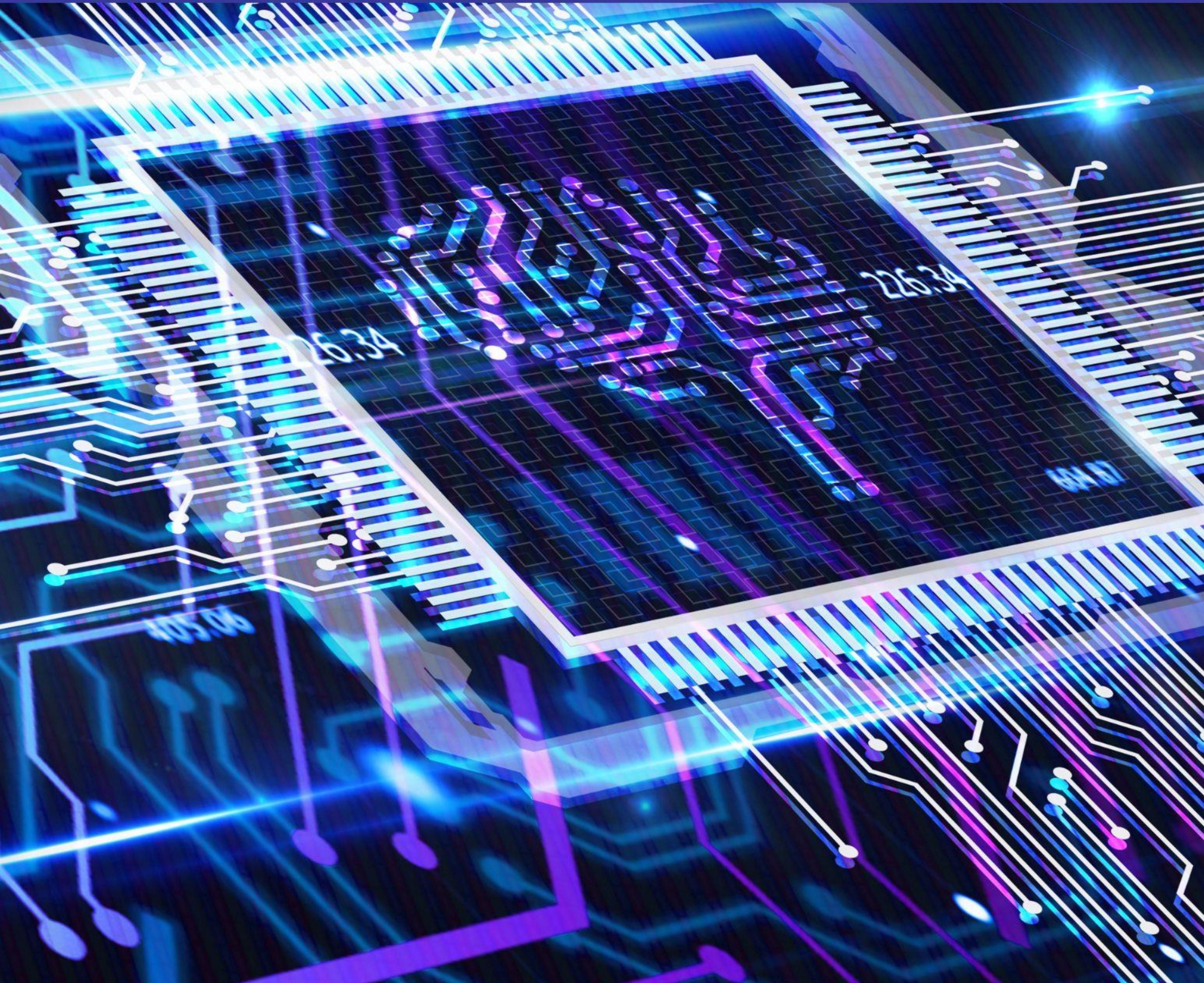
TOP TRENDS IN MACHINE LEARNING FOR 2023
Written for Rootstrap, Inc.



By Ravi Das

Introduction

In the world of automation, whether it be related to Cybersecurity, DevSecOps, or even just software development, the use of Artificial Intelligence (AI) has started to emerge as a strong tool to deliver projects on time and under budget for the client. Probably its greatest advantage is in speed. For example, depending upon the scope and the depth of the project, it can take a software developer quite a bit of time to run a batch of Quality Assurance (QA) checks to make sure that source code is free from any defects.

But with the use of AI, many of these mundane processes can be automated through various scripts, such as PHP, PERL, Python, etc. In this case, many of these defects that have been found can be self-corrected, but if it is more complex, then alerts can be sent out to the team for quick remediation. But keep in mind that AI is a very broad area of study, with other subfields in it, notably those of the following:

- Machine Learning;
- Neural Networks;
- Computer Vision.

In this whitepaper, the emphasis is on the first one.

A Formal Definition of Machine Learning

Although Machine Learning (ML) may sound complex at first, it is really not much different than AI. Here is a fundamental definition of Machine Learning:

" . . . It is an application of AI that enables systems to learn and improve from experience without being explicitly programmed. Machine learning focuses on developing computer programs that can access data and use it to learn for themselves."

(SOURCE: 1).

So, one of the key differences between Machine Learning and AI is that with the former, the ultimate objective is to have it learn on its own, with very little human intervention, if any is needed. But with the latter, the system needs to be fed a large amount of data not only for it to learn from the very beginning, but also to keep it optimized for future learning cycles as well.

Therefore, a lot of human intervention is required, not only to keep the system running on a 24 X 7 365 basis, but also checking the datasets are totally "cleansed" so that the wrong output is not calculated, and to eliminate any possibilities of statistical based skewness.

Also, AI systems work best when they are ingested with a lot of quantitative, or structured data. It does not handle qualitative data, or unstructured data as well. On the other hand, an ML system can handle both kinds of datasets quite efficiently, with smaller amounts of it. Also, one of the other primary goals of an ML system is for it to have the ability to recognize what is cleansed and uncleansed data and not to be confused with the two.

The High-Level Overview Of Machine Learning

Although Machine Learning has been around for a long time (some estimates have it as long as a couple of decades), there are number of key applications in which Machine Learning is used. Some examples of these are as follows:

1) Predictive Maintenance:

   This kind of application is typically used in supply chain, manufacturing, distribution, and the logistics sectors. For example, this is where the concept of Quality Control comes into key play. In manufacturing, you want to be able to predict how many batches of products that are going to be produced could actually become defective. Obviously, you want this number to be as low as possible. Theoretically, you do not want any type or kind of product to be defective, but in the real world, this is almost impossible to achieve. With Machine Learning, you can set up the different permutations in both the mathematical and statistical based algorithms with different permutations as to what is deemed to be a defective product or not.

2) Employee Recruiting:

   There is one common denominator in the recruitment industry, and that is the plethora of resumes that recruiters from all kinds of industries get. Consider some of these statistics:

   - Just recently, Career Builder, one of the most widely used job search portals has:

     *2.3 million jobs were posted;

     *680 unique profiles of job seekers were collected;

     *310 million resumes were collected;

     *2.5 million background checks were conducted with the Career Builder platform.

   (SOURCE: 2).

   Just imagine how long it would take a team of recruiters to have to go through all of the above. But with Machine Learning, it can all be done an in a matter of minutes, by examining for certain keywords, in order to find the desired candidates. Also, rather than having the recruiter to post each and every job entry manually onto Career Builder, the appropriate Machine Learning tool can be used to completely automate this process, thus freeing up the time of the recruiter to interview with the right candidates for the job.

3) Customer Experience:

   In the American society of today, we want to have everything right here and right now, at the snap of a finger. Not only on top of this, but we also expect to have impeccable customer service delivered at the same time. And when none of this happens, well, we have the luxury to go to a competitor to see if they can do any better. In this regard, many businesses and corporations have stated to make use of Virtual Agents. These are the little chat boxes typically found on the lower right part of your web browser. With this, you can actually communicate with somebody in order to get your questions answered or shopping issues resolved. The nice thing about these is that they are also on demand, on a 24 X 7 X 365 basis. But, in order to provide a seamless experience to the customer or prospect, many business entities are now making use of what are known as "Chat Bots". These are a much more sophisticated version of the Virtual Agent because they make use of Machine Learning algorithms. By doing this, the Chat Bot can find much more specific answers to your queries by conducting more "intelligent" based searches in the information repositories of the business or corporation. Also, many call centers are making use of Machine Learning as well. In this particular fashion, when a customer calls in, their call history, profile, and entire conversations are pulled up in a matter of seconds for the call center agent, so that they can much easily anticipate your questions, and provide you with the best levels of service that are possible.

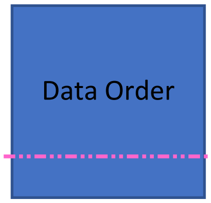
4) Finance:

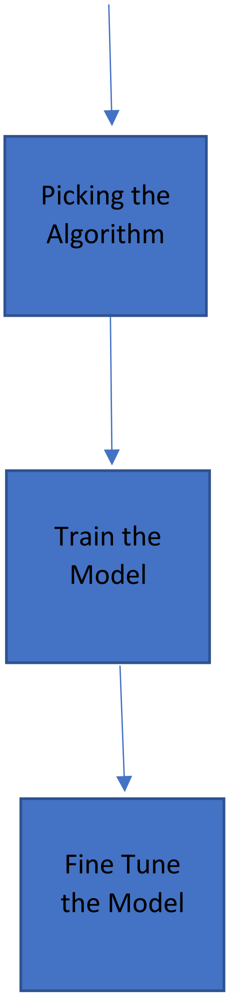
In this market segment, there is one thing that all people, especially the traders, want to do. And that is, to have the ability to predict the financial markets, and what they will do in the future, so that they can hedge their bets and make profitable trades. Although this can be sone via a manual process, it can be a very laborious and time-consuming process to achieve. Of course, we all know that the markets can move in just a matter of seconds with uncertain volatility, as we have seen recently with the Coronavirus. In fact exactly timing and predicting the financial markets with 100% accuracy is an almost impossible feat to accomplish. But this is where the role of Machine Learning can come into play. For example, it can take all of that data that is fed into it, and within a matter of seconds make more accurate predictions as to what potentially the market can do, giving the traders that valuable time to make the split-second decisions that are needed to produce quality trades. This is especially useful for what is known as "Intra Day Trading", where the financial traders try to time the market as they are open on a minute-by-minute basis.

The Machine Learning Process

When you are applying Machine Learning to a particular question that you want answered or to predict a certain outcome, it is very important to follow a distinct process in order to accomplish these tasks. In other words, you want to build an effective model that can serve well for other purposes and objectives for a subsequent time down the road. In other words, you want to train this model in a particular fashion, so that it can provide a very high degree of both accuracy and reliability.

This process is depicted below:





Each of the listed steps from above is reviewed in more detail in the next section.

Further Details Into The Machine Learning Process

1) Data Order:

   In this step, you want to make sure that the data is as unorganized and unsorted as possible. Although this sounds quite contrary, if the datasets are by any means sorted or organized in any way shape or form, the Machine Learning Algorithms that are utilized may detect this as a pattern, which you do not want to happen in this particular instance.

2) Picking the Algorithm:

   In this phase, you will want to select the appropriate Machine Learning algorithms for your model. This will be heavily examined in this part of the chapter.

3) Training the Model:

The datasets that you have will fed into the Machine Learning system, in order for it to learn first. In other words, various associations and relationships will be created and examined so that the desired outputs can be formulated. For example, one of the simplest algorithms that can be used in Machine Learning is the Linear Regression one, which is represented mathematically as follows:

$$Y = M*X + B$$

Where:

M = The slope on a graph;

B = Is the Y intercept on the graph.

4) Model Evaluation:

In this step, you will make use of a representative sample of data from the datasets, which are technically known as the “Test Data”. By feeding this initially into the Machine Learning system, you can gauge just how accurate your desired outputs will be in a test environment before you release your datasets into the production environment.

5) Fine Tune the Model:

In this last phase, you will adjust the permutations that you have established in the Machine Learning system so that it can reasonably come up with desired outputs that you are looking for.

In the next subsection, we examine the major classifications and types of Machine Learning Algorithms that are commonly used today.

The Machine Learning Algorithm Classifications

There are four major categorizations of the Machine Learning Algorithms, and they are as follows:

1) Supervised Learning:

These types of algorithms make use of what are known as “labelled data”. This simply means that each dataset has a certain label that is associated with them. In this instance, one of the key things to keep in mind is that you need to have a large amount of datasets in order to produce the dataset you are looking for when you are using algorithms based on this category. But if the datasets do not come already labelled, it could be very time consuming to create and assign a label for each and every one of them. This is the primary downside of using Machine Learning algorithms from this particular category.

2) Unsupervised Learning:

These kinds of algorithms work with data that is typically not labelled. Because of the time constraints it would take to create and assign the labels for each category (as just previously mentioned), you will have to make use of what are known as “Deep Learning Algorithms” in order to detect any unseen data trends that lie from within all of your datasets. In this regard, one of the most typical approaches that are used in this category is that of "Clustering". With this, you are merely taking all of the unlabeled datasets and use the various algorithms that are available from within this particular category is to put these datasets into various groups, which have common denominators or affiliations with them. To help out with this, there are a number of ways to do this, which are the following:

- The Euclidean Metric:

  This is a straight line between two independent datasets.

- The Cosine Similarity Metric:

  In this instance, a trigonometric function known as the "Cosine" to measure any given angles between the datasets. The goal here is to find any closeness or similarities between at least two or more independent datasets based upon their geometric orientation.

- The Manhattan Metric:

  This technique involves taking the summation of at least two or more absolute value distances from the datasets that you have.

- The Association:

  The basic thrust here is that if a specific instance occurs in one of your datasets, then it will also likely occur in the datasets that have some sort of relationship with the initial dataset that has been used.

- The Anomaly Detection:

  With this methodology, you are statistically identifying those outliers or other anomalous patterns that may exist from your datasets. This technique has found great usage in Cybersecurity, especially when it relates to filtering out for false positives from the log files that are collected from the Firewalls, Network Intrusion Devices, and Routers, as well as any behavior that may deemed to suspicious or malicious in nature.

- The Autoencoders:

  With this particular technique, the datasets that you have on hand will be formatted and put into a compressed type of format, and from that, it will be reconstructed once again. The idea behind this is to detect and find any sort of new patterns or unhidden trends that may exist from within your datasets.

- The Reinforcement Learning:

  In this instance, you are learning and harnessing the power of your datasets through a trial-and-error process, as the name of this category implies.

- The Semi-Supervised Learning:

This methodology is actually a mixture of both Supervised Learning and Unsupervised Learning. However, this technique is only used when you have a small amount of datasets that are actually labelled. Within this, there is a sub technique which is called "Pseudo-Labelling". In this regard, you literally translate all of the unsupervised datasets into a supervised one state of nature.

The Actual Machine Learning Algorithms

There are many types and kinds of both mathematical and statistical algorithms that are used in Machine Learning. In this subsection, we examine some of the more common ones, and we will do a deeper dive into them later in this chapter. Here are the algorithms:

1) The Naïve Bayes Classifier:

   The reason why this particular algorithm is called "naïve" is that the underlying assumption is that the variables in each of the datasets that you have are actually all independent from one another. In other words, the statistical occurrence from one variable in one dataset will have nothing to do whatsoever with the variables in the remaining datasets. But there is a counterargument to this which states that if this association will prove to be statistically incorrect if any of the datasets have actually changed in terms of their corresponding values.

   It should be noted that there also specific alterations, or variations to this particular algorithm, and they are as follows:

   - The Bermoulli:

     This is only used if you have binary values in your datasets.

   - The Multinomial:

     This technique is only used if the values in your datasets are discrete, in other words, they contain mathematical based absolute values.

   - The Gaussian:

     This methodology is used only if your datasets line up to a statistical based normal distribution.

   It should be noted that this overall technique is heavily used for analyzing in granular detail those datasets that have a text-based value assigned to them. When it comes to Cybersecurity, this technique proves to be extremely useful, when it comes to identifying and confirming Phishing based Emails by examining the key features and patterns in the body of the Email message, the sender address, and the content in the subject line.

2) The K-Nearest Neighbor:

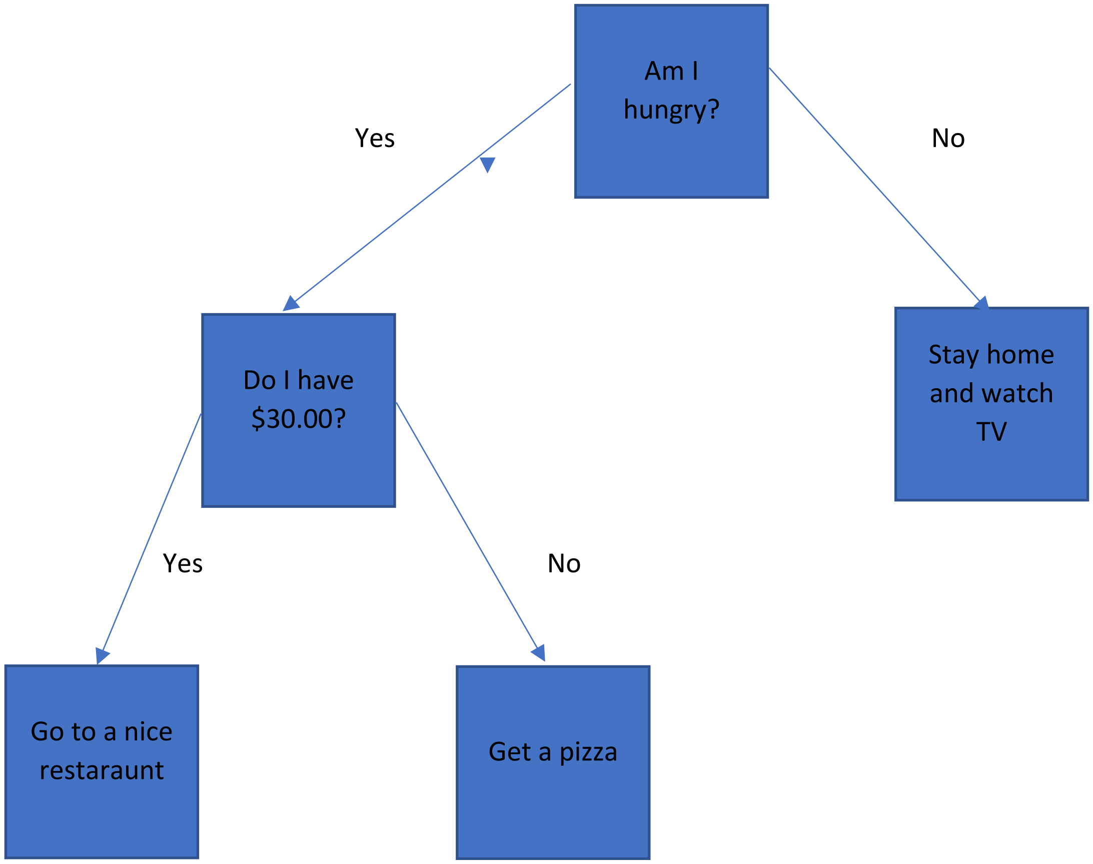
   This specific methodology is used for classifying any dataset or datasets that you have. The basic, theoretical construct of this those values that are closely related, or associated with one another in your datasets will statistically be good predictors for a Machine Learning based model. In order to use this, you need first need to compute the numerical distance between the closest values. If these values are quantitative based, you could then use the Euclidean Distance formula. But if your datasets have some sort of qualitative you could then use what is known as the “Overlap Metric”. Next, you will then have to ascertain the total number of values that are closely aligned with another. While having more of these kinds of values in your datasets could mean a much more efficient and robust Machine Learning Model, it also translates into using much more processing resources of your Machine Learning System. To help accommodate for this, you can always assign higher value statistical weights to those particular values that are closely affiliated with one another.

3) The Linear Regression:

This kind of methodology is strictly statistical based. This means that it tries to examine and ascertain the relationship between preestablished variables that reside from within your datasets. With this, a line is typically plotted, and can be further smoothened out using a technique called “Least Squares”.

4) The Decision Tree:

This methodology actually provides an alternative to the other techniques described thus far. In fact, the Decision Tree works far better and much more efficiently with non-numerical data, such as those that deal with text-based values. The main starting point of the decision is at the node, which typically starts at the top of any given chart. From this point onwards, there will be a series of decision branches that will be stemming out, thus giving it its name. The following example depicts a very simple example of a Decision Tree:



The above is of course, a very simple Decision Tree to illustrate the point. But when it comes to

5) Machine Learning, they can become very long, detailed and much more complex. One of the key advantages of using a Decision Tree is that they can actually work very well with very large datasets, and provide a degree of transparency during the Machine Leaning Model building process.

   But, on the flip side, a Decision Tree can also have its serious disadvantages as well. For example, if just one branch of it fails, it will have a negative, cascading effect in a downward effect on the other branches of the Decision Tree.

6) The Ensemble Model:

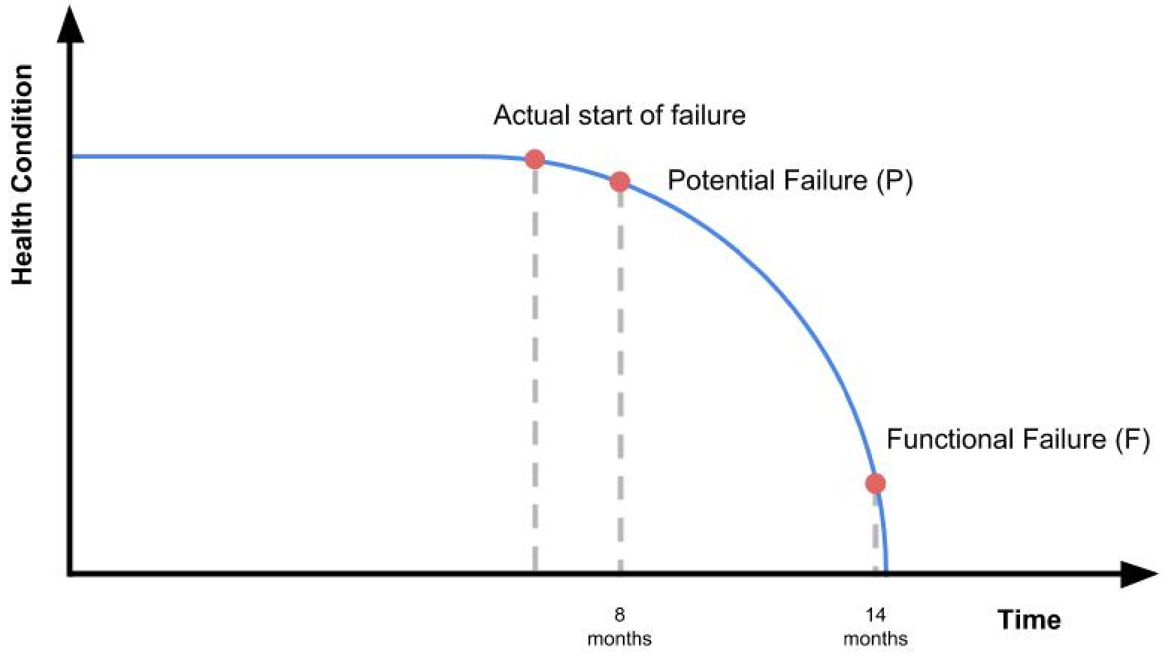
   As its name implies, this particular technique means using more than just one model, it uses a combination of what has been reviewed so far.

So far in this whitepaper, we have examined of what the technical details are in ML. The remainder of this whitepaper will now review of what the top trends of it will be going into 2023.

The IoT & Machine Learning

One of the areas in which Machine Learning is making a big splash is in the world of manufacturing. One of the main drivers behind this is the Internet of Things (IoT). In its most simple form, this is where all of the objects that we interact with are interconnected amongst one another in both the virtual and physical worlds. Given the technological advancements in this market sector, many specialized processes and even robots are now connected together to create a seamless and autonomous manufacturing process.

One area where ML is starting to be used is in the actual detection of any small irregularities in the production line which indicate a larger issue could be at hand, such as the breakdown of equipment. The ML application can create what is known as a "Potential – Failure" (PF) Graph, which is a visual demonstration when the manufacturing or robotic could potentially fail. An example of this is illustrated below:



(SOURCE: 3).

By having some idea of when a potential breakdown could happen, companies can become much more proactive in fixing their equipment, and even ordering spare parts well ahead of time should anything catastrophic happen. This is especially advantageous now when machine parts are in great shortage.

In a different world of IoT and ML, there is a strong potential for using it in Edge Computing. Simply put, Edge Computing is where you bring the IoT device closer to the point of processing and transacting data, rather than having the device connect to various sorts of APIs to connect to the central point of the database. By also embedding ML here, a company can realize some of these benefits:

*Automatic scalability can happen, which allow for the company to adapt to changes that are looming on the horizon. By using ML here, each of the IoT devices will have the "intelligence" to handle their data processing workloads.

*Data privacy will become much less of a serious issue.

*The IoT devices at the edge can theoretically make their own decisions, provided that the ML system that is overseeing them is being fed with enough real time data.

*The IoT will be even able to make its own decisions as to which ML model it wants to run to process the data that it is querying and extracting for, that is fit for its own environment.

The bottom line here is that with ML being used in Edge Computing, each of the IoT devices can be managed automatically from a central point of location, without hardly any human intervention being needed.

Natural Language Processing (NPL)

Another field where ML is starting to make a strong impact is in the area of what is known as Natural Language Processing, or "NLP" for short. A good definition of it is as follows:

"[NLP] . . . enables machines to understand the human language. Its goal is to build systems that can make sense of text and automatically perform tasks like translation, spell check, or topic classification."

(SOURCE: 4).

A perfect example of NLP are the Virtual Personal Assistants such as those of Siri and Cortana. For instance, as you are driving around in your car, they can keep track of the geographic area in which you are at, and from there, provide recommendations of restaurants, cafes, and other such establishments that fit your profile when you first initiated them. But they simply don't give text-based outputs, rather they actually talk it out to you, in the voice/language that they have been programmed to speak in.

Another area where NLP is being used extensively is that of the Chatbot. These are simply the chatting mechanisms that you see in the lower, right-hand side of your web browser. The idea of using ML here as the back engine is to provide a smart answer back to the person using it, not just canned message. For example, if you have to contact a technical support line, the goal of the chatbot would be to help provide an answer that is unique to your situation, thus providing some sort of remedy to the situation.

These intelligent based Chatbots can be used on the other side as well. For instance, if you call into the customer support number of an online merchant, your profile will come up automatically, with a summary of your previous chats, as well as your buying preferences. This helps the sales rep not only customize the call to your needs, but it makes it more efficient as well.

It is important to note at this point that NLP is a part of AI, is ML. In fact, all three of them have been used in conjunction with another, which has caused a lot of confusion. But the truth of the matter in which how AI and ML is used in NLP all have their own specific objectives. The former is used to power the reasoning process of the NLP package, whereas the latter is used to automate it so that you can get quick responses in jus a matter of a few, short seconds based on previous interactions. In this situation, the objective here is to have the NLP package learn on its own, and refine its answers to queries without any extra source code programming.

The ML Algorithms of Natural Language Processing

NLP makes use of two very powerful algorithms (which of course can be customized to fit the needs of the environment), and they are as follows:

1) Syntactic Analysis:

   This makes use of the rules of grammar to help identify with the structures of the sentences, the organization of words, and how they all relate to on another. In order to for the NLP package to digest this, some of the following techniques are used:

   - Tokenization: This is where the sentences are broken into smaller bits, which are known as "tokens".

   - Tagging: This is where grammar labels are applied to the words in the sentence, such as verbs, adjectives, pronouns, nouns, adverb, etc. By following this, the NLP package can much better understand the full scope of what is said or being asked of it.

   - Stop word removal: This is where words that are used in repetitious patterns with no real meaning to them are discarded. This helps the NLP package to stay on track in order to provide the right answer.

2) Semantic Analysis:

   This is an algorithm that focuses primarily upon discerning the actual meaning of the text. For example, each word in a sentence if first analyzed, then it builds the sentence back up again in order to gain an understanding of what the whole thing means. The specific techniques that are used here include the following:

   - Word sense disambiguation: This attempts to identify of how the context of the words are being used for the statement or question being asked. For example, if it is a help desk related issue, the NLP package will try to understand how the words are being used, and from there provide the most accurate answers possible. However, the term "accurate" is a concept that is still hotly debated in the world of ML today.

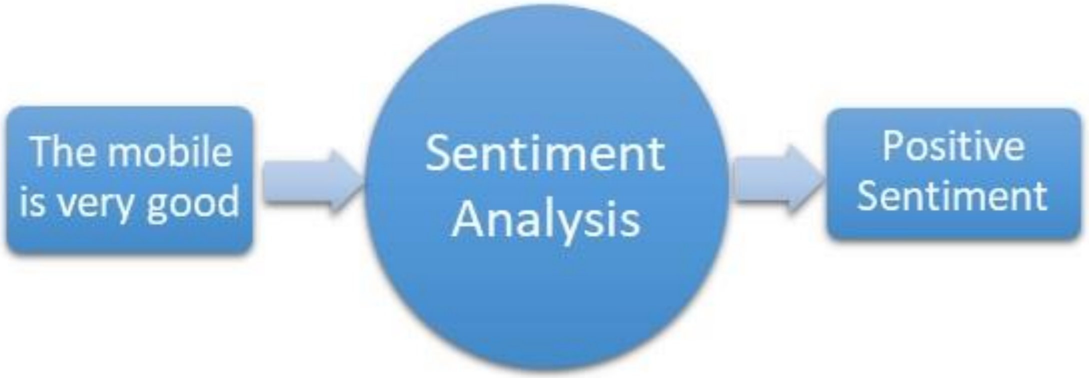
   - Relationship extraction: This makes an effort to understand how the words in a statement or query are related to one another. This technique is quite helpful in the sense that it prevents the ML package to repeatedly ask for clarification when it is presented with a query from the end user.

The Use Cases Of Natural Language Processing

Apart from being used in the Virtual Personal Assistants as described earlier, NLP has other applications as well, and are described below.

1) Sentiment Analysis:

This is where a customer's opinion or sentiments are used to help interpret their feelings about a certain product or service. This is especially useful when conducting market research about an upcoming launch:



(SOURCE: 5).

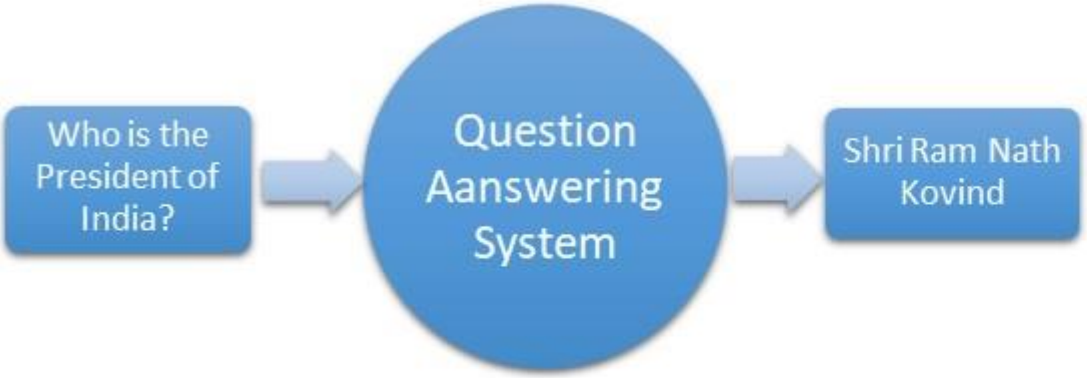
It is important to note that in this application, another set of ML algorithms are used, which are known as "Deep Learning".

2) <u>Systems that answer questions</u>:

This is where the Chatbot answers queries that are presented to it. As stated before, the goal of the ML system is to provide specific answers that are relevant to the end user, and just some general statement. These kinds of Chatbots can actually be quite complex in nature, as the are made up of three subcomponents:

- Question processing;
- Information extraction;
- Answer processing.

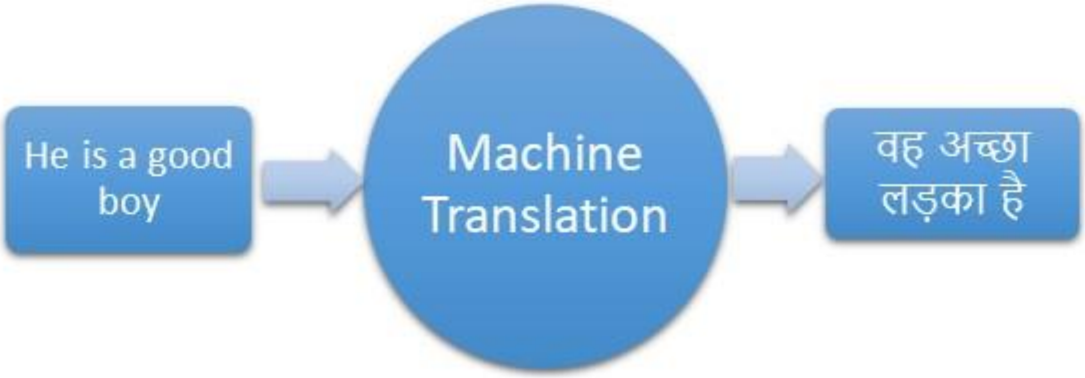
The goal here is fo the Chatbot to actually understand the question, and not just do a keyword comparison of what is actually stored in its database.



(SOURCE: 5).

3) Machine Translation:

This used most when it comes to foreign language translation, such as translating to English to Hindi an vice versa. This is area of ML is deemed to be one of the hardest things yet to fully accomplish as a perfect translation is still not yet fully possible. For example, the two primary algorithms that are being used here are those of the Recurrent Neural Network Deep Learning and the Gated Recurrent Unit, which even includes their bidirectional components.



(SOURCE: 5).

Greater Advancements In Cybersecurity

Probably one of the most widely used of ML has been in the world of Cybersecurity. This world is constantly changing on a dynamic basis, and it is simply far too difficult for any human being to keep track of it all. This is where ML is being called into play, as even AI. The following are some of the top uses of ML in Cyber at the present time:

1) Detecting threats on the networks:

Typically, this has been the role of the Firewalls, Routers, and the Network Intrusion Devices. But the can only detect threats that are based upon known signature profiles, they cannot take different pieces of data, put them together, and from there draw different types of conclusions as to what a new threat variant could look potentially look like. This is where ML can be used because it can draw upon the outcomes that have already been produced from an AI system.

2) Staying away from phony websites:

Although this kind of threat has been around for quite some time, it picked in volume rapidly once the COVID-10 pandemic hit. Although there are databases out there of blacklisted domains, ML takes this one step further in the sense that it is always on the prowl, on a real time basis constantly looking for phony websites that pop up. It does not wait for people to report them.

3) Protect the endpoints:

Even to this day, many businesses fail to protect their endpoints, which represent the starting and stopping points of the network lines of communications. Therefore, the Cyberattacker is taking full advantage of this, staying at these points until they can move further in a lateral, covert fashion. But with ML, the endpoints are now receiving more attention, thus capturing the Cyberattacker at even earlier stages in the dame.

4) Securing the Cloud:

Today, many businesses are migrating 100% into the Cloud, whether it be the AWS or Microsoft Azure. But here too, the Cyberattacker is on the prowl, and these Cloud providers are offering great tools to those account holders that make use of them. For example, Azure has what is known as Microsoft Sentinel, which also provides a real time dashboard to the IT Security team to give them a macro view of what is going on in their Cloud environment, and to help reduce onslaught of false positives.

5) Malware detection in the VPN:

For the longest time, the use of VPNs have used for employees working remotely. While they provide great levels of security, they simply are not designed to handle the capacity of the near 99% Remote Workforce that we are seeing today. Therefore, the Cyberattacker is able to take advantage of this, and deploy covert malicious payloads into the VPN. But with recent training advancements being made in ML, is now able to detect these kinds of threats and alert the IT Security team of such.

Conclusions

Overall, this whitepaper has examined in detail what ML is, and what is projected for its uses going into 2023. A future whitepaper will examine the next generation of trends of ML, which will include the following:

- The Use Of Deepfakes & Concerns of Ethics;
- Automated Coding Techniques;
- Multimodal Learning;
- Models With More Than Objective;
- The Birth of Tiny ML;
- Quantum ML;
- The Development Of Digital Twins;
- The Metaverse.

Sources

1) https://www.expert.ai/blog/machine-learning-definition/
2) “Artificial Intelligence Basics: A Non-Technical Introduction”. Tom Tauli, Apress, 2019.
3) https://www.upkeep.com/learning/p-f-curve
4) https://monkeylearn.com/blog/nlp-ai/
5) https://www.analyticsvidhya.com/blog/2021/04/role-of-machine-learning-in-natural-language-processing/
6) https://odsc.medium.com/recent-advances-in-machine-learning-with-applications-to-iot-23cb43382a17
7) https://valohai.com/blog/mlops-for-iot-and-edge/