**Isolation Forest**

**Definition**: - Isolation forest works on the principle of the decision tree algorithm. It isolates the outliers by randomly selecting a feature from the given set of features and then **randomly selecting a split value between the maximum and minimum values of the selected feature**. This random partitioning of features will produce smaller paths in trees for the anomalous data values and distinguish them from the normal set of the data.

This process continues recursively until each data point is isolated.

Isolation Forests were built based on the fact that anomalies are the data points that are “**few and different**”.

Isolation Forests are nothing but an **ensemble of binary decision trees**

Arguably, the anomalies need **fewer random partitions to be isolated compared** to the so defined normal data points in the datasets.

**The core principle**:-

The core of the algorithm is to “**isolate**” anomalies by creating decision trees over random attributes. The **random partitioning produces noticeable shorter paths for anomalies since**

* fewer instances (of anomalies) result in smaller partitions
* distinguishable attribute values are more likely to be separated in early partitioning

Hence, when a forest of random trees **collectively produces shorter path lengths for some particular points, then they are highly likely to be anomalies.**

**Process**:

1. When given a dateset, a **random sub-sample of the data** is selected and assigned to a **binary tree**.
2. Branching of the tree starts by **selecting a random feature** (from the set of **all N features**) first. And then branching is done on a **random threshold** ( **any value in the range of minimum and maximum values** of the selected feature).
3. If the value of a data point is **less than the selected threshold**, it goes to the **left branch** else to the **right**. And thus a **node is split into left and right branches**.
4. This process from **step 2** is continued recursively **till each data point** is completely isolated or till **max depth**(if defined) is reached.
5. The above steps are repeated to construct random binary trees.

**Why sub-sampling helps:**

The Isolation Forest algorithm works well when the trees are created, **not from the entire datasets, but from a sub-sampled data set**. This is very different from almost all other techniques where they thrive on data and demands more of it for greater accuracy. Sub-sampling works wonder in this algorithm because normal instances can interfere with the isolation process by being a little closer to the anomalies.



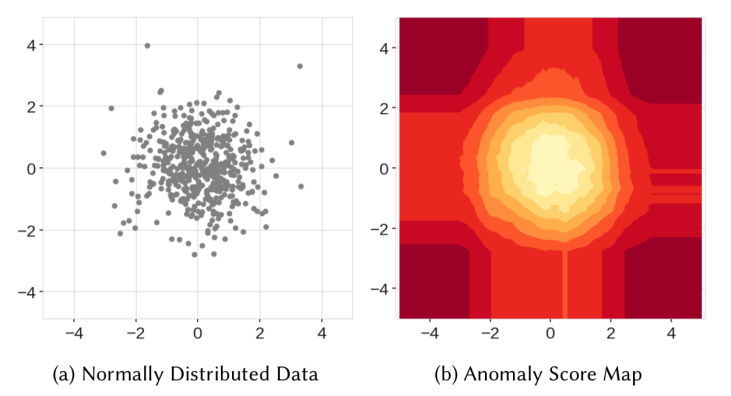
The image above shows how **sub-sampling actually makes a clear separation between normal points and anomalies.** In the original dateset, we see that normal points and very close to anomalies making detection tougher and inaccurate (with a lot of false negatives). Because of sub-sampling, we could see a clear separation of anomalies and normal instances. This makes the entire process of anomaly detection efficient and accurate.

Now, an ‘**anomaly score**’ is assigned to each of the **data points based on the depth of the tree required to arrive at that point**. This score is an aggregation of the depth obtained from each of the iTrees. An **anomaly score of -1 is assigned to anomalies and 1 to normal points based on the contamination(percentage of anomalies present in the data) parameter** provided.

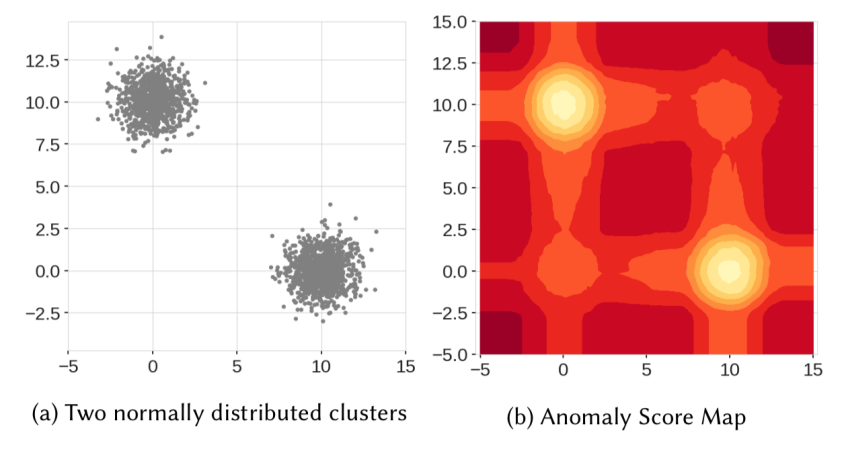
**Disadvantages:**

The final anomaly score depends on the contamination parameter, provided while training the model. This implies that we should have an idea of what percentage of the data is anomalous beforehand to get a better prediction.

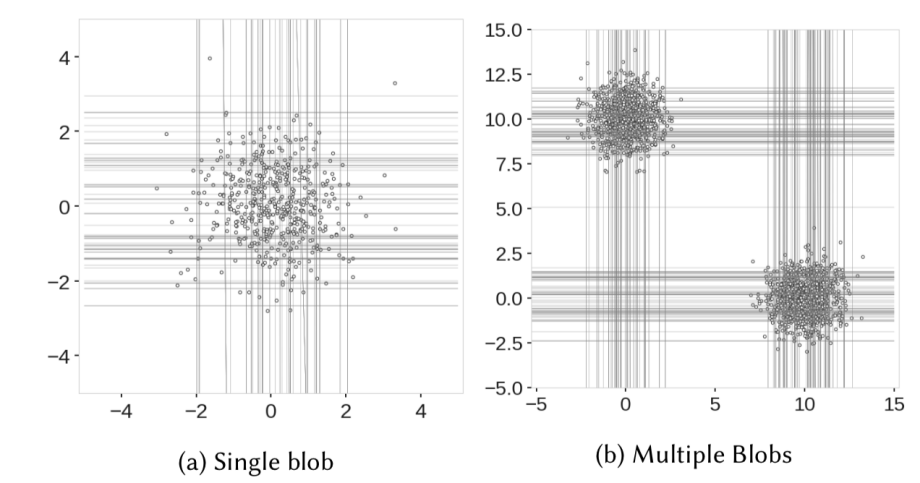
Also, the **model suffers from a bias due to the way the branching takes place**.



1. When the Isolation forest map the anomaly it will generate the score map to identify the anomaly easily.(**Score 0 is Normal and 1 is Outliers**) has some values are **light far from normal data points those data point has a near to 0 score instead of 1 which outliers.**



1. When there are two blobs of data clusters it will **automatically creates the two ghost clusters (on the top right and bottom left )** which never even **existed in the data**.



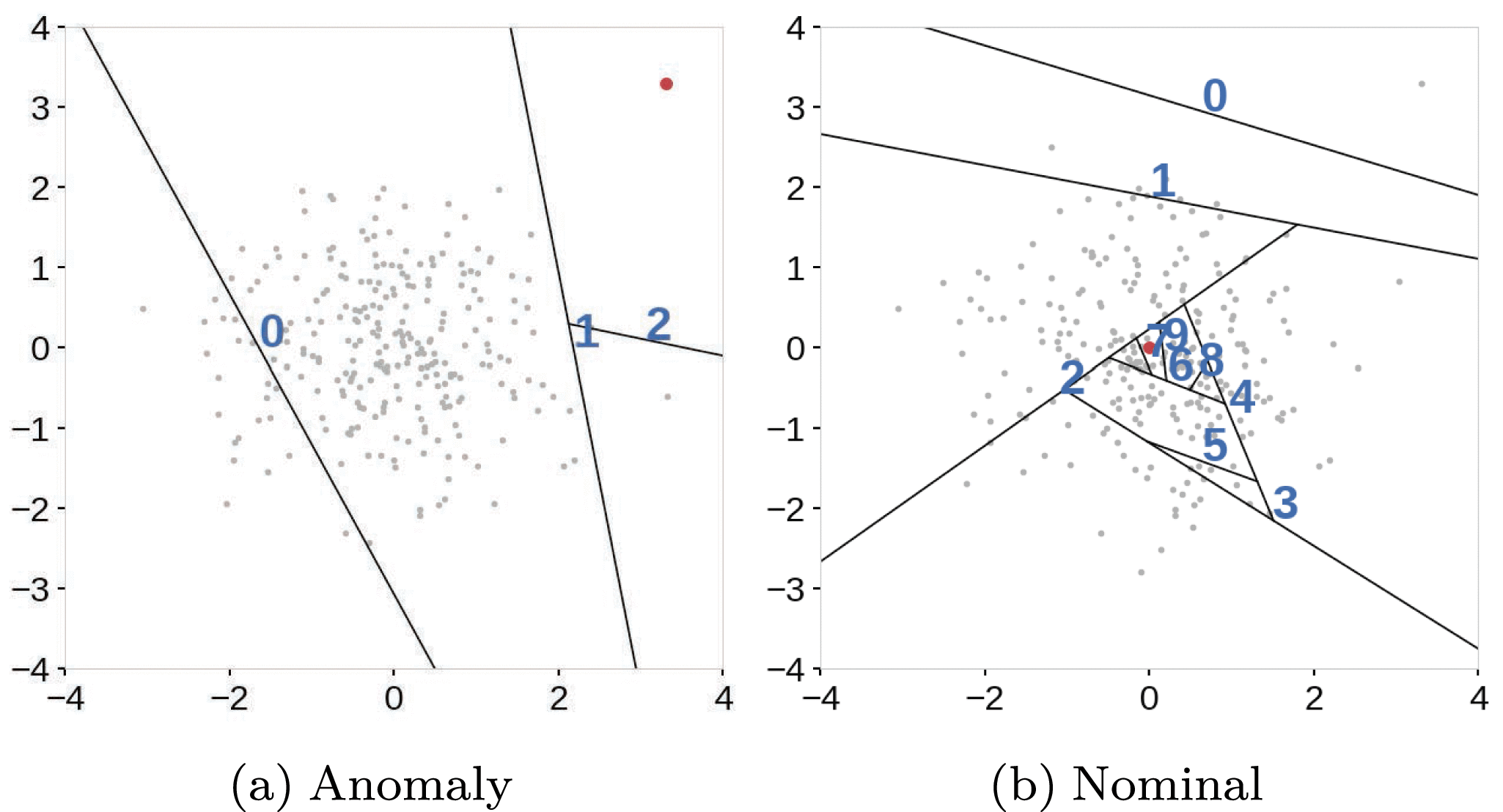
1. The reason for these behaviour The Isolation Forest divide the **decision boundaries with parallel lines (which are Vertical and Horizontal)**. This result over construction of many tree's.In that region that don’t occupy any **data points with unnecessary branch cuts.**

**To Solve this problem Extended Isolation Forest(EIF) Has came into place**.

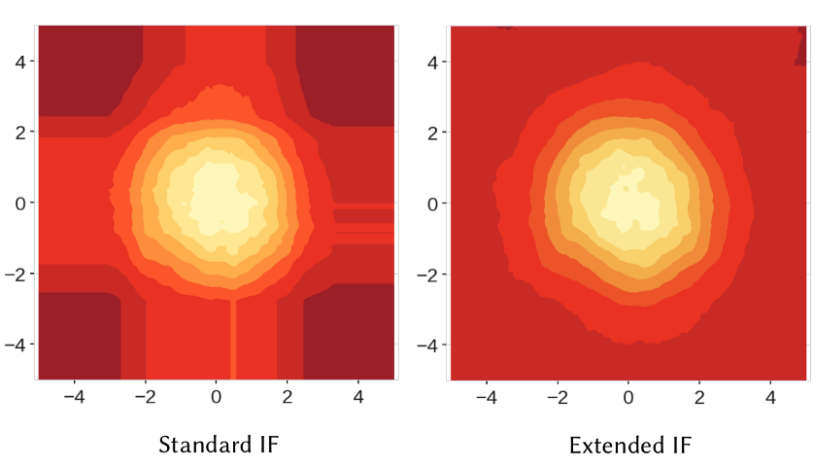
**Extended Isolation Forest: -** To overcome the Isolation Forest drawback led to the conclusion that the problem is caused by **only horizontal/vertical branch cuts**. Extended Isolation Forest addresses that issue by approaching the **problem a bit differently**. Instead of selecting a **random feature and then random value** within the range of data it selects:

* The **random slope** for the branch cuts
* **Random intercept** chosen from the range of available values from the training data

The process of choosing the **branch cuts altered for every Node**.It will choose **Slope** and **Intercept** Randomly with the help of **linear Regression(y=mx+c).**



Extended Isolation Forest generalizes well into **higher dimensions**, where instead of **straight lines** they are dealing with hyper-planes.



**compute\_paths()**: - Will give the anomaly score and the data points has to pass has **numpy array (X.values)**

The **hyper-parameters** of the EIF model are:

* **number of trees**,
* **sub-sampling size**,
* **Extension level** = value in range [0, P − 1]; where P is the number of features.

The minimum value of the hyper-parameter is 0, which corresponds to IF

behavior. The maximum is P − 1 and stands for a fully-extended EIF.The minimum value of the hyper-parameter is 0, which corresponds to IF

behavior. The maximum is P − 1 and stands for a fully-extended EIF.

* **As score is closer to 1, then it is an anomalous point**
* **As the score is closer to 0, it a normal observation**
* **A score near 0.5, indicates it doesn’t have much distinction from normal observations**

**Important parameters for Isolation Forest:**

**Number of estimators**: n\_estimators refers to the number of base estimators or trees in the ensemble, i.e. the number of trees that will get built in the forest. This is an integer parameter and is optional. The default value is 100.

**Max samples**: max\_samples is the number of samples to be drawn to train each base estimator. If max\_samples is more than the number of samples provided, all samples will be used for all trees. The default value of max\_samples is 'auto'. If 'auto', then max\_samples=min(256, n\_samples)

**Contamination**: This is a parameter that the algorithm is quite sensitive to; it refers to the expected proportion of outliers in the data set. This is used when fitting to define the threshold on the scores of the samples. The default value is 'auto'. If ‘auto’, the threshold value will be determined as in the original paper of Isolation Forest.

**Max features**: All the base estimators are not trained with all the features available in the datasets. It is the number of features to draw from the total features to train each base estimator or tree.The default value of max features is one.

**Important Methods for Isolation Forest:**

**Decision\_function()**: - Returns a score such that examples having more negative score as more anomalous.

**Predict()**: - Returns -1 for Anomalous points and 1 for normal point

**References**: -

https://towardsdatascience.com/outlier-detection-with-extended-isolation-forest-1e248a3fe97b

https://dspace.cvut.cz/bitstream/handle/10467/87988/F8-DP-2020-Valenta-Adam-thesis.pdf

https://arxiv.org/pdf/1811.02141.pdf

https://github.com/sahandha/eif

https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf?q=isolation-forest

https://towardsdatascience.com/isolation-forest-a-tree-based-algorithm-for-anomaly-detection-4a1669f9b782

https://medium.com/analytics-vidhya/understanding-isolation-forest-technique-for-anomaly-detection-261c702fa92f

https://www.analyticsvidhya.com/blog/2021/07/anomaly-detection-using-isolation-forest-a-complete-guide/

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8888179

https://www.linkedin.com/pulse/fraud-analytics-using-extended-isolation-forest-algorithm-kumar

https://medium.com/analytics-vidhya/anomaly-detection-using-isolation-forest-algorithm-8cf36c38d6f7

https://dspace.cvut.cz/bitstream/handle/10467/87988/F8-DP-2020-Valenta-Adam-thesis.pdf

https://practicaldatascience.co.uk/machine-learning/how-to-use-the-isolation-forest-model-for-outlier-detection

https://subscription.packtpub.com/book/security/9781789614671/1/ch01lvl1sec10/anomaly-detection-with-isolation-forest

**Note**: - The matter which we wrote is what we have understood from the multiple articles Feel Free to change the matter if its wrong.