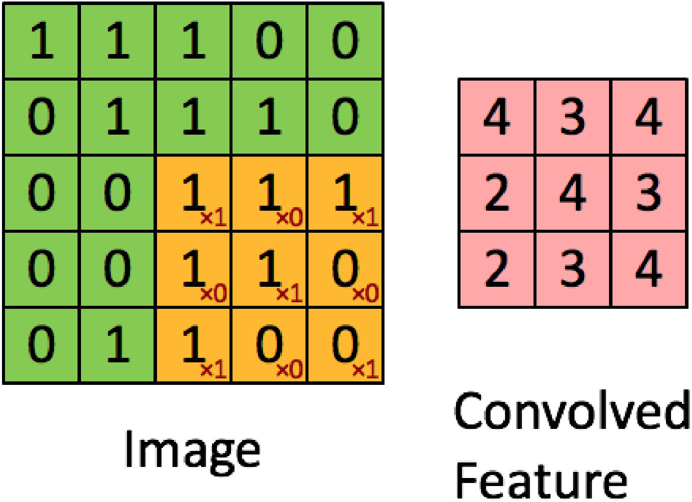


Image credit: Zhu *et* al , (2019)



1	1	1	0	0
0	1	1	1	0
0	0	$1_{\times 1}$	$1_{\times 0}$	$1_{\times 1}$
0	0	$1_{\times 0}$	$1_{\times 1}$	$0_{\times 0}$
0	1	$1_{\times 1}$	$0_{\times 0}$	$0_{\times 1}$

Image

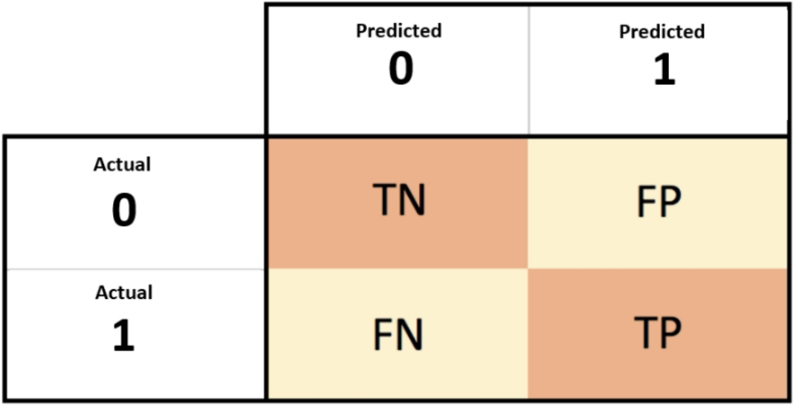
4	3	4
2	4	3
2	3	4

Convolved Feature

Image credit: Medium

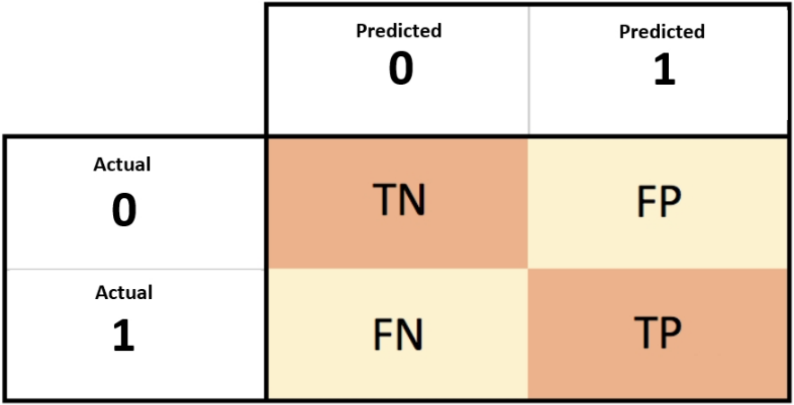
Performance Metrics

Performance Metrics Classification I



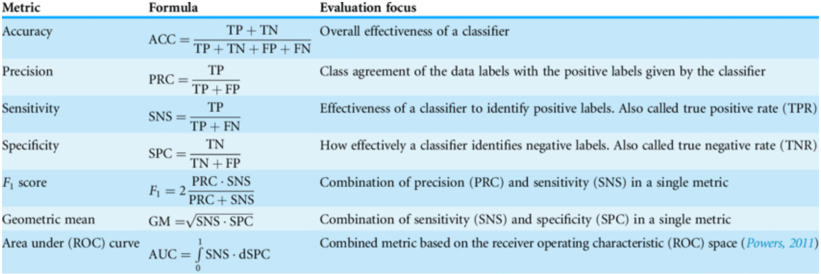
	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Performance Metrics Classification II



	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Performance Metrics Classification III



Metric	Formula	Evaluation focus
Accuracy	$\mathrm{ACC} = \dfrac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$	Overall effectiveness of a classifier
Precision	$\mathrm{PRC} = \dfrac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$	Class agreement of the data labels with the positive labels given by the classifier
Sensitivity	$\mathrm{SNS} = \dfrac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$	Effectiveness of a classifier to identify positive labels. Also called true positive rate (TPR)
Specificity	$\mathrm{SPC} = \dfrac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$	How effectively a classifier identifies negative labels. Also called true negative rate (TNR)
F_1 score	$F_1 = 2\dfrac{\mathrm{PRC} \cdot \mathrm{SNS}}{\mathrm{PRC} + \mathrm{SNS}}$	Combination of precision (PRC) and sensitivity (SNS) in a single metric
Geometric mean	$\mathrm{GM} = \sqrt{\mathrm{SNS} \cdot \mathrm{SPC}}$	Combination of sensitivity (SNS) and specificity (SPC) in a single metric
Area under (ROC) curve	$\mathrm{AUC} = \int_0^1 \mathrm{SNS} \cdot d\mathrm{SPC}$	Combined metric based on the receiver operating characteristic (ROC) space (Powers, 2011)

DEMO

Remarks

Workflow of ML

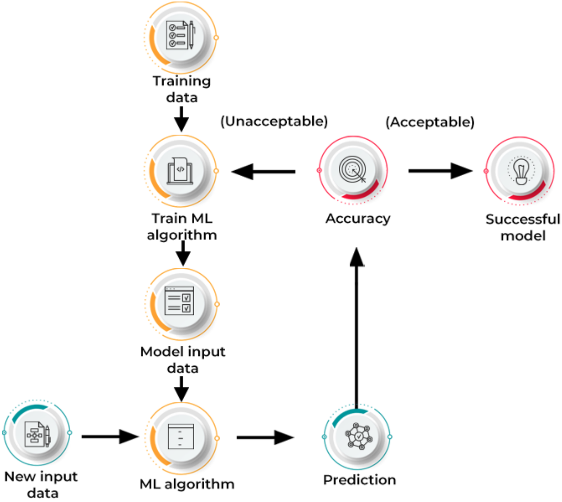


Image credit: spiceworks site

That's All Folks