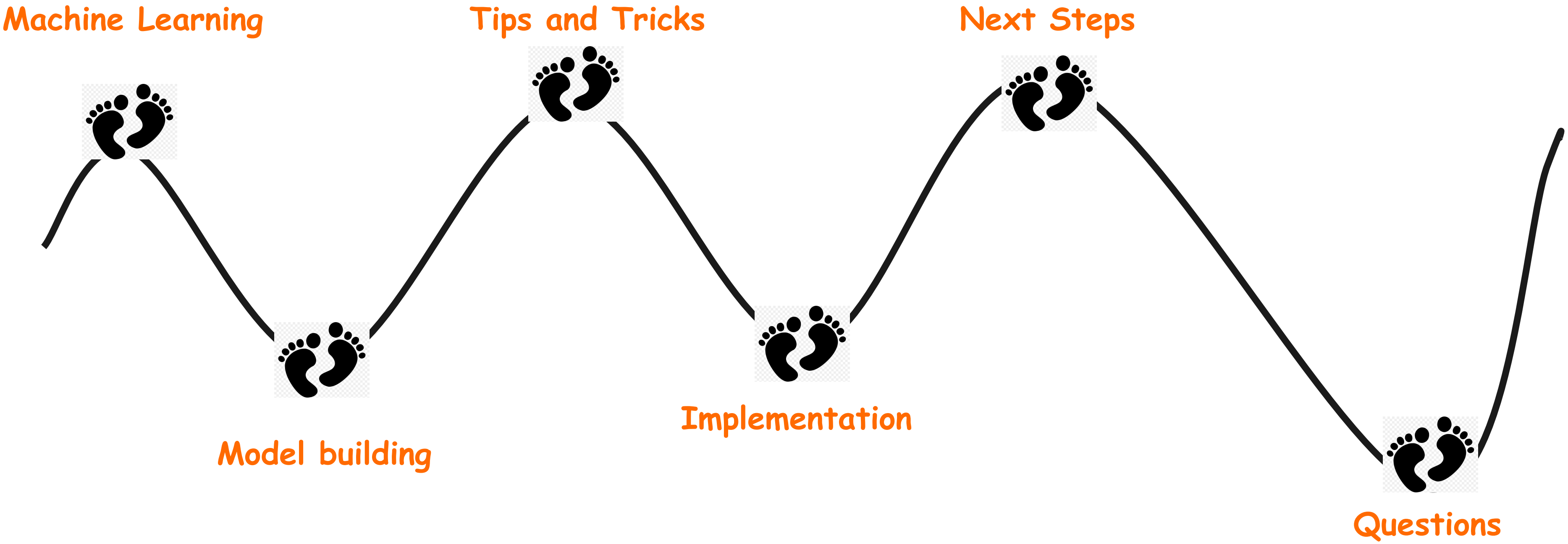
WOMEN IN DATA SCIENCE
STANFORD UNIVERSITY

greenlink

Baby Steps towards building your first ML model

Manogna Mantripragada
Data Scientist
Greenlink Analytics

Objectives



Machine Learning
Model building
Tips and Tricks
Implementation
Next Steps
Questions

Machine Learning Simplified >>>

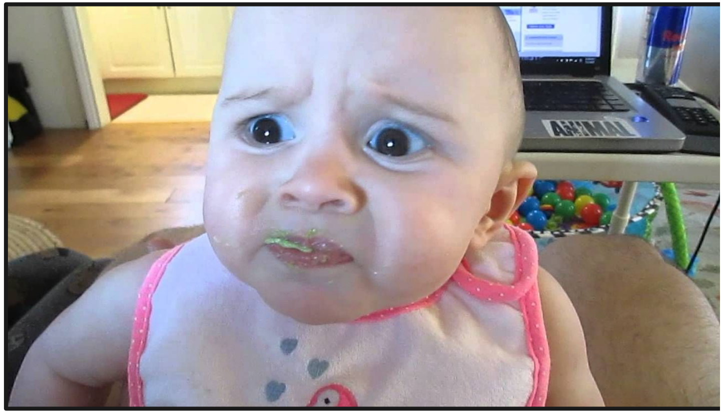
Introduction

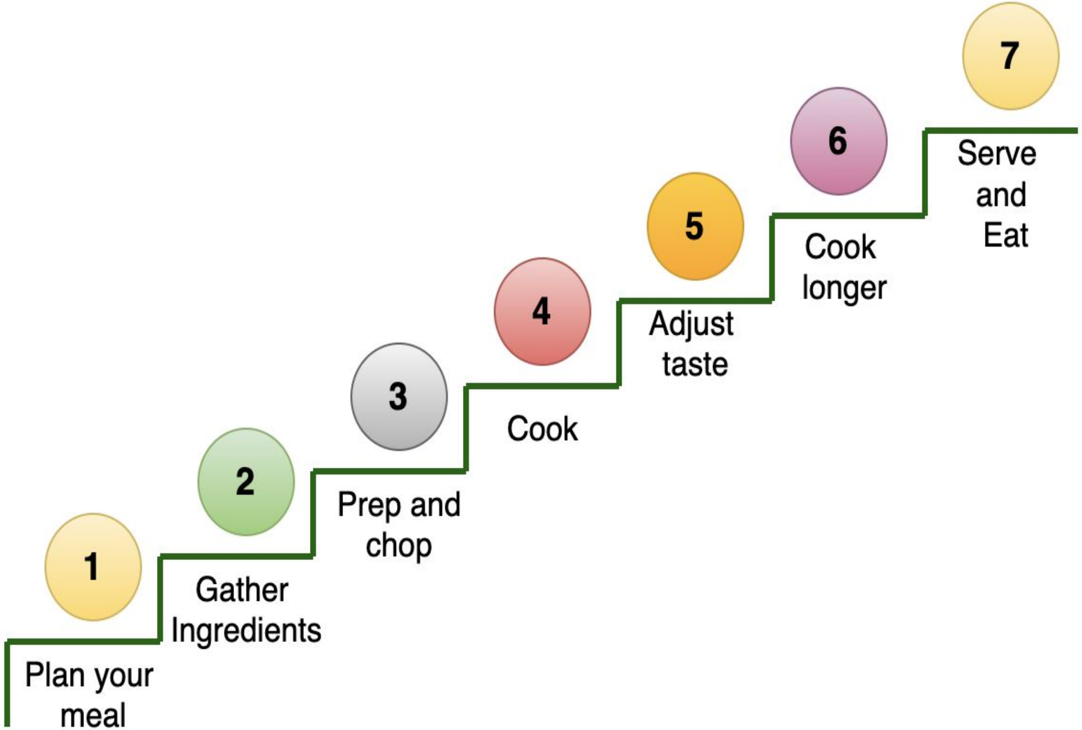
MACHINE LEARNING

CLASSIFYING

QUANTITATIVE PREDICTIONS

If building a model was like cooking…





Building it right >>>

Steps for building the model

1. Define Problem Statement
2. Gather required data
3. EDA + Preprocessing
4. Baseline/Dummy model
5. Choosing evaluation metrics
6. Candidate models training
7. Best model selection
8. Hyperparameter tuning
9. Cross validation
10. Model testing
11. Results

Problem Statement

Explore trends in energy burden in
- two states (CO and GA)
- across 4 years (2013 - 2016)

ENERGY BURDEN = $\frac{\text{Mean Household Energy Bills}}{\text{Mean Household Income}}$

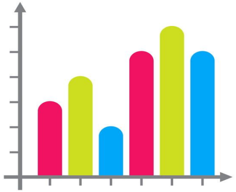
Sourced via Walt Disney Television Animation

Get Data? From where? >>>

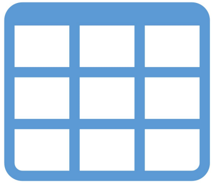
Data Gathering

Web Scraping

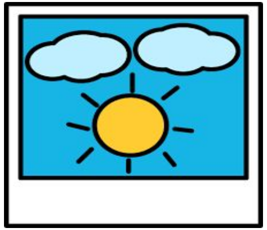
Real-time data gathering



Pre-existing data sets



Tabular



Image

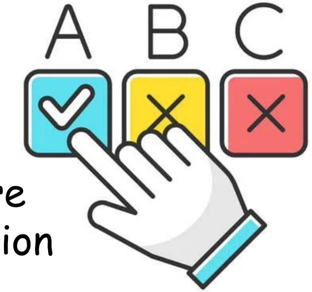
Text

Audio

EDA + Preprocessing >>>

EDA + Preprocessing

Data
Cleaning



Feature
Selection

Data
Transformation

Feature
Engineering

EDA + Preprocessing - Data Cleaning (STEP 1)

- **Dealing with Missing Values:**

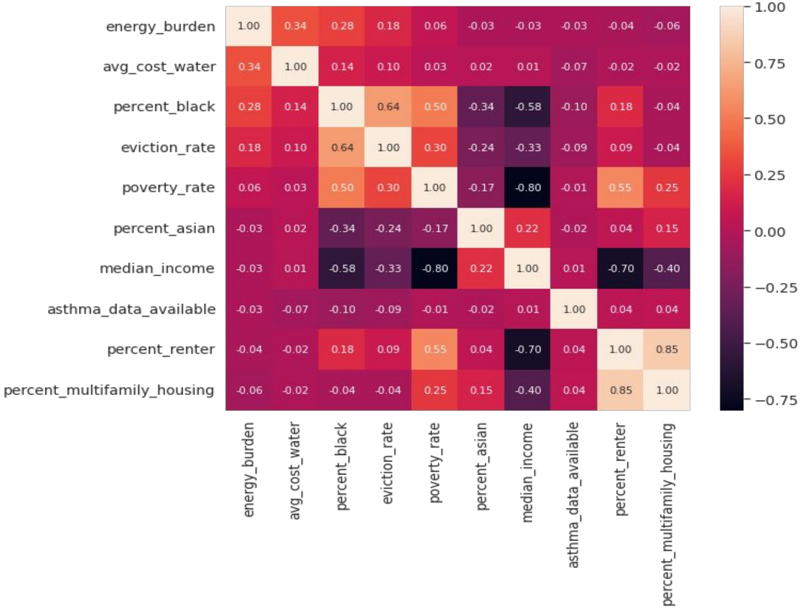
  NOTE: NO CHANGING THE DATA DISTRIBUTION!!!

  - Drop:
    - Data (rows or columns) > 70% nulls can be dropped
    - Specific Rows
  - Keep:
    - Mean/Median/Mode
    - Missing value indicator column
    - Forward fill and Backward fill - Time series model
    - Build a Regression model
      - Linear Regression for Continuous variable
      - Logistic Regression for discrete/categorical
- **Removing Duplicates**

EDA + Preprocessing - Feature Selection (STEP 2)

Practice of choosing subset features for eliminating irrelevant and redundant features:

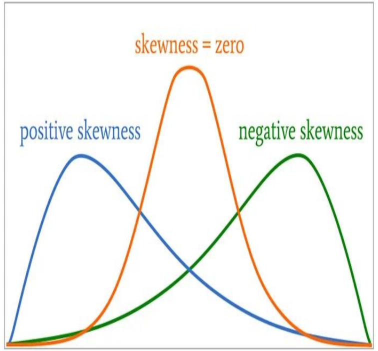
- **Correlation Matrix**

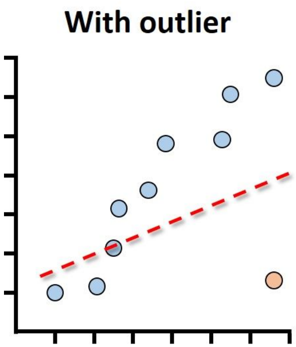


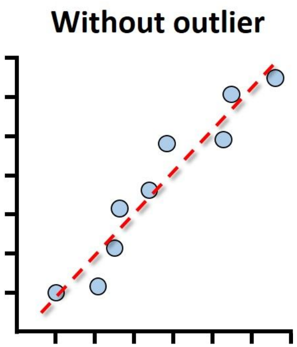
- Drop columns with high **multicollinearity**
    - Use Variance Inflation Factor (VIF)
        - (Implementation and interpretation in the colab notebook)
- ***Statistical significance - p-value***
    - P<0.05 — Significant

EDA + Preprocessing - Data Transformation (Step 3)

- **Outlier Detection**
  - Univariate analysis
  - Multivariate analysis
  - Skewness
  - Kurtosis







- **Scaling and normalizing data**
  a. Log transformations
  b. Balancing unbalanced data
     i. Oversampling
     ii. Undersampling

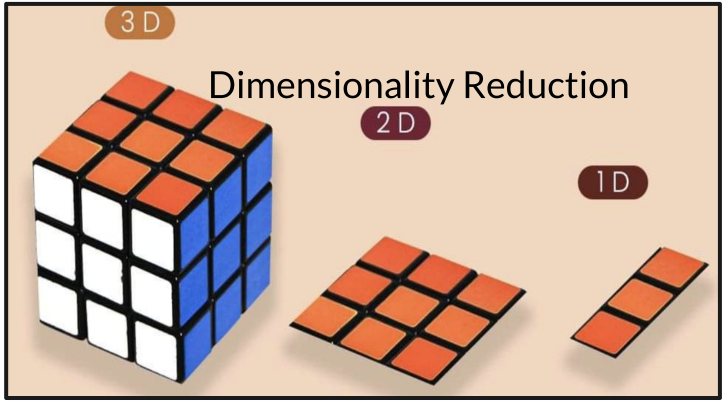
EDA + Preprocessing - Feature Engineering (Step 4)

1. One-Hot Encoding
2. Feature Creation
3. Dimensionality Reduction
    a. Eg: PCA

One-Hot Encoding

datagy.io

Island	Biscoe	Dream	Torgensen
Biscoe	1	0	0
Torgensen	0	0	1
Dream	0	1	0

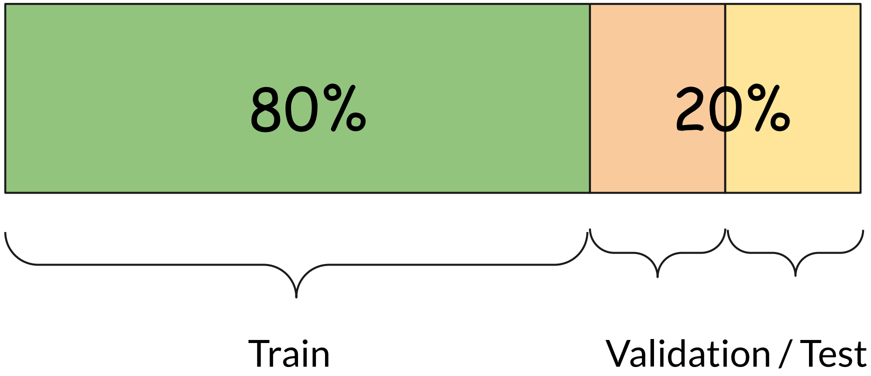


Model Building >>>

Model data + Evaluation Metrics - (STEP 1)

Train/Test Split

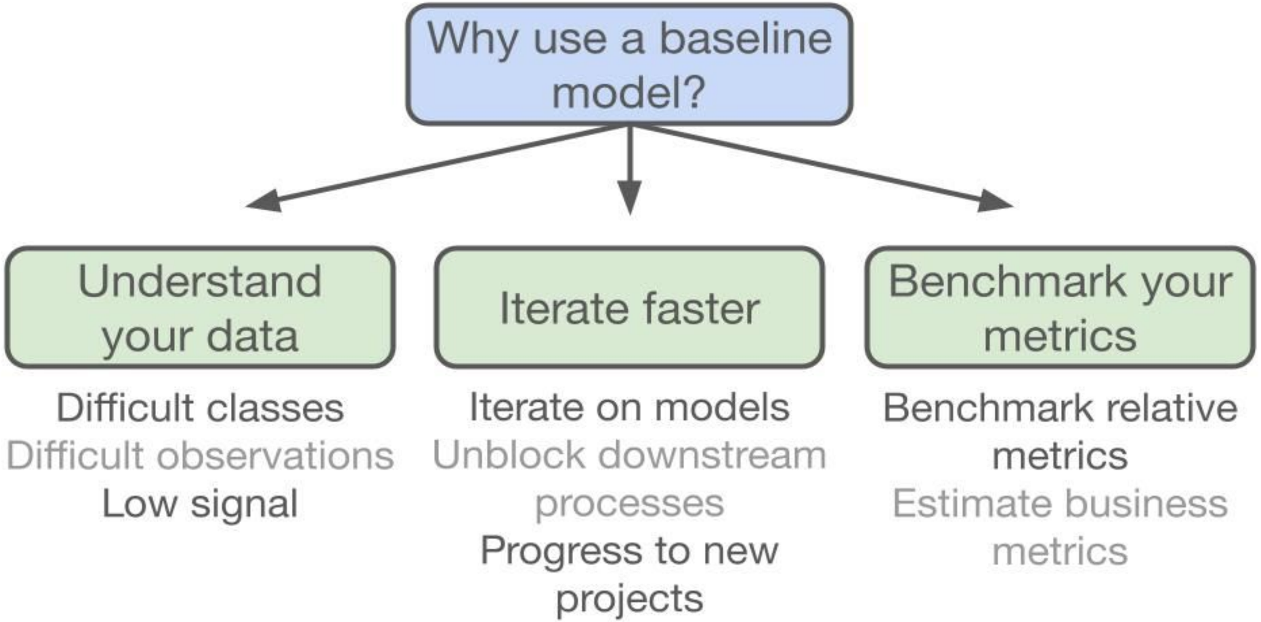
- Shuffle data (if required)
- Split Train, Validation/test



Evaluation metrics

- Set Evaluation metrics
  - Eg: R2, Accuracy, Precision
- Set evaluation error
  - Eg: MSE, MAE, Log loss

Baseline/Dummy model - (STEP 2)



Source: Crunching The Data
https://crunchingthedata.com/baseline-models-for-machine-learning/

Candidate models for training - (STEP 3)

Regression (Continuous Variable)

- Linear Regression
- Neural Networks
- Support Vector regressor
- Decision Tree Regressor
- Random Forest Regressor
- etc

Regression (Categorical Variable)

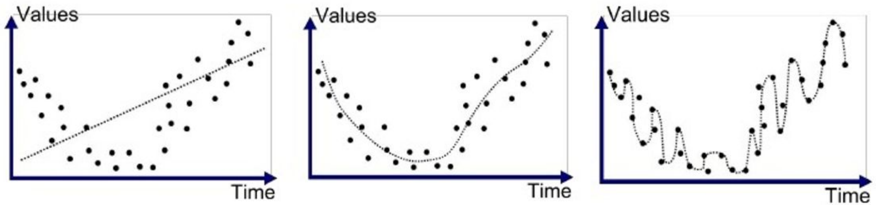
- Logistic Regression
- Neural Networks
- Support Vector Classifier
- Decision Tree Classifier
- Random Forest Classifier
- Naive Bayes Classifier, etc.

Clustering (unsupervised)

- K-means clustering
- DBSCAN clustering
- Hierarchical Clustering
- Mean-shift clustering
- Variational Autoencoders (VAEs)
- etc.

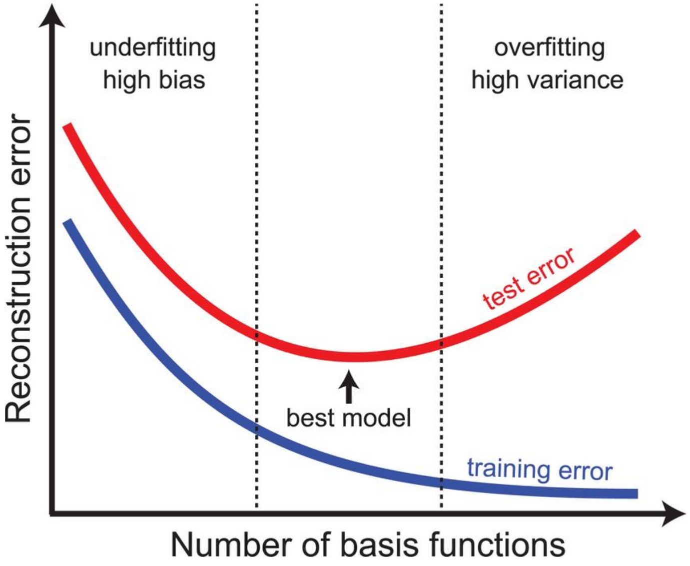
Model selection - (STEP 4)

- Select Model based on the **metrics established**, **eg:** Accuracy, R2 etc, Loss function, MSE etc)
- Bias-Variance Tradeoff
- No Overfitting



Underfitted Good Fit/Robust Overfitted

Source: Ken Hoffman Medium article
https://medium.com/swlh/machine-learning-how-to-prevent-overfitting-fdf759cc00a9



Beyeler, Michael & Rounds, Emily & Carlson, Kristofor & Dutt, Nikil & Krichmar, Jeff. (2019). Neural correlates of sparse coding and dimensionality reduction. PLOS Computational Biology. 15. e1006908. 10.1371/journal.pcbi.1006908.

Hyperparameter tuning - (STEP 5)

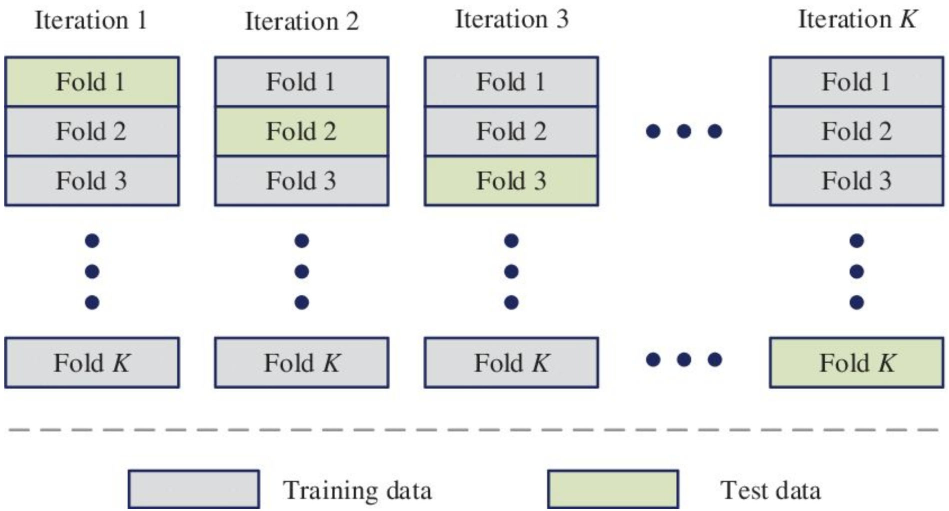
Tweak hyperparameters for the selected model to improve the performance

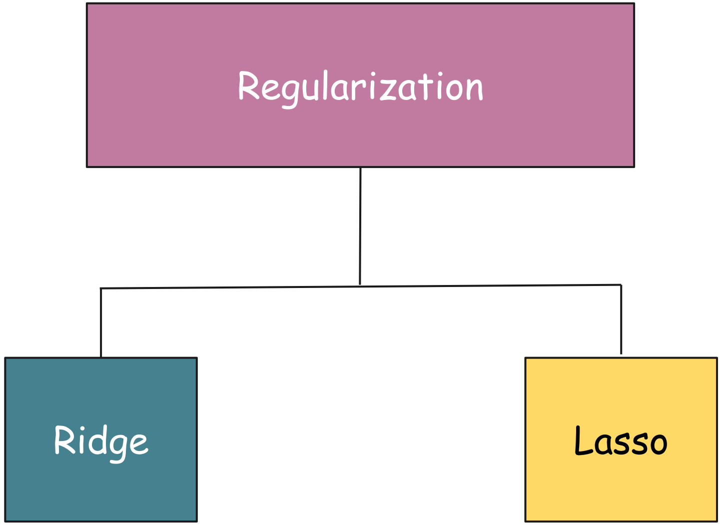
- Random Search
- Grid Search

Eg: Decision Tree: Max depth, minimum sample split, criterion, max features, etc.

Cross validation & Regularization - (STEP 6)

Ren, Qiubing & Li, Mingchao & Han, Shuai. (2019). Tectonic discrimination of olivine in basalt using data mining techniques based on major elements: a comparative study from multiple perspectives. Big Earth Data. 3. 1-18. 10.1080/20964471.2019.1572452.





Model testing - (STEP 7)

- Use the final model (after hyperparameter tuning to test on unseen data.
- Check the performance metrics for learning how well the model performed.
- If not good, go back and repeat all steps again.

Source:
https://manisha-sirsat.blogspot.com/2019/04/confusion-matrix.html

		Predicted Class		
		Positive	**Negative**	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

confusion matrix for a binary classification problem

Results

1. Tie back Results to the problem statement
2. Identify trends, exceptions etc and highlight in analysis
3. Acceptable margin of error of the model may differ
4. Use Visualizations to display results
5. Account for scaling and deployment

Source: https://www.onlc.com/blog/10-types-tableau-charts-using/

Looking back >>>

We learnt..

1. Always define your problem statement
2. Data gathering and cleaning is time consuming, but very important
3. Explore the data, visually if possible, and preprocess before training
4. Select metrics and create baseline model
5. Train and test the model
6. Displaying Results

Takeaways

1. Model building is like *building your own ice cream.*
2. Identify when ML needs to used and when not
3. Explore about Pre-built models
4. No project is a failed project - *Always* a learning from a data project
5. Get your hands Dirty with the data.

Next Steps

1. Start with an existing dataset
    a. Kaggle , Registry of open data on AWS, Awesome Public Datasets
2. Spend time on EDA and Feature Engineering
3. Try different approaches to understand how they work
4. Read Documentation
5. Towards Data Science Articles

Thankyou....

Manogna Mantripragada
LinkedIn: manogna-mantripragada
Email: mmantripragada@greenlinkanalytics.org