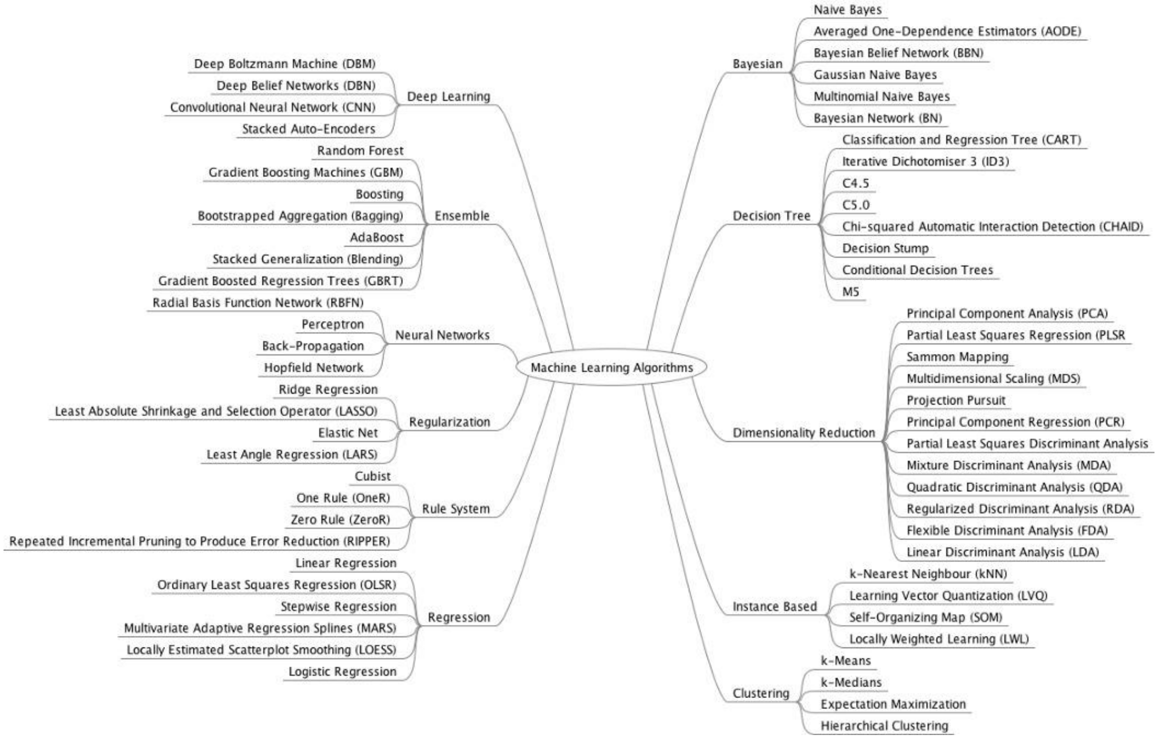
WHICH MACHINE LEARNING MODEL TO CHOOSE?

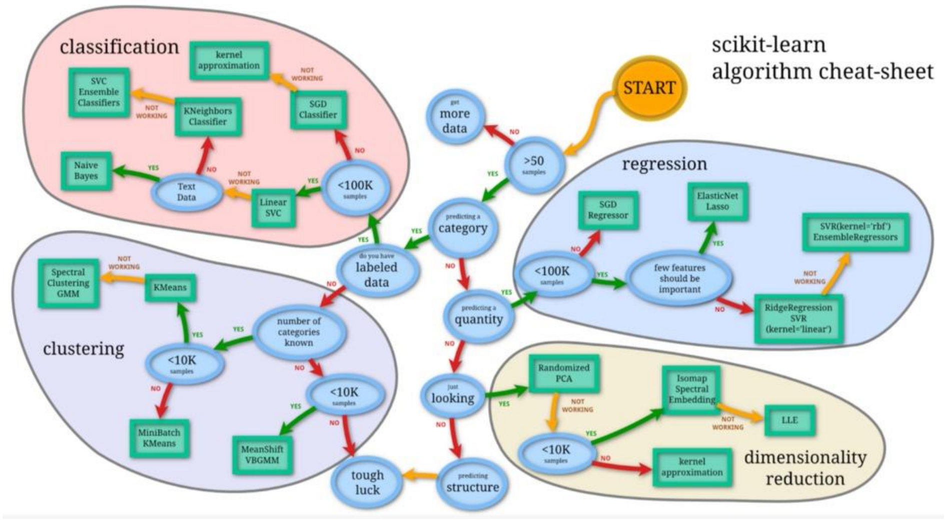
DR. ALVIN ANG

CONTENTS

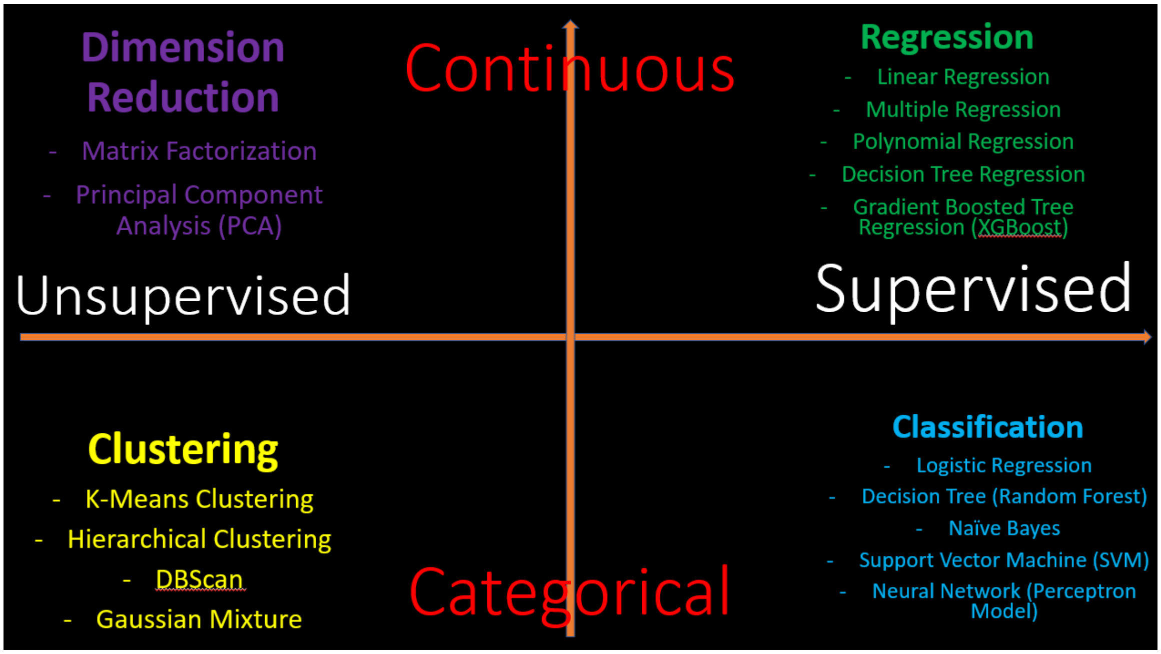
I. THERE ARE TOO MANY MODELS…

A. CONFUSING MODELS


Machine Learning Algorithms
Deep Learning
Deep Boltzmann Machine (DBM)
Deep Belief Networks (DBN)
Convolutional Neural Network (CNN)
Stacked Auto-Encoders
Ensemble
Random Forest
Gradient Boosting Machines (GBM)
Boosting
Bootstrapped Aggregation (Bagging)
AdaBoost
Stacked Generalization (Blending)
Gradient Boosted Regression Trees (GBRT)
Neural Networks
Radial Basis Function Network (RBFN)
Perceptron
Back-Propagation
Hopfield Network
Regularization
Ridge Regression
Least Absolute Shrinkage and Selection Operator (LASSO)
Elastic Net
Least Angle Regression (LARS)
Rule System
Cubist
One Rule (OneR)
Zero Rule (ZeroR)
Repeated Incremental Pruning to Produce Error Reduction (RIPPER)
Regression
Linear Regression
Ordinary Least Squares Regression (OLSR)
Stepwise Regression
Multivariate Adaptive Regression Splines (MARS)
Locally Estimated Scatterplot Smoothing (LOESS)
Logistic Regression
Bayesian
Naive Bayes
Averaged One-Dependence Estimators (AODE)
Bayesian Belief Network (BBN)
Gaussian Naive Bayes
Multinomial Naive Bayes
Bayesian Network (BN)
Decision Tree
Classification and Regression Tree (CART)
Iterative Dichotomiser 3 (ID3)
C4.5
C5.0
Chi-squared Automatic Interaction Detection (CHAID)
Decision Stump
Conditional Decision Trees
M5
Dimensionality Reduction
Principal Component Analysis (PCA)
Partial Least Squares Regression (PLSR
Sammon Mapping
Multidimensional Scaling (MDS)
Projection Pursuit
Principal Component Regression (PCR)
Partial Least Squares Discriminant Analysis
Mixture Discriminant Analysis (MDA)
Quadratic Discriminant Analysis (QDA)
Regularized Discriminant Analysis (RDA)
Flexible Discriminant Analysis (FDA)
Linear Discriminant Analysis (LDA)
Instance Based
k-Nearest Neighbour (kNN)
Learning Vector Quantization (LVQ)
Self-Organizing Map (SOM)
Locally Weighted Learning (LWL)
Clustering
k-Means
k-Medians
Expectation Maximization
Hierarchical Clustering



B. SIMPLIFIED MODEL



Dimension Reduction

- Matrix Factorization
- Principal Component Analysis (PCA)

Continuous

Regression

- Linear Regression
- Multiple Regression
- Polynomial Regression
- Decision Tree Regression
- Gradient Boosted Tree Regression (XGBoost)

Unsupervised

Supervised

Clustering

- K-Means Clustering
- Hierarchical Clustering
- DBScan
- Gaussian Mixture

Categorical

Classification

- Logistic Regression
- Decision Tree (Random Forest)
- Naïve Bayes
- Support Vector Machine (SVM)
- Neural Network (Perceptron Model)

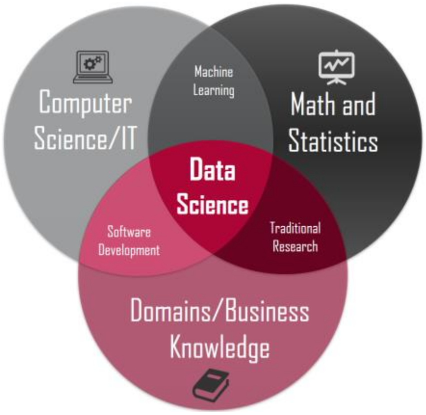
II. CONSIDERATIONS

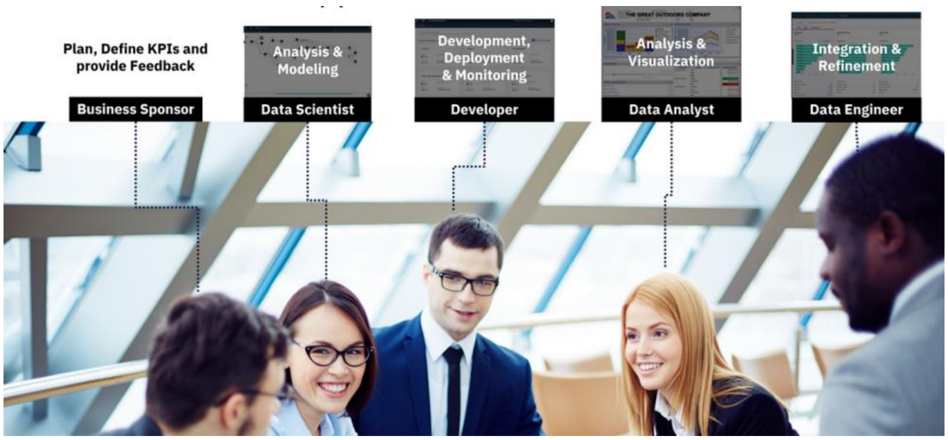
Kaggle only cares about PERFORMANCE of the model… but in REAL LIFE, there are MANY OTHER FACTORS!

In my opinion, in terms of PRIORITY:

1. BUSINESS OBJECTIVES: EXPLAINABILITY VS COMPLEXITY
2. BUSINESS CONSTRAINTS: INFERENCE TIME / TRAINING TIME / COST
3. DATASET SIZE (NUMBER OF ROWS) AND DIMENSIONALITY (NUMBER OF COLUMNS)
4. MODEL’S PERFORMANCE

III. CONSIDERATION 1: BUSINESS OBJECTIVES (EXPLAINABILITY VS COMPLEXITY)

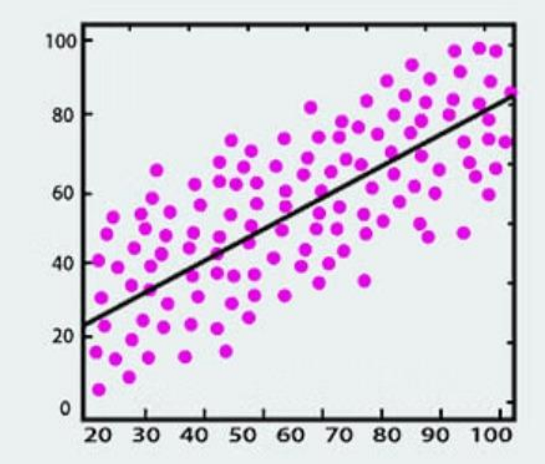
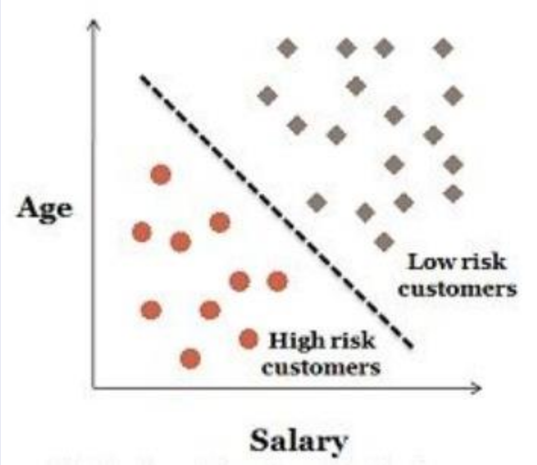
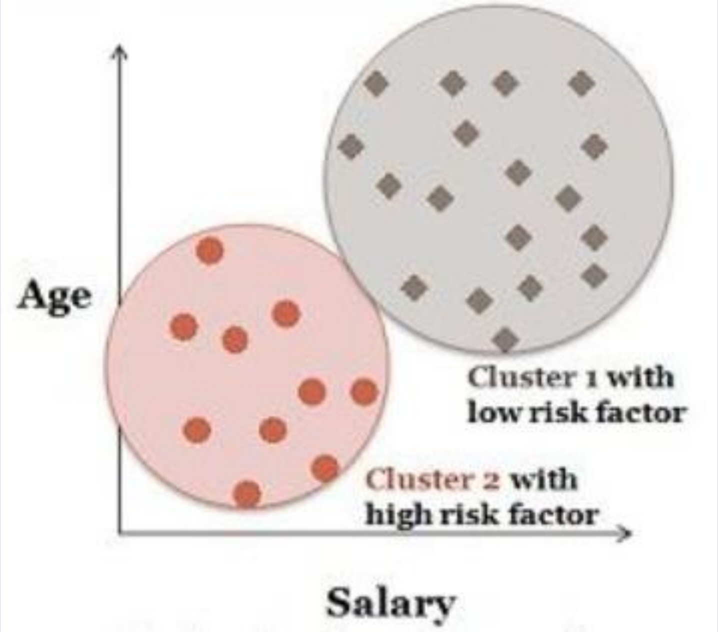




https://dr-alvin-ang.medium.com/list/the-data-science-team-29e63bae3aec

- What is the Business Objective? What are we trying to do?
- The Business Sponsor is supposed to have years of domain knowledge.
- Before selecting a ML model, we need to fully understand HIS business objectives first.
- ML Models can be Complex.
- The more Complex a model is, the harder to Explain its results.

A. ARE WE TRYING TO DO A REGRESSION / CLASSIFICATION / CLUSTERING PREDICTION?

	Regression	Classification	Clustering
Definition		Multiclass Classification • Dog • Cat • Horse • Fish • Bird • ...	 Age Cluster 1 with low risk factor Cluster 2 with high risk factor Salary
Learning	Supervised	Supervised	Unsupervised

IV. CONSIDERATION 2: BUSINESS CONSTRAINTS (INFERENCE TIME / TRAINING TIME / COST)

A. ARE WE TRYING TO DO A REGRESSION / CLASSIFICATION (SUPERVISED)?

Supervised Learning	Slow but Accurate	Fast & Easy to Explain	Fast but Hard to Explain
Regression	Random Forest Neural Network Gradient Boosting Tree (similar to Random Forest but easier to overfit)	Decision Tree Linear Regression	nil
Classification	Random Forest Neural Network Gradient Boosting Tree (similar to Random Forest but easier to overfit) Non-Linear SVM	Decision Tree Logistics Regression	Linear SVM Naïve Bayes

B. ARE WE TRYING TO DO A CLUSTERING (UNSUPERVISED)?

Unsupervised Learning	Hierarchical Clustering	Non - Hierarchical Clustering
Clustering	Hierarchical Clustering	DBScan K Means Gaussian Mixture Model

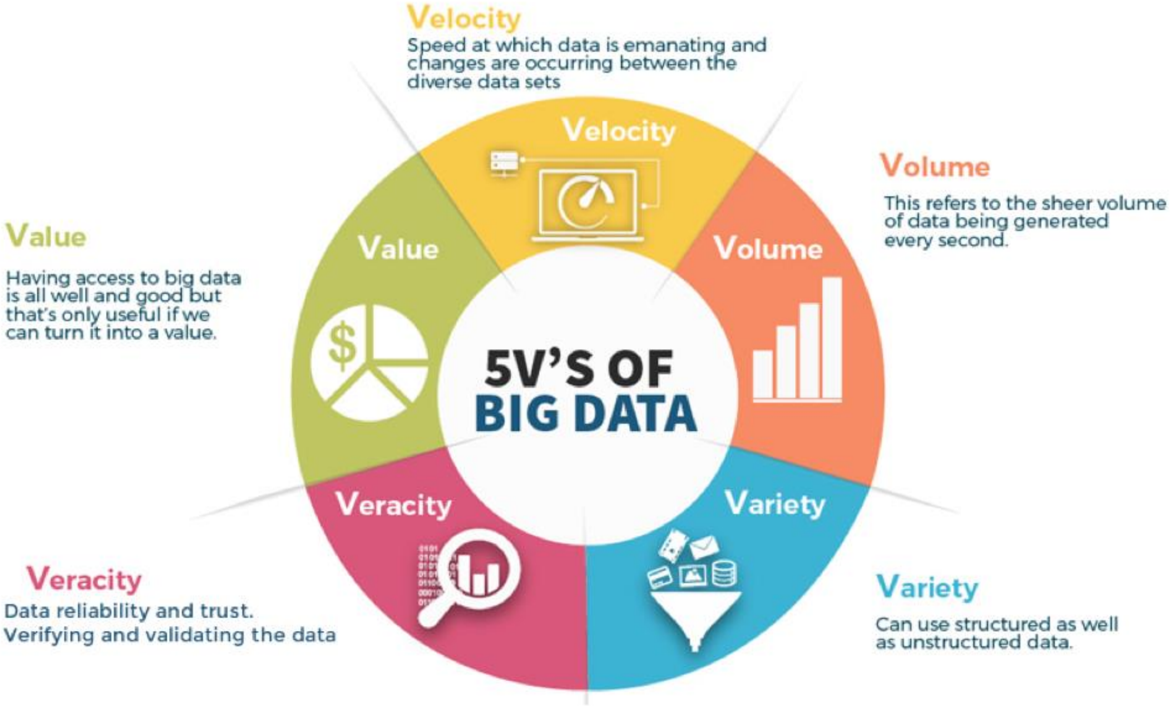
C. TRAINING TIME VS INFERENCE TIME

- Training Time
    - How long does it take to train your model?
    - The longer it takes, the more expensive the project budget.
    - What's the project budget?
- Inference Time
    - A self - driving car needs to make decision in real time.
    - It can't take too long to decide to turn left or right.
    - How much time are we given to infer from the data?

V. CONSIDERATION 3: DATASET SIZE (NUMBER OF ROWS) AND DIMENSIONALITY (NUMBER OF COLUMNS)

A. 5V'S OF BIG DATA

- How big is our dataset?
- How big is big? How small is small?
- There's no specific size / amount that can quantify "BIG DATA" (even though there's the 5Vs to describe it as shown below)



5V'S OF BIG DATA
Velocity
Speed at which data is emanating and changes are occurring between the diverse data sets
Volume
This refers to the sheer volume of data being generated every second.
Variety
Can use structured as well as unstructured data.
Veracity
Data reliability and trust. Verifying and validating the data
Value
Having access to big data is all well and good but that's only useful if we can turn it into a value.

B. BIG DATA BELONGS TO THE BIG BOYS

- I can’t tell if my Data is big or small… so how?

- It goes back to our FIRST PRIORITY: Consideration 1: Business Objectives (Explainability vs Complexity) (ctrl + click) because normally,

    - Big Data belongs to the “Big Boys” (MNCs) who have the “Big Budget” to do “Big Data Analytics”…. And like wise… “Bigger Vision” with “Bigger Timeframes”

    - Small Data belongs to the “Small Boys” (SMEs) who have “Small Budgets” and can only perform “Small Analytics”…. “Smaller Objectives” with “Smaller Timeframes”…

- Are you doing a “Big Project” or “Small Project”?

C. MY PERSONAL DEFINITION OF BIG DATA

- According to my humble personal experience, my definition of "BIG DATA" is quantified as follows:

1. EXCEEDING THE 1 MILLION ROWS OF DATA IN ONE EXCEL SPREADSHEET.

    - Why?

    - Because you can't plot graphs from TWO spreadsheets (when 1 million rows spills over to the next spreadsheet)…. So you can't get any descriptive statistics / inference done.

2. EXCEEDING THE 16,000 COLUMNS OF DATA IN ONE EXCEL SPREADSHEET.

    - Why?

    - Too many Features which boggles the mind…. And likewise, spilling over to the next spreadsheet which hints you that there's too many variables you are considering

3. POSSIBLY EXCEEDING 20MB FILE SIZE OF ONE CSV.

    - Why?

    - Because just double clicking it to open it in Excel can take forever… and moving the cursor around hangs your computer (on a basic 8 GB Ram laptop).

D. WHAT TO DO WITH "HUGE" NUMBER OF ROWS?

- Is your Project Budget (and likewise Timeframe) Big or Small?

- If Small Budget with Tight Timelines….

    - Refer to "Are we trying to do a Regression / Classification (Supervised)?" (ctrl + clickl)

    - Use only Supervised Learning methods (because its faster)… and choose only the "Fast and Easy" to explain methods…

    - But if your Data comes unlabelled… well… good luck! (the road ahead is going to be tough….HAHAHA…)… you have to use Unsupervised Methods….which might take long time and/or unsatisfactory results….

- If Big Budget with Longer Timelines….

    - Use Neural Networks.

E. WHAT TO DO WITH "HUGE" NUMBER OF COLUMNS?

- You need to perform Principal Component Analysis (PCA) for Dimension Reduction

- https://www.alvinang.sg/s/Principal-Component-Analysis-PCA-with-PySpark-by-Dr-Alvin-Ang.pdf

CONSIDERATION 4: MODEL'S PERFORMANCE

- These are the typical metrics people use to evaluate ML models:
  - Accuracy
  - F1 Score
  - ROC / AUC curve
  - PR Curve
  - All of which are found here: https://www.alvinang.sg/s/Performance-Metrics-for-Machine-Learning-Models-by-Dr-Alvin-Ang.pdf
- The most popular one being ROC / AUC curve.

CONCLUSION

- There’s no BEST model… it depends on the situation.
- People normally fixate on their favourite model… the one that they are most familiar with.
- As long as it gives them good results for their past projects, the model can be re-applied.
- The basic considerations are:
  - Business Objectives: Explainability vs Complexity
  - Business Constraints: Inference Time / Training Time / Cost
  - Dataset Size (number of Rows) and Dimensionality (number of Columns)
  - Model’s Performance

REFERENCES

https://towardsdatascience.com/considerations-when-choosing-a-machine-learning-model-aa31f52c27f3

https://towardsdatascience.com/which-machine-learning-model-to-use-db5fdf37f3dd

ABOUT DR. ALVIN ANG

Dr. Alvin Ang earned his Ph.D., Masters and Bachelor degrees from NTU, Singapore. He is a scientist, entrepreneur, as well as a personal/business advisor. More about him at www.AlvinAng.sg.