APOSTOLOS MAVROGIANNAKIS

CHARALAMBOS VARSAMIS

Fraud detection

WITH MACHINE LEARNING

Why Machine Learning?

- SUPER FAST
- ACCURATE
- CHEAP

GRAPH ANALYSIS METRICS

- **Pagerank:** determine a rough estimate of how important the node is.
- **Closeness Centrality:** how close and central a node is to other nodes.
- **Eigenvector Centrality:** centrality for a node based on the centrality of its neighbors.

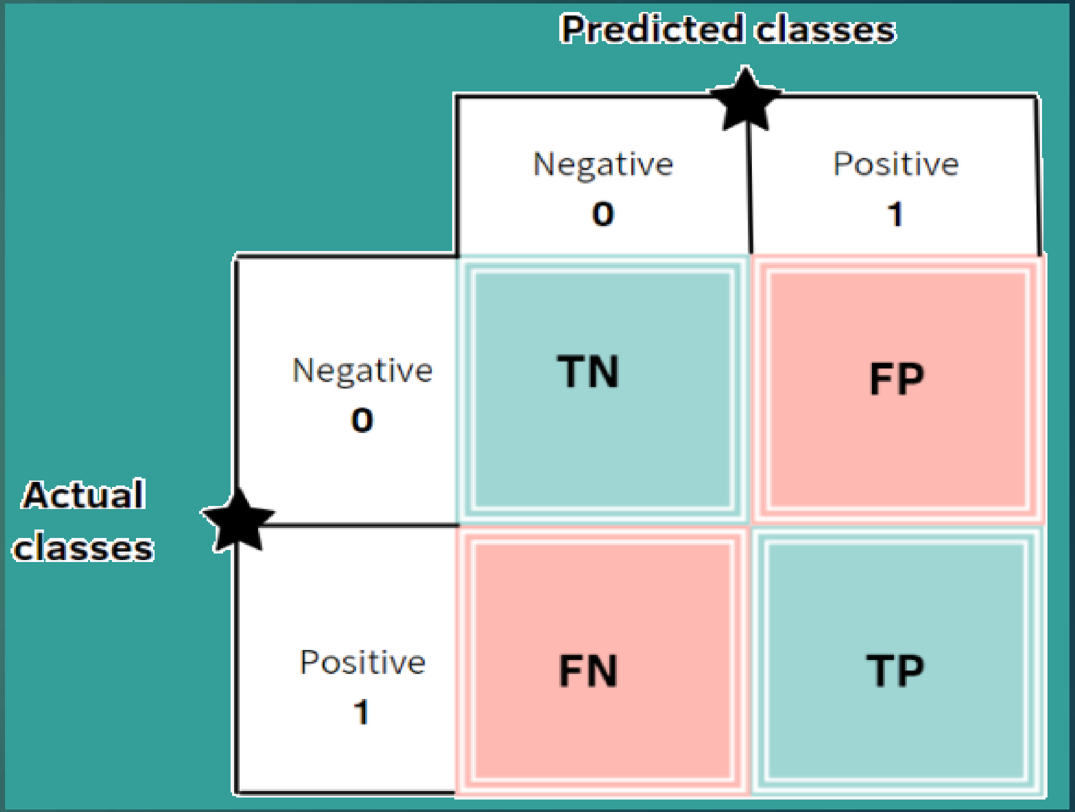
Important Terminologies

- **Confusion Matrices**

- **ROC Curve & AUC score**

Confusion Matrix

- Allows visualisation of the performance of an algorithm
- True Positives
- False Positives
- True Negatives
- False Negatives



		Predicted classes	
		Negative 0	Positive 1
Actual classes	Negative 0	TN	FP
	Positive 1	FN	TP

ROC & AUC

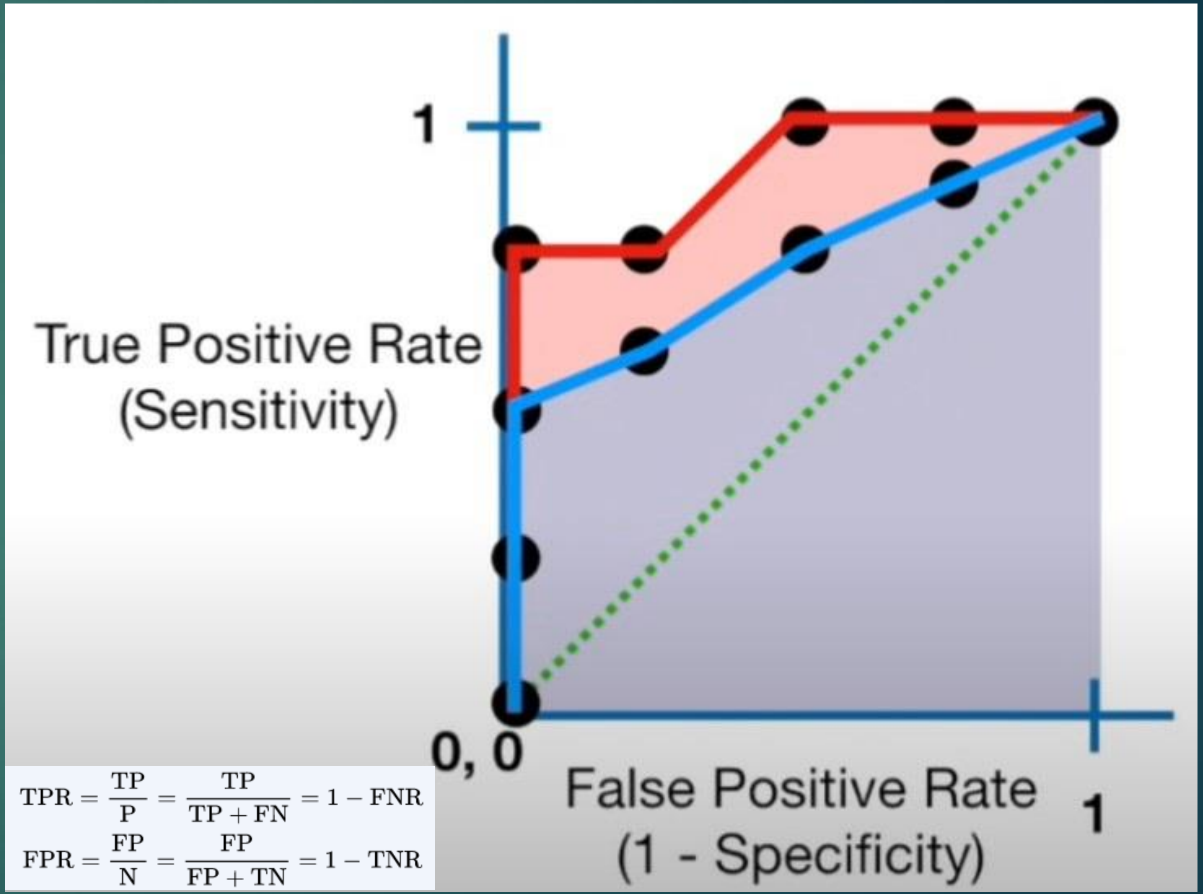
ROC

- True Positive Rate against False Positive Rate

- Summary of Confusion Matrices for each threshold

AUC

- Easy to compare ROC Curves



$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = 1 - \mathrm{FNR}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} = 1 - \mathrm{TNR}$$

Supervised Learning
&
Unsupervised Learning

Supervised learning

- Random Forest
- Adaptive Boosting
- Logistic Regression

RANDOM FOREST

- Create a bootstrap dataset

Original

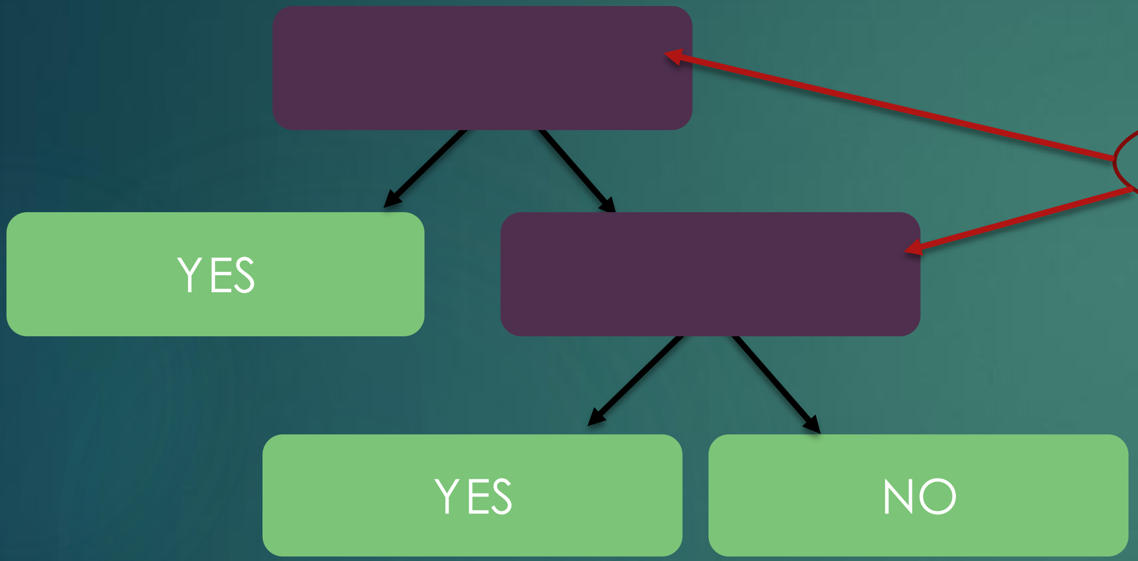
FEATURE 1	FEATURE 2	FEATURE 3	FEATURE 4
YES	NO	YES	YES
NO	YES	YES	NO
YES	NO	YES	NO
NO	YES	NO	YES

Bootstrap

FEATURE 1	FEATURE 2	FEATURE 3	FEATURE 4
YES	NO	YES	YES
NO	YES	YES	NO
YES	NO	YES	NO
YES	NO	YES	NO

RANDOM FOREST

- Create Decision Tree with random columns

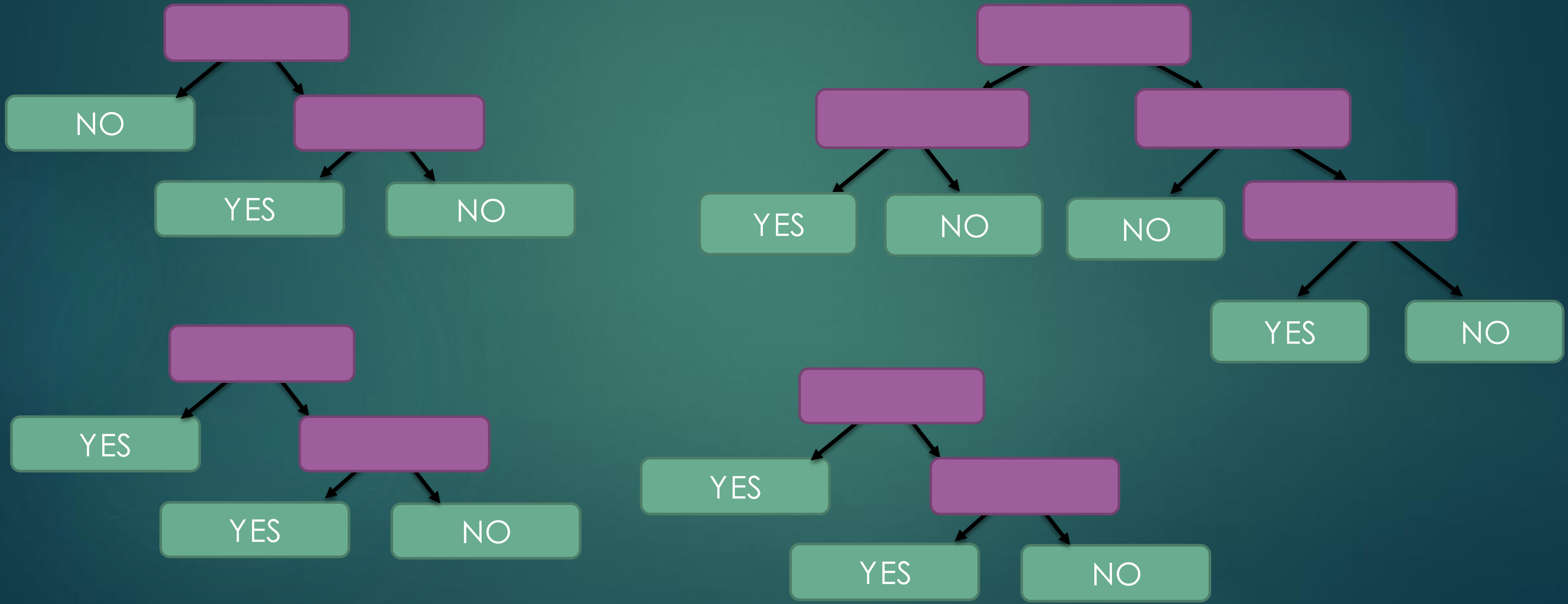


Bootstrap Dataset

FEATURE 1	FEATURE 2	FEATURE 3	FEATURE 4	FRAUD
YES	NO	YES	YES	YES
NO	YES	YES	NO	NO
YES	NO	YES	NO	NO
YES	NO	YES	NO	NO

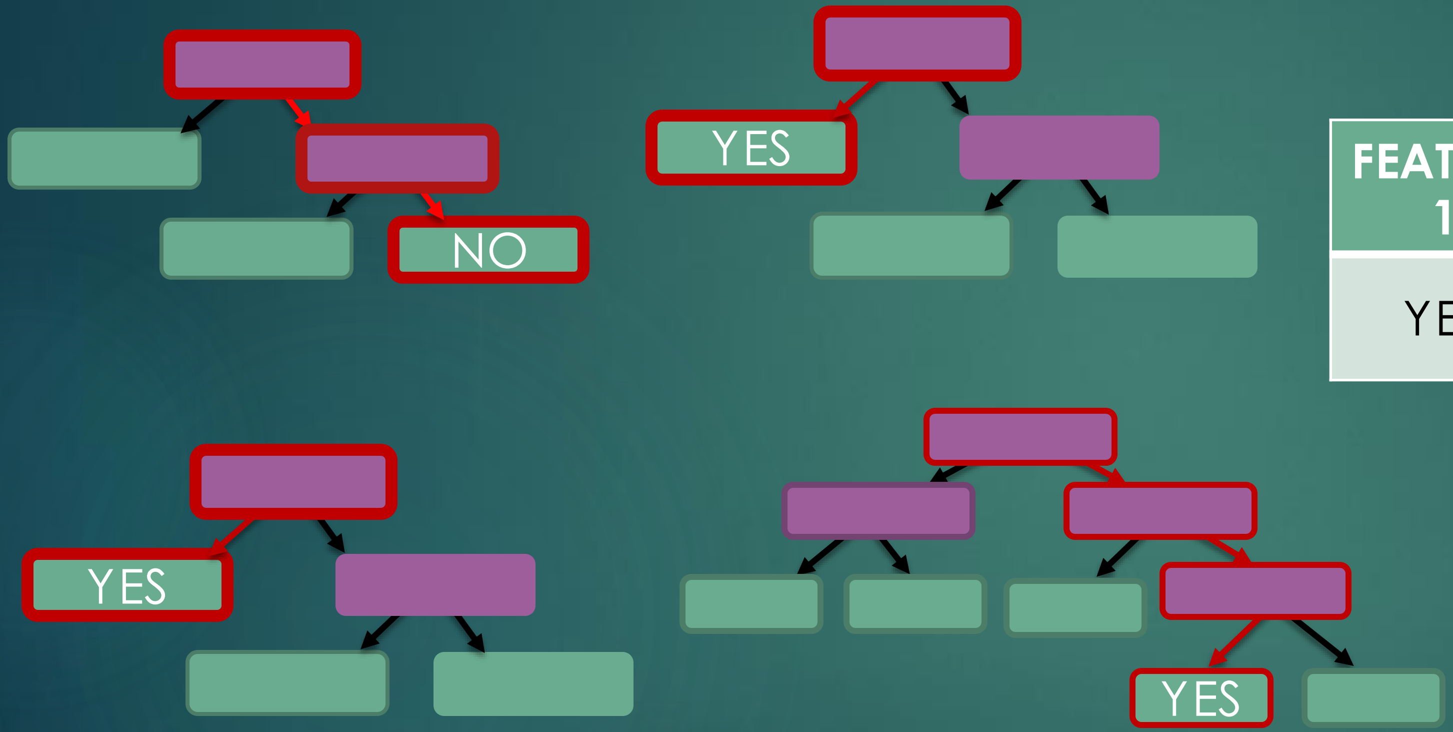
RANDOM FOREST

- Create Multiple Decision Trees



RANDOM FOREST

- Take each row and run through the Decision Trees we created



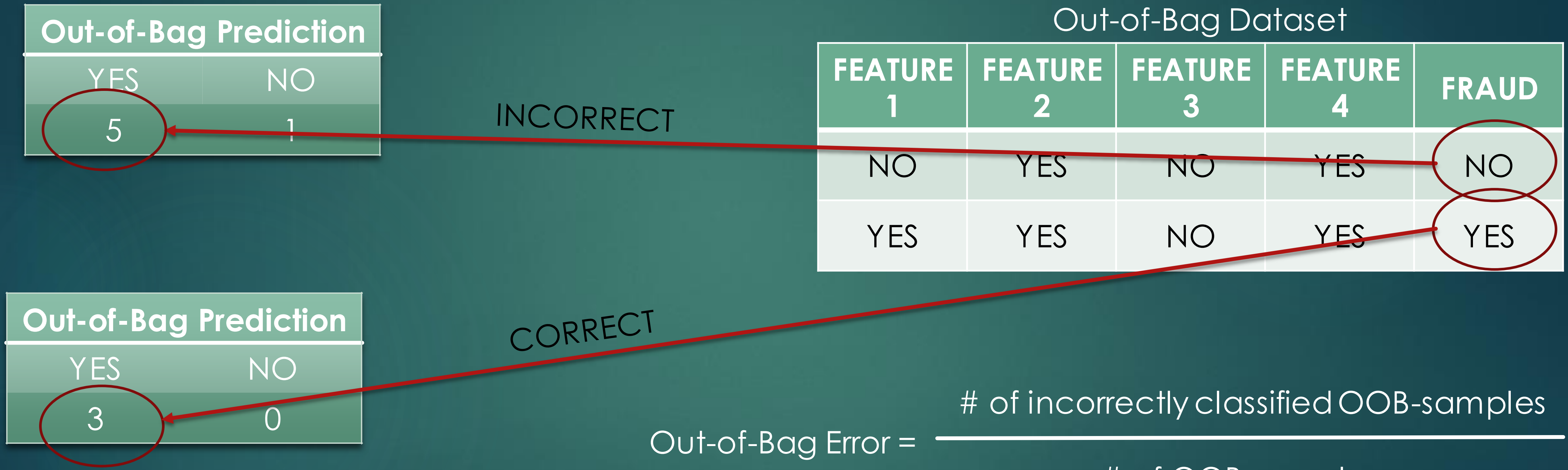
Bootstrap Dataset

FEATURE 1	FEATURE 2	FEATURE 3	FEATURE 4	FRAUD
YES	NO	YES	YES	**YES**

FRAUD	
YES	NO
3	1

RANDOM FOREST

- To evaluate Random Forest we need to calculate an Error over Out-of-Bag samples



Out-of-Bag Prediction	
YES	NO
5	1

INCORRECT

Out-of-Bag Dataset

FEATURE 1	FEATURE 2	FEATURE 3	FEATURE 4	FRAUD
NO	YES	NO	YES	NO
YES	YES	NO	YES	YES

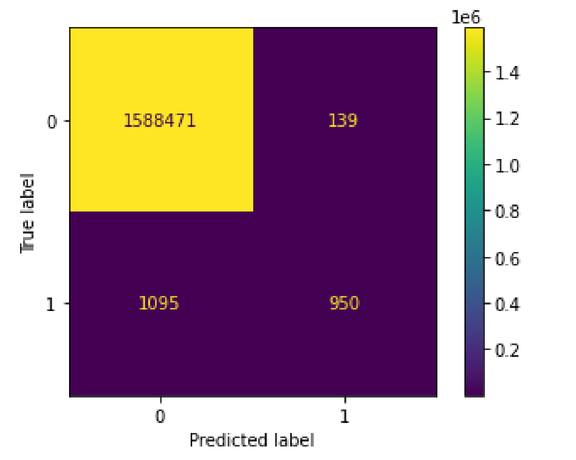
CORRECT

Out-of-Bag Prediction	
YES	NO
3	0

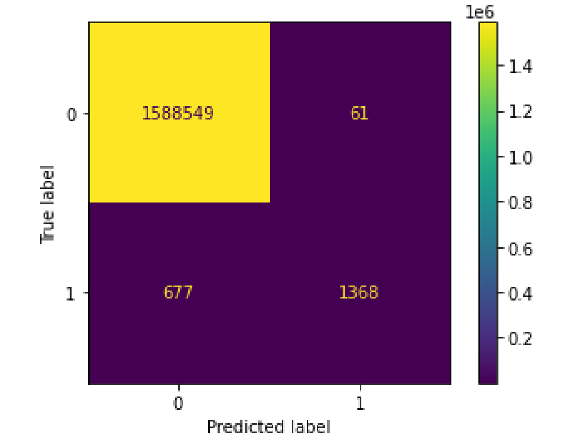
$$\text{Out-of-Bag Error} = \frac{\text{\# of incorrectly classified OOB-samples}}{\text{\# of OOB-samples}}$$

OVERIEW OF SUPERVISED MODELS

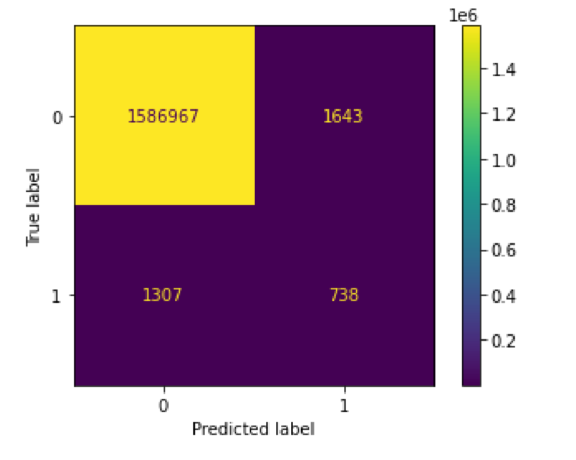
Adaptive Boosting

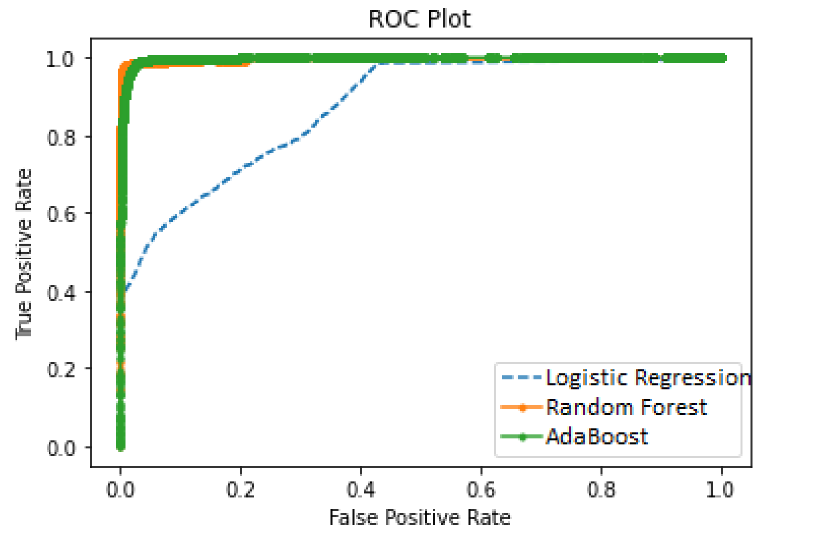


Random Forest



Logistic Regression





ROC Curve

Unsupervised Learning

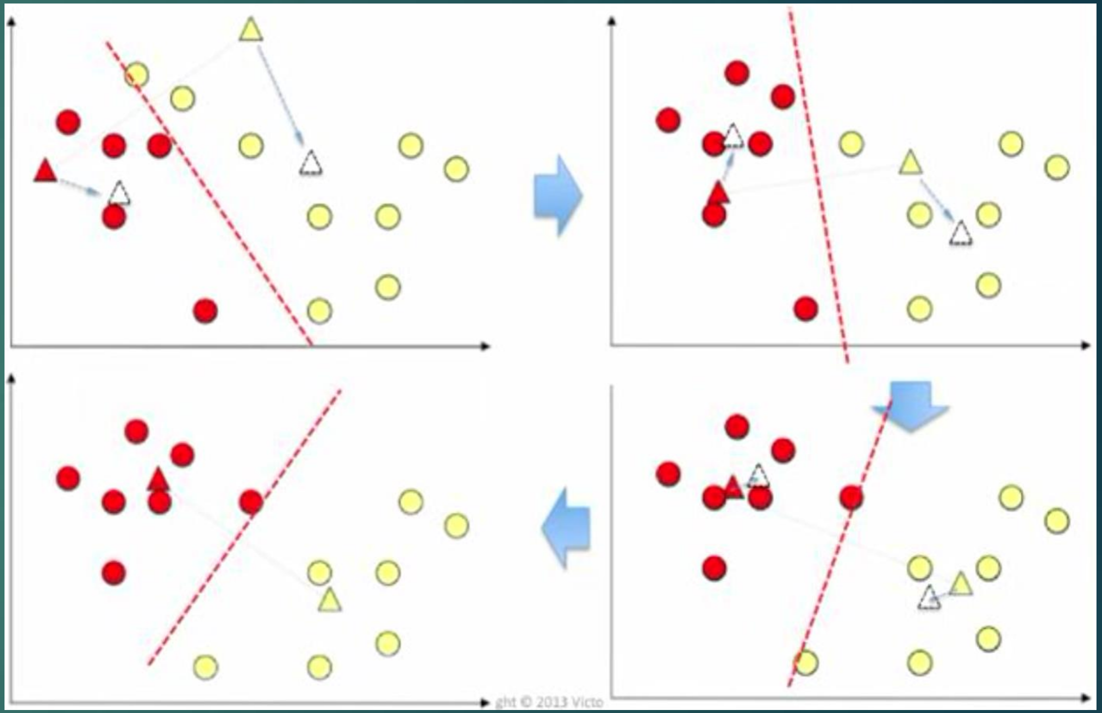
- K-Means Clustering
- Isolation Forest

Why Unsupervised Learning ?

K-Means Clustering

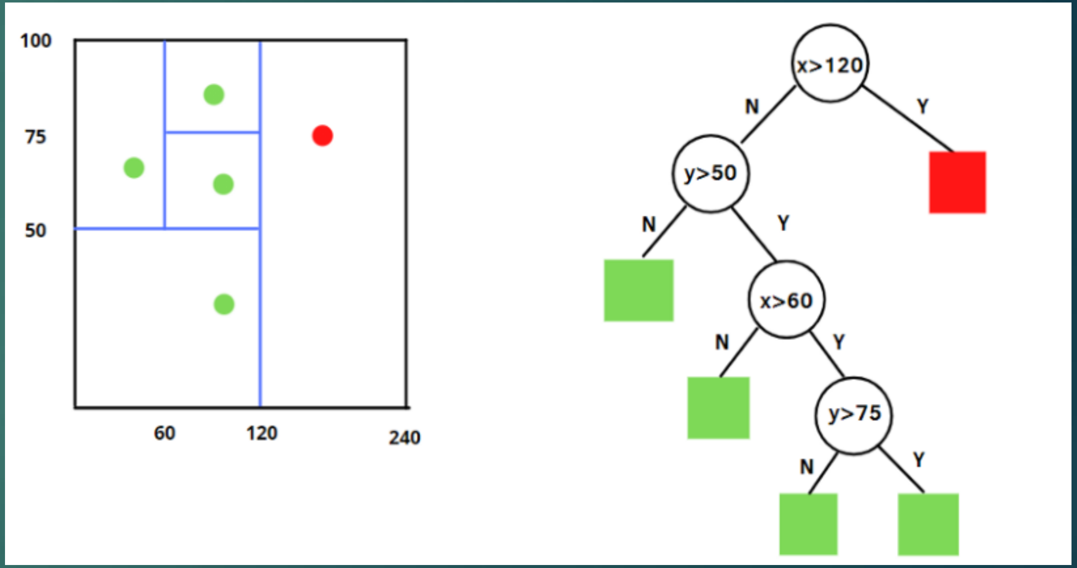
K-means is a centroid-based algorithm.

1. Choose the number of clusters k
2. Select k random points from the data as centroids
3. Assign all the points to the closest cluster centroid
4. Recompute the centroids of newly formed clusters
5. Repeat steps 3 and 4

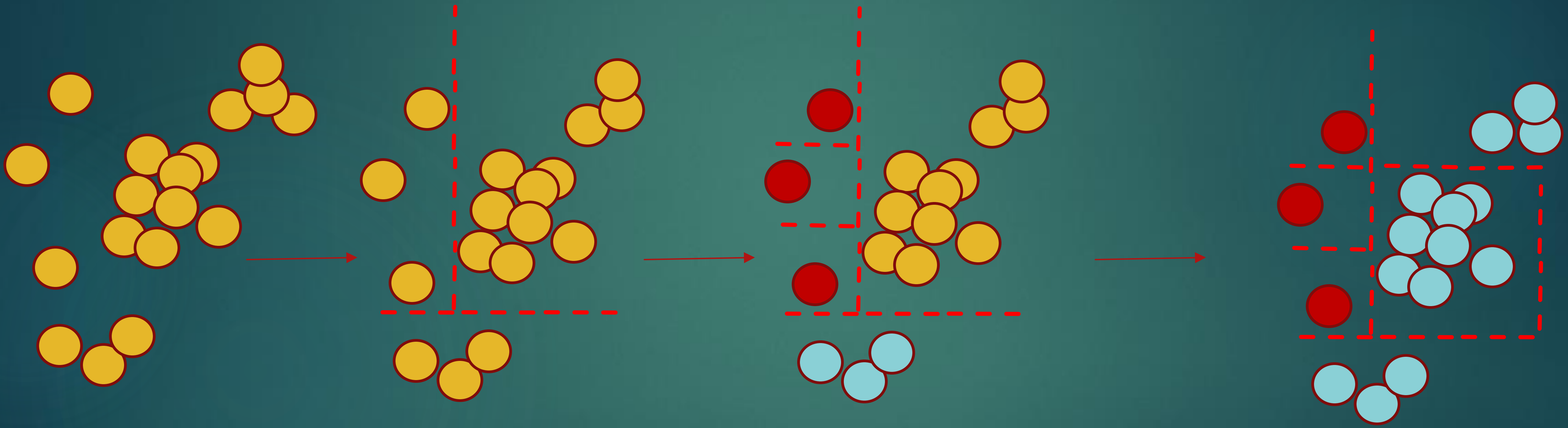


Isolation Forest

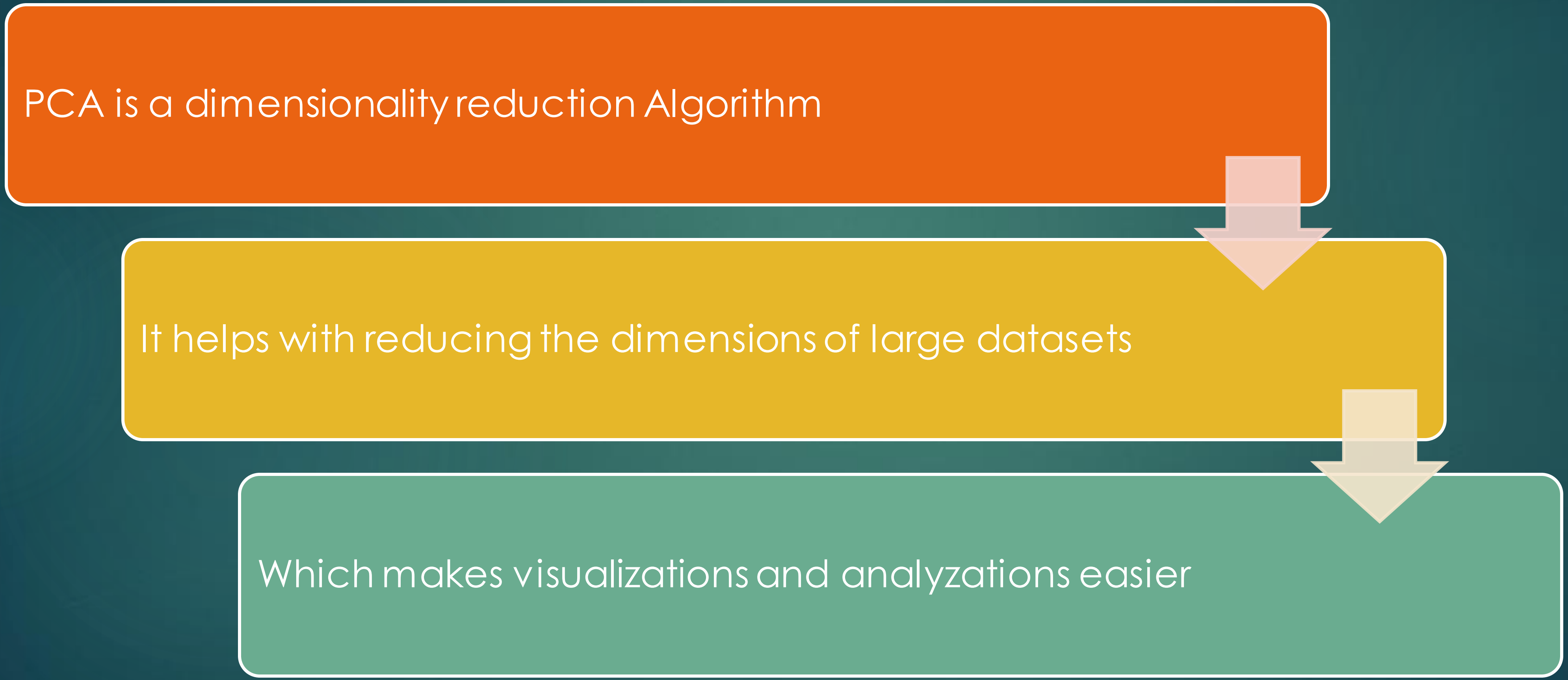
- Isolation forest is a machine learning algorithm for anomaly detection.
- Isolation Forest is based on the Decision Tree algorithm
- How does it detect anomalies?
- F(x) = P('Anomaly' | G(x))



Isolation Forest



Principal Component Analysis



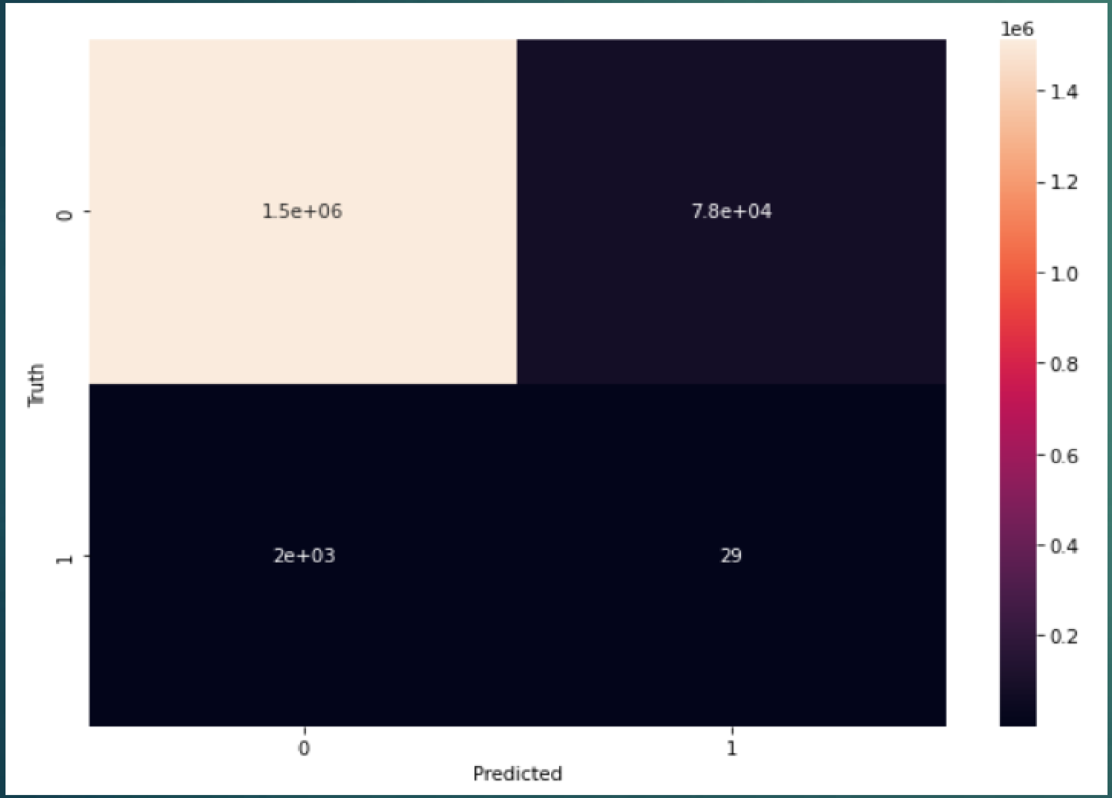
PCA is a dimensionality reduction Algorithm

It helps with reducing the dimensions of large datasets

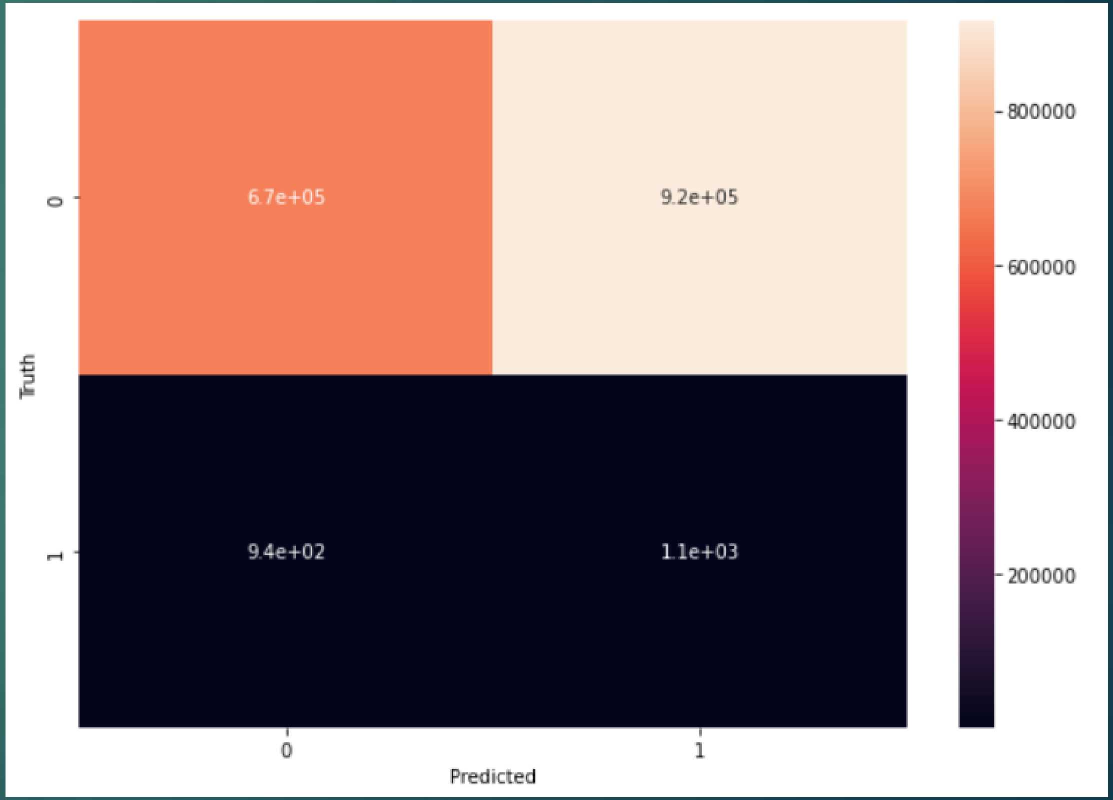
Which makes visualizations and analyzations easier

OVERIEW OF UNSUPERVISED MODELS

K-Means Clustering



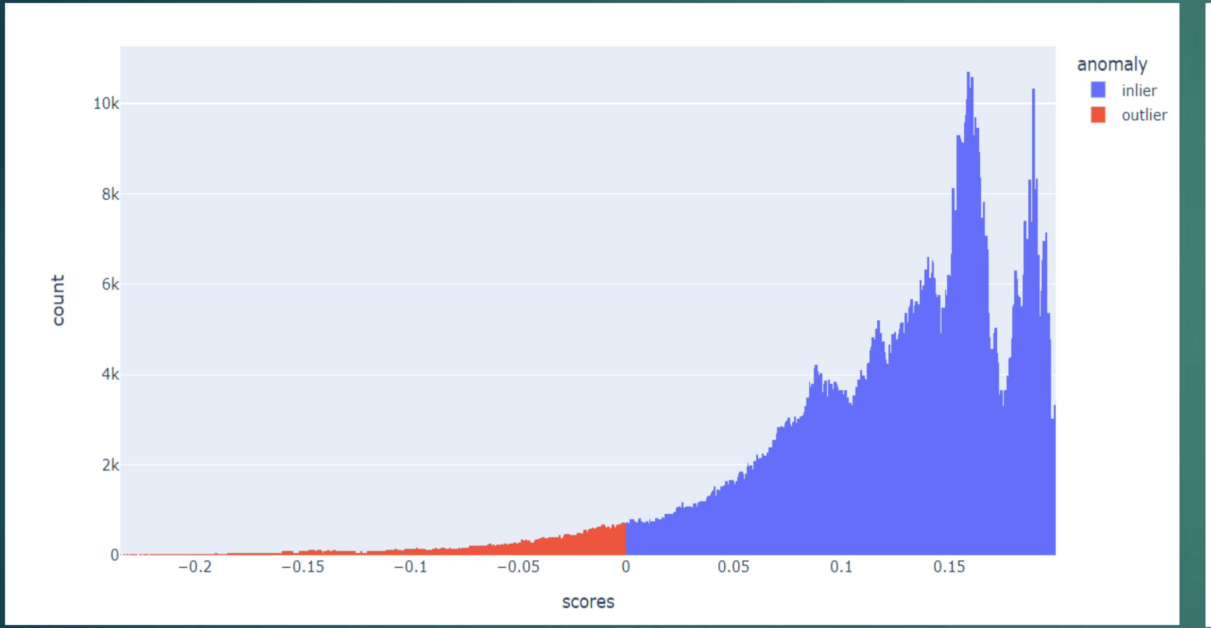
Without PCA



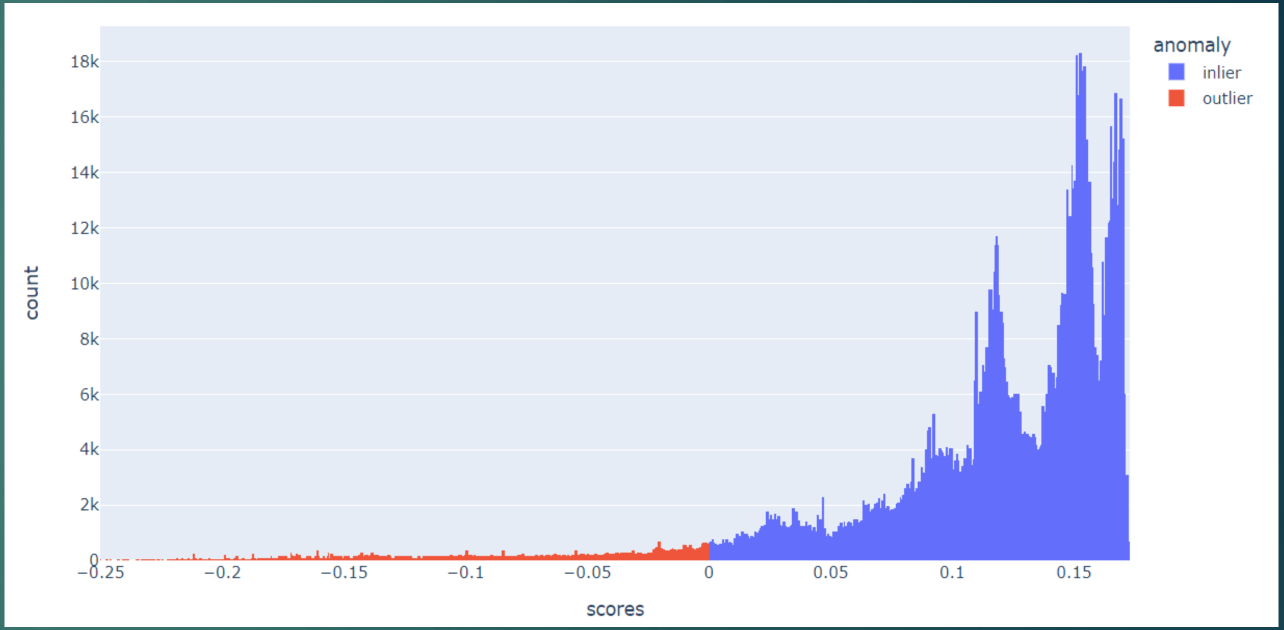
With PCA

OVERIEW OF UNSUPERVISED MODELS

Isolation Forest



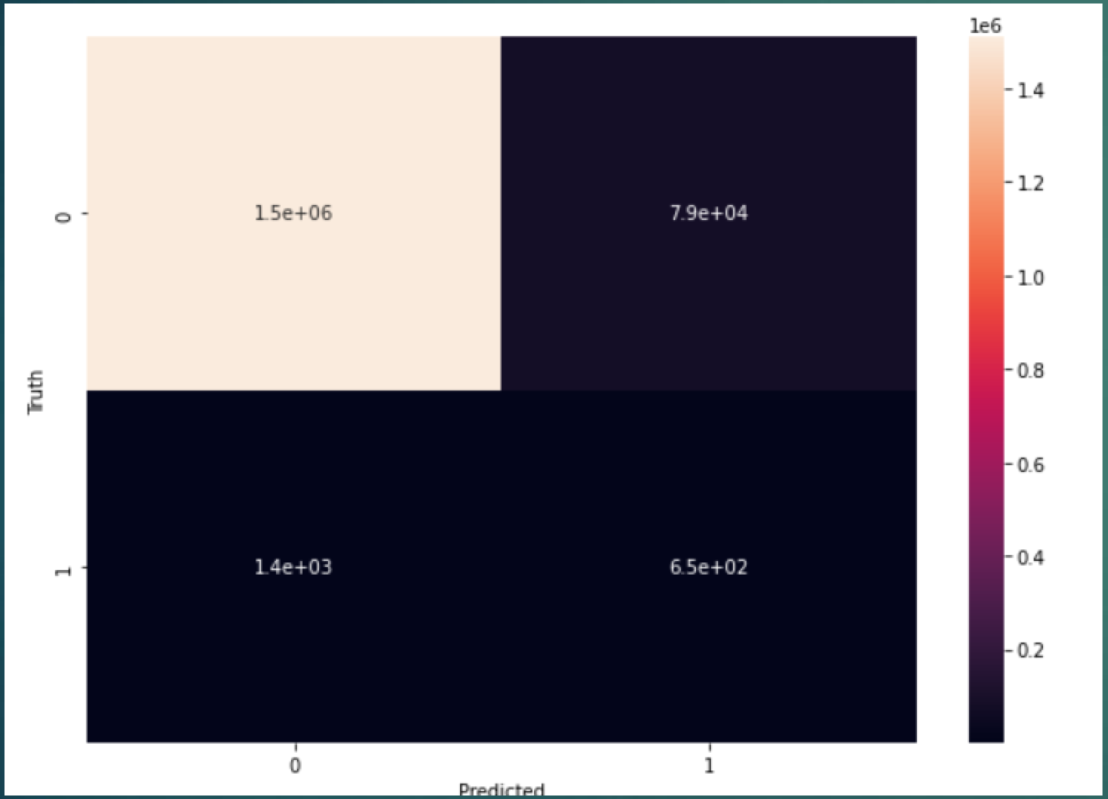
Without PCA



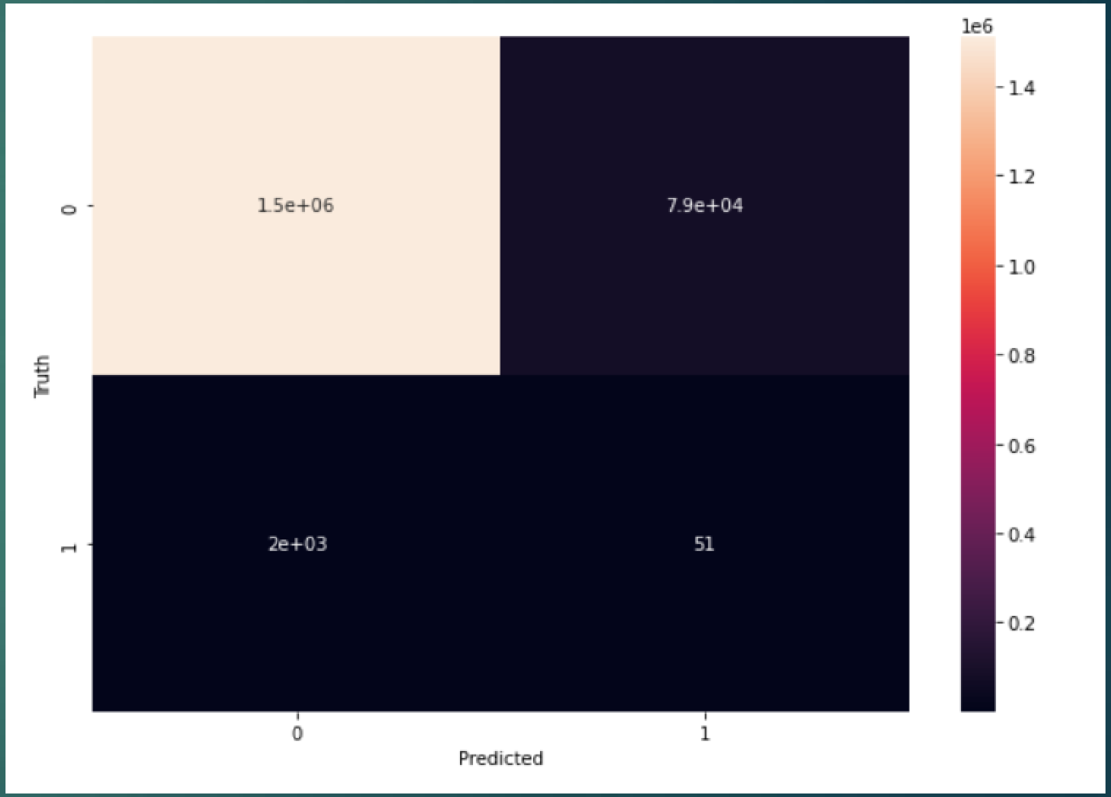
With PCA

OVERIEW OF UNSUPERVISED MODELS

Isolation Forest



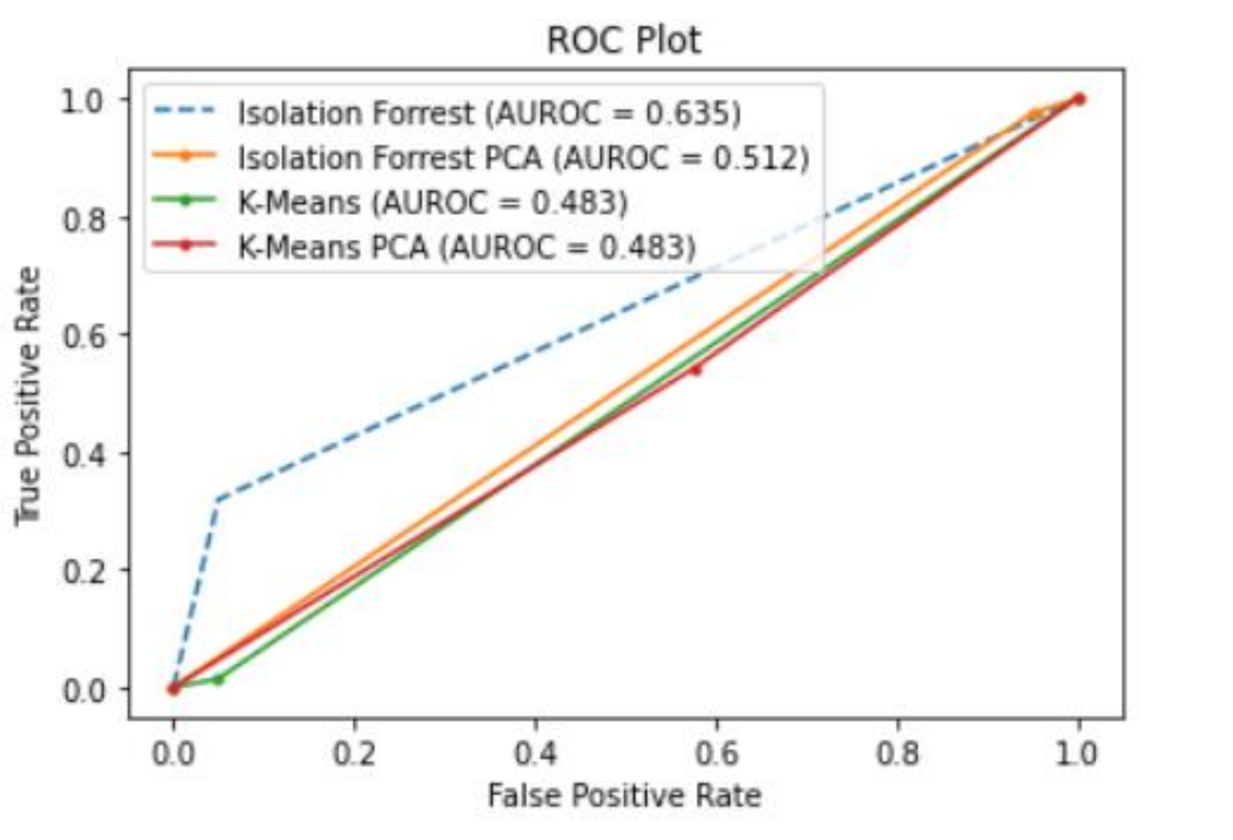
Without PCA



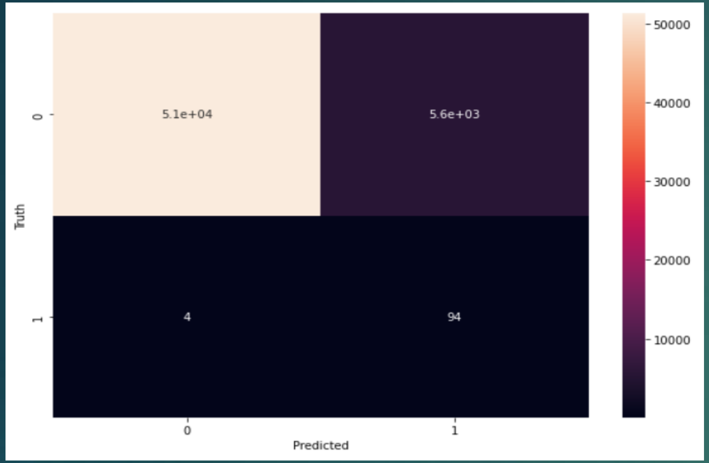
With PCA

OVERIEW OF UNSUPERVISED MODELS

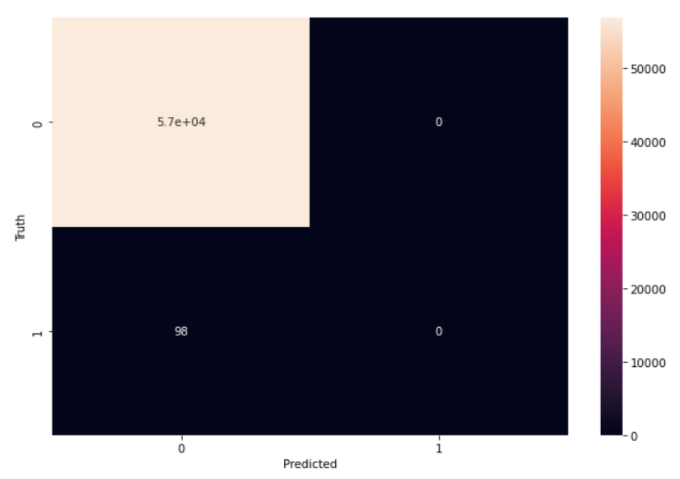
ROC Curve



OVERIEW OF UNSUPERVISED MODELS

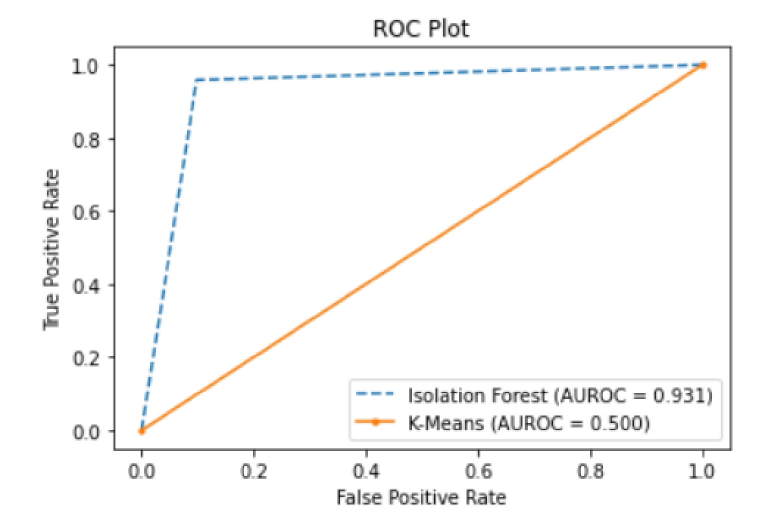


Isolation Forest



K-Means

ROC Plot



Questions ?