Anomaly Detection with Python Outlier Detection Isolated Forest

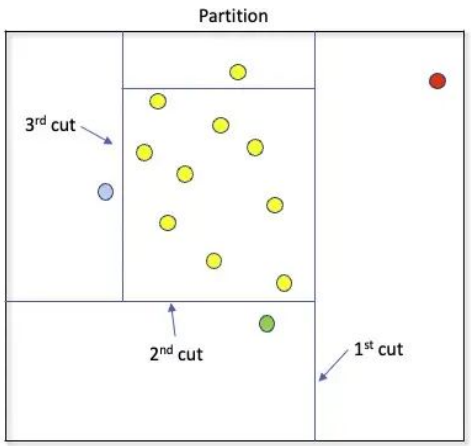
What Is the Isolate Forest?

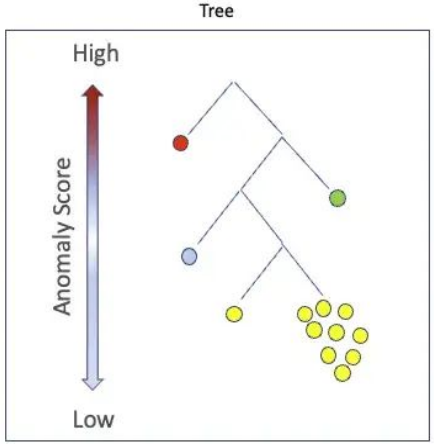
Many outlier detection methods profile the norm data points first, then identify those observations that do not conform to the patterns of the normal data. ***IForest*** identifies anomalies directly. It applies a **tree** structure to isolate every observation.
Anomalies will be the data points first to be singled out; whereas normal points tend to hide deep in the tree.
Each **tree** is the *Isolation Tree* or ***iTree***.

Isolation Data using Partition Map or a Tree

It takes only one “cut” to separate the red dot from the others. The second cut is for the green dot, and the third cut is for the blue dot, and so on. The more cuts it takes to separate a dot, the deeper it is in the tree.

The inverse of the number of cuts is the **Anomaly Score**.





An **iTree** is a **Binary** tree

- An **iTree** is a **binary** tree, where each node in the tree has exactly zero or two daughter nodes. An **Tree** starts to grow until one of the conditions is met: The end node has only one data point.

- The **Anomalies** are closer to the **root**

iTree Algorithm

iTree algorithm is different from the **decision tree** algorithm because iTree does not use a target variable to train the tree. It is an **unsupervised** learning method.

The goal of **IForest** is to assign an outlier score to each observation

It randomly selects *any number of rows* and *any number of columns* to create tables such as (1), (2), and (3)

An **iTree** is built for each table to render outlier scores. There are ***N*** tables, there will be ***N* iTrees**. An observation can have up to ***N* scores.**

IForest computes the arithmetic mean of the scores to get the final score.

$A_{ij} =$

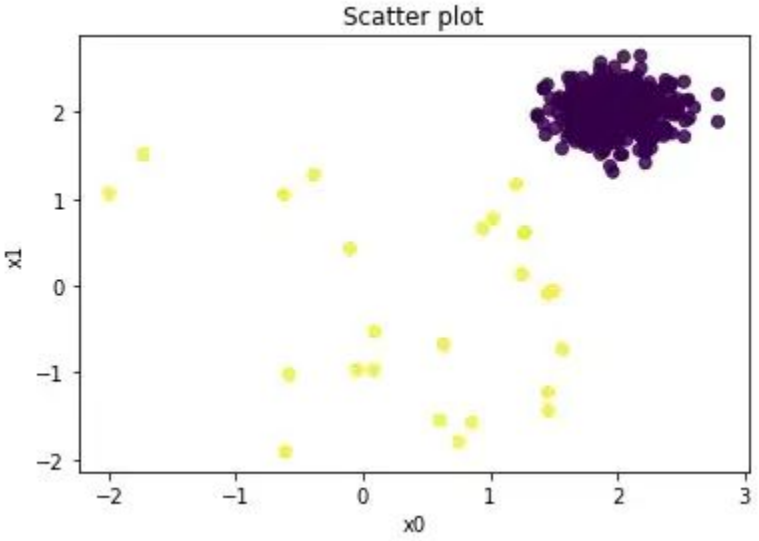
(1)					(2)				
0	3	1	0	2	3	2	1	1	3
1	1	0	0	7	1	2	2	3	3
1	2	2	0	0	3	3	1	2	2
9	8	9	10	6	7	9	10	7	9
3	2	2	1	2	3	2	1	6	0
17	41	14	25	33	29	46	21	36	10
4	1	1	5	2	2	5	0	3	4
5	0	1	6	2	0	0	0	1	5
1	6	3	3	4	6	2	0	1	1
1	2	2	4	1	1	3	0	2	2
		(3)							

⇦ The 6th obs.

Modeling Procedure

- A mock of dataset with six variables and 500 observations was generated. The percentage of outliers is set to 5% with “contamination=0.05.”

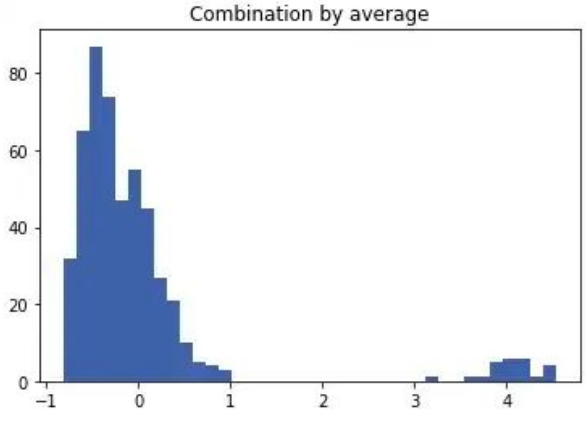
- The yellow points are the outliers and the purple points are the normal data points.



Determine a Reasonable Threshold for the Model

- The threshold is to be determined by the histogram of the outlier scores.
- It identifies 25 data points to be the outliers.

	Group	Count	Count %	0	1	2	3	4	5	Anomaly_Score
0	Normal	475	95.0	2.00	2.01	2.01	1.99	2.01	1.98	-0.21
1	Outlier	25	5.0	0.45	-0.21	-0.47	-0.23	-0.03	-0.06	3.96



Summary

- IForest directly and explicitly isolates anomalies.
- Isolation Forest does not use any distance measures to detect anomalies, it is fast and suitable for large data sizes and high-dimensional problems.