***1.Explain the term machine learning, and how does it work? Explain two machine learning applications in the business world. What are some of the ethical concerns that machine learning applications could raise?***

Ans->Machine means our Computer. Computer made of silicon chips which is help us to build and ,or etc gate for logical operation. When we import machine learning algo for example linear regression we give our pc y=mx+c now with the help of and ,or gates machine found best value of m and c its called learning.

Machine Learning in finance helps in portfolio management, algorithmic trading, loan underwriting, and fraud detection. However, future applications of Machine Learning in finance will include Chatbots and other conversational interfaces for security, customer service, and sentiment analysis.

* Cost to innovation.
* Harm to physical integrity.
* Lack of access to public services.
* Lack of trust.
* “Awakening” of AI.
* Security problems.
* Lack of quality data.
* Disappearance of job

***2. Describe the process of human learning:***

i. Under the supervision of experts🡪Teacher teach his students(spoon feeding). Here teacher is an expert and he supervised students.(Supervised machine learning)

ii. With the assistance of experts in an indirect manner🡪Learning from an institute who just provide study material. Basically they just give us the pattern how to study.( unsupervised learning)

iii. Self-education🡪As it says completely depend on yourself.( Reinforcement learning )

***3. Provide a few examples of various types of machine learning.***

Ans🡪 Supervised machine learning, unsupervised learning, Reinforcement learning.

***4. Examine the various forms of machine learning.***

Ans🡪 Supervised learning is a machine learning approach that’s defined by its use of labeled datasets. These datasets are designed to train or “supervise” algorithms into classifying data or predicting outcomes accurately. Using labeled inputs and outputs, the model can measure its accuracy and learn over time.

Supervised learning can be separated into two types of problems when data mining: classification and regression:

Classification problems use an algorithm to accurately assign test data into specific categories, such as separating apples from oranges. Or, in the real world, supervised learning algorithms can be used to classify spam in a separate folder from your inbox. Linear classifiers, support vector machines, decision trees and random forest are all common types of classification algorithms.

Regression is another type of supervised learning method that uses an algorithm to understand the relationship between dependent and independent variables. Regression models are helpful for predicting numerical values based on different data points, such as sales revenue projections for a given business. Some popular regression algorithms are linear regression, logistic regression and polynomial regression.

Unsupervised learning uses machine learning algorithms to analyze and cluster unlabeled data sets. These algorithms discover hidden patterns in data without the need for human intervention (hence, they are “unsupervised”).

Unsupervised learning models are used for three main tasks: clustering, association and dimensionality reduction:

Clustering is a data mining technique for grouping unlabeled data based on their similarities or differences. For example, K-means clustering algorithms assign similar data points into groups,

where the K value represents the size of the grouping and granularity. This technique is helpful for market segmentation, image compression, etc.

Association is another type of unsupervised learning method that uses different rules to find relationships between variables in a given dataset. These methods are frequently used for market basket analysis and recommendation engines, along the lines of “Customers Who Bought This Item Also Bought” recommendations.

Dimensionality reduction is a learning technique used when the number of features (or dimensions) in a given dataset is too high. It reduces the number of data inputs to a manageable size while also preserving the data integrity. Often, this technique is used in the preprocessing data stage, such as when autoencoders remove noise from visual data to improve picture quality.

In reinforcement learning, developers devise a method of rewarding desired behaviors and punishing negative behaviors. This method assigns positive values to the desired actions to encourage the agent and negative values to undesired behaviors. This programs the agent to seek long-term and maximum overall reward to achieve an optimal solution.

These long-term goals help prevent the agent from stalling on lesser goals. With time, the agent learns to avoid the negative and seek the positive. This learning method has been adopted in artificial intelligence (AI) as a way of directing unsupervised machine learning through rewards and penalties.

***5. Can you explain what a well-posed learning problem is? Explain the main characteristics that must be present to identify a learning problem properly.***

Ans🡪 Well Posed Learning Problem – A computer program is said to learn from experience E in context to some task T and some performance measure P, if its performance on T, as was measured by P, upgrades with experience E.

Any problem can be segregated as well-posed learning problem if it has three traits –

Task

Performance Measure

Experience

Certain examples that efficiently defines the well-posed learning problem are –

1. To better filter emails as spam or not

Task – Classifying emails as spam or not

Performance Measure – The fraction of emails accurately classified as spam or not spam

Experience – Observing you label emails as spam or not spam

2. A checkers learning problem

Task – Playing checkers game

Performance Measure – percent of games won against opposer

Experience – playing implementation games against itself

3. Handwriting Recognition Problem

Task – Acknowledging handwritten words within portrayal

Performance Measure – percent of words accurately classified

Experience – a directory of handwritten words with given classifications

4. A Robot Driving Problem

Task – driving on public four-lane highways using sight scanners

Performance Measure – average distance progressed before a fallacy

Experience – order of images and steering instructions noted down while observing a human driver

5. Fruit Prediction Problem

Task – forecasting different fruits for recognition

Performance Measure – able to predict maximum variety of fruits

Experience – training machine with the largest datasets of fruits images

6. Face Recognition Problem

Task – predicting different types of faces

Performance Measure – able to predict maximum types of faces

Experience – training machine with maximum amount of datasets of different face images

7. Automatic Translation of documents

Task – translating one type of language used in a document to other language

Performance Measure – able to convert one language to other efficiently

Experience – training machine with a large dataset of different types of languages

***6. Is machine learning capable of solving all problems? Give a detailed explanation of your answer.***

Machine learning is now seen as a silver bullet for solving all problems, but sometimes it is not the answer.

“If a typical person can do a mental task with less than one second of thought, we can probably automate it using AI either now or in the near future.”

— Andrew Ng

Most people reading this are likely familiar with machine learning and the relevant algorithms used to classify or predict outcomes based on data. However, it is important to understand that machine learning is not the answer to all problems. Given the usefulness of machine learning, it can be hard to accept that sometimes it is not the best solution to a problem.

In this article, I aim to convince the reader that there are times when machine learning is the right solution, and times when it is the wrong solution.

Machine learning, a subset of artificial intelligence, has revolutionalized the world as we know it in the past decade. The information explosion has resulted in the collection of massive amounts of data, especially by large companies such as Facebook and Google. This amount of data, coupled with the rapid development of processor power and computer parallelization, has now made it possible to obtain and study huge amounts of data with relative ease.

Nowadays, hyperbole about machine learning and artificial intelligence is ubiquitous. This is perhaps rightly so, given the potential for this field is massive. The number of AI consulting agencies has soared in the past few years, and, according to a report from Indeed, the number of jobs related to AI ballooned by 100% between 2015 and 2018.

As of December 2018, Forbes found that 47% of business had at least one AI capability in their business process, and a report by Deloitte projects that a penetration rate of enterprise software with AI built-in, and cloud-based AI development services, will reach an estimated 87 and 83 percent respectively. These numbers are impressive — if you are planning to change careers anytime soon, AI seems like a pretty good bet.

So it all seems great right? Companies are happy and, presumably, consumers are also happy — otherwise, the companies would not be using AI.

It is great, and I am a huge fan of machine learning and AI. However, there are times when using machine learning is just unnecessary, does not make sense, and other times when its implementation can get you into difficulties.

Limitation 1 — Ethics

Machine learning, a subset of artificial intelligence, has revolutionalized the world as we know it in the past decade. The information explosion has resulted in the collection of massive amounts of data, especially by large companies such as Facebook and Google. This amount of data, coupled with the rapid development of processor power and computer parallelization, has now made it possible to obtain and study huge amounts of data with relative ease.

It is easy to understand why machine learning has had such a profound impact on the world, what is less clear is exactly what its capabilities are, and perhaps more importantly, what its limitations are. Yuval Noah Harari famously coined the term ‘dataism’, which refers to a putative new stage of civilization we are entering in which we trust algorithms and data more than our own judgment and logic.

Whilst you may find this idea laughable, remember the last time you went on vacation and followed the instructions of a GPS rather than your own judgment on a map — do you question the judgment of the GPS? People have literally driven into lakes because they blindly followed the instructions from their GPS.

The idea of trusting data and algorithms more than our own judgment has its pros and cons. Obviously, we benefit from these algorithms, otherwise, we wouldn’t be using them in the first place. These algorithms allow us to automate processes by making informed judgments using available data. Sometimes, however, this means replacing someone’s job with an algorithm, which comes with ethical ramifications. Additionally, who do we blame if something goes wrong?

The most commonly discussed case currently is self-driving cars — how do we choose how the vehicle should react in the event of a fatal collision? In the future will we have to select which ethical framework we want our self-driving car to follow when we are purchasing the vehicle?

If my self-driving car kills someone on the road, whose fault is it?

Whilst these are all fascinating questions, they are not the main purpose of this article. Clearly, however, machine learning cannot tell us anything about what normative values we should accept, i.e. how we should act in the world in a given situation. As David Hume famously said, one cannot ‘derive an ought from an is’.

Limitation 2 — Deterministic Problems

This is a limitation I personally have had to deal with. My field of expertise is environmental science, which relies heavily on computational modeling and using sensors/IoT devices.

Machine learning is incredibly powerful for sensors and can be used to help calibrate and correct sensors when connected to other sensors measuring environmental variables such as temperature, pressure, and humidity. The correlations between the signals from these sensors can be used to develop self-calibration procedures and this is a hot research topic in my research field of atmospheric chemistry.

However, things get a bit more interesting when it comes to computational modeling.

Running computer models that simulate global weather, emissions from the planet, and transport of these emissions is very computationally expensive. In fact, it is so computationally expensive, that a research-level simulation can take weeks even when running on a supercomputer.

Good examples of this are MM5 and WRF, which are numerical weather prediction models that are used for climate research and for giving you weather forecasts on the morning news. Wonder what weather forecasters do all day? Run and study these models.

Running weather models is fine, but now that we have machine learning, can we just use this instead to obtain our weather forecasts? Can we leverage data from satellites, weather stations, and use an elementary predictive algorithm to discern whether it is going to rain tomorrow?

The answer is, surprisingly, yes. If we have knowledge of the air pressures around a certain region, the levels of moisture in the air, wind speeds, and information about neighboring points and their own variables, it becomes possible to train, for example, a neural network. But at what cost?

Using a neural network with a thousand inputs to determine whether it will rain tomorrow in Boston is possible. However, utilizing a neural network misses the entire physics of the weather system.

Machine learning is stochastic, not deterministic.

A neural network does not understand Newton’s second law, or that density cannot be negative — there are no physical constraints.

However, this may not be a limitation for long. There are multiple researchers looking at adding physical constraints to neural networks and other algorithms so that they can be used for purposes such as this.

Limitation 3 — Data

This is the most obvious limitation. If you feed a model poorly, then it will only give you poor results. This can manifest itself in two ways: lack of data, and lack of good data.

Lack of Data

Many machine learning algorithms require large amounts of data before they begin to give useful results. A good example of this is a neural network. Neural networks are data-eating machines that require copious amounts of training data. The larger the architecture, the more data is needed to produce viable results. Reusing data is a bad idea, and data augmentation is useful to some extent, but having more data is always the preferred solution.

If you can get the data, then use it.

Lack of Good Data

Despite the appearance, this is not the same as the above comment. Let’s imagine you think you can cheat by generating ten thousand fake data points to put in your neural network. What happens when you put it in?

It will train itself, and then when you come to test it on an unseen data set, it will not perform well. You had the data but the quality of the data was not up to scratch.

In the same way that having a lack of good features can cause your algorithm to perform poorly, having a lack of good ground truth data can also limit the capabilities of your model. No company is going to implement a machine learning model that performs worse than human-level error.

Similarly, applying a model that was trained on a set of data in one situation may not necessarily apply as well to a second situation. The best example of this I have found so far is in breast cancer prediction.

Mammography databases have a lot of images in them, but they suffer from one problem that has caused significant issues in recent years — almost all of the x-rays are from white women. This may not sound like a big deal, but actually, black women have been shown to be 42 percent more likely to die from breast cancer due to a wide range of factors that may include differences in detection and access to health care. Thus, training an algorithm primarily on white women adversely impacts black women in this case.

What is needed in this specific case is a larger number of x-rays of black patients in the training database, more features relevant to the cause of this 42 percent increased likelihood, and for the algorithm to be more equitable by stratifying the dataset along the relevant axes.

If you are skeptical of this or would like to know more, I recommend you look at this article.

Limitation 4 — Misapplication

Related to the second limitation discussed previously, there is purported to be a “crisis of machine learning in academic research” whereby people blindly use machine learning to try and analyze systems that are either deterministic or stochastic in nature.

For reasons discussed in limitation two, applying machine learning on deterministic systems will succeed, but the algorithm which not be learning the relationship between the two variables, and will not know when it is violating physical laws. We simply gave some inputs and outputs to the system and told it to learn the relationship — like someone translating word for word out of a dictionary, the algorithm will only appear to have a facile grasp of the underlying physics.

For stochastic (random) systems, things are a little less obvious. The crisis of machine learning for random systems manifests itself in two ways:

P-hacking

Scope of the analysis

P-hacking

When one has access to large data, which may have hundreds, thousands, or even millions of variables, it is not too difficult to find a statistically significant result (given that the level of statistical significance needed for most scientific research is p < 0.05). This often leads to spurious correlations being found that are usually obtained by p-hacking (looking through mountains of data until a correlation showing statistically significant results is found). These are not true correlations and are just responding to the noise in the measurements.

This has resulted in individuals ‘fishing’ for statistically significant correlations through large data sets, and masquerading these as true correlations. Sometimes, this is an innocent mistake (in which case the scientist should be better trained), but other times, it is done to increase the number of papers a researcher has published — even in the world of academia, competition is strong and people will do anything to improve their metrics.

Scope of the Analysis

There are inherent differences in the scope of the analysis for machine learning as compared with statistical modeling — statistical modeling is inherently confirmatory, and machine learning is inherently exploratory.

We can consider confirmatory analysis and models to be the kind of thing that someone does in a Ph.D. program or in a research field. Imagine you are working with an advisor and trying to develop a theoretical framework to study some real-world system. This system has a set of pre-defined features that it is influenced by, and, after carefully designing experiments and developing hypotheses you are able to run tests to determine the validity of your hypotheses.

Exploratory, on the other hand, lacks a number of qualities associated with the confirmatory analysis. In fact, in the case of truly massive amounts of data and information, the confirmatory approaches completely break down due to the sheer volume of data. In other words, it simply is not possible to carefully lay out a finite set of testable hypotheses in the presence of hundreds, much less thousands, much less millions of features.

Therefore and, again, broadly speaking, machine learning algorithms and approaches are best suited for exploratory predictive modeling and classification with massive amounts of data and computationally complex features. Some will contend that they can be used on “small” data but why would one do so when classic, multivariate statistical methods are so much more informative?

ML is a field which, in large part, addresses issues derived from information technology, computer science, and so on, these can be both theoretical and applied problems. As such, it is related to fields such as physics, mathematics, probability, and statistics but ML is really a field unto itself, a field which is unencumbered by the concerns raised in the other disciplines. Many of the solutions ML experts and practitioners come up with are painfully mistaken…but they get the job done.

Limitation 5 — Interpretability

Interpretability is one of the primary problems with machine learning. An AI consultancy firm trying to pitch to a firm that only uses traditional statistical methods can be stopped dead if they do not see the model as interpretable. If you cannot convince your client that you understand how the algorithm came to the decision it did, how likely are they to trust you and your expertise?

As bluntly stated in “Business Data Mining — a machine learning perspective”:

“A business manager is more likely to accept the [machine learning method] recommendations if the results are explained in business terms”

These models as such can be rendered powerless unless they can be interpreted, and the process of human interpretation follows rules that go well beyond technical prowess. For this reason, interpretability is a paramount quality that machine learning methods should aim to achieve if they are to be applied in practice.

The blossoming -omics sciences (genomics, proteomics, metabolomics and the like), in particular, have become the main target for machine learning researchers precisely because of their dependence on large and non-trivial databases. However, they suffer from the lack of interpretability of their methods, despite their apparent success

***7. What are the various methods and technologies for solving machine learning problems? Any two of them should be defined in detail.***

Regression

Classification

Clustering

Dimensionality Reduction

Ensemble Methods

Neural Nets and Deep Learning

Transfer Learning

Reinforcement Learning

Natural Language Processing

Word Embeddings

Regression

Regression methods fall within the category of supervised ML. They help to predict or explain a particular numerical value based on a set of prior data, for example predicting the price of a property based on previous pricing data for similar properties.

The simplest method is linear regression where we use the mathematical equation of the line (y = m \* x + b) to model a data set. We train a linear regression model with many data pairs (x, y) by calculating the position and slope of a line that minimizes the total distance between all of the data points and the line. In other words, we calculate the slope (m) and the y-intercept (b) for a line that best approximates the observations in the data.

Let’s consider a more a concrete example of linear regression. I once used a linear regression to predict the energy consumption (in kWh) of certain buildings by gathering together the age of the building, number of stories, square feet and the number of plugged wall equipment. Since there were more than one input (age, square feet, etc…), I used a multi-variable linear regression. The principle was the same as a simple one-to-one linear regression, but in this case the “line” I created occurred in multi-dimensional space based on the number of variables.

The plot below shows how well the linear regression model fit the actual energy consumption of building. Now imagine that you have access to the characteristics of a building (age, square feet, etc…) but you don’t know the energy consumption. In this case, we can use the fitted line to approximate the energy consumption of the particular building.

Note that you can also use linear regression to estimate the weight of each factor that contributes to the final prediction of consumed energy. For example, once you have a formula, you can determine whether age, size, or height is most important.

Linear Regression Model Estimates of Building’s Energy Consumption (kWh).

Regression techniques run the gamut from simple (like linear regression) to complex (like regularized linear regression, polynomial regression, decision trees and random forest regressions, neural nets, among others). But don’t get bogged down: start by studying simple linear regression, master the techniques, and move on from there.

Classification

Another class of supervised ML, classification methods predict or explain a class value. For example, they can help predict whether or not an online customer will buy a product. The output can be yes or no: buyer or not buyer. But classification methods aren’t limited to two classes. For example, a classification method could help to assess whether a given image contains a car or a truck. In this case, the output will be 3 different values: 1) the image contains a car, 2) the image contains a truck, or 3) the image contains neither a car nor a truck.

The simplest classification algorithm is logistic regression — which makes it sounds like a regression method, but it’s not. Logistic regression estimates the probability of an occurrence of an event based on one or more inputs.

For instance, a logistic regression can take as inputs two exam scores for a student in order to estimate the probability that the student will get admitted to a particular college. Because the estimate is a probability, the output is a number between 0 and 1, where 1 represents complete certainty. For the student, if the estimated probability is greater than 0.5, then we predict that he or she will be admitted. If the estimated probabiliy is less than 0.5, we predict the he or she will be refused.

The chart below plots the scores of previous students along with whether they were admitted. Logistic regression allows us to draw a line that represents the decision boundary.

Logistic Regression Decision Boundary: Admitted to College or Not?

Because logistic regression is the simplest classification model, it’s a good place to start for classification. As you progress, you can dive into non-linear classifiers such as decision trees, random forests, support vector machines, and neural nets, among others.

***8. Can you explain the various forms of supervised learning? Explain each one with an example application.***

1. Regression

In regression, a single output value is produced using training data. This value is a probabilistic interpretation, which is ascertained after considering the strength of correlation among the input variables. For example, regression can help predict the price of a house based on its locality, size, etc.

In logistic regression, the output has discrete values based on a set of independent variables. This method can flounder when dealing with non-linear and multiple decision boundaries. Also, it is not flexible enough to capture complex relationships in datasets.

2. Classification

It involves grouping the data into classes. If you are thinking of extending credit to a person, you can use classification to determine whether or not a person would be a loan defaulter. When the supervised learning algorithm labels input data into two distinct classes, it is called binary classification. Multiple classifications means categorizing data into more than two classes.

3. Naive Bayesian Model

The Bayesian model of classification is used for large finite datasets. It is a method of assigning class labels using a direct acyclic graph. The graph comprises one parent node and multiple children nodes. And each child node is assumed to be independent and separate from the parent.

Decision Trees

A decision tree is a flowchart-like model that contains conditional control statements, comprising decisions and their probable consequences. The output relates to the labelling of unforeseen data.

In the tree representation, the leaf nodes correspond to class labels, and the internal nodes represent the attributes. A decision tree can be used to solve problems with discrete attributes as well as boolean functions. Some of the notable decision tree algorithms are ID3 and CART.

4. Random Forest Model

The random forest model is an ensemble method. It operates by constructing a multitude of decision trees and outputs a classification of the individual trees. Suppose you want to predict which undergraduate students will perform well in GMAT – a test taken for admission into graduate management programs. A random forest model would accomplish the task, given the demographic and educational factors of a set of students who have previously taken the test.

5. Neural Networks

This algorithm is designed to cluster raw input, recognize patterns, or interpret sensory data. Despite their multiple advantages, neural networks require significant computational resources. It can get complicated to fit a neural network when there are thousands of observations. It is also called the ‘black-box’ algorithm as interpreting the logic behind their predictions can be challenging.

6. Support Vector Machines

Support Vector Machine (SVM) is a supervised learning algorithm developed in the year 1990. It draws from the statistical learning theory developed by Vap Nick.

SVM separates hyperplanes, which makes it a discriminative classifier. The output is produced in the form of an optimal hyperplane that categorizes new examples. SVMs are closely connected to the kernel framework and used in diverse fields. Some examples include bioinformatics, pattern recognition, and multimedia information retrieval.

***9. What is the difference between supervised and unsupervised learning? With a sample application in each region, explain the differences.***

| **Parameters** | **Supervised machine learning technique** | **Unsupervised machine learning technique** |
| --- | --- | --- |
| Process | In a supervised learning model, input and output variables will be given. | In unsupervised learning model, only input data will be given |
| Input Data | Algorithms are trained using labeled data. | Algorithms are used against data which is not labeled |
| Algorithms Used | Support vector machine, Neural network, Linear and logistics regression, random forest, and Classification trees. | Unsupervised algorithms can be divided into different categories: like Cluster algorithms, K-means, Hierarchical clustering, etc. |
| Computational Complexity | Supervised learning is a simpler method. | Unsupervised learning is computationally complex |
| Use of Data | Supervised learning model uses training data to learn a link between the input and the outputs. | Unsupervised learning does not use output data. |
| Accuracy of Results | Highly accurate and trustworthy method. | Less accurate and trustworthy method. |
| Real Time Learning | Learning method takes place offline. | Learning method takes place in real time. |
| Number of Classes | Number of classes is known. | Number of classes is not known. |
| Main Drawback | Classifying big data can be a real challenge in Supervised Learning. | You cannot get precise information regarding data sorting, and the output as data used in unsupervised learning is labeled and not known. |

***10. Describe the machine learning process in depth.***

What is Machine Learning?

Machine Learning: Machine Learning (ML) is a highly iterative process and ML models are learned from past experiences and also to analyze the historical data. On top, ML models are able to identify the patterns in order to make predictions about the future of the given dataset.

machine learning testimonials

Why is Machine Learning Important?

Since 5V’s are dominating the current digital world (Volume, Variety, Variation Visibility, and Value), so most of the industries are developing various models for analyzing their presence and opportunities in the market, based on this outcome they are delivering the best products, services to their customers on vast scales.

machine learning 5Vs

What are the major Machine Learning applications?

Machine learning (ML) is widely applicable in many industries and its processes implementation and improvements. Currently, ML has been used in multiple fields and industries with no boundaries. The figure below represents the area where ML is playing a vital role.

machine Learning Applications

Where is Machine Learning in the AI space?

Just have a look at the Venn Diagram, we could understand where the ML in the AI space and how it is related to other AI components.

As we know the Jargons flying around us, let’s quickly look at what exactly each component talks about.

machine Learning in AI space

How Data Science and ML are related?

realtion with Data science

Machine Learning Process, is the first step in ML process to take the data from multiple sources and followed by a fine-tuned process of data, this data would be the feed for ML algorithms based on the problem statement, like predictive, classification and other models which are available in the space of ML world. Let us discuss each process one by one here.

machine learning process

Machine Learning – Stages: We can split ML process stages into 5 as below mentioned in the flow diagram.

Collection of Data

Data Wrangling

Model Building

Model Evaluation

Model Deployment

Identifying the Business Problems, before we go to the above stages. So, we must be clear about the objective of the purpose of ML implementation. To find the solution for the given/identified problem. we must collect the data and follow up the below stages appropriately.

Identify the stages

Collection of Data

Data collection from different sources could be internal and/or external to satisfy the business requirements/problems. Data could be in any format. CSV, XML.JSON, etc., here Big Data is playing a vital role to make sure the right data is in the expected format and structure.

data collection

Data Wrangling and Data Processing: The main objective of this stage and focus are as below.

Data Processing (EDA):

Understanding the given dataset and helping clean up the given dataset.

It gives you a better understanding of the features and the relationships between them

Extracting essential variables and leaving behind/removing non-essential variables.

Handling Missing values or human error.

Identifying outliers.

The EDA process would be maximizing insights of a dataset.

Feature engineering:

Handling missing values in the variables

Convert categorical into numerical since most algorithms need numerical features.

Need to correct not Gaussian(normal). linear models assume the variables have Gaussian distribution.

Finding Outliers are present in the data, so we either truncate the data above a threshold or transform the data using log transformation.

Scale the features. This is required to give equal importance to all the features, and not more to the one whose value is larger.

Feature engineering is an expensive and time-consuming process.

Feature engineering can be a manual process, it can be automated

machine learning steps

Training and Testing:

The training data is used to make sure the machine recognizes patterns of the data, cross-validation of data is used to ensure better accuracy and

the efficiency of the algorithm which is used to train the machine.

Test data is used to see how well the machine can predict new answers based on its training.

The train-test split procedure is used to estimate the ML performance of algorithms when they are used to make predictions on data that is not

used to train the model.

train and test

Training

Training data is the data set on which you train the model.

Train data from which the model has learned the experiences.

Training sets are used to fit and tune your models.

Testing

Test data is the data which is used to check if the model has

learnt good enough from the experiences it got in the train data set.

Test sets

are “unseen” data to evaluate your models.

Train data: It trains our machine learning algorithm

Test data: After the training the model, test data is used to test its efficiency and performance of the model

The purpose of the random state in train test split: Random state ensures that the splits that you generate are reproducible. The random state that you provide is used as a seed to the random number generator. This ensures that the random numbers are generated in the same order.

data split

Data Split into Training/Testing Set

We used to split a dataset into training data and test data in the machine learning space.

The split range is usually 20%-80% between testing and training stages from the given data set.

A major amount of data would be spent on to train your model

The rest of the amount can be spent to evaluate your test model.

But you cannot mix/reuse the same data for both Train and Test purposes

If you evaluate your model on the same data you used to train it, your model could be very overfitted. Then there is a question of whether models can predict new data.

Therefore, you should have separate training and test subsets of your dataset.

MODEL EVALUATION: Each model has its own model evaluation mythology, some of the best evaluations are here.

Evaluating the Regression Model.

Sum of Squared Error (SSE)

Mean Squared Error (MSE)

Root Mean Squared Error (RMSE)

Mean Absolute Error (MAE)

Coefficient of Determination (R2)

Adjusted R2

Evaluating Classification Model.

Confusion Matrix.

Accuracy Score.

AUC and ROC.|

Deployment of an ML-model simply means the integration of the finalized model into a production environment and getting results to make business decisions.

a. Make brief notes on any two of the following:

i. MATLAB is one of the most widely used programming languages.

ii. Deep learning applications in healthcare

***iii. Study of the market basket***

Market Basket Analysis, also known as Affinity Analysis, is a modeling technique based on the theory that if you buy a certain group of items, you’re more likely to purchase another group of items. For example, someone purchasing peanut butter and bread is far more likely to also want to purchase jelly. However, not all relationships are as immediately obvious. Foreknowledge of consumer behavior can increase sales and give the retailer a significant edge against competitors. Strictly speaking, Market Basket Analysis is just one application of association analysis techniques, although many online articles and tutorials may confuse the two. To put it in perspective of other machine learning techniques I’ve written about before, Market Basket Analysis is an unsupervised learning tool that requires little in the way of feature engineering and a limited amount of data cleaning and preparation. In practice, insights gleaned from Market Basket Analysis can be further explored with other AI or data science tools.

Despite its ability to uncover hidden patterns, Market Basket Analysis is relatively easy to explain and doesn’t require knowledge of advanced statistics or calculus. However, there are a few terms and conventional notations to review. First, the notions of cause and effect are referred to as antecedent and consequent**. In the example I mentioned previously, peanut butter and bread are the antecedent and jelly is the consequent**. The formal notation for this relationship would be {Peanut Butter, Bread} -> {Jelly} indicating that there’s a connection between these items. Also take note that both antecedents and consequents can consist of multiple items.

There are three important mathematical measures required for Market Basket Analysis: Support, Lift and Confidence. Support represents the number of times antecedents appear together in the data. To simplify the example, imagine the following relationship: {Peanut Butter} -> {Grape Jelly}. Given 100 customers (and one transaction per customer), consider the following scenario:

* 15 customers bought Peanut Butter
* 13 bought Grape Jelly
* 11 bought Peanut Butter and Grape Jelly

Support represents the number of times items appear in a transaction together, which in this example is 11 out of 100, or 0.11. To use statistical terms, there’s a probability of 11 percent that any given transaction will include both Peanut Butter and Grape Jelly. Confidence takes the value of Support (.11) and divides it by the probability of a transaction of having Grape Jelly alone, equating to a value of 0.846. This means that nearly 85 percent of the time that Grape Jelly was purchased, it was purchased along with Peanut Butter. Finally, there’s Lift, which takes Confidence (0.846) and divides it by the probability of Peanut Butter. This equate to 5.64 (rounded to two decimal places).

All this might be clearer in a simple chart, as shown in **Figure 1**.

**Figure 1 Support Confidence and Lift Values**

|  |  |  |
| --- | --- | --- |
| **TABLE 1** | | |
| Measure | Formula | Value |
| Support | P(Peanut Butter & Grape Jelly) | .011 |
| Confidence | Support / P(Grape Jelly) | 0.846 |
| Lift | Confidence / P(Peanut Butter) | 5.64 (rounded) |

***iv. Linear regression (simple)***

**Linear Regression** is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.  


Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.  
In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

**Hypothesis function for Linear Regression :**  

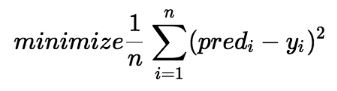

While training the model we are given :  
**x:** input training data (univariate – one input variable(parameter))  
**y:** labels to data (supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ1 and θ2 values.  
**θ1:** intercept  
**θ2:** coefficient of x

Once we find the best θ1 and θ2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

**How to update θ1 and θ2 values to get the best fit line ?**

**Cost Function (J):**  
By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the θ1 and θ2 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).





Cost function(J) of Linear Regression is the **Root Mean Squared Error (RMSE)** between predicted y value (pred) and true y value (y).

[Gradient Descent](https://www.geeksforgeeks.org/gradient-descent-in-linear-regression/)**:**  
To update θ1 and θ2 values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the model uses Gradient Descent. The idea is to start with random θ1 and θ2 values and then iteratively updating the values, reaching minimum cost.

***11. Make a comparison between:-***

1. ***Generalization and abstraction***

abstraction reduces complexity by hiding irrelevant detail, generalization reduces complexity by replacing multiple entities which perform similar functions with a single construct.

1. ***Learning that is guided and unsupervised***

Under the supervision of experts🡪Teacher teach his students(spoon feeding). Here teacher is an expert and he supervised students.(Supervised machine learning)

With the assistance of experts in an indirect manner🡪Learning from an institute who just provide study material. Basically they just give us the pattern how to study.( unsupervised learning)

1. ***Regression and classification***

In regression we predict continuous values. Model evaluation done with the help of R2,adjr2,RMSE

In classification we predict multi/binary class.Model evaluation done by confussion matric,auc-roc.