Anomaly Detection on the Hypothyroidism Dataset

Daniele Montagnani

21 January 2024

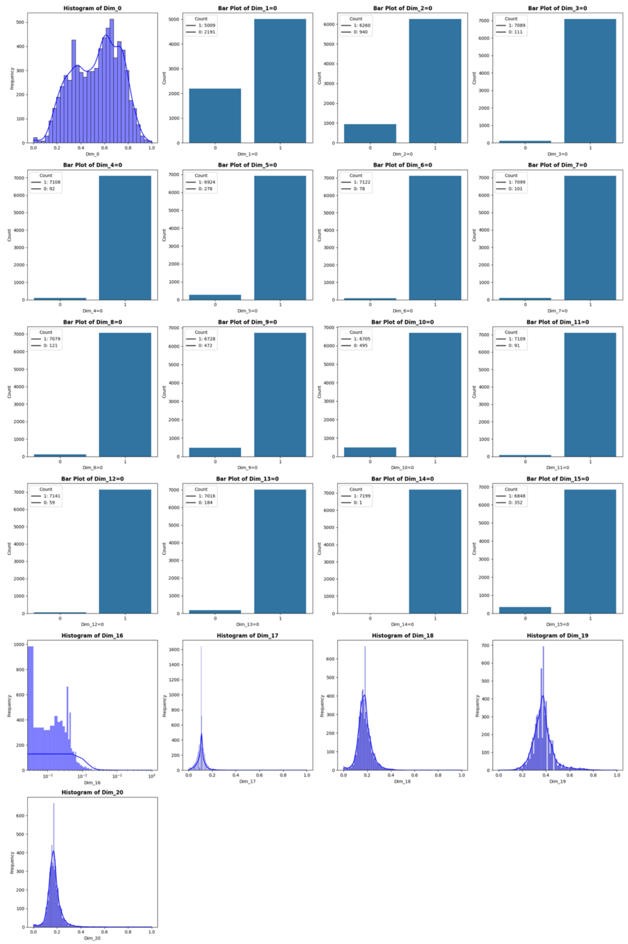
Introduction

Aim of the project: perform an anomaly detection analysis on the Hypothyroidism dataset

Completed Tasks:

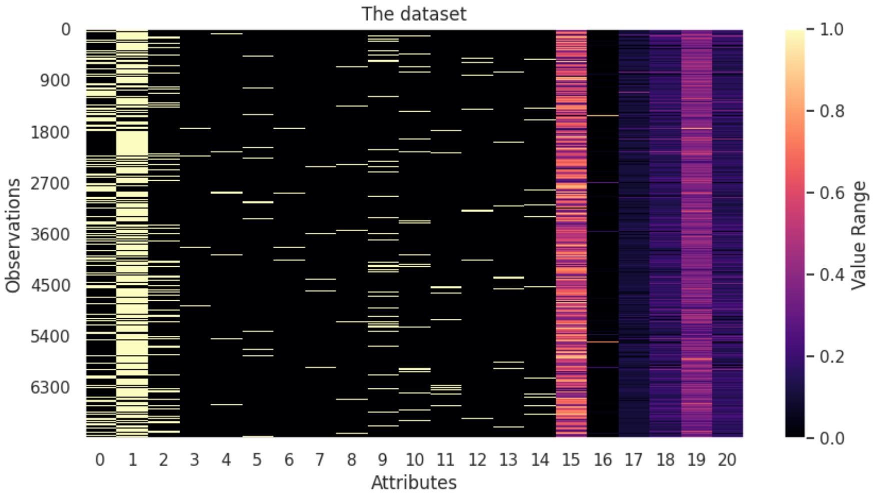
- Dataset Exploration and Pre-processing
- Implementation of a Custom Distance
- Clustering Analysis
- Anomaly Detection Analysis
- Final Decision on Outliers

Data Pre-processing



Dataset Exploration → Considerations → Actions

- 7200 observations across 21 dimensions (15 binary, 6 continuous)
- Data scaled, missing data had already been imputed
- Sparse binary variables (above 20% unbalance) → swap values
- One categorical variable remaining → obtain a one-hot encoding
- 14th binary variable highly uninformative → drop it



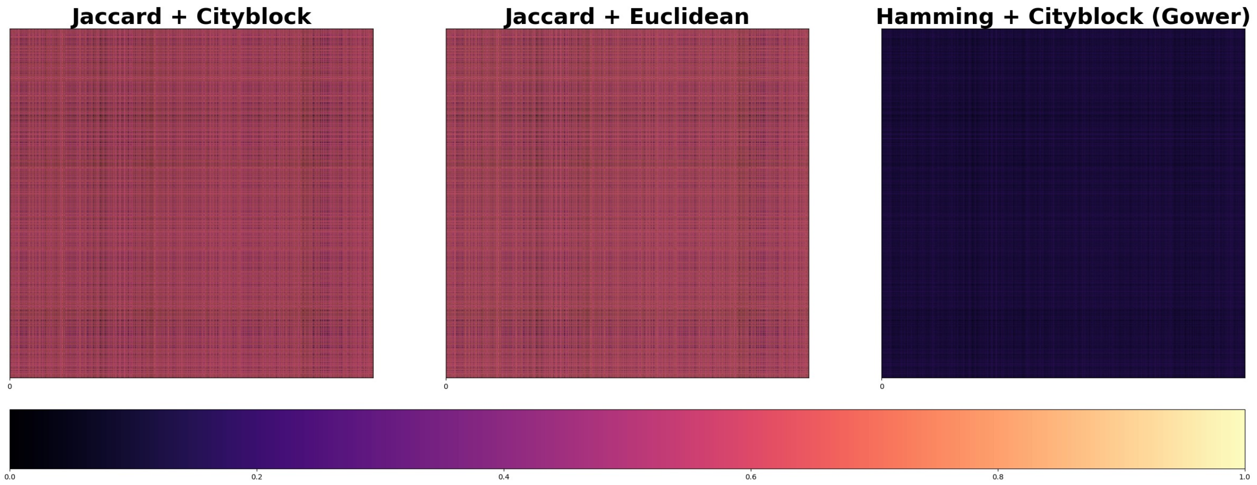
A Custom Distance

Mixed Data → Gower Distance
However, sparse data → Jaccard Coefficient

One-hot encoded categorical → Jaccard = Simple Matching Coefficient

Two options:

- Jaccard on Binary, Manhattan on Continuous
- Jaccard on Binary, Euclidean on Continuous



Jaccard + Cityblock

Jaccard + Euclidean

Hamming + Cityblock (Gower)

```
Jaccard + Cityblock
Mean: 0.44124299099318737
Standard Deviation: 0.2714828152660941
Min: 0.0
Max: 0.8712898421052632
25th, 50th, 75th Percentiles: [0.34613361 0.4763099  0.70427542]

Jaccard + Euclidean
Mean: 0.4539256086094625
Standard Deviation: 0.2716854661405964
Min: 0.0
Max: 0.9008808331453159
25th, 50th, 75th Percentiles: [0.34748397 0.48483049 0.71315514]

Hamming + Cityblock (Gower)
Mean: 0.105127404193917 49
Standard Deviation: 0.0631685032724041
Min: 0.0
Max: 0.46613354887218045
25th, 50th, 75th Percentiles: [0.0579326  0.1100686  0.15371565]
```

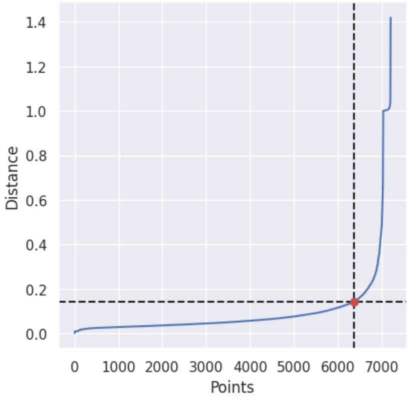
Clustering and Clustering Based Anomaly Detection

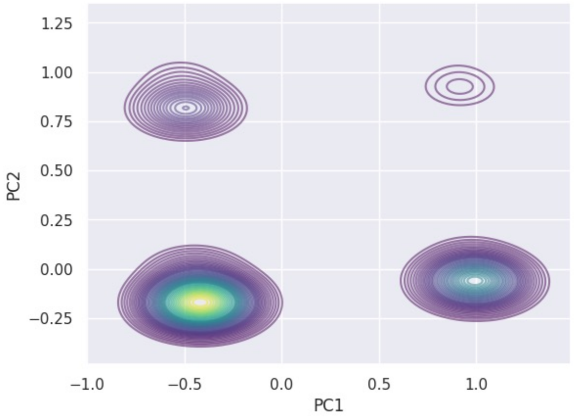
PCA

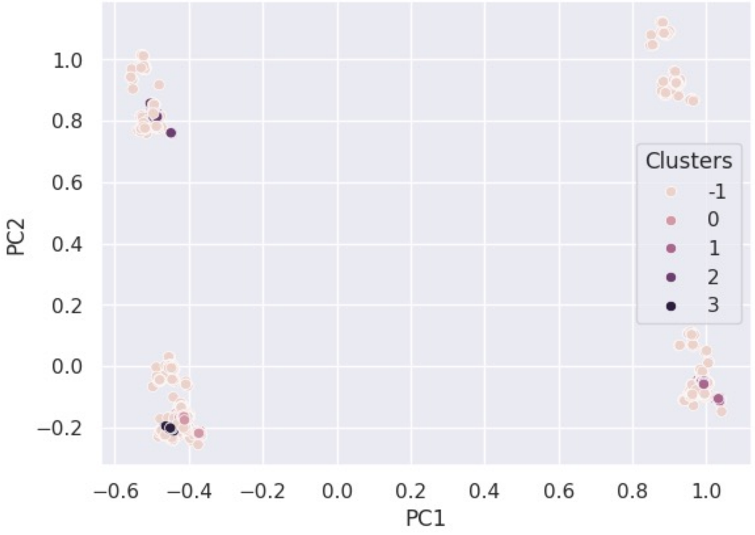
- To get an idea of possible clusters
- Used density estimation to identify small clusters

DBSCAN

- Elbow method → epsilon
- Highly aggressive outlier detection
- Given by a high *min_samples* parameter
- Silhouette score: 0.6429





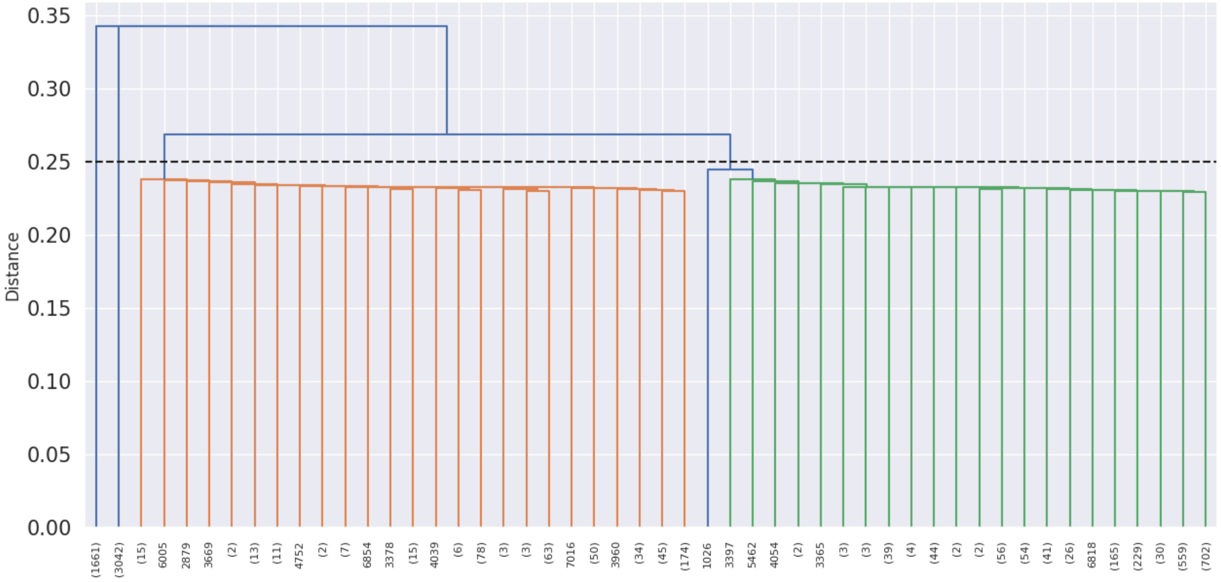


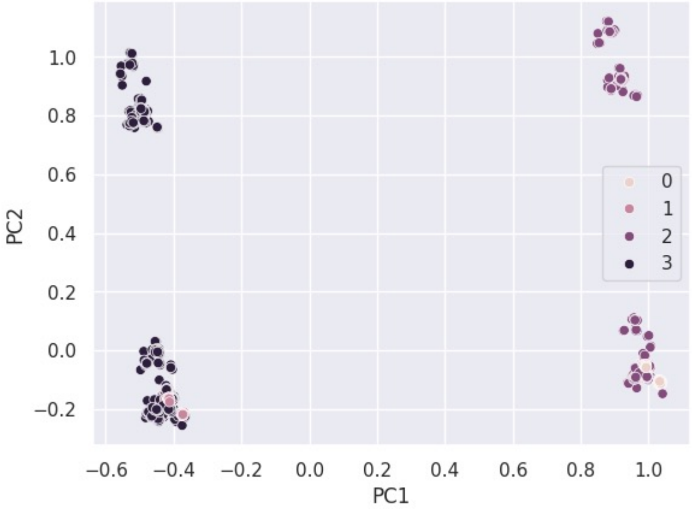
Cluster	Observations
-1	1674
0	3042
1	1661
2	559
3	264

Clustering and Clustering Based Anomaly Detection

Hierarchical

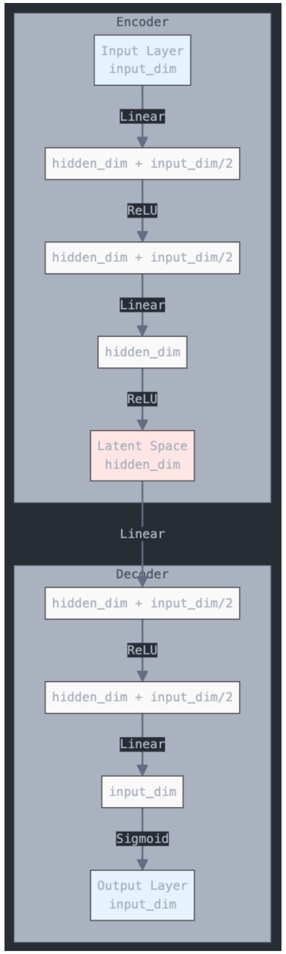
- Well separated → single linkage agglomeration
  - Cut to get 4 clusters
  - Silhouette score 0.5734





Cluster	Observations
0	1661
1	3042
2	530
3	1967

Anomaly Detection: Reconstruction Based



Encoder

Input Layer
input_dim

Linear

hidden_dim + input_dim/2

ReLU

hidden_dim + input_dim/2

Linear

hidden_dim

ReLU

Latent Space
hidden_dim

Linear

Decoder

hidden_dim + input_dim/2

ReLU

hidden_dim + input_dim/2

Linear

input_dim

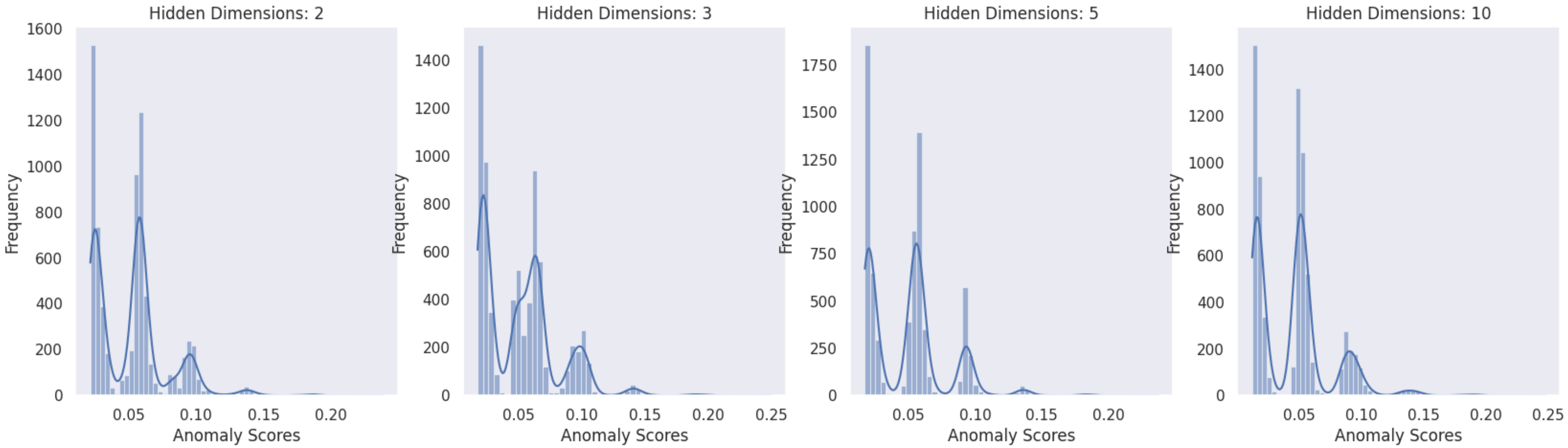
Sigmoid

Output Layer
input_dim

Autoencoder:

Train an encoder to learn a latent representation of the data
Train a decoder to reconstruct every observation from their latent representation
The model will invest its costrained resources to represent normal data
Therefore, the reconstruction error of every observation becomes a measure of its exceptionality

- Early stopping → regularization → more meaningful scores



Hidden Dimensions: 2

Hidden Dimensions: 3

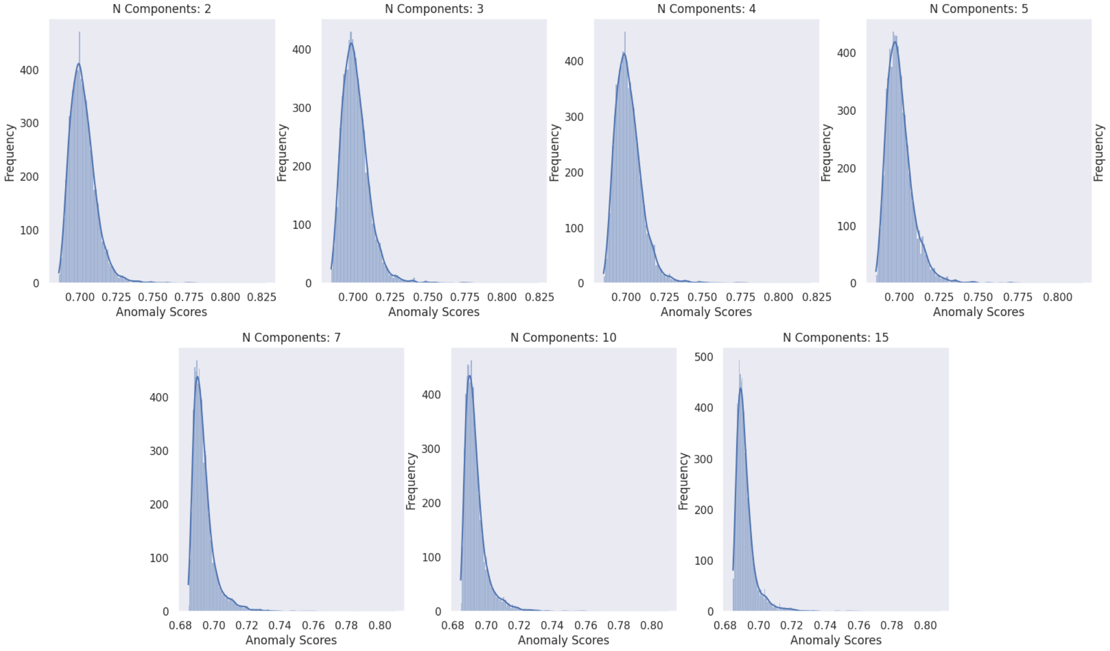
Hidden Dimensions: 5

Hidden Dimensions: 10

Anomaly Detection: Reconstruction Based

PCA

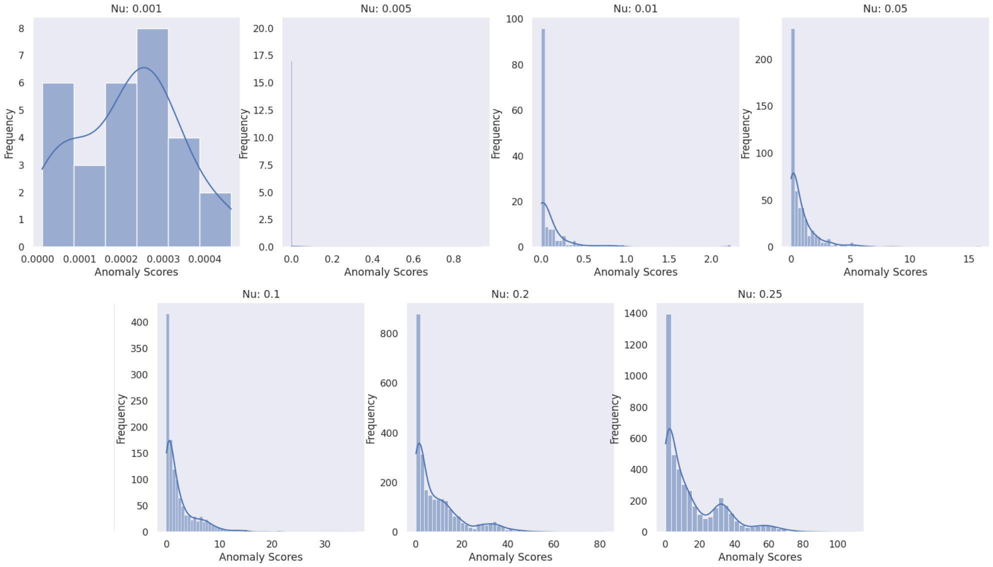
- Tried different number of components
- Resonally well-behaved scores
- Min-max scaled into [0,1] range



Anomaly Detection: “Isolation Based”

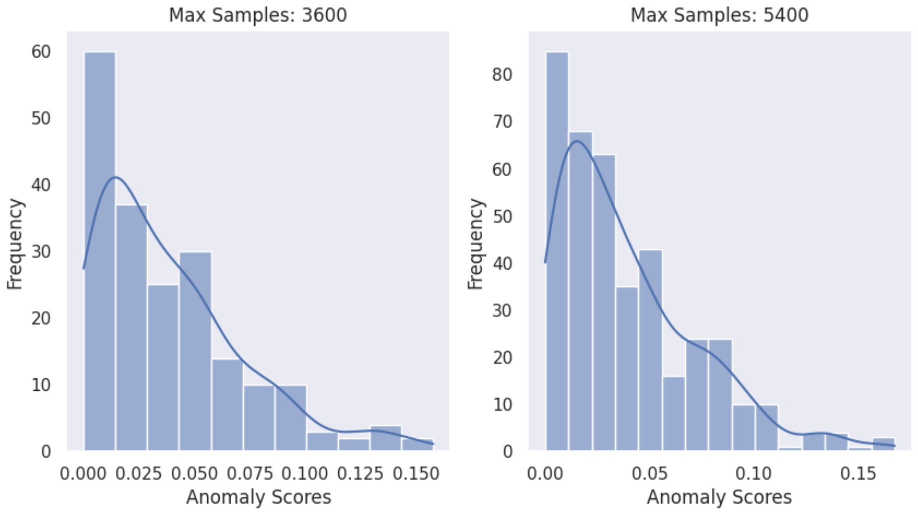
One class support vector machine

- Nu = expected proportion of outliers
- Nu + Score behavior suggestive on the real number of outliers
- Custom kernel
- Well behaved apart from nu = 0.001, 0.005, 0.25



Isolation Forest

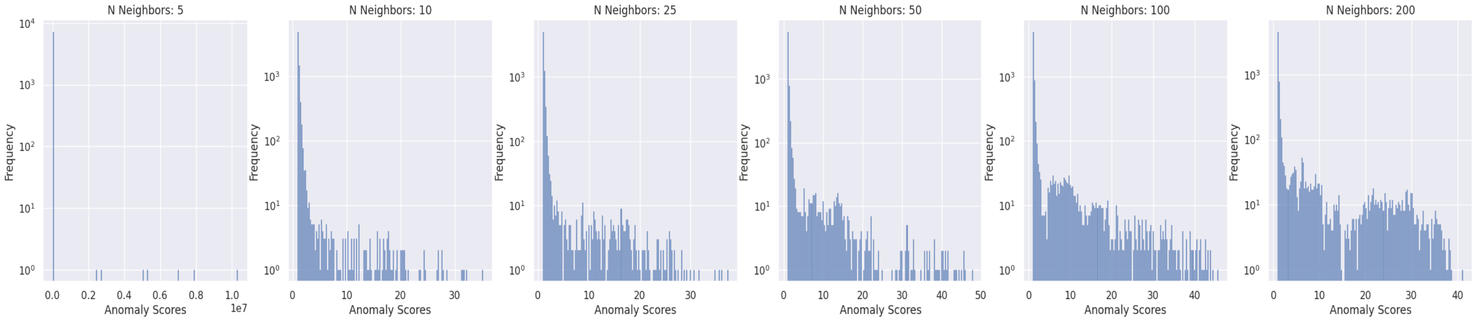
- Resonally well-behaved scores
- Did not explore many hyperparameters
- High number of max samples



Anomaly Detection: Proximity Model Free

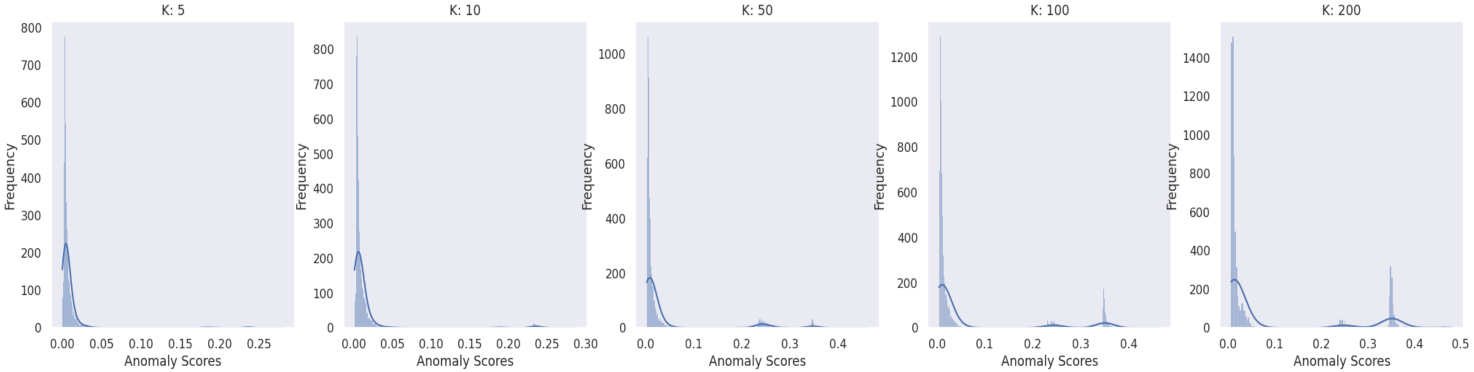
Kth Nearest Neighbor

- Resonally well-behaved scores (log scale)
- Ks = [5, 10, 25, 50, 100, 200]



Local Outlier Factor

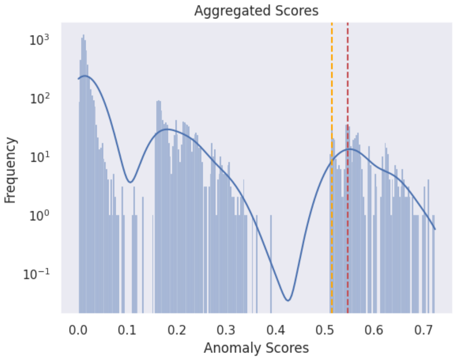
- Resonally well-behaved scores
- Ks = [5, 10, 25, 50, 100, 200]



Final Decision

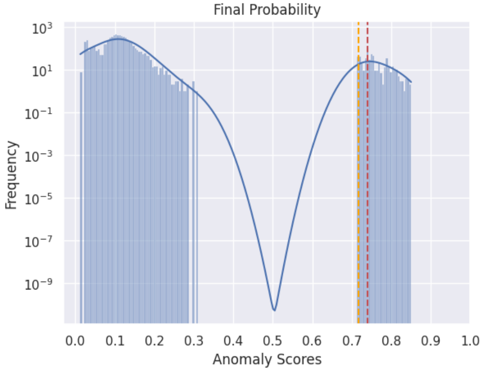
Final Scores

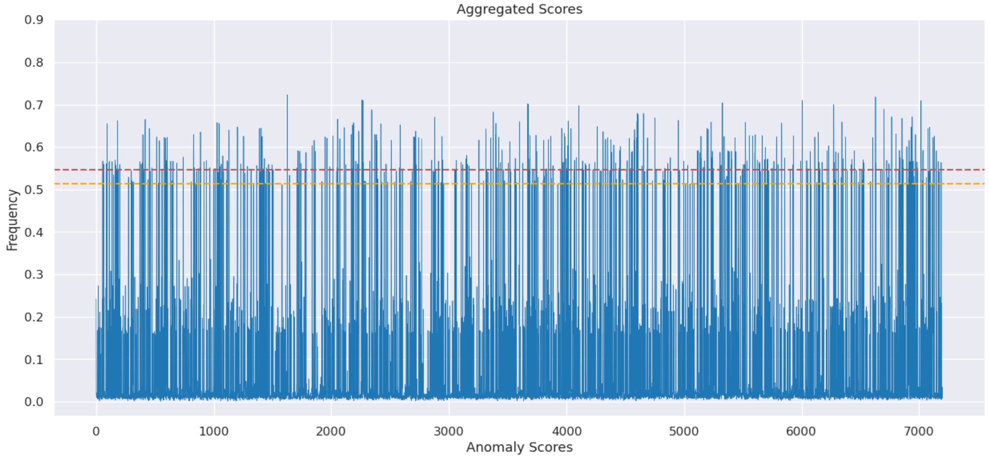
- Uniform aggregation at the single method level
- Weighted aggregation within a family of methods
- Final weighted aggregation

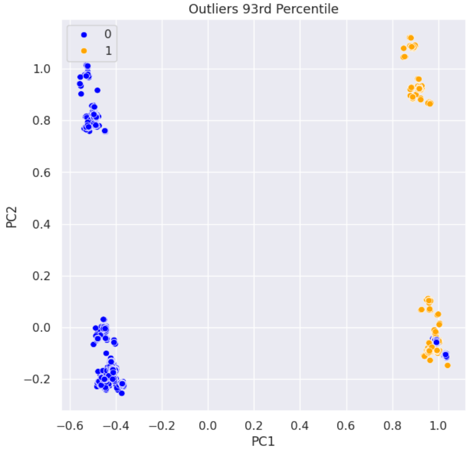


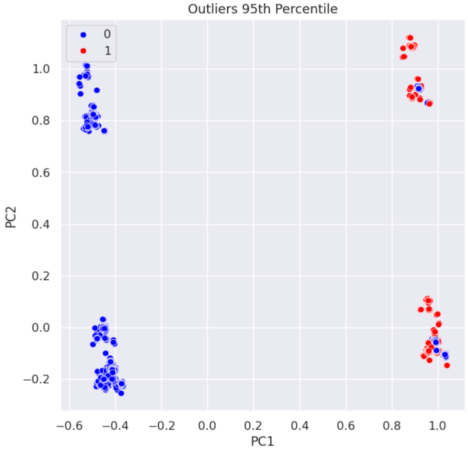
Final Decision

- Outliers: 5% or 7%
- Visual Inspection
- Avg aggregation → confidence artifacts
- Increase confidence with a stretch / contraction in appropriate ranges









Conclusion

In conclusion:

- Lots of degrees of freedom
- Outliers' decision highly influenced by the clustering decision

- In restrospect, would I reduce the degrees of freedom? No, quite the opposite
- I'd increase them: weighted distance, more methods
- But, I'd maintain a more neutral stance wrt methods

- Furthermore, I'd explore more ensambling methods
- And, be more surgical about it
- Logic → Ensambling method, for example, conjunction → Multiplication, Disjunction → Max
- Averaging is kind of a blunt middle ground solution

"there are two types of statisticians: those who know what decision problem they are solving and those who don't"