ANOMALY DETECTION

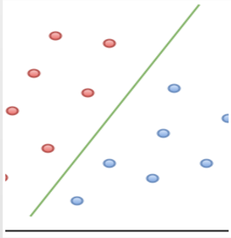
Rodwel Mupambirei
Consulting Actuary

PyData Bristol
18 July 2019

OUTLINE

- Why I care about anomalies in my models.
- What methods are currently used to detect anomalies – and their effectiveness (or lack of).
- Deep dive into a method that actually works.
- Incorporating anomaly detection in pipelines

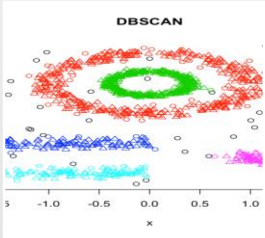
ANOMALY DETECTION METHODS



Model Based

SVM

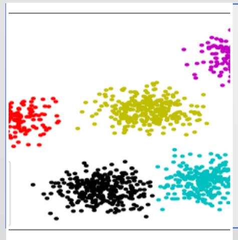
Statistical



Density Based

DBSCAN

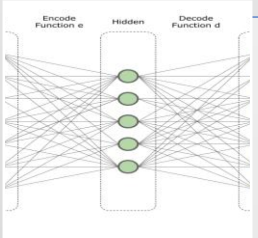
LOF



Distance based

K-means

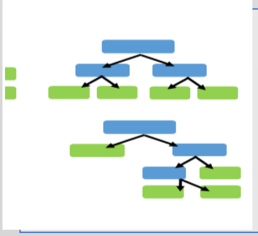
KNN



Network based

Autoencoder

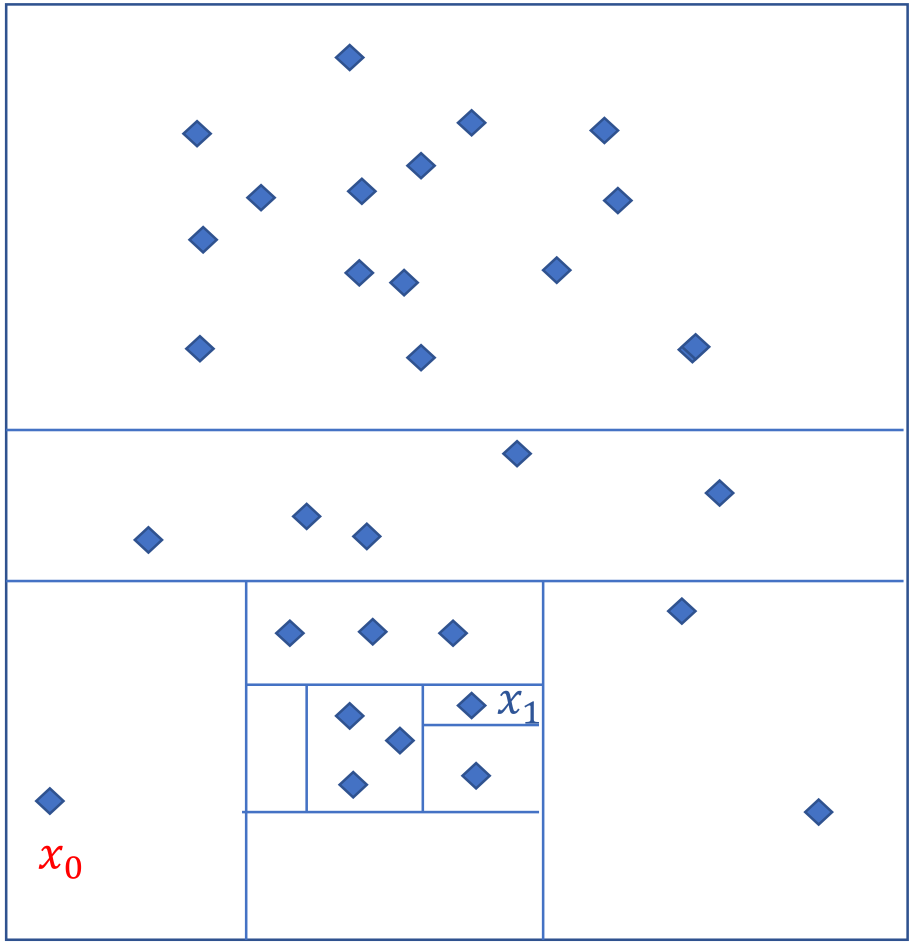
Rules Based

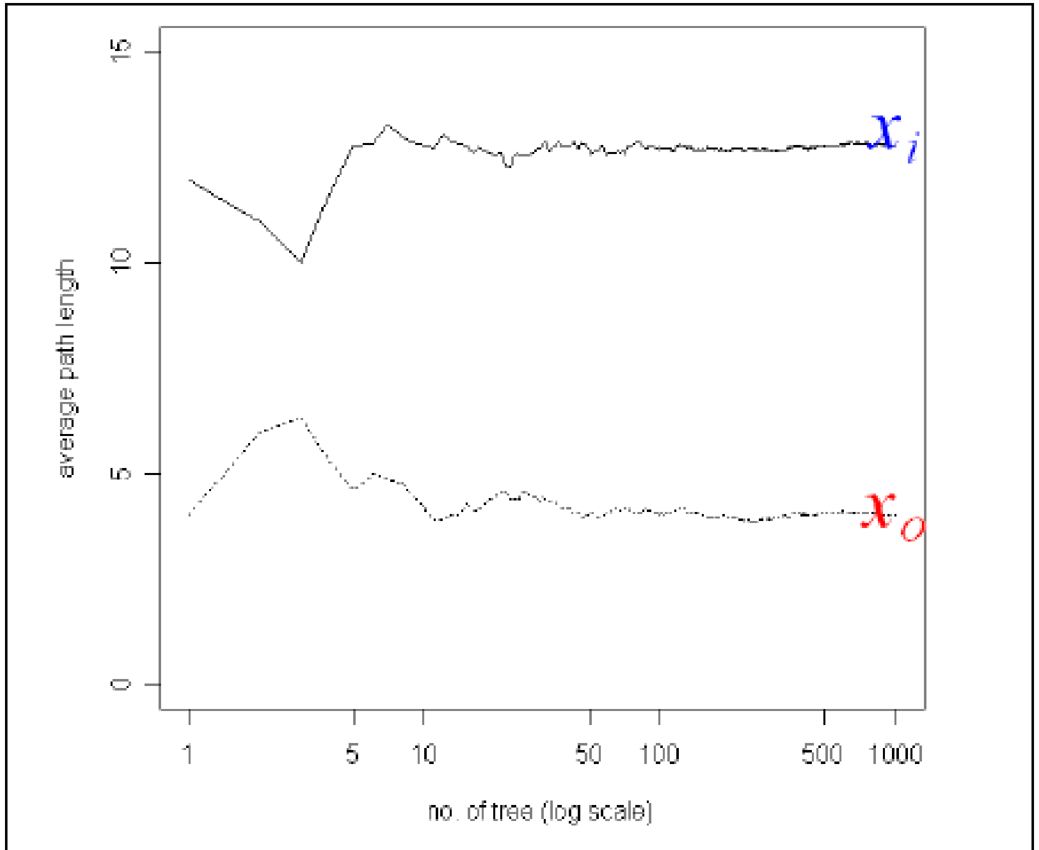


Trees

Liu

GROWING ISOLATION FORESTS



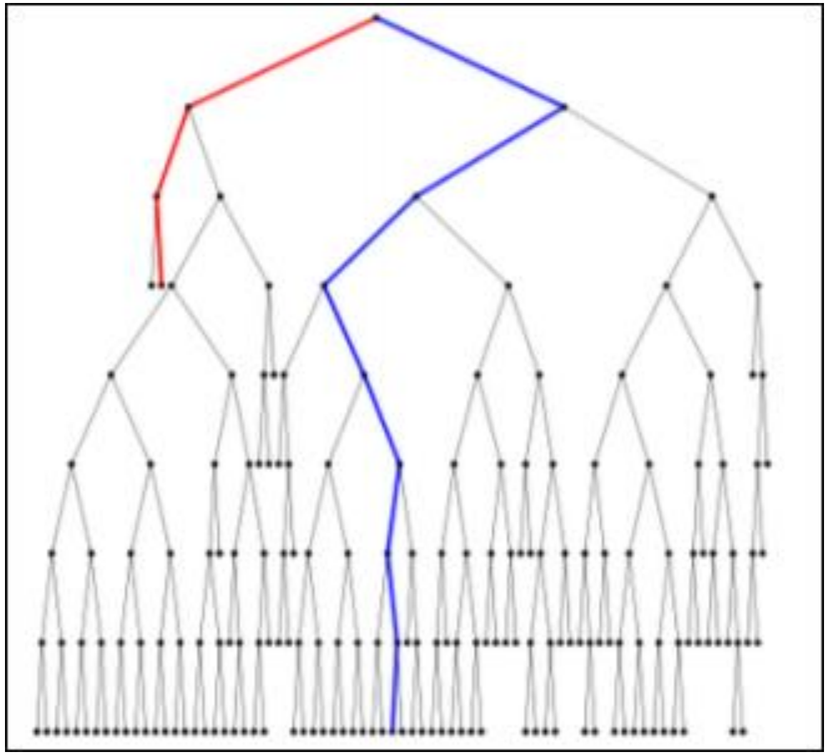


(c) Average path lengths converge

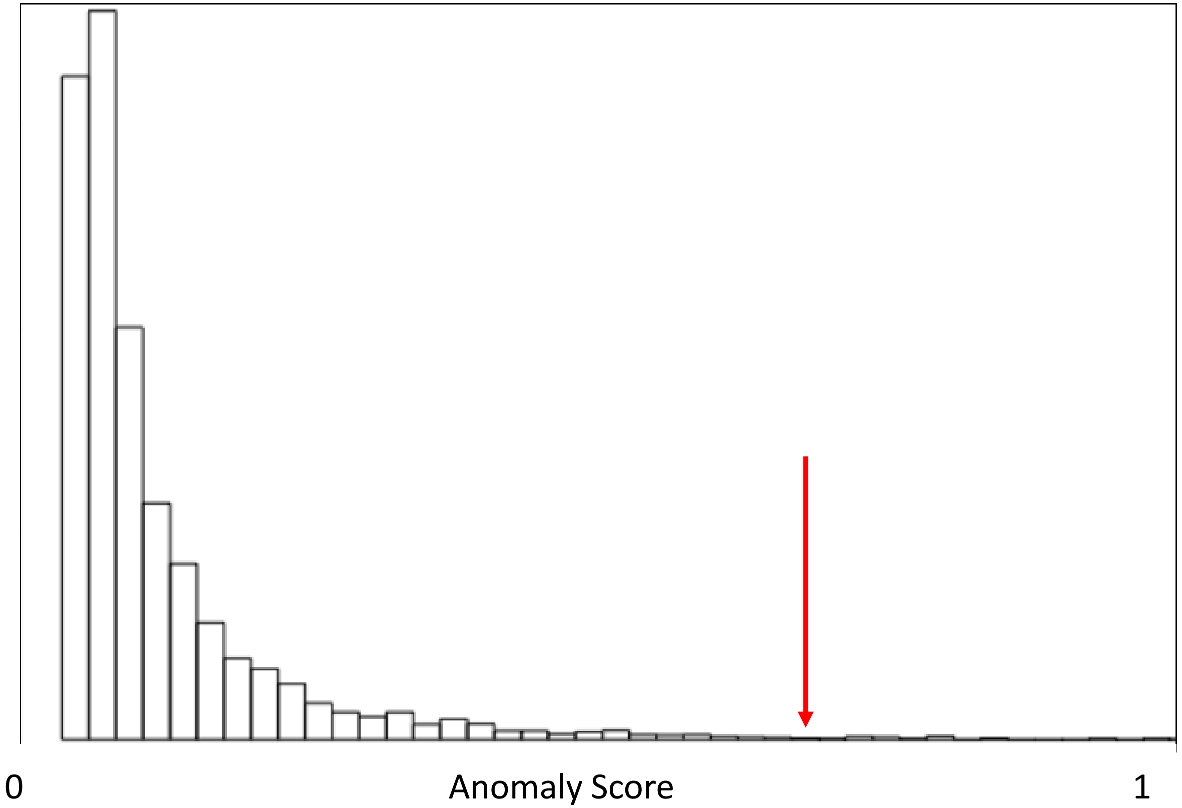
Liu, F. T., Ting, K. M., & Zhou, Z-H. (2012). Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data*, *6*(1), 1 - 39. https://doi.org/10.1145/2133360.2133363

PREDICTION ISOLATION PROCESS

Anomaly Score Distribution

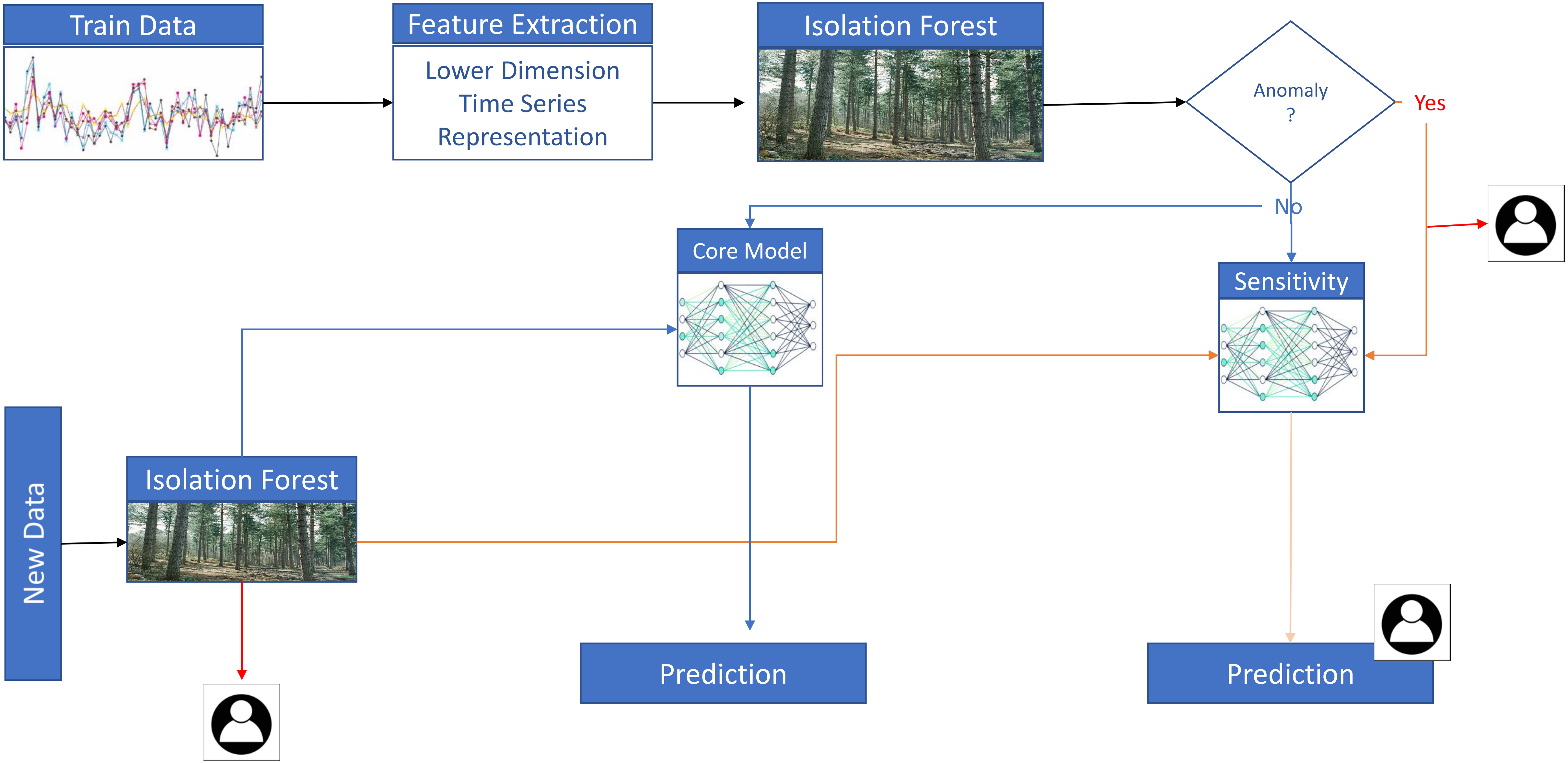


Anomaly Score Distribution



Sahand Hariri and Matias Carrasco Kind Extended Isolation Forest for Anomaly Detection
https://github.com/sahandha/eif

PIPELINE



Train Data

Feature Extraction

Lower Dimension Time Series Representation

Isolation Forest

Anomaly ?

Yes

No

Core Model

Sensitivity

New Data

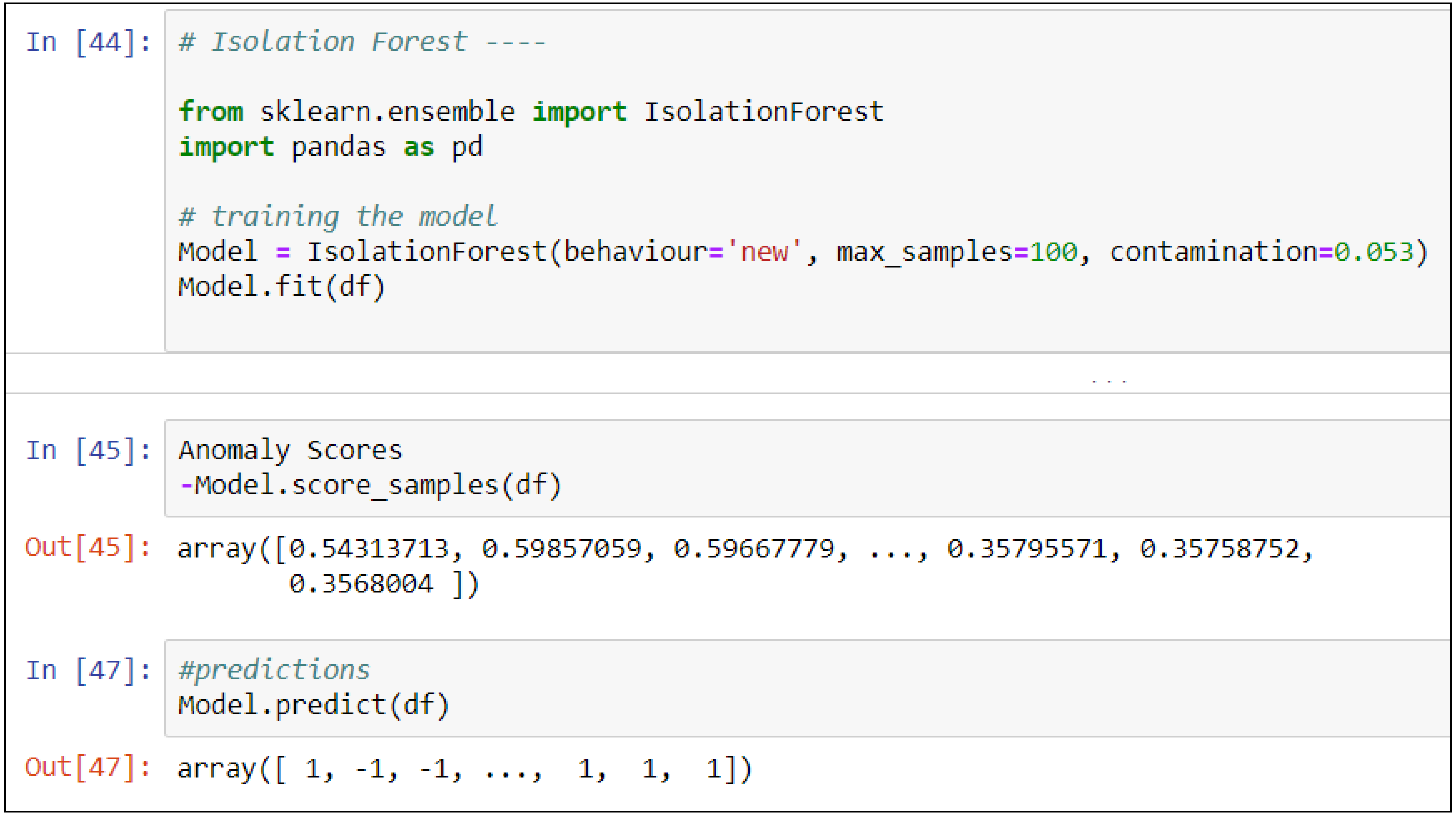
Isolation Forest

Prediction

Prediction

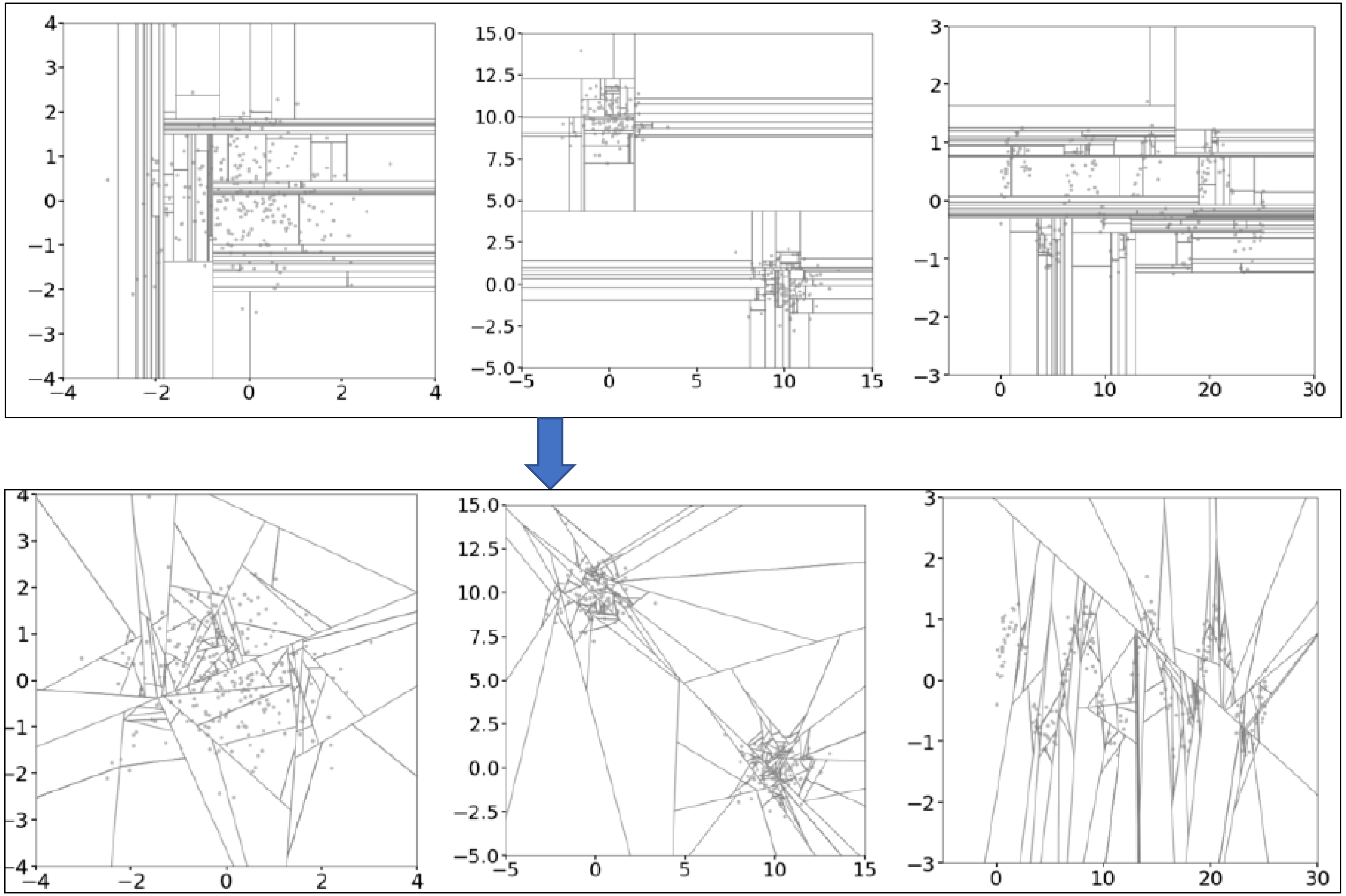
IMPLEMENTATION

scikit-learn implementation of Isolation Forest



```
In [44]: # Isolation Forest ----

         from sklearn.ensemble import IsolationForest
         import pandas as pd

         # training the model
         Model = IsolationForest(behaviour='new', max_samples=100, contamination=0.053)
         Model.fit(df)

                                                    ...

In [45]: Anomaly Scores
         -Model.score_samples(df)

Out[45]: array([0.54313713, 0.59857059, 0.59667779, ..., 0.35795571, 0.35758752,
                0.3568004 ])

In [47]: #predictions
         Model.predict(df)

Out[47]: array([ 1, -1, -1, ...,  1,  1,  1])
```

EXTENDED ISOLATION FOREST (eIF)



IF splits the data using a single feature at each split

EIF uses vectors rather that a single feature in the partitioning

Sahand Hariri and Matias Carrasco Kind Extended Isolation Forest for Anomaly Detection
https://github.com/sahandha/eif

PIPELINE

I have hopefully convinced you that:

You should be using isolation forests to make your models more robust

Isolation Forest are an efficient method for identifying anomalies

Isolation Forest based anomaly detection can be integrated in Machine Leaning Pipelines