User Guide to Machine Learning

and data transformation in the cloud

MATILLION

Table of Contents

2 Introduction

3 What is Machine Learning?

4 Getting Started with Machine Learning

5 The Method of Learning
5 Supervised
5 Unsupervised
5 Semi-Supervised
5 Reinforcement

6 Recognizing the Right Group for Your Task
6 Classification
6 Regression
6 Ranking
6 Clustering

8 Learning Models
8 Linear Regression
8 Decision Trees
9 Support Vector Machines
9 K-Means Clustering
10 Hierarchical Clustering
10 Neural Networks

11 Machine Learning Technologies

12 Machine Learning as a Service

13 Data Transformation for Machine Learning

15 Guidelines for Machine Learning

16 Conclusion

Introduction

Industry experts, competitors and even your customers are talking about machine learning and artificial intelligence. The terms, while used widely and interchangeably, are often misunderstood and carry a narrow definition. Both machine learning and artificial intelligence have distinct and practical applications for your business - not only driverless cars!

Machine learning is the process of building and training models to process data. In this capacity, your models are learning from your data to make predictions. Artificial intelligence, consequently, uses these learnings to make a computer or technology stack act more human, apply learnings in an automated manner. In this way, machine learning allows computer systems to learn from data and make decisions without being explicitly programmed to do so.

This ebook will focus on the study of machine learning and the implementation of statistical models to make predictions or decisions based on patterns in your data.

This level of advanced technology, means your business can process, and more importantly, understand data faster allowing you to run more effective marketing campaigns, make your logistics operations more efficient, and significantly outpace your competitors.

In this ebook we will cover in more detail what, specifically, machine learning is and what are common business use cases that can be improved with machine learning.

What is Machine Learning?

Alan Turing, the father of machine learning, wrote a paper in 1950 called “Computing Machinery and Intelligence”. The question Turing considered was “Can machines think?” This led to experimenting with a machine’s ability to act as a human with intelligent action. Today we know it as the **Turing Test.** Building on Turing’s initial work, Arthur Lee Samuel, a pioneer in computer gaming and artificial intelligence, coined the term “machine learning” in 1959 and developed the famous computer checkers program. Evidently, machine learning has been around for decades as so demonstrated by Turing and Samuel’s work. So why is the term trending now in a modern technological era of cloud computing?

Perhaps the advancements in technology have enabled data scientists to realize those theoretical discoveries more profoundly than before. With the increasing amount of accessible data and the cost of high powered computing becoming more affordable, data scientists no longer need to rely on small, thoroughly curated datasets. Instead, large and even unorganized datasets can be used with thousands of parameters to train algorithms and generate predictions.

Based on these modern workloads, machine learning is understood to be a form of artificial intelligence and mainly refers to computers that can learn and improve their analysis on data over time without reprogramming their core logic. Related to machine learning, deep learning is a subset of machine learning involving artificial neural networks which is inspired by the function and structure of a brain.

> **Understanding machine learning**
>
> **"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E."**
>
> *(Mitchell, T. (1997). Machine Learning. McGraw Hill. p. 2)*
> *http://profsite.um.ac.ir/~monsefi/machine-learning/pdf/Machine-Learning-Tom-Mitchell.pdf*

In traditional programming, the engineer will write a set of instructions telling the computer exactly what to do. With machine learning, specifically with neural networks, the engineer no longer needs to give it a set of instructions. Instead, the machine is designed to learn the instructions from a given dataset.

Getting Started with Machine Learning

Machine learning has taken a massive leap in adoption over recent years and many businesses have already started to plan or have already developed machine learning models. However, despite the evolution of machine learning products, some companies still stumble over many barriers. Two of those barriers involve siloed and anonymized data.

> **"AI companies tend to organize the data better. So putting data in a centralized data warehouse makes it more efficient for engineers or software to exploit that data. Instead of federated or distributed data sets, we like to bring it together because it's like gunpowder. You put a lot into it to make a big bang."**
>
> *Andrew Ng (Cofounder of Coursera, AI Fund, and Landing.AI)*
> *https://www.mckinsey.com/featured-insights/artificial-intelligence/how-artificial-intelligence-and-data-add-value-to-businesses*

Regardless of the age, structure, or size of an organization, they may all experience the problems of siloed and / or anonymized data at some point. This could be due to having many separated systems or having a siloed structure of departments. Whatever the reason may be, this naturally resulted in valuable data being kept separated and bringing that data together became a challenge.

Sometimes there can also be legal and regulatory implications which hinder the use of certain data, such as Personally Identifying Information (PII). In this situation, the sensitive data will be anonymized at the data preparation stage making it difficult to consolidate the data because the entities won't necessarily match.

Nevertheless, carefully evaluating your business objectives and aligning it with analytics will help you to identify where machine learning fits within your data strategy.

The Method of Learning

When it comes to machine learning, you'll find there are different types of learning algorithms and each one has a unique characteristic for a specific use case. We'll begin by describing the two most common methods of learning before introducing two newer approaches.

Supervised

Supervised learning consists of a labelled dataset with features. This is then fed into the learning algorithm during the training process where it will work out the relationship between the selected features and the labels. The learning outcome will then be used to classify new unlabelled data.

Unsupervised

With unsupervised learning we tend not to know what the correct answer will be therefore the dataset is unlabelled. Rather, it is expected to discover patterns that suggest natural groupings in the data by itself. An example scenario for this type of learning could be a marketing team looking for buying patterns using historical transactional data. The answer isn't obvious. Moreover, there could be many combinations of the correct answer dependent on what variables are factored in.

Semi-Supervised

What happens when you come across a large dataset that is only partially labelled? You either go through the process of labelling the rest of the data or you can try deploying a semi-supervised learning algorithm.

Many real world machine learning problems fall into this zone as it's generally too expensive and time consuming to label your whole dataset for a fully supervised learning approach. On the other hand, an unsupervised learning approach may be unnecessary. Combining the two learning methods, therefore, should in theory give you the best of both worlds. Furthermore, many machine learning researchers have found that using both labelled and unlabelled data can produce considerable improvement in the learning accuracy when compared to unsupervised learning alone.

Reinforcement

Reinforcement learning is a sophisticated style of learning inspired by game theory and behavioural psychology. This method of learning usually involves an agent, the machine that is making the action, and an interpreter. The agent will be exposed to an environment where it will execute an action and then the interpreter will either reward or punish the agent dependent on the success of that action. The goal for the agent is to find the best way to maximize the rewards by iteratively interacting with the environment in different ways. The only thing that the data scientist will provide in this type of learning is a method for the agent to quantify its performance.

This approach is already being used by many companies developing robotics and self driving vehicles. However, it's usually deployed alongside other learning techniques such as supervised learning, creating an ensemble learning model. This is because it becomes difficult to apply reinforced learning to scenarios where the environment, actions and rules are variable. That being said you may have come across reinforcement learning through projects involving Atari games. In Atari games, all three variables are stable; environment, actions, and rules, making it an ideal scenario for applying reinforcement. Nonetheless, this learning method holds the closest resemblance to human learning.

Selecting a Group for Your Task

Now that you understand what machine learning is, you can get started. The first step in machine learning is identifying what type of predictive analytics you'll want to carry out and finding a suitable algorithm group for your task. In this section we outline the primary algorithms commonly used for machine learning.

Classification

Classification algorithms are used to identify the class that an object from a dataset would belong to. These tasks usually fall into three different categories; binomial classification, multi-class classification and anomaly detection.

Binomial classification can be used for scenarios where the object will fall into one of the two classifications e.g. identifying whether an email is junk mail or flagging a fraudulent insurance claim. Whereas **multi-class classification** focuses on scenarios where there are many classifications, such as identifying the type of an animal; dog, cat, or horse, or figuring out what type of product a customer is likely to buy; a tablet, laptop, or desktop.

Anomaly detection is a type of single class classification algorithm where the only goal is to find outliers in your dataset or unusual objects that appear outside of the normal distribution. This can be used in events such as flagging fraudulent transactions, medical problems, or malfunctioning equipment.

Regression

Regression algorithms involve continuous data, usually used in predicting a dependent variable based on the relationship between your independent variables. Some examples of regression models include financial forecasting, weather forecasting, or predicting product demand.

Ranking

Ranking algorithms determine the relative importance of objects in connection with other objects in the dataset. The page rank algorithm is probably the most well known example as it's extensively used by Google on their search engine results page. Other scenarios include movie recommendations, based on viewing patterns, or hotel recommendations based on popularity.

Clustering

Unlike classification or regression, some clustering algorithms are not used to predict a specific value. Rather, the algorithm is expected to explore the data and organize the objects into groups called clusters. Cluster analysis is mostly used in unsupervised learning techniques and deals with more complex scenarios such as where to place an emergency ward within a particular geography based on regions that are more accident-prone.

MATILLION

The better your data, the more valuable your Machine Learning

Matillion can help you transform your data to train your models

GET A DEMO

Give your Machine Learning models what they need - good, clean data!

- *60+ data source connectors to access your data*
- *Transformation components including, Calculator, Aggregate, Join, Filter, Convert, Transpose, and many more, to help you clean your data*
- *Map internal codes to user-friendly names*
- *SQL, Bash, and Python scripting components to support your efforts*
- *Use Change Data Capture (CDC) to connect to databases, automatically and incrementally load new and changed data*

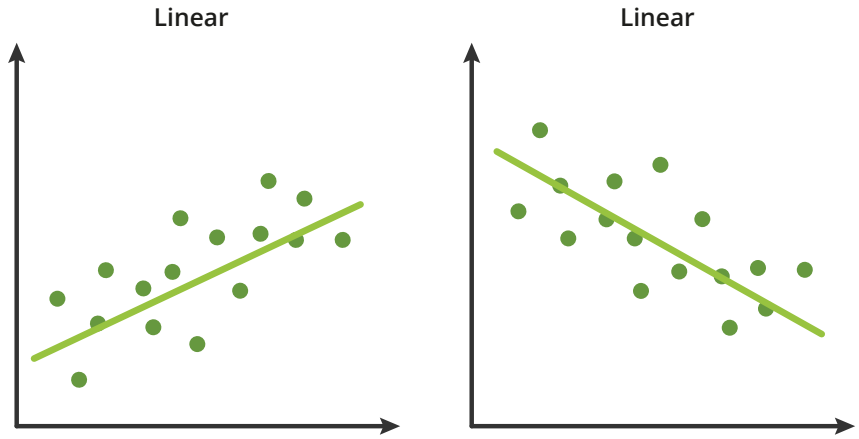
matillion.com/get-a-demo/

Learning Models

Linear Regression *(Figure a)*

Linear Regression is a model very familiar to statisticians! This model has also been applied to machine learning, as a standard method for showcasing relationships between a dependent variable and independent variable when the independent variable changes.

Method of Learning: Supervised

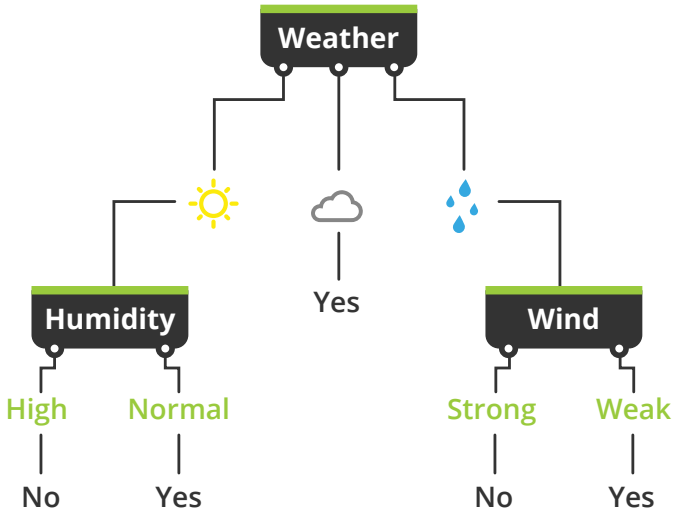


(Figure a)

Decision Trees *(Figure b)*

This type of algorithm has high interpretability and handles outliers and missing observations well. It is possible to have multiple decision trees working together to create a model known as ***ensemble trees***; *random forest* and *gradient boosting* are examples of this type of model. Ensemble trees have the ability to increase prediction and accuracy whilst decreasing overfitting to some extent.

Method of Learning: Supervised

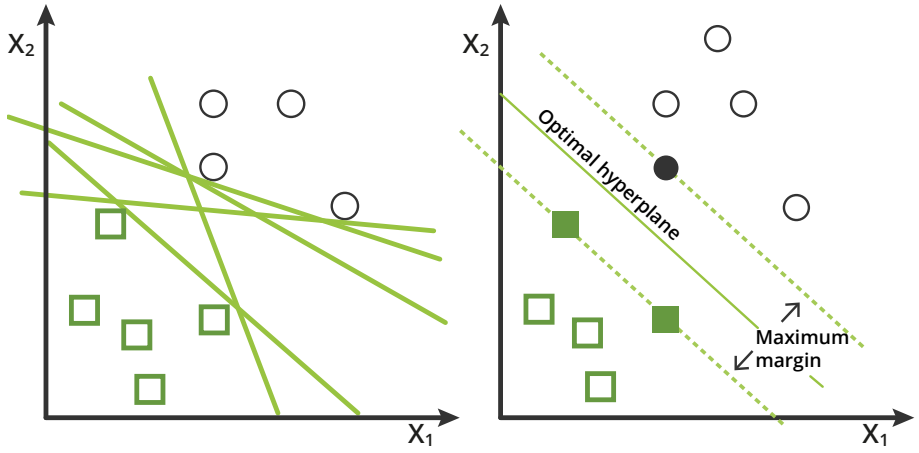


(Figure b)

Support Vector Machines *(Figure c)*

A typical algorithm that is used for classification, but can also be turned into a regression algorithm, is Support Vector Machines (SVM). SVM can bring greater accuracy when it comes to classification problems by finding the optimal hyperplane, a division between the different data classes. To find the optimal hyperplane, the algorithm will draw multiple hyperplanes between the classes. Then, the algorithm will calculate the distance from the hyperplane to the closest vector points, commonly referred to as the margins. It'll then choose to use the hyperplane which produces the greatest margin; the optimal hyperplane. Finally, it'll utilize the optimal hyperplane in the classification process.

Method of Learning: Supervised

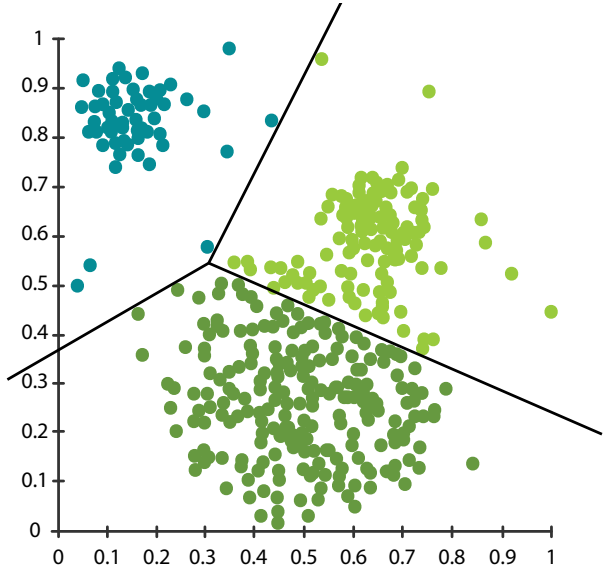


(Figure c)

K-Means Clustering *(Figure d)*

K-Means clustering is used for finding similarities between data points and categorizing them into a number of different groups, K being the number of groups.

Method of Learning: Unsupervised



(Figure d)

Hierarchical Clustering *(Figure e)*

Hierarchical clustering creates a known number of overlapping clusters of different sizes along a hierarchical tree to form a classification system. This type of clustering can be achieved through various methods with the most common methods being *agglomerative* and *divisive*.

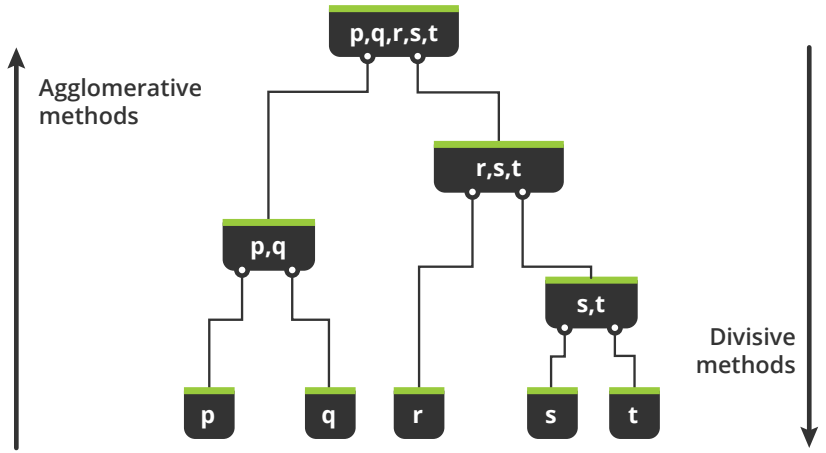
The agglomerative approach is a bottom-up method which consists of all objects starting within their own respective clusters. These clusters of objects are then joined together by taking the two most similar clusters and merging them. Conversely, divisive clustering takes a top-down approach where all the objects start in the same cluster and are then divided into two separate clusters through an algorithmic process similar to K-Means. The splitting process is repeated until the desired number of clusters is achieved.

Method of Learning: Unsupervised

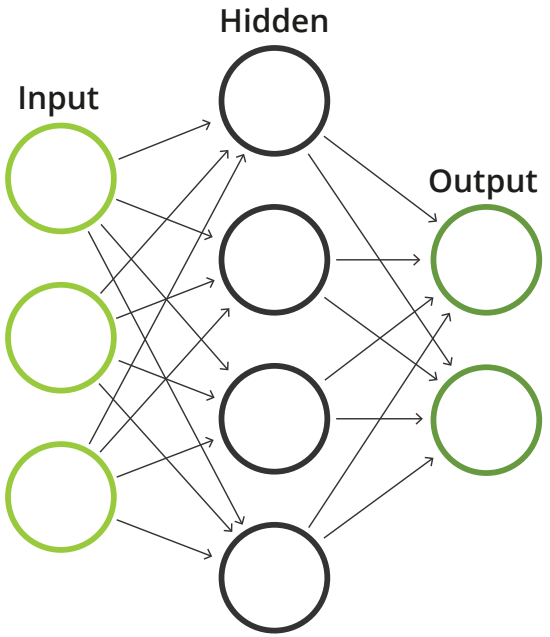
> **Curse of dimensionality**
>
> *An increase in processing power and training data is required as the dimensions of data increase.*

Neural Networks *(Figure f)*

Neural networks are highly associated with robotics and neuroscience which naturally makes it the most exciting algorithm to explore. Neural networks, specifically artificial neural networks, consists of three layers; an input layer, an output layer and one or many hidden layers which are used to detect patterns in the data. It does this by assigning a weight to a neuron inside the hidden layer each time it processes a set of data.



(Figure e)



(Figure f)

Machine Learning Technologies

Machine learning was an extremely difficult task in the past and was only approached by individuals who were highly capable in mathematical disciplines which it is built upon, such as linear algebra, probability theory, calculus, and statistics. Today, we have a wide range of frameworks and services that abstract the complexity and simplify the process of building, training and running your own machine learning models. These advancements have made it possible for many businesses to not only build out their machine learning capabilities but to also automate their operations and build a smarter business.

In this section, we'll explore the most popular framework and services used for machine learning.

Tensorflow

Tensorflow is a widely used library for machine learning and other algorithms involving heavy numerical computations. It's an open source framework created by the Google Brain team. In fact, Google uses the framework themselves to incorporate machine learning in their own applications such as Google Photos, Google Voice Search, etc. Therefore, if you've been using any Google products then you've probably been using Tensorflow indirectly.

To build machine learning models in Tensorflow you need to know how to use Python. Python is a general purpose programming language that has gradually become more popular with data scientists because of it's dedicated libraries for data analysis and predictive modelling. Building applications with the Tensorflow framework requires Python while the execution of those applications are actually performed in C++. This is because C++ is a lower level language compared to Python which means it'll give you a performance advantage when training or running your models.

The framework can train and run deep neural networks for many applications such as image recognition, handwriting recognition, natural language processing, etc. Best of all, you can run these models on virtually any platform - from a mobile device or a local machine to a cluster of computers in the cloud. Not to mention, if you happen to own a Google Cloud Platform (GCP) account you can even run Tensorflow using Google's Tensor Processing Units (TPU) which is specifically designed for Tensorflow to accelerate machine learning workloads. Alternatively, it can also be deployed on AWS using a service called Sagemaker as well as on Azure through a containerization method.

When it comes to machine learning, Tensorflow offers a lot of convenience for data scientists. However, if you're looking for an even greater level of abstraction then you might want to consider Machine Learning as a Service.

> Other machine learning libraries include **Apache MXNET, Microsoft CNTK and Pytorch.** The library you choose will likely be dependent on your data strategy and resource available. *Pandas, scikit-learn numpy, matplotlib* may also be used to support your machine learning efforts.

Machine Learning as a Service

Machine Learning as a Service (MLaaS) is a service provided by many cloud vendors. It's main objective is to enable the user to get started with building their machine learning models immediately without having to worry about the supporting components. In some cases, you don't even need to write any code or understand any of the learning algorithms that were discussed in the previous chapters. It's a great way of accelerating your machine learning project, allowing you to yield valuable insights from predictions with a relatively small team.

In this section we're going to focus on the three major cloud vendors and discuss some of the machine learning services they have to offer.

Google Cloud Platform

BigQuery ML is an extension of Google's BigQuery product. It provides a quick and easy way to build machine learning models using structured and semi-structured data that is stored in BigQuery. In addition to this, everything is written in SQL so your typical SQL analyst will be able to build machine learning models without having to learn a programming language. Moreover, BigQuery ML automates most of the process for you by doing things such as automatically splitting your data into training and test sets, tuning the learning rate and standardizing numerical features. The beauty of this product is that it will leverage the power of BigQuery by keeping and running everything including the machine learning models inside the cloud data warehouse itself.

Amazon Web Services

Amazon SageMaker was released in 2017 as a successor to the Amazon Machine Learning service (Amazon ML). Similar to Amazon ML, SageMaker users can build, train and deploy machine learning models in the cloud. The difference is that it offers a wider variety of pre-built learning models and increased flexibility for adding custom models, through integrating SageMaker with external machine learning libraries such as Tensorflow. This new arrival provided enough freedom for those experienced data scientists whilst offering a concise solution to those who are just starting off.

Microsoft Azure

Azure Machine Learning Studio (ML Studio) is a MLaaS similar to the products mentioned above. The platform presents you with a graphical drag and drop user interface where you'll build out your machine learning process by dragging components from the resource panel onto a canvas. You'll then connect these components up via their nodes, effectively creating a workflow. Almost all operations can be completed through the user interface including data preprocessing, data exploration, choosing your machine learning methods and validating your results.

Machine learning with ML Studio will entail a steeper learning curve because of the amount of control it gives you and, although no coding skills are required, it is possible to make use of Python and R through the built in modules. Like it's competitors it also provides a range of pre-built machine learning models selectable in the form of a component. The service supports around 100 methods that addresses classification (binary and multi-class), anomaly detection, regression, recommendation, and text analysis.

Data Transformation for Machine Learning

“Garbage in, Garbage out”

Data in the real world can be really messy and in most cases some sort of data cleansing needs to be performed prior to any data analysis. However, this can be a daunting task. Without the right technology stack in place, data transformation is time consuming expensive, and tedious. Nevertheless, this is a critical step as it ensures maximum data quality which increases the accuracy of predictions. Based on our customers’ experiences, these are common data transformations required before data can be processed within machine learning models.

Matillion has the ability to connect to databases, automatically and incrementally load new data, and regularly refresh data loads when changes are detected via Change Data Capture (CDC). This means your machine learning models will always have access to the most up-to-date data. www.matillion.com

Remove Unused and Repeated Columns

Hand picking the data that you specifically need will not only improve the speed at which your model trains but also helps when you come to analyze it.

Change Data Types

Using the correct data types helps to save memory usage. It can also be a requirement, such as making numerical data an integer, in order for calculations to be performed against it.

Handle Missing Data

At some point you'll come across incomplete data and resolving it can vary depending on the dataset. For example, if the missing value doesn't render its associated data useless then you may want to consider imputation. Imputation is the process of replacing the missing value with a simple placeholder, or another value, based on some kind of assumption. Otherwise, if your dataset is large enough then there is a likelihood that you can remove the data without incurring any substantial loss to your statistical power. However, proceed with caution as you may inadvertently create a bias in your model. On the other hand, not treating the missing data can also skew your results.

Remove String Formatting and Non-Alphanumeric Characters

This involves removing characters like line breaks, carriage returns, white spaces at the beginning and the end of values, currency symbols, etc. In addition, you may also want to consider word-stemming as part of this process. Although removing formatting and other characters makes the sentence less readable for humans, this approach helps the algorithm to better digest the data.

Convert Categorical Data to Numerical

This step isn't always necessary but a lot of machine learning models require categorical data to be in a numerical format. This means converting values such as yes and no into 1 and 0. However, be cautious not to accidentally create order to unordered categories such as converting mr, miss and mrs into 1, 2 and 3.

Convert Timestamps

You may encounter timestamps in all types of formats. In this case it's a good idea to define a specific date/time format and convert all timestamps to the defined format.

This list is not exhaustive and only acts as a simple guideline to get you started. There are other factors you may want to consider such as being aware of outliers. You may or may not want to remove them from your dataset depending on the training model you use. For example, not removing the outliers may skew your training results but if it's an anomaly detection algorithm then it makes sense to include it.

Guidelines for Machine Learning

Throughout the sections above we've looked at what machine learning is, an introduction into the algorithms, cloud services for machine learning and how to prepare your data for a machine learning project. In this final section we'll finish with a few guidelines inspired by rules put together by a group of Google engineers to guide you through your machine learning project.

The original documentation, "Rules of Machine Learning: Best Practices for ML Engineering" published by Martin Zinkevich is a long document containing a total of 43 'rules'. We have picked out a few key guidelines to summarize each phase of your machine learning project.

Source: https://developers.google.com/machine-learning/guides/rules-of-ml/

Guideline 1 Machine Learning Preliminaries
Don't hesitate to launch a product without machine learning! If the use case becomes too complex, it may be beneficial to move forward with a machine learning approach, this could be easier to develop and maintain in the long run.

When the time comes to start with machine learning, have your preliminaries in place. Whether you're using MLaaS or hand coding, focus on your infrastructure first. Establish an infrastructure with a simple model and ensure that it works before proceeding to the next step. With this setup you'll want to add tests at every stage to examine your infrastructure, data and machine learning models. Furthermore, you'll want some kind monitoring to detect model degradation and failures.

During the process of development be cautious not to lose any useful data or ignore heuristics - it may be possible to turn them into features for your machine learning model.

Guideline 2 Feature Engineering
Once you are set up and running with initial machine learning there's a lot of low hanging fruit. You're now ready to add and remove various features from your model. Start with the directly observable features, leaving your learned features, features that are generated by the model itself or an external system, for future updates. Business objectives are likely to evolve over time along with new features appearing so plan to iterate. Ensure that you're not adding features that make the model too complex for future editions. The most important point in this phase is to frequently test your model to measure how it performs on training, validation, testing and production data. Track how they differ and go back to try and minimise the difference.

Guideline 3 Slowed Growth, Optimization Refinement, and Complex Models
The guideline we offer is about debugging and refining your model. At some point you may find that the performance of your model is starting to decline. In this situation you'll want to revisit your business objectives to ensure that both the model and the objectives are still aligned. From there you may want to consider adding new features from a different source or introducing a new level of sophistication by implementing ensemble models. Ensemble models combine the scores from different models and can usually generate better results. However, the rule of thumb is to keep things simple. Each model should either be an ensemble only taking the input of other models, or a base model taking many features, but not both. If you have models on top of other models that are trained separately, combining them can result in bad behavior.

Conclusion

Machine learning can help your business process and understand data insights faster - empowering data-driven decisions to be made across your organization. For machine learning to be successful, however, your models will need to consume clean data sets. As the quality of your data increases, you can expect the quality of our insights to increase as well. Transforming data for analysis can be challenging based on the growing volume, variety and velocity of big data. This challenge will need to be overcome to unlock the potential of your data and to mobilize your business to move faster and outpace competitors. When you are ready for machine learning, Matillion's purpose-built data transformation for cloud data warehouses can help you increase the ROI on your data, transforming your data so it is machine learning ready!

About Matillion

Matillion provides industry-leading data transformation products for cloud data warehouses. Delivering a true end-to-end data transformation (not just data preparation or movement of data 'as-is' from one location to another), Matillion provides an instant-on experience to get you up and running in just a few clicks, a pay-as-you-go billing model to cut out lengthy procurement processes, and an intuitive user interface to minimize technical pain and speed up time to results. Matillion is available globally for Amazon Redshift, Snowflake, and Google BigQuery on leading cloud infrastructures.

Find out more at **www.matillion.com**

MATILLION

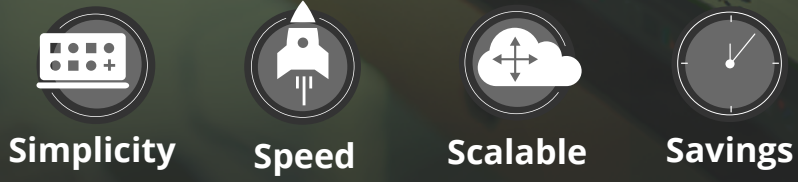
Simplicity. Speed. Scalability. Savings.

Get a demo to see first-hand the power of Matillion data transformation

GET A DEMO

Purpose-built data transformation for cloud data warehouses

- *Our intuitive UI and approach to data transformation makes complex tasks simple*
- *We deliver the fastest time to value, from launch to development to production*
- *Built to take advantage of the power and features of Amazon Redshift, Snowflake, and Google BigQuery*
- *Pay-as-you-go with no long term commitments*



matillion.com/get-a-demo/