An Introduction to Machine Learning

September 2021

Authors

Serena Li
Quantitative Associate,
Systematic Global Equities

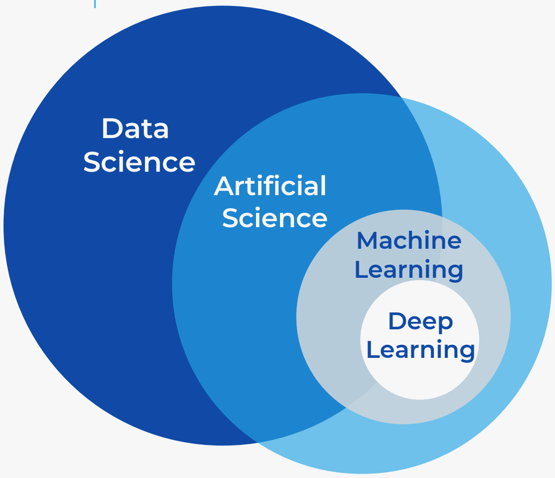
Christina Watson
Equity Intern,
Girls Who Invest Scholar

Sumali Sanyal, CFA
Managing Director, Senior Portfolio
Manager, Systematic Global Equities

Cameron McLennan, CFA
Director, Portfolio Manager,
Systematic Global Equities

Machine Learning is a broad term used to describe the study of computer algorithms[1] that can improve automatically through experience and by the use of data. Machine learning algorithms build a model based on sample data, known as "training data.[2] in order to make predictions or decisions without being explicitly programmed to do so. In its simplest form, Machine Learning is using data to answer questions. As it collects and processes data, it learns to siphon pertinent insights to produce results to solve complex questions. The machine is trained using large amounts of data and algorithms and learns how to perform the task during the process. Machine Learning is a subset of Artificial Intelligence (AI). Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, computer vision, and investment management. (**Chart 1** and **Table 1**)

Chart 1



Source: Cambridge Finsights

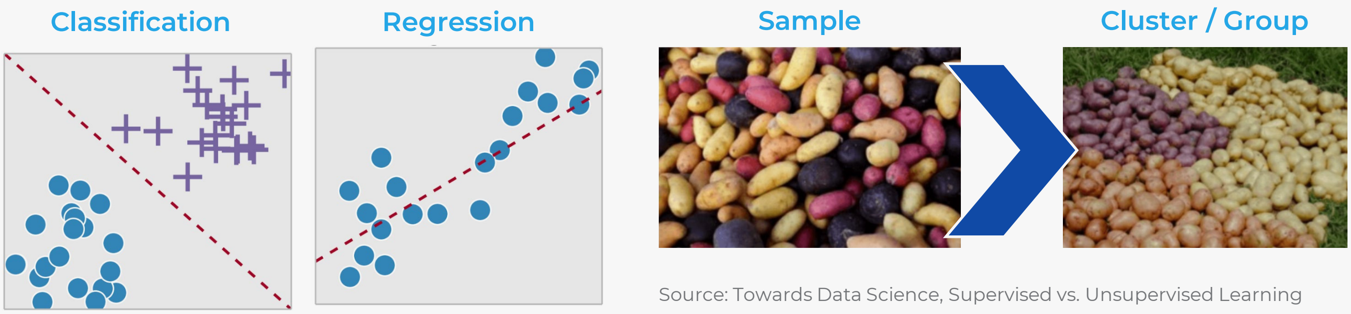
Table 1 Machine Learning Models

	ML Algorithm Type	
Variables	**Supervised (Target Variable)**	**Unsupervised (No Target Variable)**
Continuous	**Regression** • Linear; Penalized Regression/LSSO • Logistic • Classification and Regression Tree (CART) • Random Forest	**Dimensionality Reduction** • Principal Components Analysis (PCA) **Clustering** • K-Means • Hierarchical
Categorical	**Classification** • Logit • Support Vector Machine (SVM) • K-Nearest Neighbor (KNN) • Classification and Regression Tree (CART)	**Dimensionality Reduction** • Principal Components Analysis (PCA) **Clustering** • K-Means • Hierarchical
Continuous or Categorical	Neural Networks Deep Learning Reinforcement Learning	Neural Networks Deep Learning Reinforcement Learning

Source: Kathleen DeRose, CFA and Christophe Le Lannou. "Quantitative Methods, Machine Learning", CFA Institute, 2019.

Data researchers and data scientists categorize Machine Learning algorithms into supervised and unsupervised learning. Supervised learning is when an algorithm learns with labelled data that have correlating outputs to specific inputs. The algorithm learns to determine the relationship between inputs and outputs in the data. There are two types of supervised learning techniques: regression and classification. The former is used to predict numeric, or continuous value such as stock prices, returns, etc., and the latter is used to classify discrete, or categorical outputs. Unsupervised learning, on the other hand, does not have labeled outputs, so the goal is to estimate the natural structure and cluster data within the inputs given.[3] Unsupervised learning models are utilized for three main tasks—clustering, association, and dimensionality reduction.[4] (**Chart 2**)

Chart 2 | **Classification vs. Regression** **Unsupervised Learning**



Source: Towards Data Science, Supervised vs. Unsupervised Learning

In general, Machine Learning models uncover more complicated relationships between inputs and target variables than linear models and are often more effective in terms of handling high correlations among factors. Some of the most commonly used Machine Learning models are discussed below.

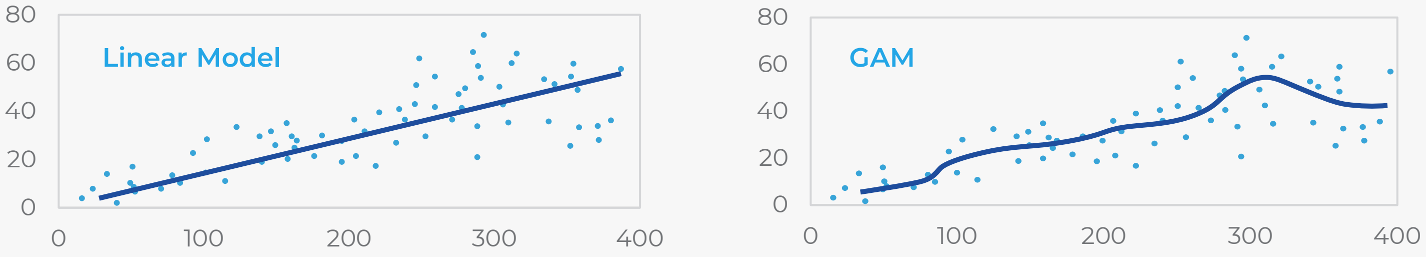
I. Regression & Generalized Additive Models (GAM)

Linear regression is a simple but attractive model which attempts to model the relationship between the features and the output by fitting a linear function to the data. The Capital Asset Pricing Model (CAPM), for example, is a widely used linear regression model in the finance industry as it predicts asset prices by taking a number of factors and finding out the relationships between them. It is an interpretable model since investors can easily explain why some factors are effective and how the model works, which is crucial for stock investments.

Ridge regression and Lasso (least absolute shrinkage and selection operator) are enhanced forms of linear regression model. They can both improve the prediction accuracy and interpretability of the results by picking relevant features that will be useful in the model.[5] These methods are particularly useful when the variables are highly correlated with each other.

However, the world is not always linear. GAM is a generalized linear model which predicts the target variable using a sum of flexible functions. By allowing nonlinear relationships between features and the output, it potentially provides higher prediction power but also maintains the interpretability of linear models. (**Chart 3**)

Chart 3 | **Linear Regression vs. GAM: Fitting a Straight Line vs. a More Flexible Function**

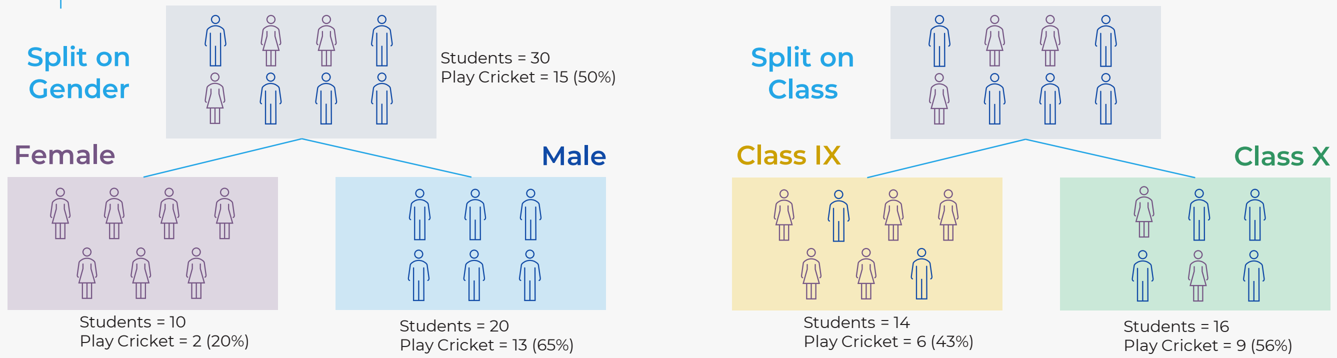


Source: Interpretable Machine Learning, Christoph Molnar

II. Decision Tree-based Models

Decision tree is a supervised learning algorithm that uses a set of splitting rules to segment the characteristics of data in order to make predictions for a given target, which can either be quantitative or qualitative. This type of model is simple and efficient to interpret as it can be easily visualized.[6] (**Chart 4**)

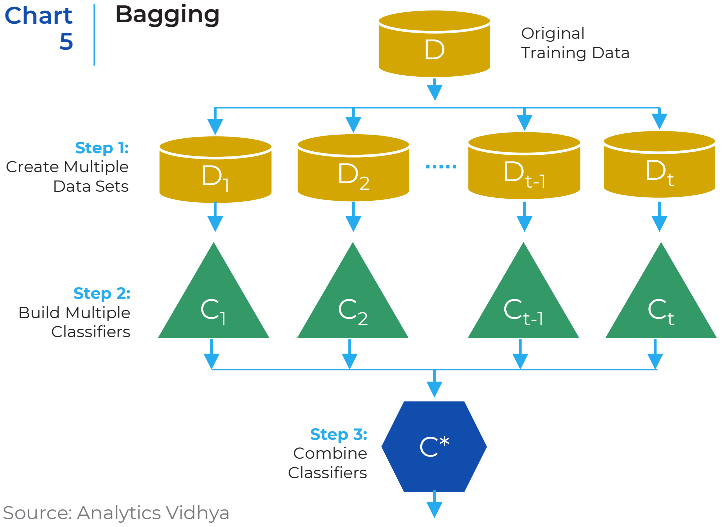
Chart 4 | **Tree-Based Model Example**



Source: Analytics Vidhya, Tree Based Algorithms

One criticism about decision tree is that it is unstable and may not have enough prediction power. This led to the creation of the bagging method. Bagging (or Bootstrap Aggregation) improves the accuracy by taking repeated samples from the data and combining results from all individual trees. In this case, even if one tree gives inaccurate prediction, after averaging the outputs the model is able to generate more stable outputs. Random Forest is a popular method that improves from bagged decision trees.[7] It slightly alters the algorithm to decorrelate the trees so that the predictions of different trees are not highly correlated. (**Chart 5**)

Chart 5 | **Bagging**


D
Original Training Data
Step 1: Create Multiple Data Sets
D_1
D_2
.....
D_{t-1}
D_t
Step 2: Build Multiple Classifiers
C_1
C_2
C_{t-1}
C_t
Step 3: Combine Classifiers
C*

Source: Analytics Vidhya

XGBoost has quickly gained popularity in recent years since its introduction. It is a tree-based ensemble ML algorithm that uses a boosting framework. Boosting is a method that evolved from the decision tree algorithm, where the trees grow sequentially, each using the information from previous trees. The algorithm learns from a large number of trees instead of a single large tree, and combines the information from all trees, thus significantly improving the prediction power. Based on this scheme, XGBoost enhances the algorithm and optimizes the system to generate strong results in a short time.

The common characteristics for both bagging and boosting techniques is the idea of ensemble, meaning they both generate predictions from small, weak trees and aggregate them to become a strong prediction.

III. Clustering

Clustering is a machine learning technique which groups unlabeled data based on how similar they are. The most commonly used example is K-means clustering. It looks for a fixed number (k) of clusters to aggregate similar data points, which refers to the number of centroids needed in a data set.[8] The algorithm then allocates every data point to the nearest cluster and tries to keep the centroids as small as possible. (**Chart 6**)

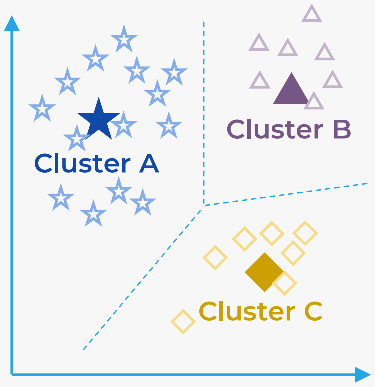
IV. Dimensionality Reduction and Principal Component Analysis (PCA)

While more data generally yields more accurate results, it can also negatively impact the performance of machine learning algorithms and make it difficult to visualize data. Dimensionality reduction is a technique used when the number of variables, or dimensions is too high. PCA is one of the dimensionality reduction algorithms. It combines the input variables in a specific way called "principal components" and brings out strong patterns in a data set.[9] The new components are all independent from each other, which benefits the linear regression model. (**Chart 7**)

V. Deep Learning

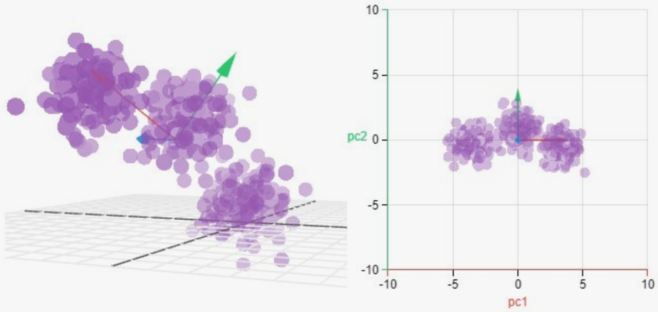
More advanced forms include Deep Learning (DL) Models and Natural Language Programing (NLP). D L is the technique that use complicated layers (often multi-layers) of nonlinear processing nodes to process information and provide more accurate outputs.[10] It can either learn in supervised or unsupervised manners. For example, artificial neural networks (ANN) learn patterns differently from classical Machine Learnings algorithms by mimicking the biological networks of human brains. One advantage over the classical Machine Learning approaches is that DL models process raw inputs that come in sequences as well as improve the prediction accuracy. However, the black box nature of the DL models along with the higher demand for data makes it challenging to be used for real world investing. (**Chart 8**)

Chart 6 | K-Means Clustering



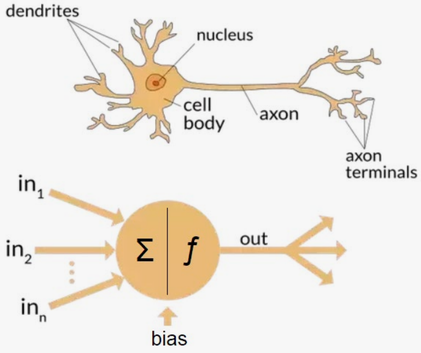
Source: Introduction to K-means Clustering, Dileka Madushan

Chart 7 | **PCA Example:** Bring Three-Dimension Problem to Two Dimensions with Transformation



Source: Principal Component Analysis, Victor Powell

Chart 8 | **Deep Learning:** Mimic Human Brains



Source: The difference between artificial and biological neural networks, Richard Nagyfi

VI. Natural Language Programing (NLP)

Natural language processing (NLP) is an AI-based solution that uses machines to extract and analyze vast amounts of text data, enabling the extraction of unique insights from the information. For example, sorting out information from financial statements manually was an almost impossible mission, but with NLP techniques investors can extract positive or negative tones in the disclosure. Another popular use is to obtain information in the daily news that could potentially affect short-term price movements.

Conclusion

In this article, we discussed the concepts of artificial intelligence, the classification of machine learning techniques, as well as different types of machine learning algorithms. What differentiates machine learning from traditional statistical models is that it enables the computers to learn from large datasets, even improve themselves to generate accurate predictions without being explicitly programmed.[11] In general, machine learning algorithms can be categorized into supervised learning and unsupervised learning based on whether the dataset contains labeled target variables. Supervised learning algorithms include techniques such as linear regression, generalized additive model, decision tree, random forest and boosting models. Unsupervised learning algorithms are used to perform tasks such as dimensionality reduction and clustering datasets. More advanced forms in AI include deep learning algorithms and natural language processing.

In the second part of the article, we will discuss specific uses of machine learning in investment management, the benefits of using these models, and the challenges inherent in their use.

References:

[1] *"Algorithm." Wikipedia, Wikimedia Foundation, 15 Sept. 2021, en.wikipedia.org/wiki/Algorithm.*

[2] *"Training, Validation, and Test Sets." Wikipedia, Wikimedia Foundation, 12 Aug. 2021, en.wikipedia.org/wiki/Training,_validation,_and_test_sets#training_set.*

[3] *Devin Soni. "Supervised vs. Unsupervised Learning." Medium, Towards Data Science, 21 July 2020, towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d.*

[4] *By: IBM Cloud Education. "What Is Unsupervised Learning?" IBM, www.ibm.com/cloud/learn/unsupervised-learning.*

[5] *"Lasso (Statistics)." Wikipedia, Wikimedia Foundation, 16 Sept. 2021, en.wikipedia.org/wiki/Lasso_(statistics).*

[6] *Morde, Vishal. "Xgboost Algorithm: Long May She Reign!" Medium, Towards Data Science, 8 Apr. 2019, towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d.*

[7] *Brownlee, Jason. "Bagging and Random Forest Ensemble Algorithms for Machine Learning." Machine Learning Mastery, 2 Dec. 2020, machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/.*

[8] *Garbade, Dr. Michael J. "Understanding k-Means Clustering in Machine Learning." Medium, Towards Data Science, 12 Sept. 2018, towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1.*

[9] *Vicapow. "Principal Component Analysis Explained Visually." Explained Visually, setosa.io/ev/principal-component-analysis/.*

[10] *Carolyn Deng, CFA. "Ai Investment Primer: Laying the Groundwork (Part i)." Toptal Finance Blog, Toptal, 25 Oct. 2018, www.toptal.com/finance/market-research-analysts/ai-investment-primer.*

[11] *Pant, Ayush. "Introduction to Machine Learning for Beginners." Medium, Towards Data Science, 22 Jan. 2019, towardsdatascience.com/introduction-to-machine-learning-for-beginners-eed6024fdb08.*

This report is neither an offer to sell nor a solicitation to invest in any product offered by Xponance® and should not be considered as investment advice. This report was prepared for clients and prospective clients of Xponance® and is intended to be used solely by such clients and prospects for educational and illustrative purposes. The information contained herein is proprietary to Xponance® and may not be duplicated or used for any purpose other than the educational purpose for which it has been provided. Any unauthorized use, duplication or disclosure of this report is strictly prohibited.

This report is based on information believed to be correct, but is subject to revision. Although the information provided herein has been obtained from sources which Xponance® believes to be reliable, Xponance® does not guarantee its accuracy, and such information may be incomplete or condensed. Additional information is available from Xponance® upon request. All performance and other projections are historical and do not guarantee future performance. No assurance can be given that any particular investment objective or strategy will be achieved at a given time and actual investment results may vary over any given time.