Day 1 & 2: Modeling Overview and Industry Applications

Session Overview

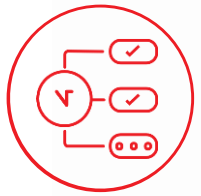
Modeling Overview and Industry Applications

This is an introductory session on modeling overview, which will help you understand the application of modeling in different aspects of the business domain like:

- Insurance
- Banking
- Healthcare
- Procurement

In this session, we will cover modeling terminologies and the different types of modeling such as:

- Small data and large data
- Use cases in various business domains
- What is Model, How it is Build, and Associated Terminologies



Session Outcomes

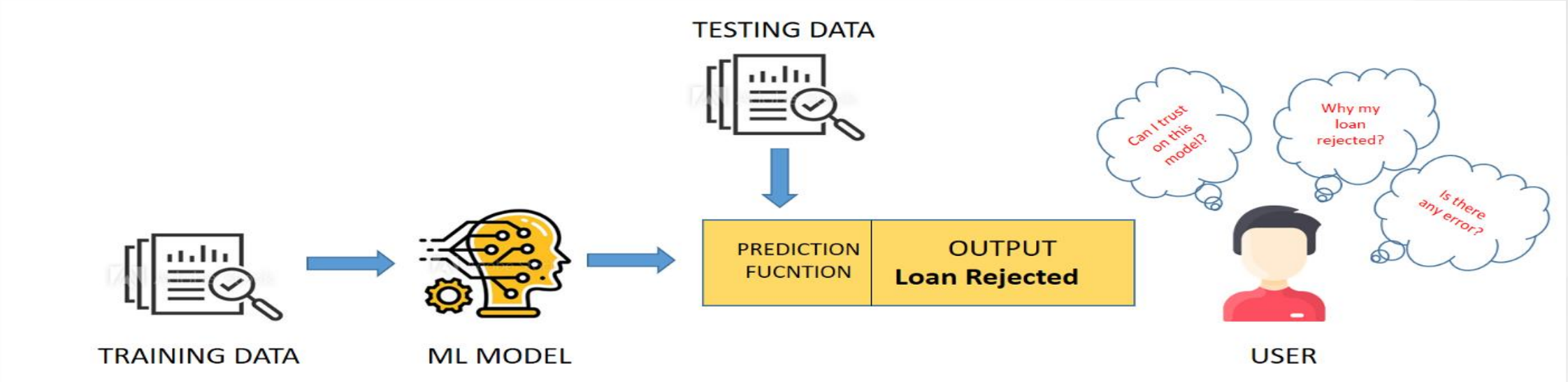
Modeling Overview and Industry Applications

On completion of the session, you should be able to:

- Describe statistical models
- Discuss about machine learning and large data sets
- Identify different types of machine learning – supervised, unsupervised, reinforcement learning
- Describe the real-world applications of machine learning
- Identify the methods used in machine learning such as train, test, validation, prediction etc.

Machine Learning Models

- Assume there is an activity in a bank to determine whether a customer is eligible for a loan or not. To examine and give the approval, bank employees were entrusted to do it manually. Now, to make the process faster, we are moving to data-driven decision-making algorithms, which are called **models**.
- Previously, bank employees handled loan processing, but now they rely on **machine learning models**, which are a set of algorithms that learn patterns in pre-historical customer records and predict whether a new customer is eligible for a loan.



What Is a Model?

Assume you want a machine to determine whether a patient is diabetic. You must collect many observations, including patient records and a variety of data. Let's have a look at the data.

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1

Now, if the model is learning these several patterns in a random manner, then the algorithm will be able to predict whether, under the specified conditions, the patient will have diabetes or not. And it becomes very easy for the healthcare department to make data-driven decisions.

Hero

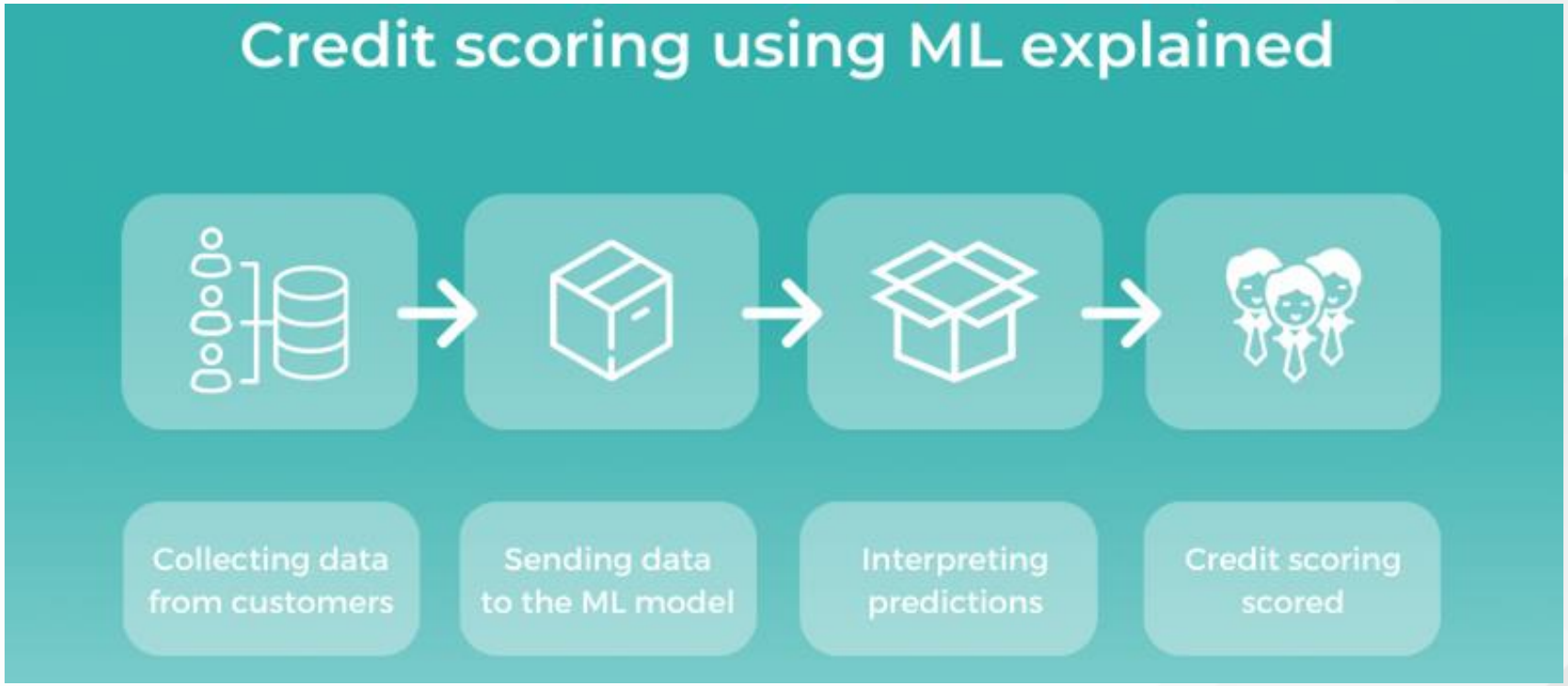
Implementation of the modeling Techniques

- These modeling techniques are implemented in Python. But this can also be developed in other languages too.
- The reason we prefer Python to implement is because it is a popular and general-purpose programming language.
- We can write machine learning algorithms using Python, and they work well.
- Python is popular among data scientists is because it has a diverse variety of modules and libraries already implemented that make our lives more comfortable..

Some Recent Applications of Modeling Techniques

Credit Scoring

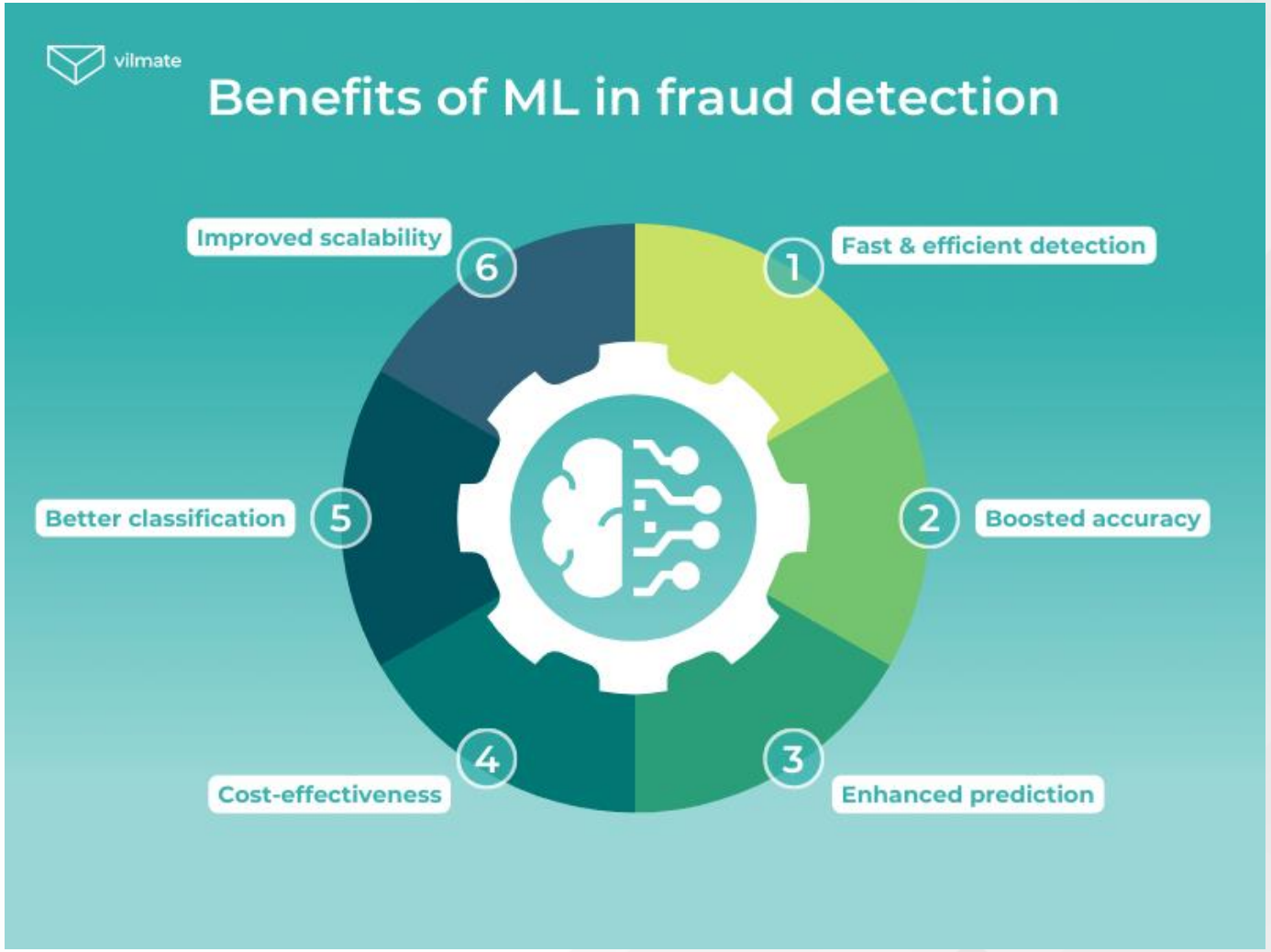
- Analytics is one of the biggest advantages of machine learning technology in the financial industry. And the fields of decision-making and credit scoring are getting the most benefits from it.
- Banking institutions and other fintech firms use machine learning algorithms to perform their operations on the ML-based credit scoring system.



Source

Security, Risk Management, and Fraud Detection

- The three biggest concerns for fintech companies are fraud, security, and safety.
- One should be perfectly aware of when and what to buy, what to sell, and when to hold in the market. The analytics feature in AI and ML makes these tasks a lot easier.



Source

Quantitative and Algorithmic Trading

Building statistical models based on historical data can lead to more profitable investments faster than competitors.

Source

Statistical Models

- Statistical modeling refers to the data science process of applying statistical analysis to datasets.
- A statistical model is a mathematical relationship between one or more random variables and other non-random variables.

source

Why Statistical Model Is Used on Small Data?

- Statistical models work with finite data sets and a reasonable number of observations.
- Increasing the data might lead to overfitting (when statistical models fit against their training data).
- On the contrary, machine learning models need vast amounts of data to learn and perform intelligent actions.

Small Dataset Example

Height	Hair	Eyes	Class
short	blonde	blue	+
short	dark	blue	-
tall	dark	brown	-
tall	blonde	brown	-
tall	dark	blue	-
short	blonde	brown	-
tall	red	blue	+
tall	blonde	blue	+

When Should You Use Statistical Modeling?

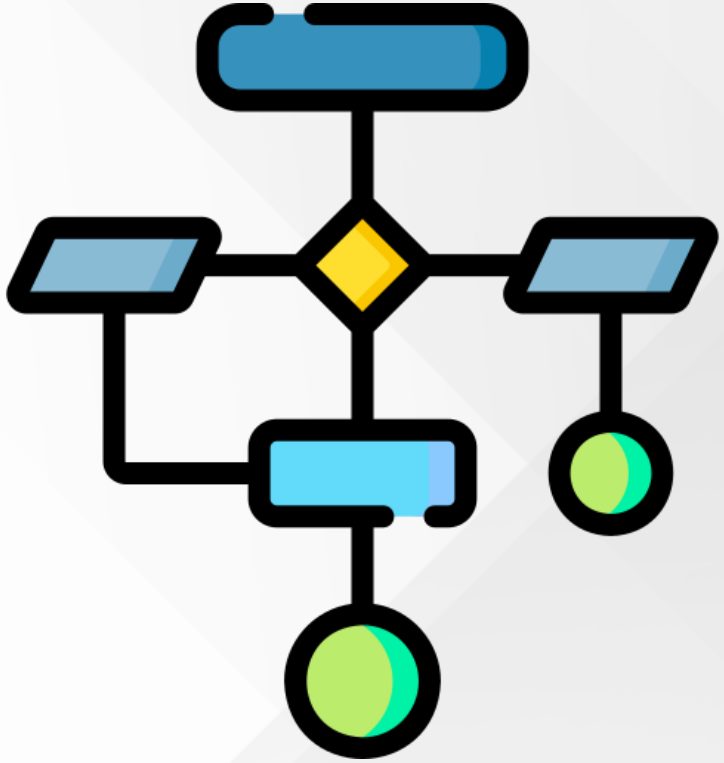
- When data volume isn’t too large.
- Errors and uncertainties in prediction are reasonable.
- Independent variables have fewer, pre-specified interactions.
- When it requires high interpretability.
- When it requires the isolation of the effects of a small number of variables.

Example

- If a content marketing agency wants to build a model to track an audience's journey, it will prefer a statistical model because, rather than accuracy, the interpretation is important.
- It would help to develop an engagement strategy based on business domain knowledge.

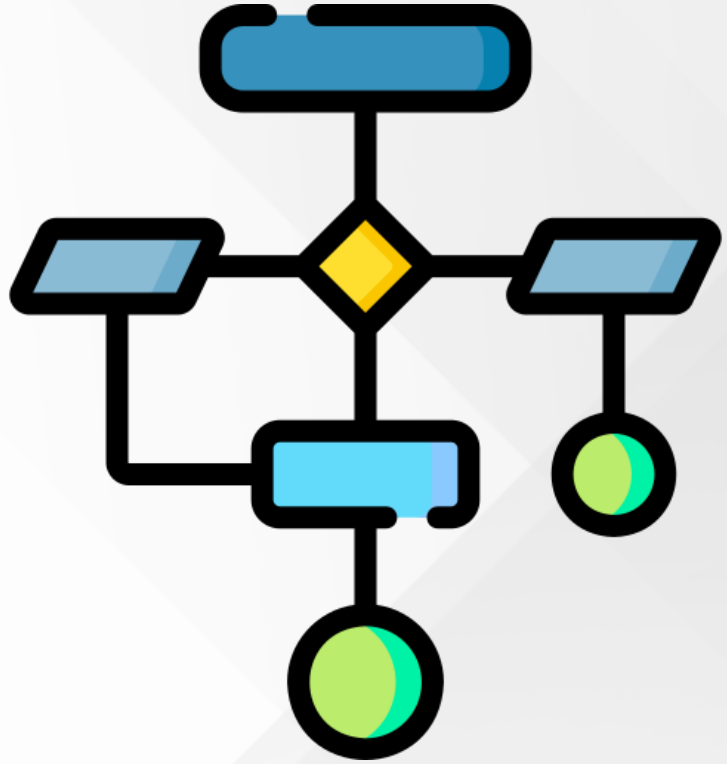
Machine Learning

- Machine learning models are computer programs that are used to recognize patterns in data or make predictions.
- Models are created from algorithms, which are trained using either labeled, unlabeled, or mixed data.
- Different machine learning algorithms are suited to different goals, such as classification or prediction modeling, so data scientists use different algorithms as the basis for different models.



Decision-Making and Credit Scoring in ML

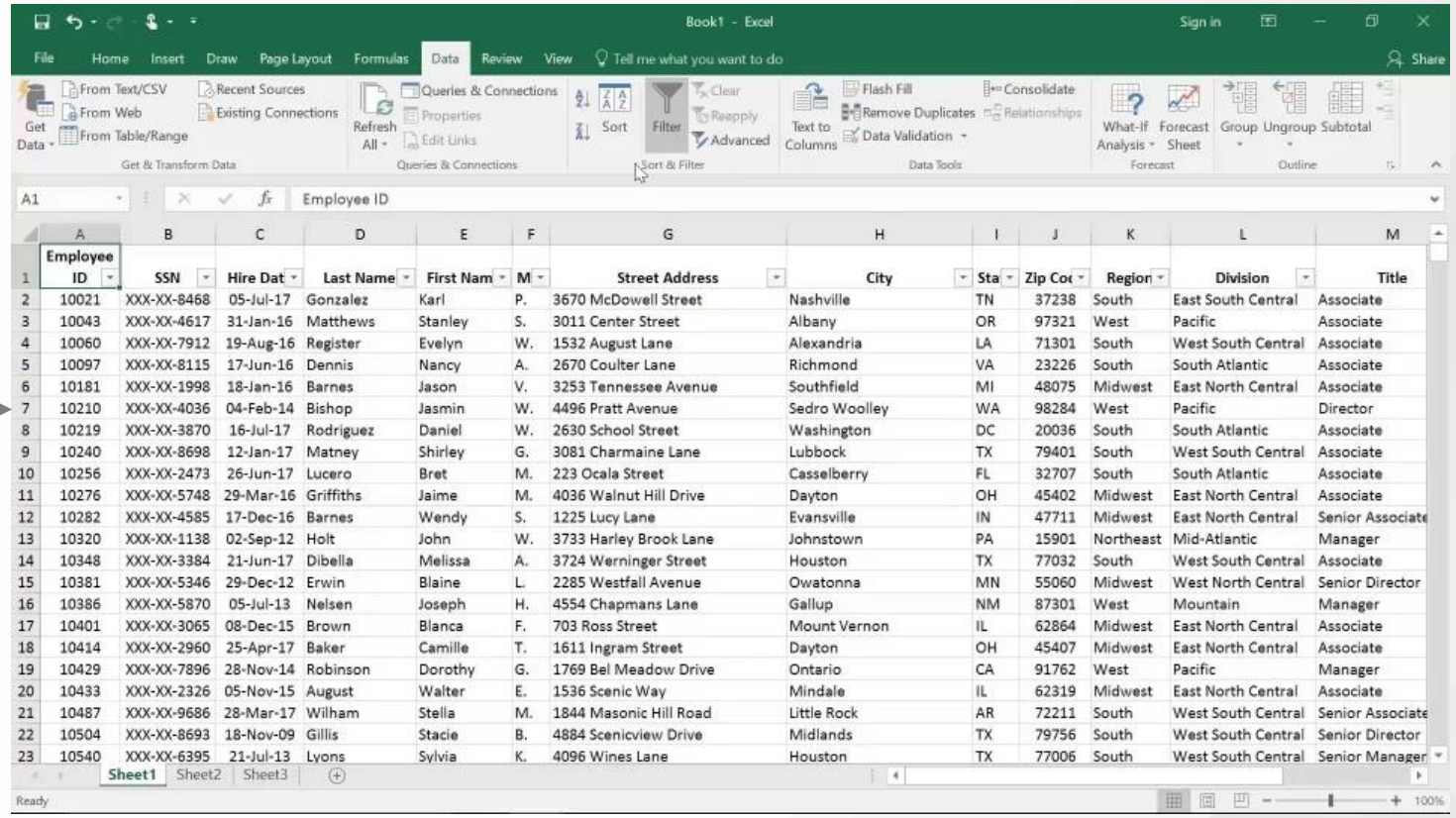
- Analytics is one of the biggest advantages that machine learning technology provides in the financial industry. And the fields of decision-making and credit scoring are getting the most benefits from it.
- Banking institutions & other fintech firms use machine learning algorithms to render their operations on the ML Based credit scoring system.



Why Is a Machine Learning Model Applied to Large Amounts of Data?

More complex algorithms always require a larger amount of data. If your project needs standard ML algorithms that use structured learning, a smaller amount of data will be enough. Even if you feed the algorithm more data than it needs, the results won't improve drastically.

Large Dataset Example

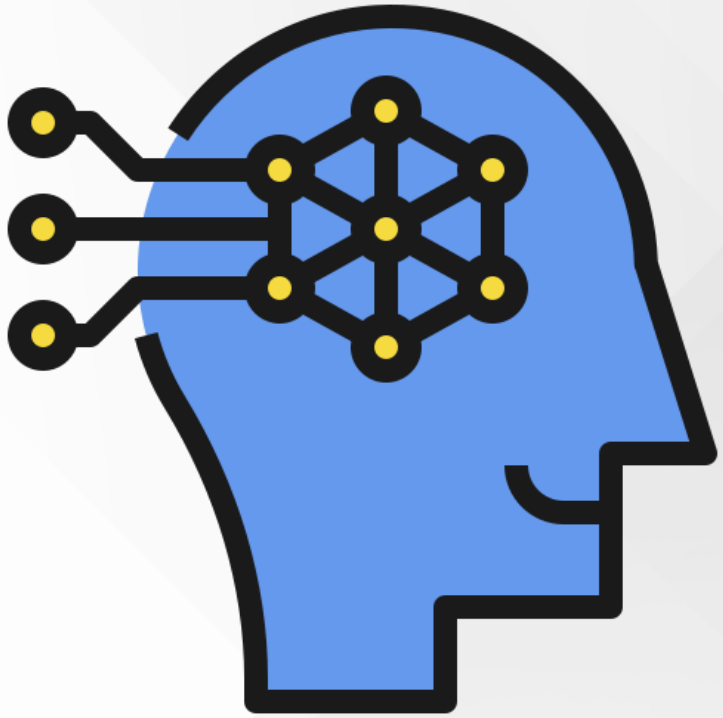


Book1 - Excel

Employee ID	SSN	Hire Dat	Last Name	First Nam	M	Street Address	City	Sta	Zip Co	Region	Division	Title
10021	XXX-XX-8468	05-Jul-17	Gonzalez	Karl	P.	3670 McDowell Street	Nashville	TN	37238	South	East South Central	Associate
10043	XXX-XX-4617	31-Jan-16	Matthews	Stanley	S.	3011 Center Street	Albany	OR	97321	West	Pacific	Associate
10060	XXX-XX-7912	19-Aug-16	Register	Evelyn	W.	1532 August Lane	Alexandria	LA	71301	South	West South Central	Associate
10097	XXX-XX-8115	17-Jun-16	Dennis	Nancy	A.	2670 Coulter Lane	Richmond	VA	23226	South	South Atlantic	Associate
10181	XXX-XX-1998	18-Jan-16	Barnes	Jason	V.	3253 Tennessee Avenue	Southfield	MI	48075	Midwest	East North Central	Associate
10210	XXX-XX-4036	04-Feb-14	Bishop	Jasmin	W.	4496 Pratt Avenue	Sedro Woolley	WA	98284	West	Pacific	Director
10219	XXX-XX-3870	16-Jul-17	Rodriguez	Daniel	W.	2630 School Street	Washington	DC	20036	South	South Atlantic	Associate
10240	XXX-XX-8698	12-Jan-17	Matney	Shirley	G.	3081 Charmaine Lane	Lubbock	TX	79401	South	West South Central	Associate
10256	XXX-XX-2473	26-Jun-17	Lucero	Bret	M.	223 Ocala Street	Casselberry	FL	32707	South	South Atlantic	Associate
10276	XXX-XX-5748	29-Mar-16	Griffiths	Jaime	M.	4036 Walnut Hill Drive	Dayton	OH	45402	Midwest	East North Central	Associate
10282	XXX-XX-4585	17-Dec-16	Barnes	Wendy	S.	1225 Lucy Lane	Evansville	IN	47711	Midwest	East North Central	Senior Associate
10320	XXX-XX-1138	02-Sep-12	Holt	John	W.	3733 Harley Brook Lane	Johnstown	PA	15901	Northeast	Mid-Atlantic	Manager
10348	XXX-XX-3384	21-Jun-17	Dibella	Melissa	A.	3724 Werninger Street	Houston	TX	77032	South	West South Central	Associate
10381	XXX-XX-5346	29-Dec-12	Erwin	Blaine	L.	2285 Westfall Avenue	Owatonna	MN	55060	Midwest	West North Central	Senior Director
10386	XXX-XX-5870	05-Jul-13	Nelsen	Joseph	H.	4554 Chapmans Lane	Gallup	NM	87301	West	Mountain	Manager
10401	XXX-XX-3065	08-Dec-15	Brown	Blanca	F.	703 Ross Street	Mount Vernon	IL	62864	Midwest	East North Central	Associate
10414	XXX-XX-2960	25-Apr-17	Baker	Camille	T.	1611 Ingram Street	Dayton	OH	45407	Midwest	East North Central	Associate
10429	XXX-XX-7896	28-Nov-14	Robinson	Dorothy	G.	1769 Bel Meadow Drive	Ontario	CA	91762	West	Pacific	Manager
10433	XXX-XX-2326	05-Nov-15	August	Walter	E.	1536 Scenic Way	Mindale	IL	62319	Midwest	East North Central	Associate
10487	XXX-XX-9686	28-Mar-17	Wilham	Stella	M.	1844 Masonic Hill Road	Little Rock	AR	72211	South	West South Central	Senior Associate
10504	XXX-XX-8693	18-Nov-09	Gillis	Stacie	B.	4884 Scenicview Drive	Midlands	TX	79756	South	West South Central	Senior Director
10540	XXX-XX-6395	21-Jul-13	Lyons	Sylvia	K.	4096 Wines Lane	Houston	TX	77006	South	West South Central	Senior Manager

When Should You Use Machine Learning?

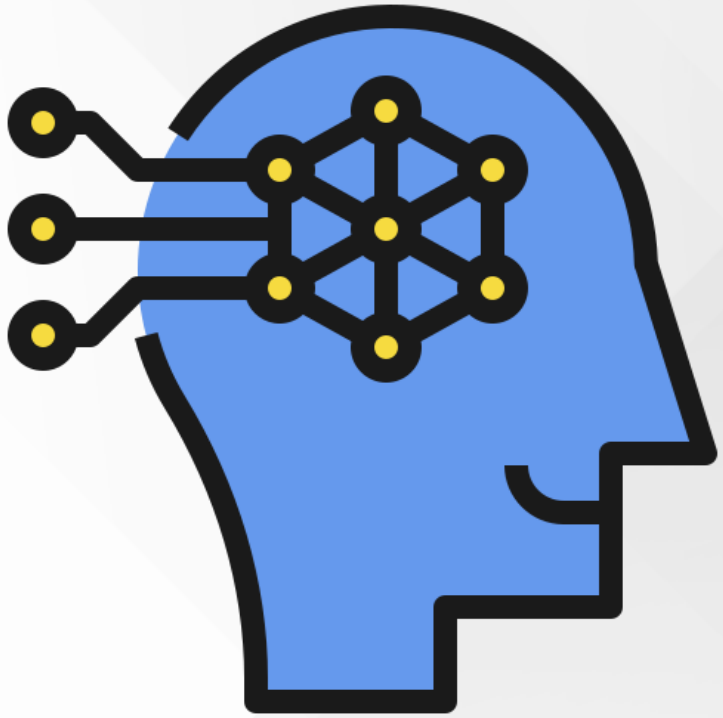
- Training learning algorithms on infinite data replications
- The goal is to get overall predictions, not relationships between variables.
- Estimating uncertainties in forecasts isn’t essential.
- The effect of any variable doesn’t need to be isolated.



Example

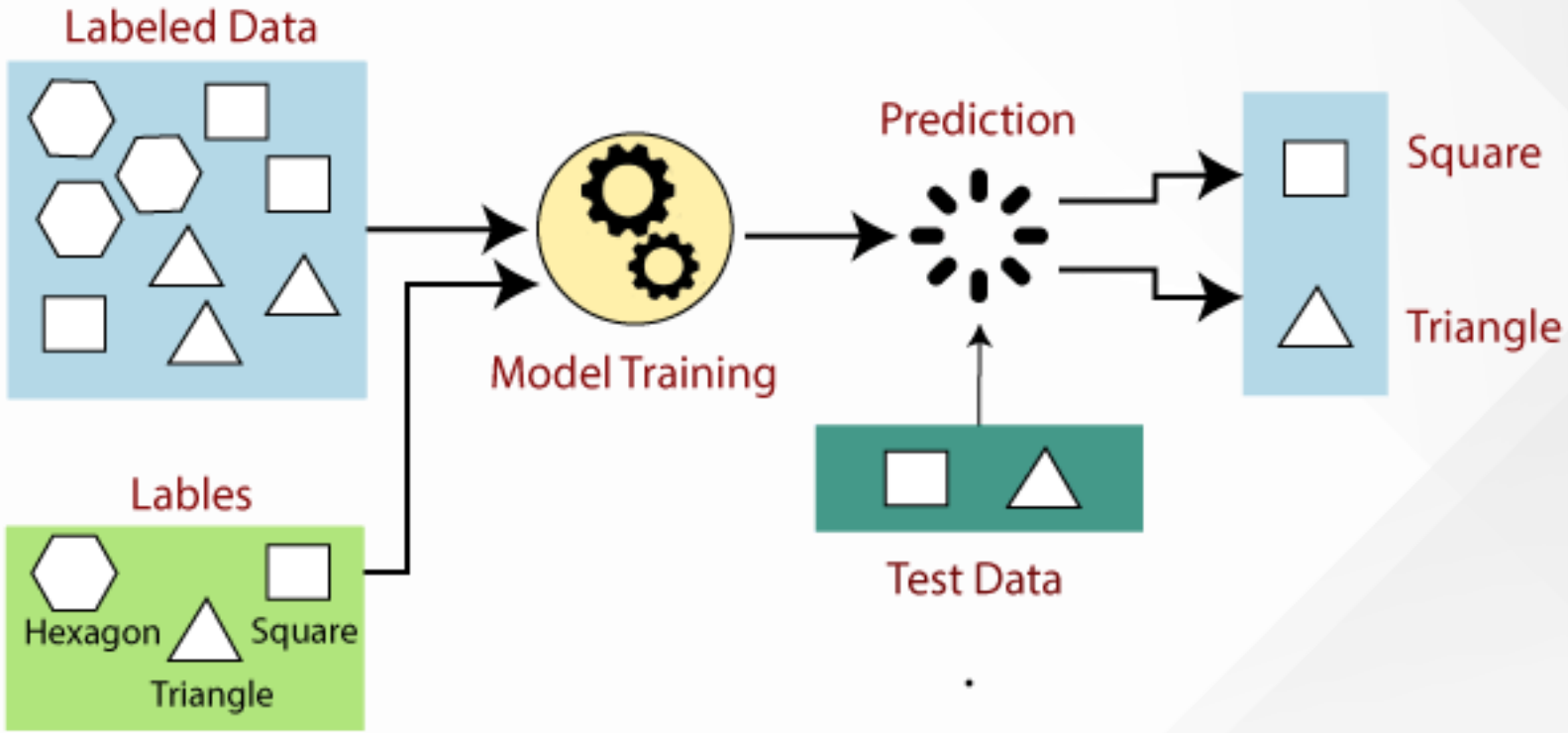
When an e-commerce website such as Amazon wants to recommend products based on previous purchases, they need a powerful recommendation engine. Here, the need for predictive accuracy is more important than the model's interpretability.

In this case, we use a machine learning model.



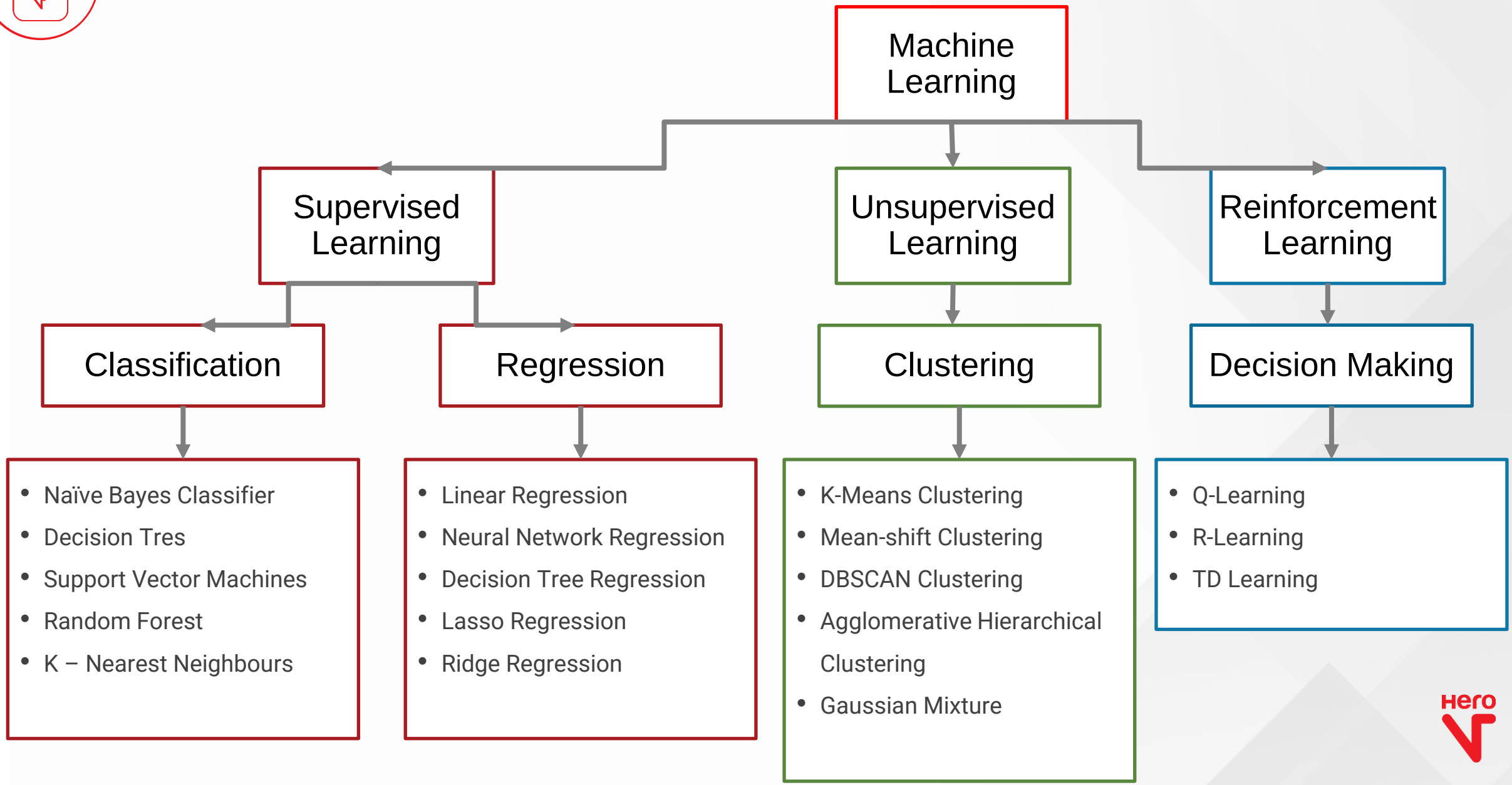
Mechanism of Machine Learning

- Machine learning is a subfield of artificial intelligence (AI).
- Machine Learning aims to understand the structure of data and fit that data into models that can be understood and utilized by people.



Source

Types of Machine Learning



Industry Specific Applications - Machine Learning

Machine Learning Applications in Finance

- Analysts use machine learning in the financial services sector to automate trading activities, detect fraud, and provide financial advising services to their clients.

Machine Learning Applications in Business

- Machine learning offers businesses an extensive number of ways to boost their effectiveness, efficiency, and offerings. Chatbots, for example, allows businesses to provide faster, more flexible customer service without employing a call center or making customers wait on hold for the next available representative.

Machine Learning Applications in Retail

- As we've already mentioned, machine learning can be incredibly helpful in understanding and decreasing customer churn (i.e., the rate at which a business loses customers each year), which is a large focus point for many retail companies.

Real-World Examples: Machine Learning

Image recognition

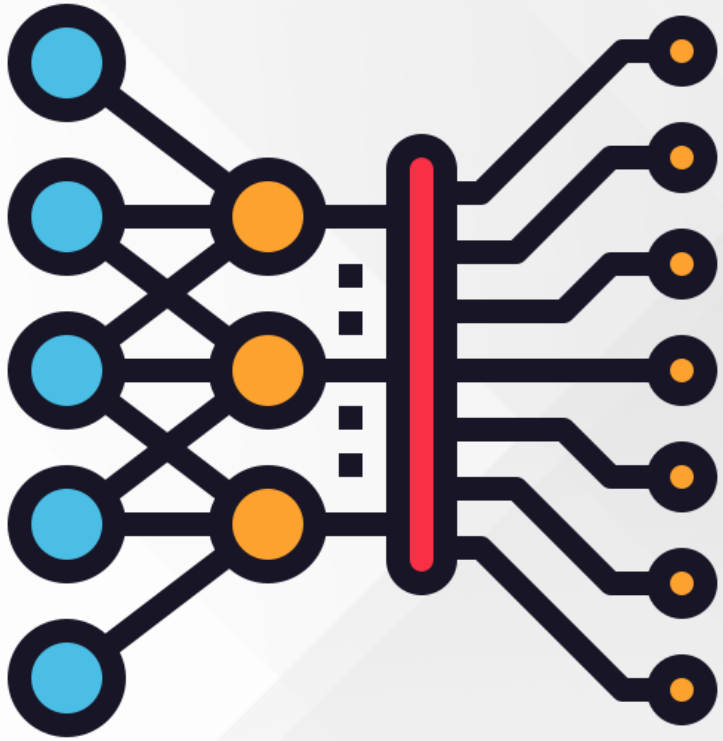
Image recognition is a well-known and widespread example of machine learning in the real world. It can identify an object as a digital image based on the intensity of the pixels in black-and-white or color images.

Speech recognition

Machine learning can translate speech into text. Certain software applications can convert live voice and recorded speech into a text file.

Medical diagnosis

Machine learning can help with the diagnosis of diseases. Many physicians use chatbots with speech recognition capabilities to discern patterns in symptoms.



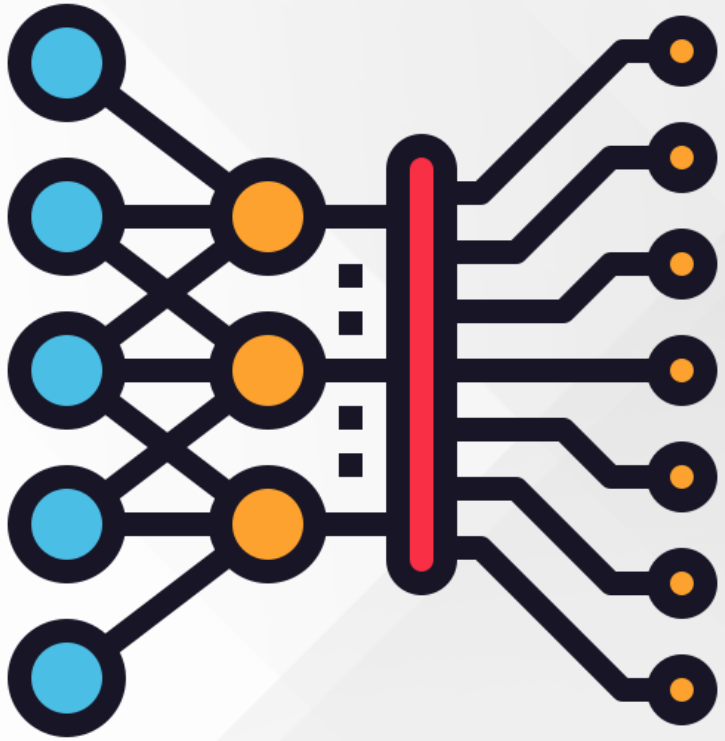
Real-World Examples: Machine Learning

Statistical arbitrage

Arbitrage is an automated trading strategy used in finance to manage a large volume of securities.

Predictive analytics

Machine learning can classify available data into groups, which are then defined by rules set by analysts.



Quiz

What is the difference between supervised learning and unsupervised learning?

1. Dataset with target labels.
2. Dataset without target labels.

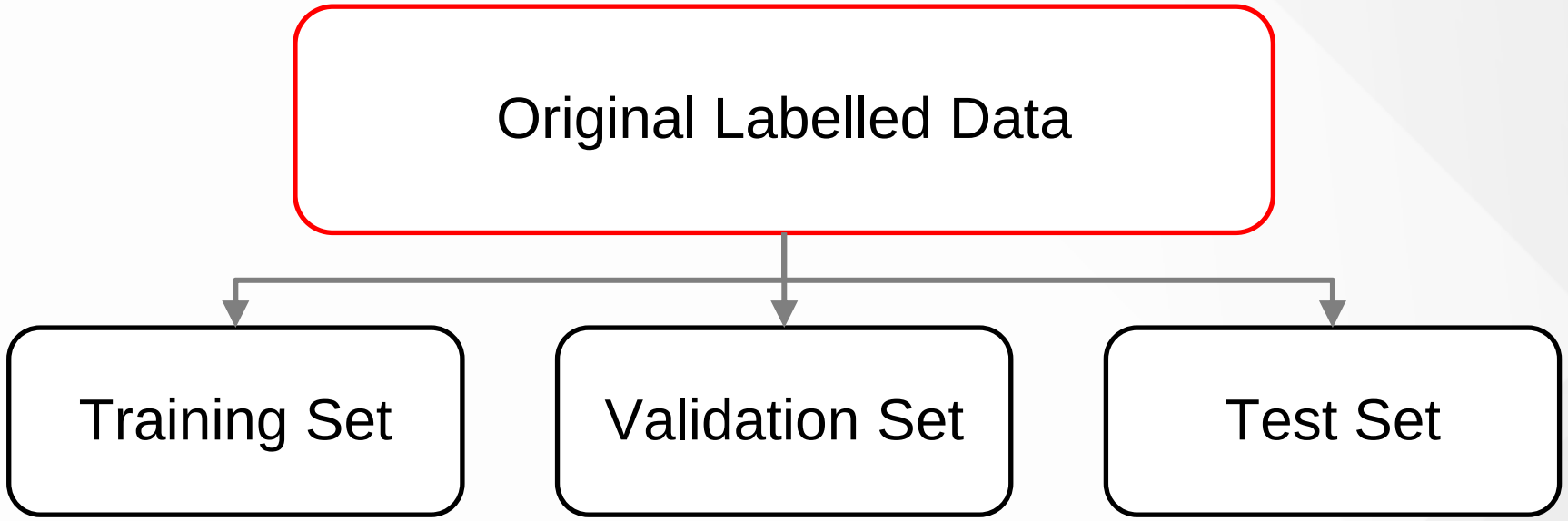
Machine learning model requires small data in complex algorithms.

1. True
2. False

Statistical models work on _____ data/ observations.

Methods of Machine Learning

Data Splitting Techniques



Methods of Machine Learning

Data Splitting Techniques

- **Training set**: The data you will use to train your model will be fed into an algorithm that generates a model. This model maps inputs to outputs.
- **Validation set**: This is smaller than the training set and is used to evaluate the performance of models with different hyperparameter values. It's also used to detect overfitting during the training stages.
- **Test set**: This set is used to get an idea of the final performance of a model after hyperparameter tuning. It's also useful to get an idea of how different models (SVMs, neural networks, random forests, etc.) perform against each other.

Random Selection

Random Selection

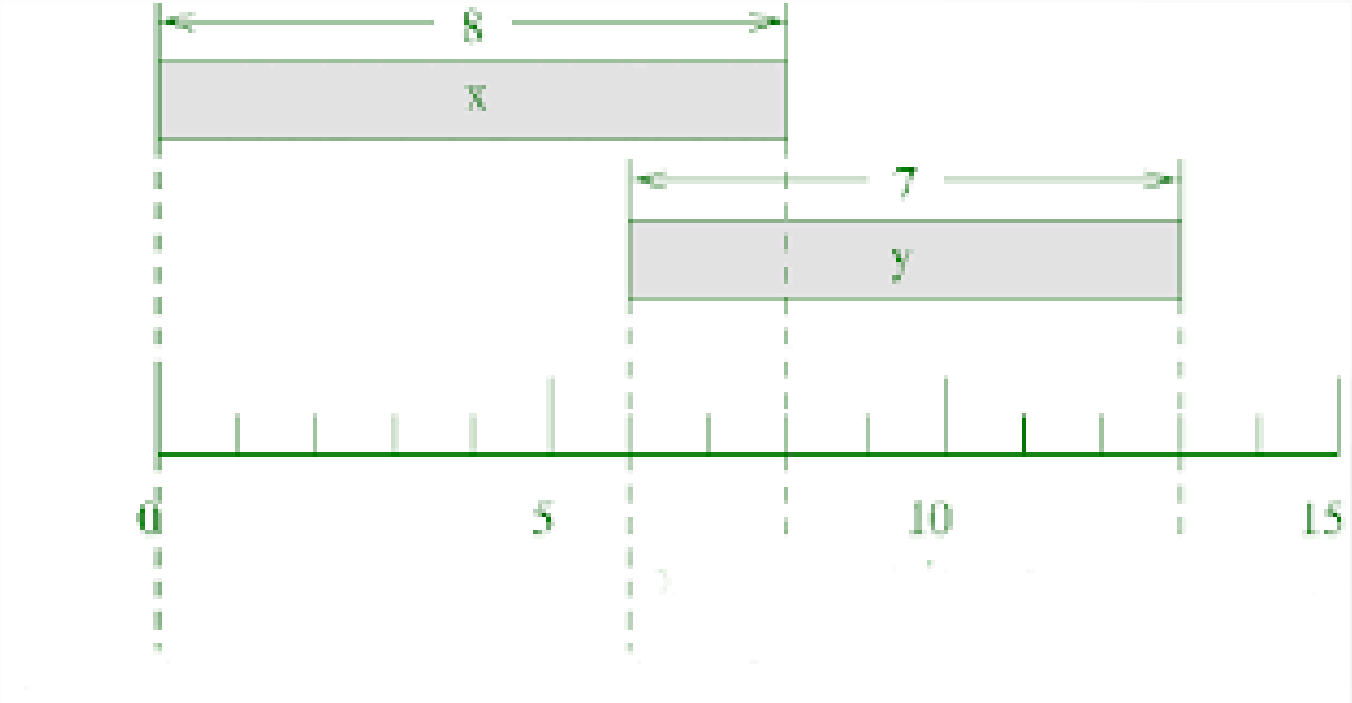
A random sample is a subset of a statistical population, whereas the subset should represent the entire population accurately and equally. To ensure that the sample is random, it is important that the population from which the data comes is stable.

Source

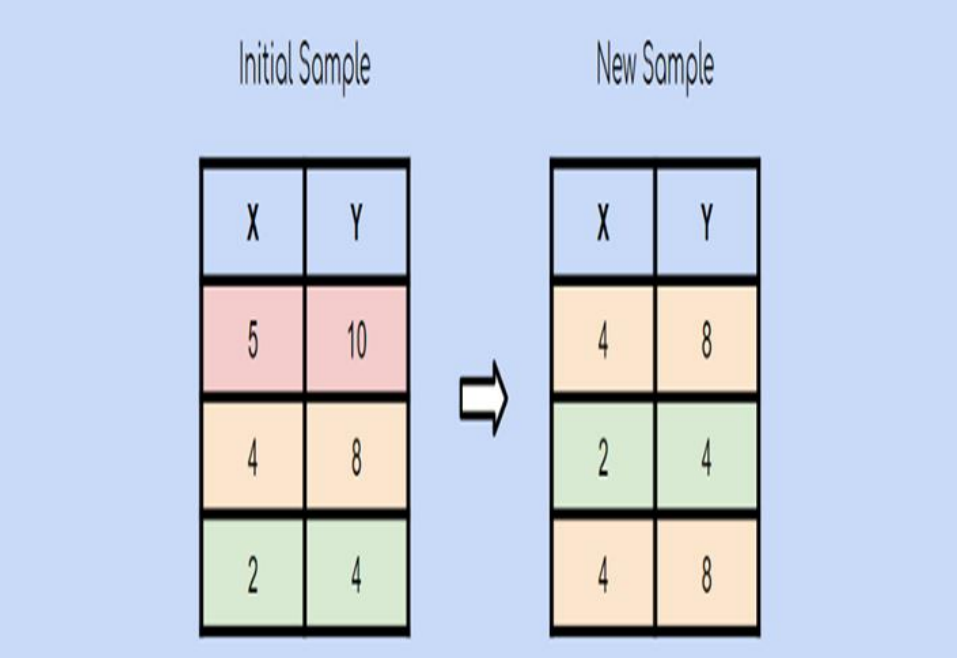
Temporal Selection

Temporal Selection

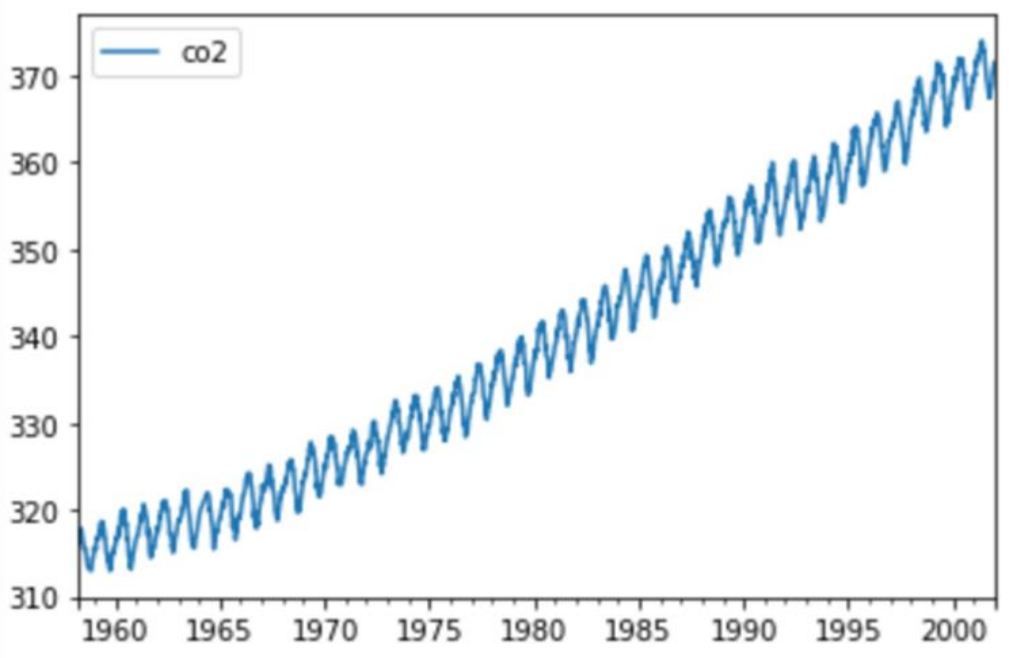
Temporal data is simply data that represents a state in time, such as the land-use patterns of Hong Kong in 1990 or the total rainfall in Honolulu on July 1, 2009. Temporal data is collected to analyze weather patterns and other environmental variables, monitor traffic conditions, study demographic trends, and so on.



Understanding Random and Temporal Selection in Machine Learning



Random Data Selection is a subset of data from the entire population that aids in inputting data into models with some random effect to understand the prediction in all possible ways.



Temporal data selection is time series-driven, which means it aids in the preparation of data sets that change over time.

Random Seeds in Machine Learning

A random seed is used to ensure that results are reproducible. In other words, using this parameter ensures that anyone who runs your code again will get the same results. Reproducibility is an extremely important concept in data science and other fields.

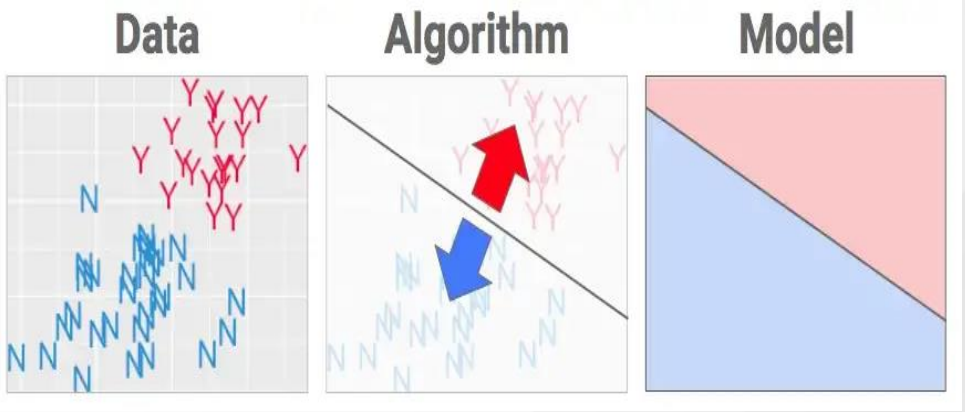
There are two common tasks where they are used:

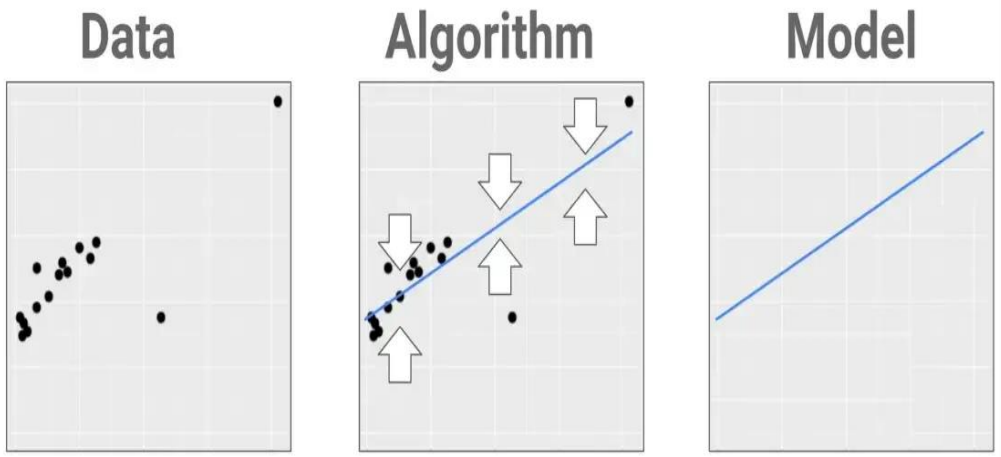
- **Splitting data into training, validation, and test sets**: Random seeds ensure that the data is divided the same way every time the code is run.
- **Model training**: Algorithms such as random forest and gradient boosting are non-deterministic (for a given input, the output is not always the same) and so require a random seed argument for reproducible results.

Classification & Regression in Machine Learning

Classification

What we have here is a dataset that is labeled with two classes (Y and N). In other words, the algorithm must put a fence in the data to best separate the Ys from the Ns. Our goal is to create a model that can be used to classify new examples correctly.



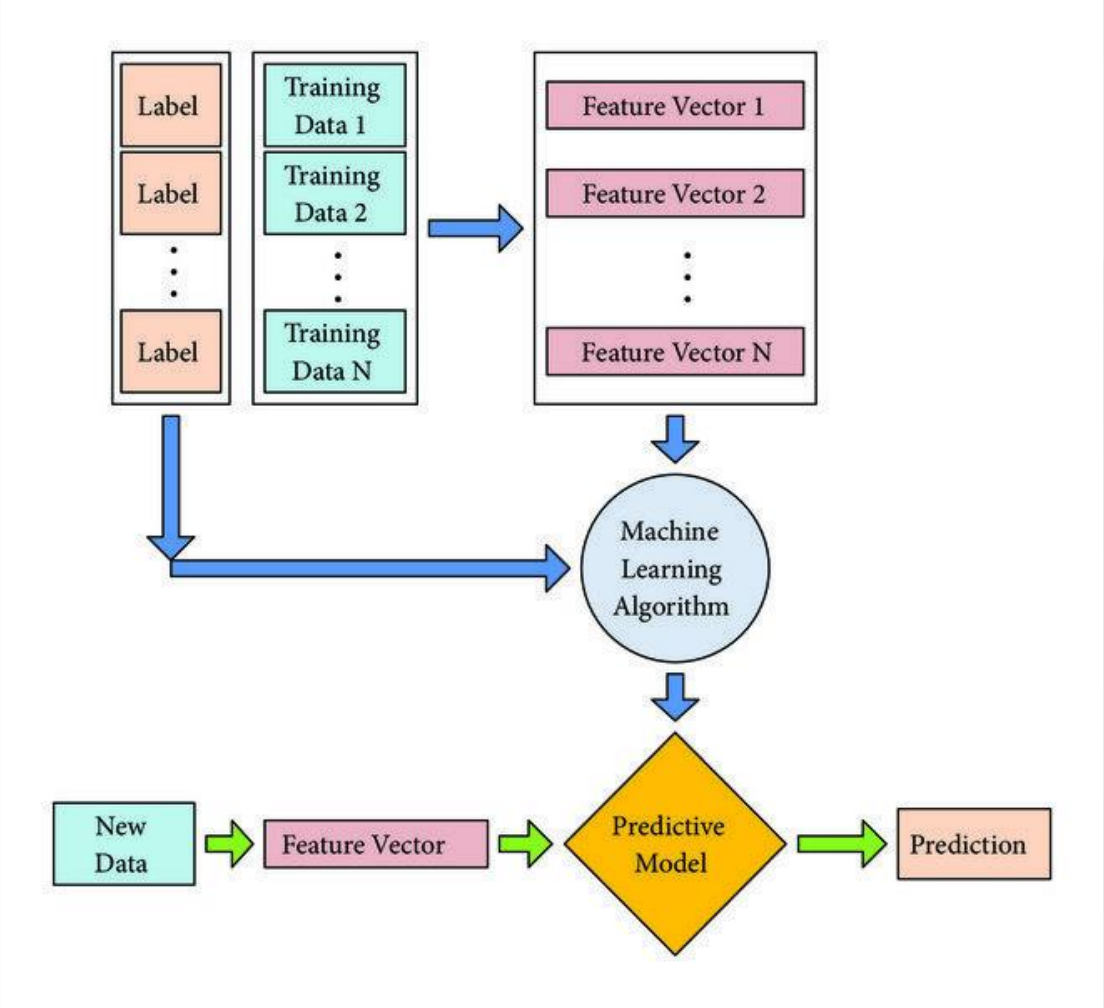


Regression

It is often used as a prediction where the model tries to predict any variables that are continuous in nature, such as demand forecasting, sales forecasting, etc.

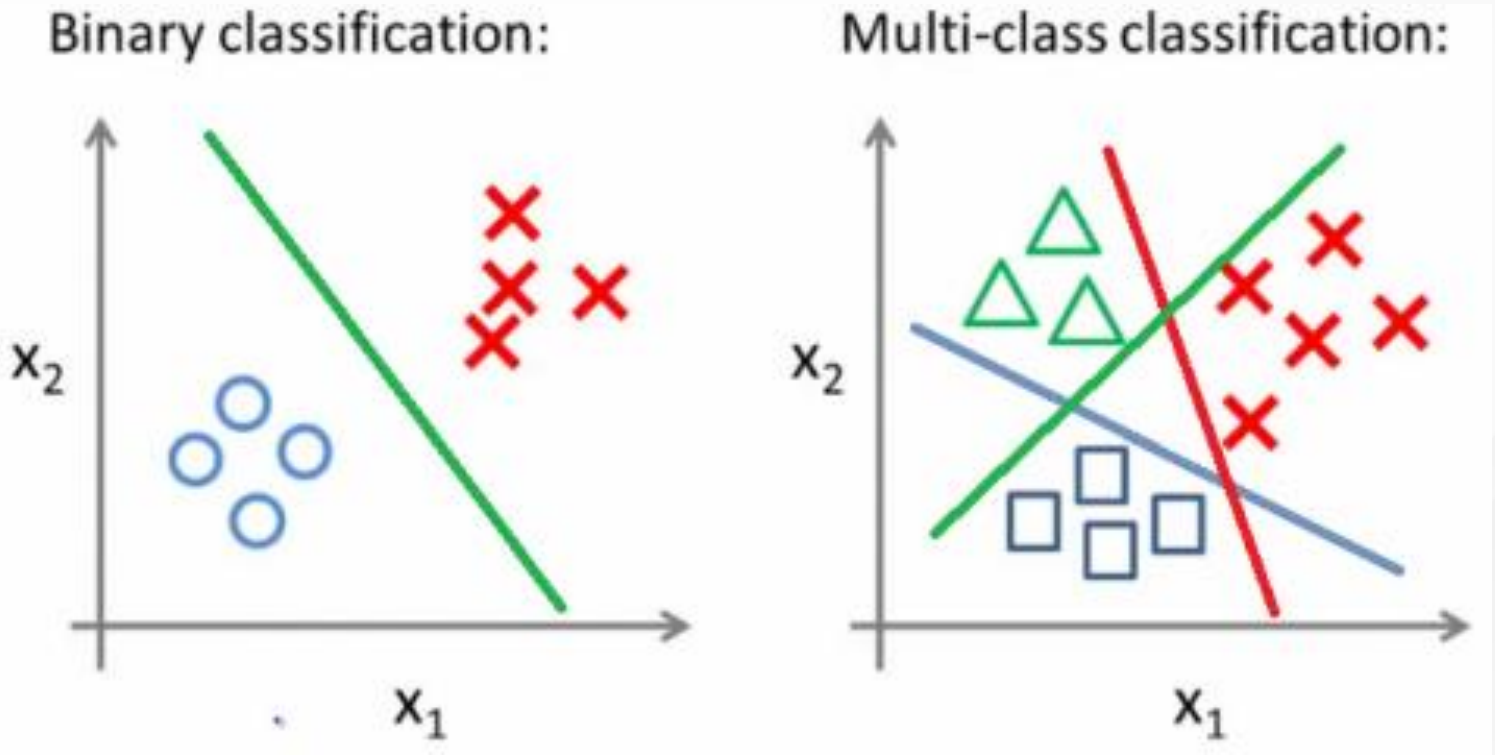
Prediction in ML Algorithm

- **Prediction** essentially means to predict a future outcome, like what is accomplished in machine learning.
- It refers to the output of an algorithm after training on a historical data set.
- After learning the historical data set, prediction makes it possible to forecast the outcome of a new data set.



Multiclass Classification

- Classification is the process of assigning any new data point to a set of categories (sub-populations) based on a mapping function.
- This function is derived from a training set of data where category membership is known for each observation.



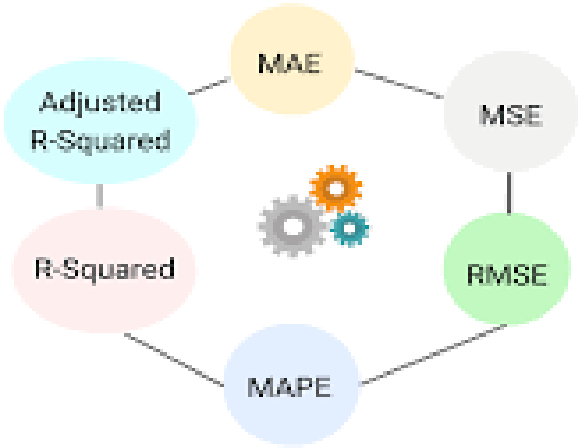
Evaluation Metrics of Classifier’s Model Performance

Model evaluation is the process of using different evaluation metrics to understand a machine learning model's performance as well as its strengths and weaknesses.

Source

Evaluating metrics of Regression Model's Performance

- Regression refers to predictive modeling problems that involve predicting a numeric value. It is different from classification, which involves predicting a class label.
- Unlike classification, you cannot use classification accuracy to evaluate the predictions made by a regression model.
- Instead, you must use error metrics specifically designed for evaluating predictions made on regression problems.



source

Quiz

A bank hires a data analyst to analyze the pre-historical dataset of customer records to determine whether the customers will churn. What will his or her modeling methodology be for the dataset?

1. Classification
2. Regression
3. Clustering
4. Reinforcement learning

Quiz

A large retail store company wants to analyze the sales forecasting for the 52 outlets it has in the Northeast region. Each store generates approximately 3 GB of data per day; as a data scientist, what modeling technique would you choose if you were tasked with sales forecasting?

1. Statistical modeling
2. Deep learning and modeling
3. Modeling with Machine Learning and Deep Learning

Thank You!

Copyright © HeroX Private Limited, 2022. All rights reserved.