CLUSTERING

Life comes in clusters, clusters of solitude, then a cluster when there is hardly time to breathe.

– May Sarton (Eleanor Marie Sarton)
Poet, Novelist. Memoirist
1912-1995

Overtime, this growing tendency of like marrying like will only reinforce clustering and geographic sorting along class lines, giving the emerging map of social, economic, and cultural segregation even greater permanence.

-Richard Florida
Urban Studies - Social & Economic Theory. Prof. @ Univ. of Toronto.
1957 - present

LEARNING GOALS

- Clustering use cases
- General knowledge of common clustering algorithms
- implement k-Means using SKLearn
- Make use of clusters later down the pipeline.
- Handle outliers using IQR
- More practice scaling
- Strategies for Missing values
- Plotting Clusters

Machine Learning

SUPERVISED

CLASSIFICATION

TARGET:

Discrete
Categorical

SUB-METHODOLOGIES:

Time series
Anomaly detection
NLP

ALGORITHMS:

Logistic Regression
SVM
Decision Tree
Random Forest
KNN
Neural Network

REGRESSION

TARGET:

Continuous
numeric, ordered

SUB-METHODOLOGIES:

time series
anomaly detection

ALGORITHMS:

OLS (linear regression)
GLM (tweedie regressor)
Polynomial Regression
Support Vector Regressor
Decision Tree Regressor
Neural Networks
LASSO, LARS

REINFORCEMENT

SUB-METHODOLOGIES

Anomaly Detection
NLP
Recommender Systems

ALGORITHMS:

Collaborative Filtering
Content Filtering
Neural Networks

UNSUPERVISED

CLUSTERING

SUB-METHODOLOGIES:

Anomaly Detection
NLP

ALGORITHMS:

K-Means
DBSCAN
Hearchical

DIMENSIONALITY REDUCTION

PCA
t-SNE
factor Analysis

CLUSTERS & CAREERS

Maggie, Customer Intel Analyst
Rackspace C. 2010

Lanham Napier, CEO
Rackspace

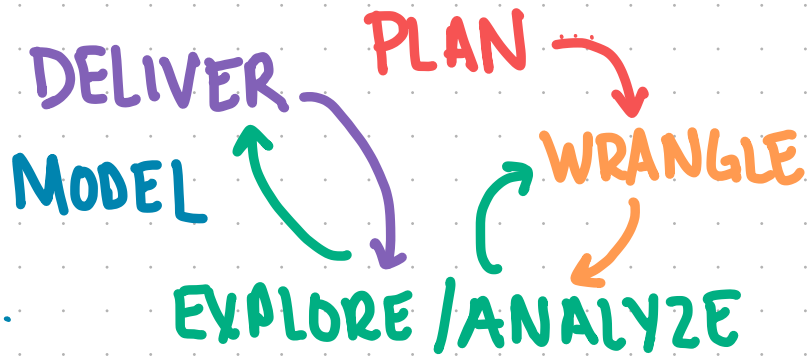
WHO ARE OUR CUSTOMERS?

Why? are we meeting customer needs? support strategy for scaling support? target customer acquisition? fire customers?

Define the Question

- A report w/ visualizations exhibiting the top discovered usage patterns (clusters) with useful labels applied
- use NPS to see if one cluster is sign. more engaged than others.
- Add info. about MRR + # of accounts.
- Any specific customer use cases I can discover within each cluster.

How do our customers use our cloud hosting product?
Can we identify <10 common patterns of number of active servers daily over time, across accounts such that when we know a customer follows pattern *, we then know what to expect from customer. So when usage deviates from the expected, THAT'S when we know to take action, but not everytime a customer, any customer spins down all their servers.



Will not do predictive modeling yet. We must assess the usefulness of the clusters created before taking that step.

GOAL

- Identify the most common usage patterns of cloud hosting by clustering accounts based on their patterns of usage measured by number of slices.
- Define, label those patterns discovered.

How?

- use K-means, No time series.
- find ways to represent the usage patterns & attributes seen over time. through summary stats & other creative summarizing features.

Details to consider:

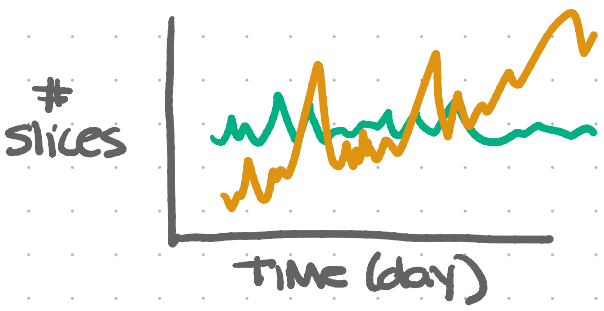
- Group by account ID.
- Start w/ history over last 3 months
- Exclude accounts whose 1st server was spun up < 4 months ago.
- # of servers in day = # of distinct server IDs 'active' status at any point in the day.

Challenges

- hundreds of millions of records
- indexing not ideal
- tens of tables, handful of databases
- permission issues in ability to write
- no unique customer identifier
- accounts may become stale but not closed.

RAW DATA to NAMED CLUSTERS

Acct # 112358, 235813



Acct.	Date	Count
112358	10/09	5
	10/08	6
	⋮	⋮
	7/11	3
112358	7/10	4
235813	10/09	10
	10/08	9
	⋮	⋮
	7/11	2
235813	7/10	1

90 days / Acct
10k Accounts

⇒ Reshape 1 Row / Account

Account	7/10	7/11	⋯	10/8	10/9
112358	4	3	⋯	6	5
235813	1	2	⋯	9	10

Why?
★ Ensure every account has an entry on every day. i.e. fill missing days w/ 0.
& 1 Row per account.

Account	σ	μ	CV (σ/μ)	% b	UBB	LBB	slope	day/day Δ	% Δ	MA	zero-use days	Cluster
112358												4
235813												3

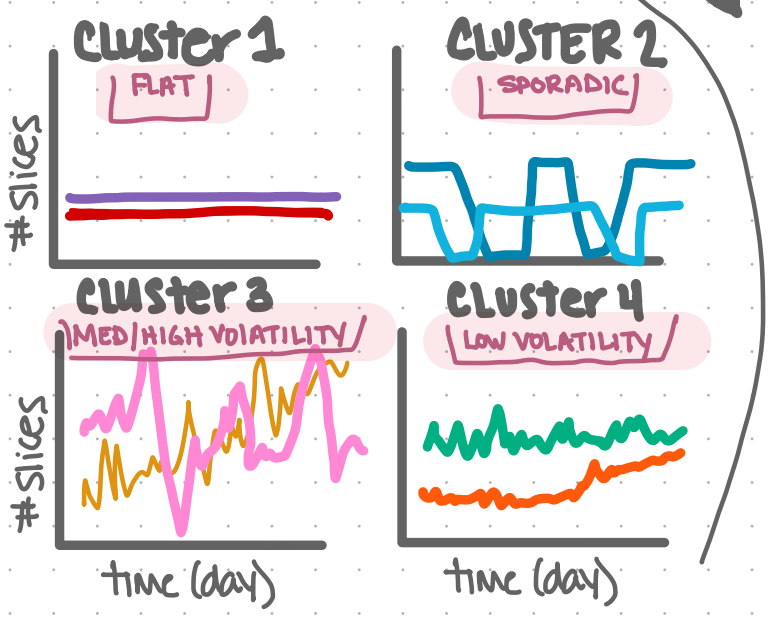
<data is sampled (train, validate, test) & scaled>

Cluster the observations using k-Means

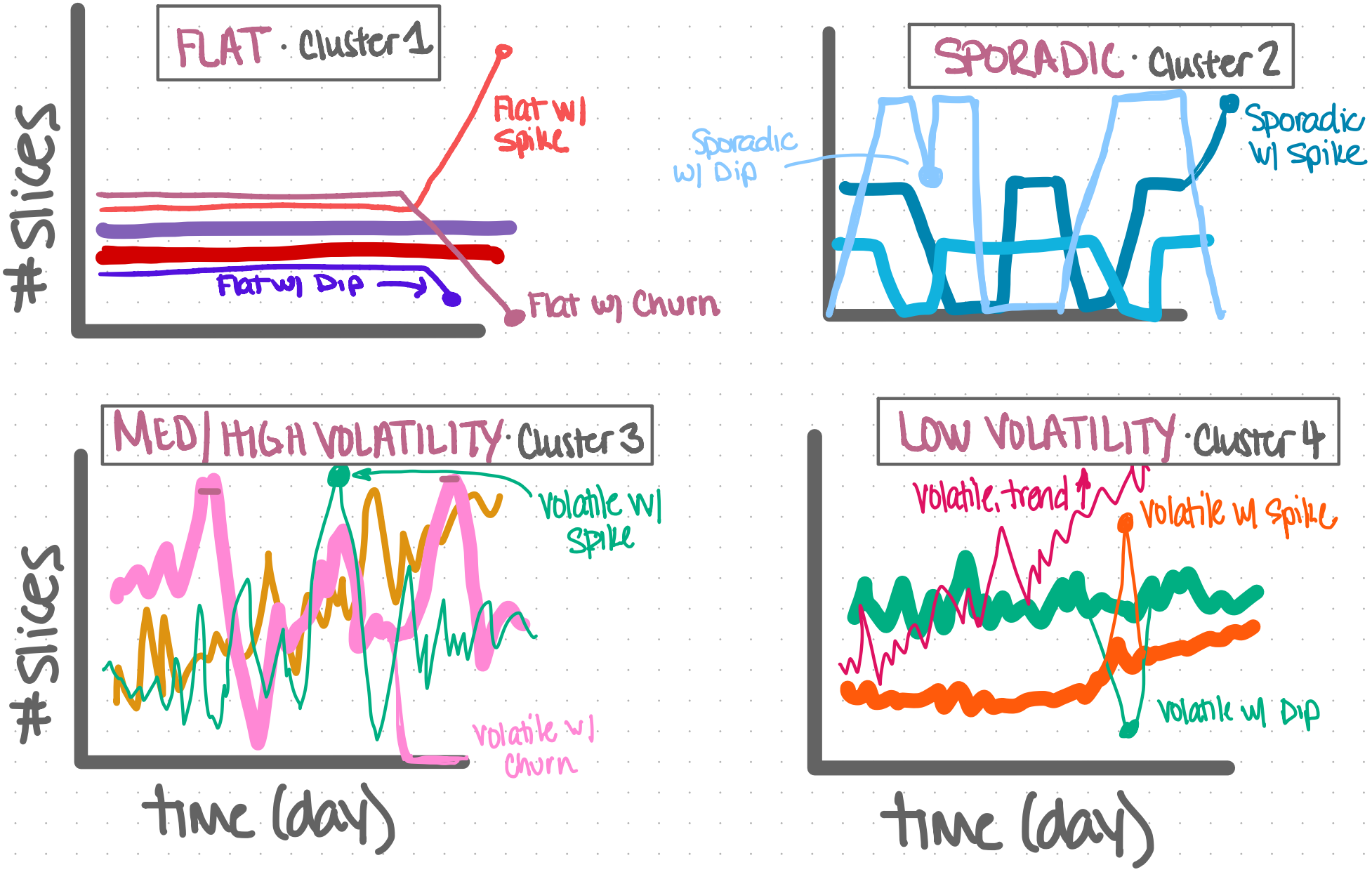
- A cluster ID is returned (an integer)
- That ID means nothing in/of itself
- All it is is a link that ties all the accounts in that cluster together.
- The ID could be "pink" if it wanted!

Explore, Understand, Analyze the clusters

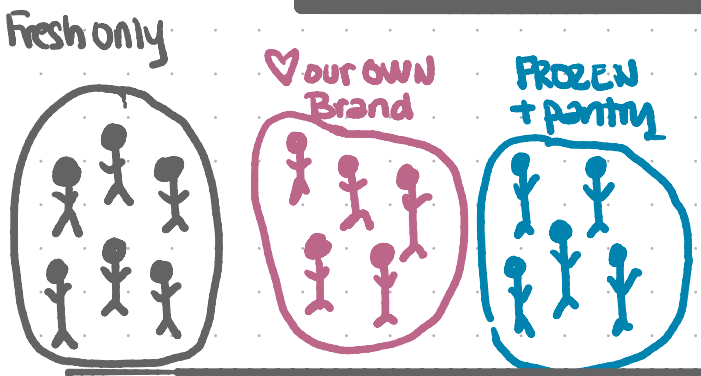
- do they make sense?
- do I need to refine features?
- More / Less clusters?
- Descriptive names of clusters?



THE CLUSTERS



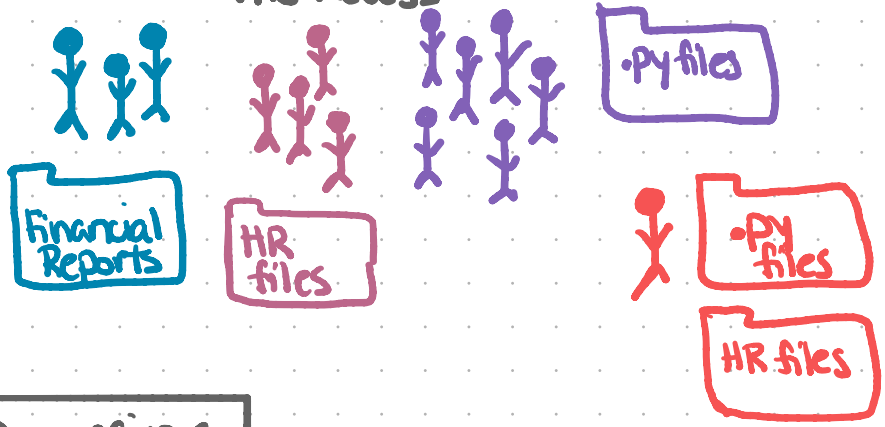
When is Clustering Useful?



Marketing, Customer Segmentation

Anomaly Detection

File Access

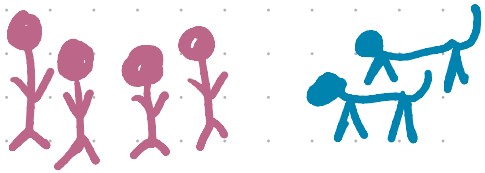


Document Classification

Support	Survey response.. 12:00
Uptime	Survey response 12:01
Support	Survey response.. 1:05
Price	Survey Response.. 1:11
⋮	

→ THEMES of responses

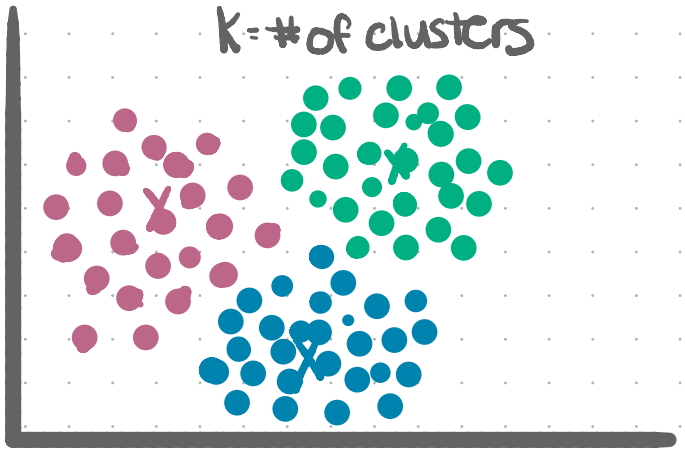
Image Processing



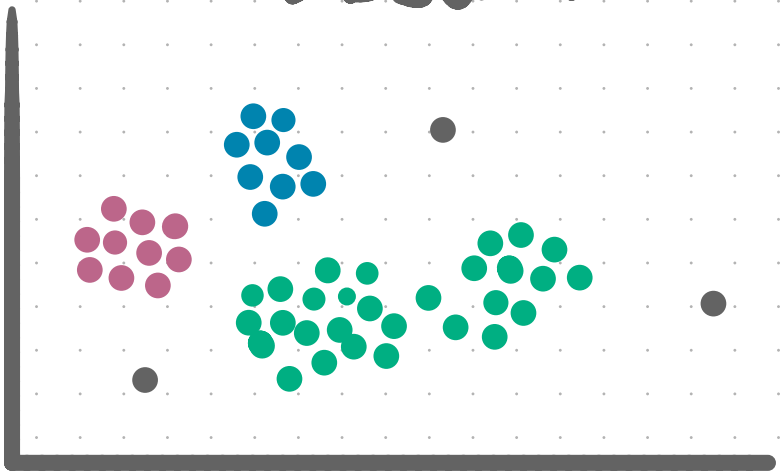
What are the different subjects of these photos?

CLUSTERING ALGORITHMS

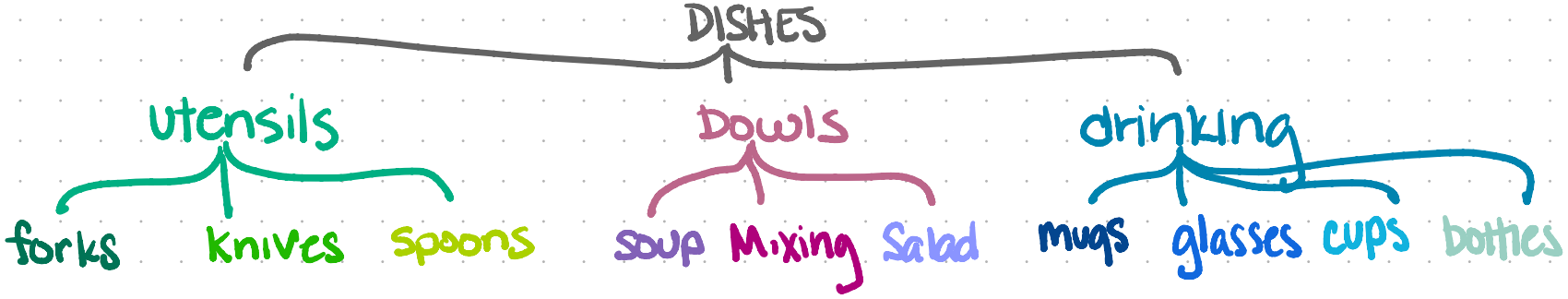
K-Means



DBSCAN

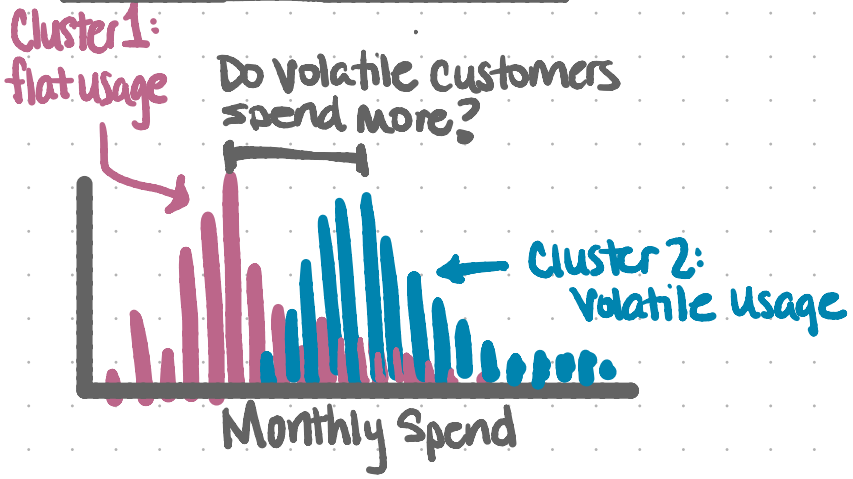


Hierarchical



I HAVE CLUSTERS. NOW WHAT?

EXPLORE & ANALYZE CLUSTER GROUPS



TURN CLUSTERS INTO FEATURES

target | features

Account	Churn	tenure	MRR	is_flat	is_volatile	is_sporadic
112358	1	12	100	0	1	0
123581	0	7	200	0	1	0
235813	0	9	10	1	0	0
358132	1	15	20	0	0	1

MODEL EACH CLUSTER SEPARATELY

PREDICTING CHURN

Model 1 - flat cluster

`.fit(X_train_flat, y_train_flat)`

Model 2 - volatile cluster

`.fit(X_train_volatile, y_train_volatile)`

Model 3 - sporadic cluster

`.fit(X_train_sporadic, y_train_sporadic)`

TURN CLUSTERS INTO LABELS

Classify NEW customers' usage patterns using labels identified through clustering

target

Account	cluster	σ	μ	%Δ	Δd/d
112358	volatile				
123581	volatile				
235813	flat				
358132	sporadic				

Machine Learning

SUPERVISED

CLASSIFICATION REGRESSION

TARGET:

Discrete
Categorical

Continuous
numeric, ordered

SUB-METHODOLOGIES:

time series
Anomaly detection
NLP

time series
anomaly detection

ALGORITHMS:

Logistic Regression
SVM-classifier
Decision Tree
Random Forest
KNN
Neural Network
Naive Bayes

OLS (linear regression)
GLM (tweedie regressor)
Polynomial Regression
Support Vector Regressor
Decision Tree Regressor
Neural Networks
LASSO, LARS
Holt-Winters
Auto Regression

Anomaly Detection
Module

REINFORCEMENT

SUB-METHODOLOGIES

Anomaly Detection
NLP
Recommender Systems

ALGORITHMS:

Collaborative Filtering
Content Filtering
Neural Networks

UNSUPERVISED

CLUSTERING DIMENSIONALITY REDUCTION

SUB-METHODOLOGIES:

Anomaly Detection
NLP

ALGORITHMS:

K-Means
DBSCAN
Hearchical

PCA
t-SNE
factor Analysis