Business framework:

7 steps to solve machine learning problems.

1.Identify and understand the business problem,

2. Data and Data types

* + acquisition of data, try understand nature of data source (primary or secondary data source) and data collection techniques,
  + Understand different data types whether its quantitative or qualitative- if its qualitative whether its continous or discrete
    - text, images, videos and audios fall under qualitative. ML algorithms dont understand qualitative and hence needs to be convered into quantitative.
    - Discrete types: binary, ordinal or categorical
    - something that is countable or where we can calculate their frequency is discrete. something that is measurable or numeric in format is continous data type.
    - certain models allow data types that are continous and some would allow both. in some cases need conversion of discrete to continous.
    - Continous types:
      * Within continous we have further types such as ratio scale or interval scale.
      * ML algorithm does not work with interval scale as it does not allow treating 0 as absolute zero. i.e 0 degree celsius has a real value but in case of interval scale it does not hold value while ratio does allow.

3. preparing/pre processing the data for further analysis - cleaning and transformation of data i.e missing data, outliers,

* + Can be split into types: Cleaning and Transformation of data.
  + there are model that have hyper parameters to be tuned like neural networks,support vector machines and there are others where there are no hyper parameters to worry about.
  + But before executing the model we need to clean and transform the data.
  + Cleaning:
    - step1: Handling missing information
    - Various actions that can be performed: missing entire row or the column, impute or replace.
  + step2: imputing outliers. - First step is detection of outliers followed by decision of how to impute those.
    - Detection:Univariate (Boxplot-IQR for continous variable), Bivariate (scatter plot), Multivariate (Residuals in linear regression. difference between actual and predicted. High residuals are outliers)
      * Multivariate outliers: Residual analysis can enable assessing if residuals are normal and if not identify the outliers. USing cooks method (frequently used distance method for residual analysis) if certain resiudals are outliers or not.
    - External bias can create bias . like in shops one id is created to capture many customers transaction as customer level id cannot be created if customers dont divulge their info. and hence sales get reported under one dummy id creating extreme outliers.
  + Decision:
  + Remove, retain or impute
  + Make use of measures of central tendency(mean, median, mode) or prediction models or special cases of imputations.
  + Using mode can be used for discrete variables. but it may not fit every time i.e like imputing missing values for genders with mode.In such a case special type of imputing might come into picture.
  + Special cases: like for time series we use forward fill or backward fill to impute missing closing balance for a given day with opening balance for next day etc.
  + prediction: K nearest neighbour (K-nn model) is most frequently used. It tries to understand average/median/mode behaviour of neighbouring record and uses that for imoutations

4. Performing transformations on data as per requirement

* statistical (when magniture of data in columns needs to be standarized) and non statistical (transform qualitative data into quantitative or discrete data to continous)
* certain models need distance calculation between rows. in such a case unscaled data can create impact during distance calculations.
  + - Statitical: Z score scaling (Standard normal distribution) or min max normalization. for models that are distance and weight based that will need this type of data pre processing.
      * such type of transformation is not only required for distance based model but also required for weight based model.
      * one of the important weight based model is neural networks.
      * Clustering model and K-nn(frequency based algorithm) are distance based.
    - Non statistical - this refers to data conversion to continous variable.
      * Types: Qualitative to quantitative (many ways including mathamatical way), Discrete to continous
      * Discrete has few types:
      * For categorical like city names can be converted to flag (like using get\_dummies in python).
      * For binary, can go with categorical approach i.e Gender: Female/Male
      * For ordinal: If ranks are adopted it can be directly used as continous, but for textual values we can transform to numeric orders

5.data visualization - sample analysis (figure out measure of central tendency or dispersion) - understand patterns of sample

* + - steps 5, 6 and 7 need business requirement understanding. Not all these steps would be required for some problem.
  + Visualization is only sample analysis (Figuring out measures of central tendency and dispersion).
  + Some BI tools extend beyond visualization but includes modelling like Tableau, R, PowerBI etc but stays at basic models right now.
  + When it comes to viz we only look at sample analysis primarily and not the population.

6.modelling - supervised, unsupervised

* + Different types of techniques, models and learning models exists when it comes to modelling.
  + 3 different steps: Model selection, Model building and Model evaluation.
    - Model selection can be of different ways: Choosing different models, running multiple models and selecting the best or average of those.
    - Parameters for models include hyper parameter, input parameters etc.
    - Strength of the model for the given dataset and model choice to be evaluated.
  + Categories of ML models to be understood: Descriptive, Inferential and Predictive modelling.
    - Descriptive: is all about data visualization (Central tendency, dispersion)
    - Inferential or predictive modelling: generally study of population using samples.
    - however interfential is statistial analysis(study of means: z test, t test, ANOVA, Test of proportions: Chi square test, test of normality(Shapiro), Test of variances(Levene's test),Hypothesis testing). there are many more models.
      * Chi square test can be used for 1 discrete variable or 2 discrete variables. for 1 discrete its called test of proportion, for 2 discrete variable its called test of independence.
    - hypothesis testing falls under inferential category and not predictive category. however slight deviation is that ANOVA test is performed as a part of linear regression model.
    - For predictive modelling: is study of unknown observation using known (sample) observations.

7.insights or decisions to be taken based on the model outcome.

Note: Without variance there is no machine learning. but extreme variance can be a problem in which case there could be outliers.

Data mining techniques (Prediction models in detail).

------------------------------------------------------

Data mining is applied to very large data sets. i.e big data

there are prediction models that comes under data mining such as supervised and unsupervised.

* supervised: regression (predicting continous variable) and classification (predicting categorical variable). has dependent variable and independent variable.
* unsupervised: association rules and clustering.

Supervised

* regression and classification techniques: Decision tree, random forest, neural networks (popularly used techniques).
* Decision tree itself has several models.
* Random forest contains multiple decision trees. i.e forest of trees
* neural networks is the preferred method for big data or large datasets.
* regression model will be used if the prediction of the business problem has to be continous. i.e how much sales in future?
* classification model will be used if the prediction of the business problem has to be binary or discrete. i.e whether particular customer is likely to churn or not churn which is a binary prediction.

unsupervised: clusering and association rules. There is no dependent variable or label to be predicted.

* clustering: hierarchical and non hierarchical clusterings are predominately used.. clustering objective is to find patterns of similarities and dissimilarities using distance based calculations. there are other methods too.
* association rules is probabality based also called market basket analysis or transaction based data. this falls under category of recommendations. i.e calculating probability that customers buy milk given they buy bread.
* unsupervised models cannot go through training and testing as it does not dependent variable while we can do for supervised training.

Differences between supervised and unsupervised.

1. dep and ind variables

2. train and test

3. model evaluation methods are directly available under supervised whereas for unsupervised the methods are qualitative or indirect in nature.

* strength of model evaluation lies in supervised whereas for unsupervised it is only to find simiarities and dissimilarities.

1.1 Clustering types and distance measures

Clustering is a part of Unsupervised learning. It is the technique of grouping objects, with heterogeneity between

groups and homogeneity within the groups. It can follow Agglomerative, Divisive or Partitioning approach.

Distance calculations are done to find similarity and dissimilarity in Clustering problems.

Distance between the clusters should be greater than that of within the clusters. (SSB > SSW). objective with which clustering techniques work.

cluster techniques need data preprocessing.

if distance is smaller they are similar and if larger they are dissimilar.

unsupervised learing happens when we dont have target label.

clustering can also be used as preliminary step for the supervised learning as we will be able to identify a new cluster/label column that is discrete variable which can be newly created on the dataset to proceed with supervised learning using tecniques like classification technique.

One example clustering use case: in order to cluster web pages (with objective of optimizing it), textual data can be converted to quantitative information followed using NLP followed by clustering techniques to identify similar pages and group them.

clustering vs PCA: clustering is finding similarities or dissimiarities between rows based on their distance,

wherease PCA is about dissimilarity between columns using variance.

In PCA grouping of information(group similar columns into one called dimension) is objective, in clustering labelling of information with particular cluster is the objective.

In PCA we reduce dimensions but in clustering we dont reduce the rows on which it is getting applied to.

Distancing types: Eucledian distance, Manhattan distance (city block), Chebyshev distance, minkowski distance (toggles between eucledian and manhattan based on p value)

metrics intended real valued vector spaces: for continuos data types with decimal points.

for two dimensional : Haversine metrics can be used. i.e for example for distance between latitude , longitude measures

for integer valued continous dataspace: hamming, canberra, braycurtis metrics. not necessary this the metrics for integer. we can use metrics from real valued vector spaces.

for boolean valued vector spaces: jaccard is popular for boolean. others are can be found in reference link

referencefor different distances:

https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.DistanceMetric.html

clustering types: Agglomerative, divisive, partitioning.

agglomerative clustering: starts with maximum clusters by each row and ends with one. advantage of this type is helps to identify even small size clusters whereas divisive is applicable for large datasets

Divisive: opposite of agglomerative clustering. for 1000 rows of data starts with 1 cluster and ends with 1000 clusters.

partition clustering: directly partitions the data by as many clusters. K means clustering. it does vector paritioning wherein each partition is a cluster . intra cluster distance is minimized and inter cluster distance is maximized using vector paritioning technique.

1.2 Hierarchical clustering

A type of Clustering approach where records are sequentially grouped to create clusters, based on distance between records and distance between clusters.

Clusters or records with closest distance will get merged step by step after computing distance between all the clusters or record units.

At step 0 its easy to find distance between records using distances mentioned earlier. but from step 1 onwards distances should be found between clusters using linkages.

Linkages: computes distance between 2 clusters. Single linkage, complete linkage, average linkage,centroid, ward.

Single Linkage method: defined as shortest distance between any two points across the clusters.

complete linkage: opposite of single. distance between two fathest points across clusters.

average linkage: set of all averages of observations across clusters (2 at a time) and average is taken finally which becomes the distance between those 2 clusters.

Centroid linkage: Distance between center of two clusters. central point to be computed.

Wards linkage: similar to group average and centroid distance.

Uses within cluster variances and increase in within cluster variance as a factor to identify the merges in the agglomerative procedure.

measure the distance between each observation within the clusters with respective centroid and add them up. before adding square them to avoid signs cancelling each other. finally sum has to be divided by DF for that cluster.

in Wards linkage when two clusters are merged and found the variance against the resulting centroid which is nothing but within cluster variance for the merged cluster. This way within cluster variance is compared with multiple merges to compare increase in within cluster variance.

the cluster whose mergers with least resulting within cluster variance will succeed.

strength: number of clusters can be arrived based on running the model and then choose desired number of clusters based on dendrogram by cutting the cluster tree at optimal position.

it corresponds to meaning taxonomy.

weakness: not good with too large datasets. also highly sensitive to outliers.

dendrogram: tree like diagram. x axis is rows in the datasets. y axis determines distance between the records/clusters.

: greater the height in Y higher the dissimilarity across records/clusters.

Scaling needs to be assessed for the dataset ahead of proceeding with hierarchical clustering.

1.4 K Means Clustering

It is a non-hierarchical clustering approach where we specify the number of clusters needed as output lets say, k.

It's a non hierarchical approach. widely used in large datasets as opposed to hierarchical clusters.

It starts with creating random K clusters and its centroid followed by assigning each record to one of the predetermined k cluster according to their distance from each such cluster.

Approach is to minimize the measure of dispersion within the cluster.

Means in K means refers to averaging of data. i.e centroid. so initially the model assigns centroid randomly to each of the clusters in K cluster.

This is followed by evaluating the distance of each observation from the randomly computed centroid to reprocess the observations to be aligned to the cluster to their distances are actually shorter.

Using the new clusters that are formed, actual centroids are recalculated for new clusters and original centroids are no more the centroids.

Again the distance between each observations across the clusters are computed w.r.t new centroids and accordingly some of the observations could be realigned to the right cluster based on their distance.so centroid recalc happens and so on until clustering could be finalized properly without any need for reassignment of cluster and recalculation of centroid.

Unlike hierarchical approach which has multiple options for computing distance, K means only uses equilidean distance since concept of centroid assignment works well with equilident distance as measure of calculation.

While cluster creation is based on mathematical distance, cluster validation step involves being able to interpret clusters in a way it makes business sense.

Z vs min max: if variance is large for a column and small for another we can go for min max. if variance between columns are more or less the same but magnitude is different we can go for Z score scaling.

1.5 Silhoutte score for K-means clustering

Indirect model evaluation techniques which we can verify once clustering procedures are completed namely the K-means model which is distance based.

Once clustering procedures namely K means that is distance based is completed there are indirect model evaluation technique that can be used to verify.

Silhoute scoring technique is one such. It validates if the mapping of observations to clusters are correct or not.

This includes computation of distance between each observations distance between itself and every established centroid to determine if there are any other centroids outside current cluster whose distance could be the shortest for the given observation.

silhoute width=(b-a)/max(a,b) which if returns positive value the observation is valid in its alignment to its cluster where b and a are distance between the observation with neighbouring centroid and current centroid respectively.

silhoute width ranges between -1 and +1.

if silhoute width=0 then the values are equistant from both centroids and it means they are not actually seperable.

if average of silhoute width of all the observations of the cluster w.r.t another cluster is positive then the cluster is well seperated, if negative its a blunder and if zero is not actually seperable.

1.6 Sil score and wssplot- traintest and overfitting

wss plot/distortion plot helps to know how many clusters are needed as output in K-means clustering.

WSS plot stands for within sum of squares plot. also known as distortion plot or error plot.

This operates with the notion that within sum of squares (variance) will be high for K=1 but when k begins to increase the sum of within sum of squares across clusters will start reducing from K=1.

However as keep adding K there could be a point where the drop of WSS may not be significant.

WSS=sum of square of differences between centroid and each observation for each observation within a cluster

the plot has Total WSS in y and K in x

Note: K nearest neighbours is a supervised model.