How Isolation Forest works ?

How it works ?

The working mechanism of Isolation Forest can be described as follows:

1. **Recursive Partitioning :** This is the core process of building a single isolation tree. The algorithm repeatedly and randomly selects a feature and a split value to partition the data. A data point is sent to the left or right branch based on whether its value is above or below the threshold.

2. **Isolation Tree Construction :** The recursive partitioning continues until every data point is isolated in its own node or a maximum depth is reached. Anomalies, being outliers, are typically isolated much faster and closer to the root of the tree, resulting in a shorter path length.

3. **Ensemble Creation :** This step involves repeating the first two steps to create an "ensemble" of multiple isolation trees, which collectively form the "forest." Using multiple trees helps to make the results more robust and reliable.

4. **Anomaly Score Calculation :** In this final step, the algorithm calculates an **anomaly score** for each data point. This score is based on the average path length it took to isolate the point across all the trees in the forest. Points with a **shorter average path length** have a **higher anomaly score**, indicating they are more likely to be an anomaly.

Table structure of the reference dataset

	name	books	tv_shows	video_games
0	Aaliyah	0.5	4.6	4.9
1	Abigail	0.0	4.5	4.8
2	Addison	0.5	4.5	5.0
3	Adeline	3.5	4.5	6.6
4	Alana	2.8	3.8	5.6
5	Alexander	5.8	4.6	6.9
6	Alivia	4.2	4.5	6.7
7	Amara	3.2	4.5	5.6
8	Amelia	0.0	4.6	4.9
9	Annabelle	4.0	4.1	6.0

1. name : student name
2. books : time spend reading books weekly
3.tv_shows : time spend watching tv shows weekly
4. video_games : time spend playing video games weekly

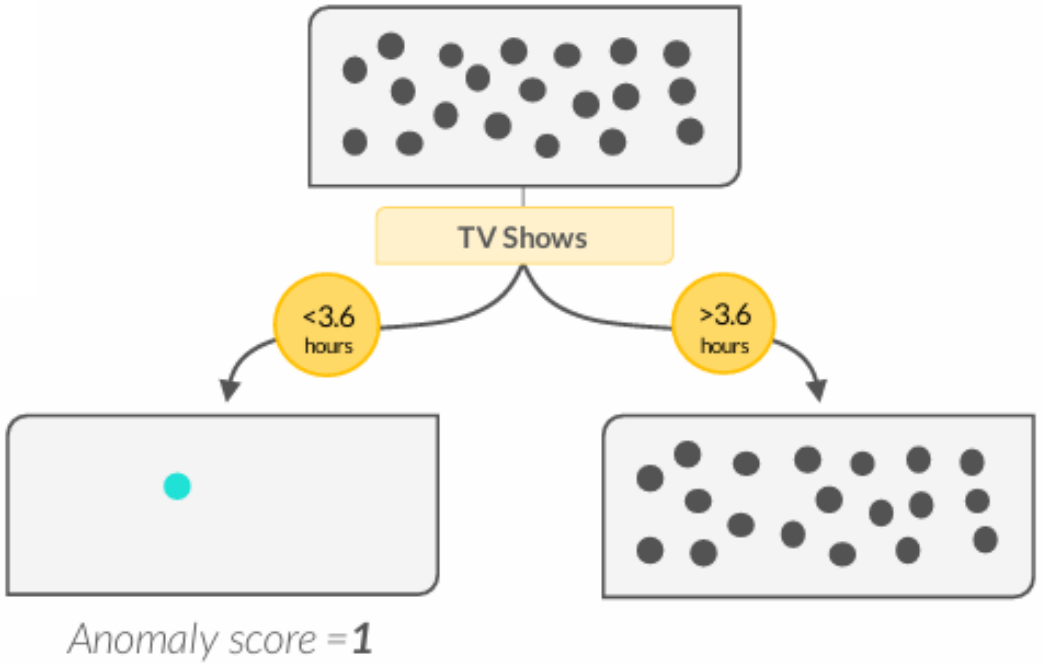
Note : we are visualizing top 10 rows of the wider dataset

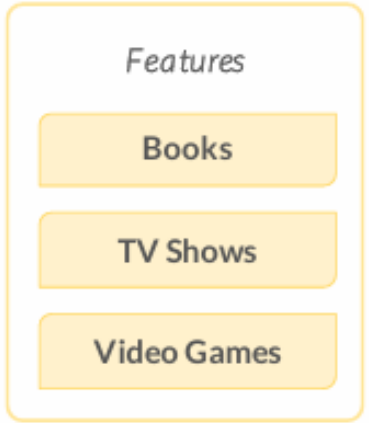
Step 1 - Recursive Partitioning

Step 1 - Recursive Partitioning

This is the core process of building a single isolation tree. The algorithm repeatedly and randomly selects a feature and a split value to partition the data. A data point is sent to the left or right branch based on whether its value is above or below the threshold.

Step 1 - Randomly select a feature from the data and split the observations using a random threshold





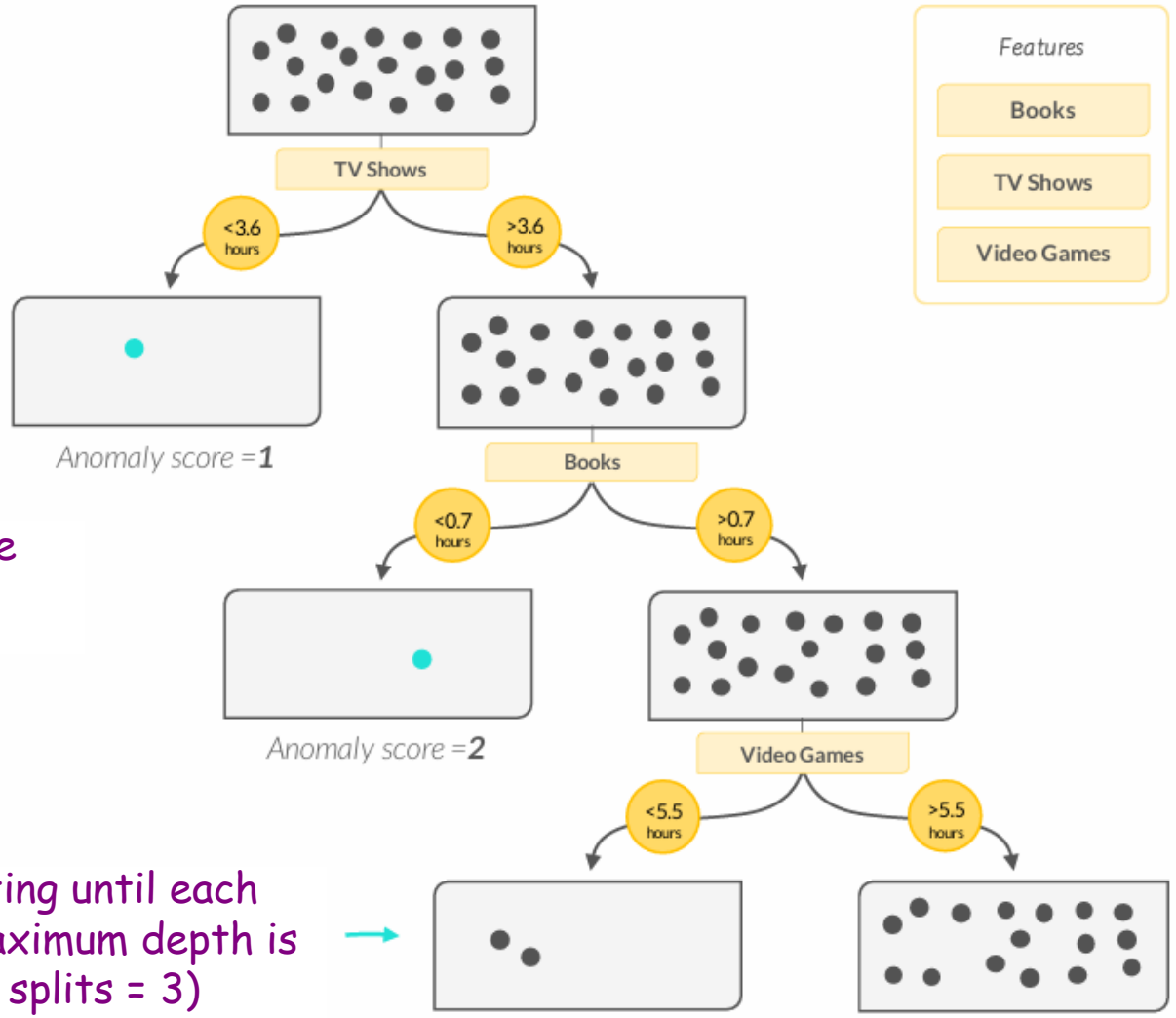
Step 2 – Isolation Tree Construction

Step 2 – Isolation Tree Construction

The recursive partitioning continues until every data point is isolated in its own node or a maximum depth is reached. Anomalies, being outliers, are typically isolated much faster and closer to the root of the tree, resulting in a shorter path length.

Step 2 - Continue splitting using random features and thresholds until every data point is isolated or a max depth is reached

These two students are likely anomalies



These would keep splitting until each point is isolated or a maximum depth is reached (i.e. number of splits = 3)

Step 3 – Ensemble Creation

Step 3 – Ensemble Creation

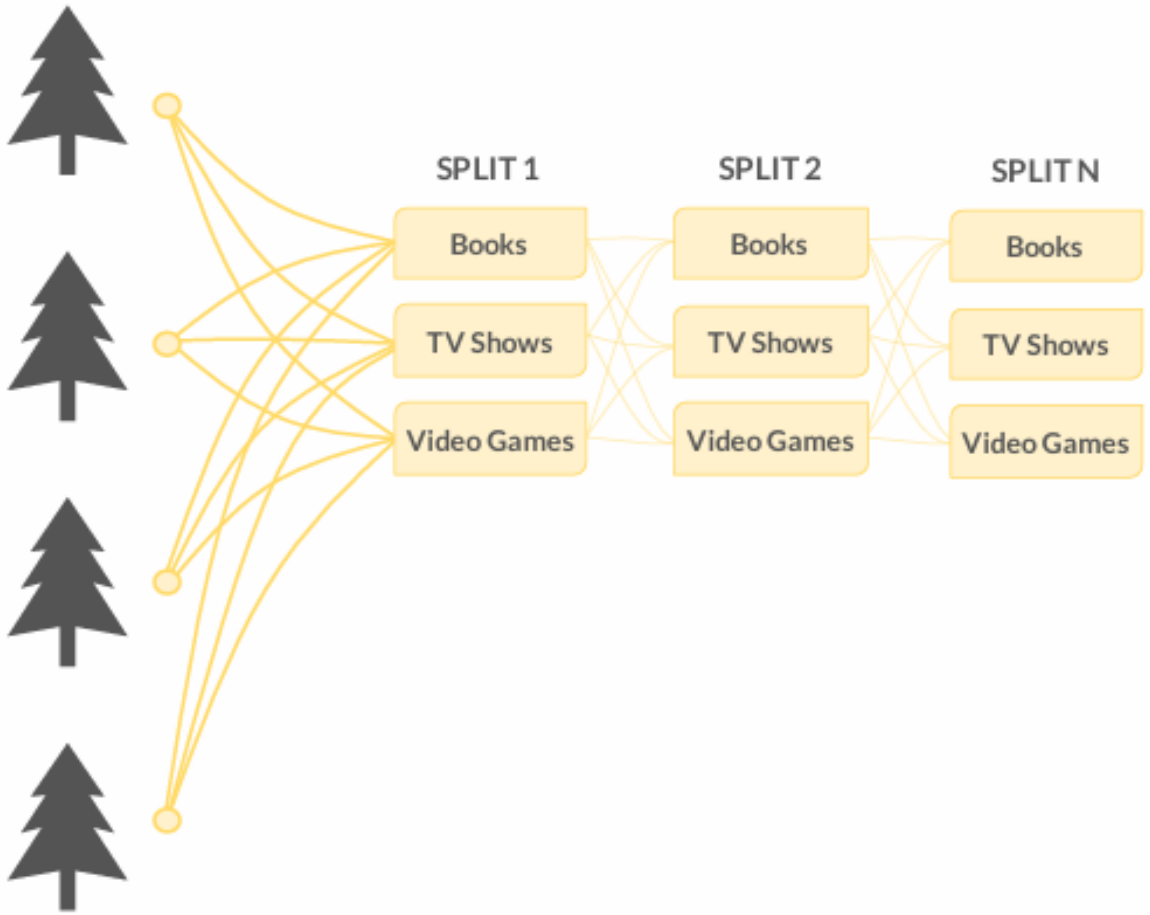
This step involves repeating the first two steps to create an "ensemble" of multiple isolation trees, which collectively form the "forest." Using multiple trees helps to make the results more robust and reliable.

Step 3 - Repeat using different features and splits to create multiple trees

Step 3 - Repeat using different features and splits to create multiple trees

Step 3 - Repeat using different features and splits to create multiple trees

Step 3 - Repeat using different features and splits to create multiple trees

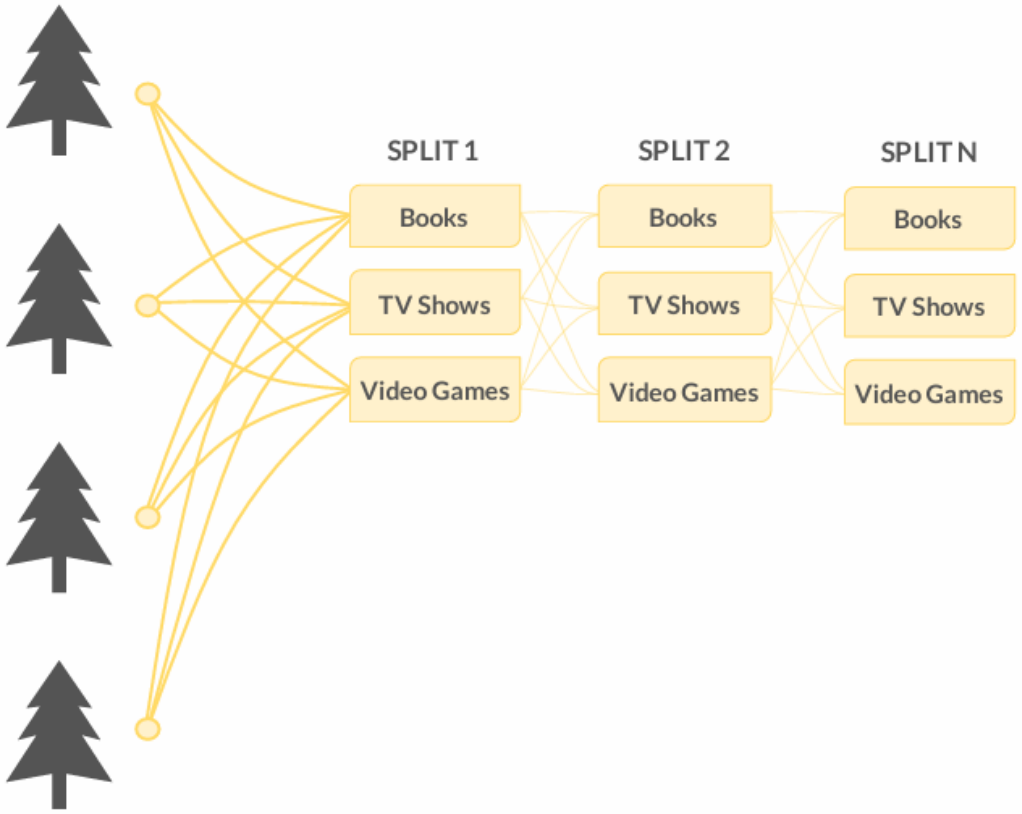


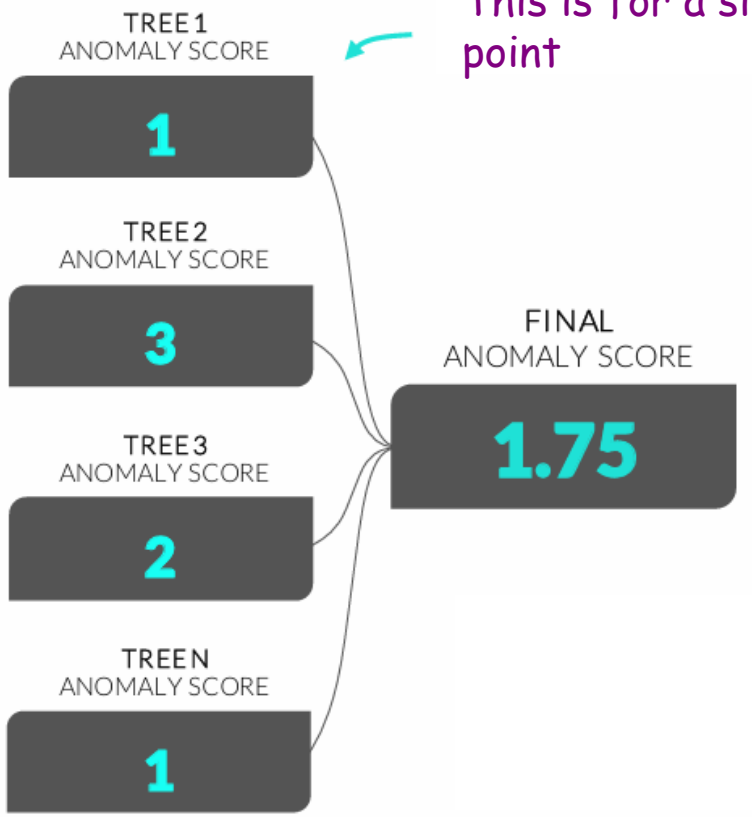
Step 4 - Anomaly Score Calculation

Step 4 - Anomaly Score Calculation

In this final step, the algorithm calculates an **anomaly score** for each data point. This score is based on the average path length it took to isolate the point across all the trees in the forest. Points with a **shorter average path length** have a **higher anomaly score**, indicating they are more likely to be an anomaly.

Step 4 - Calculate the anomaly score for each observation by averaging the number of splits it took to isolate (path length) in each tree





This is for a single data point

Since the score is so low, this value is likely an anomaly!

LET’S GO