Anomaly Detection Project

Mixed Data-type Dataset

Aly Valiev

University of Milano-Bicocca

July 18, 2024

Outline

General Info About the Dataset

- The dataset contains 7200 objects with 21 attributes.
- Six continuous attributes.
- Fifteen binary attributes.
- Continuous values lie in the range from 0 to 1.
- There is no information about the attributes and how they are related to real life.
- For this project, we treat the data just as arrays of numbers and don't make conclusions about their meaning.

Continuous Attributes

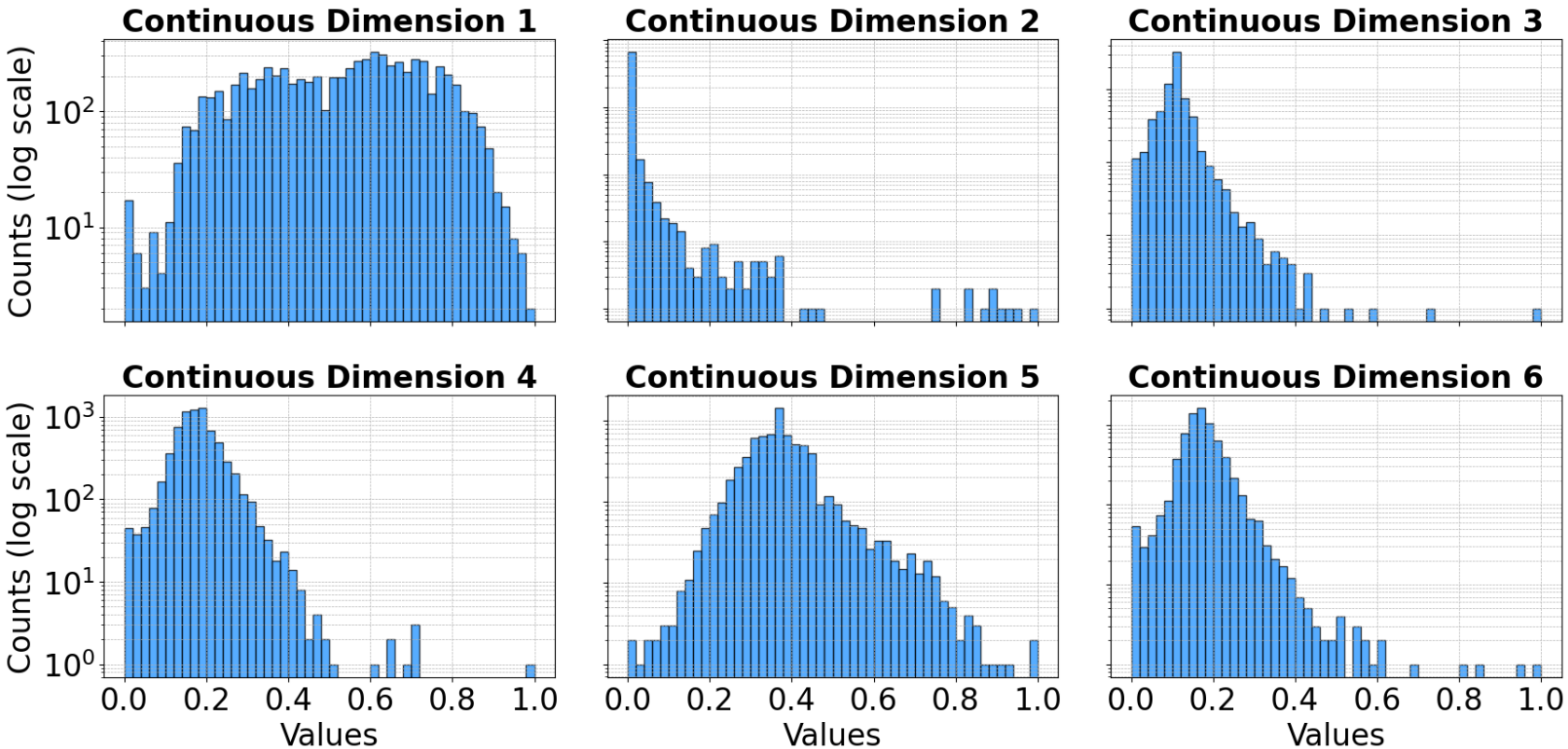


Figure: Distribution of Continuous values on a log scale.

Binary Attributes

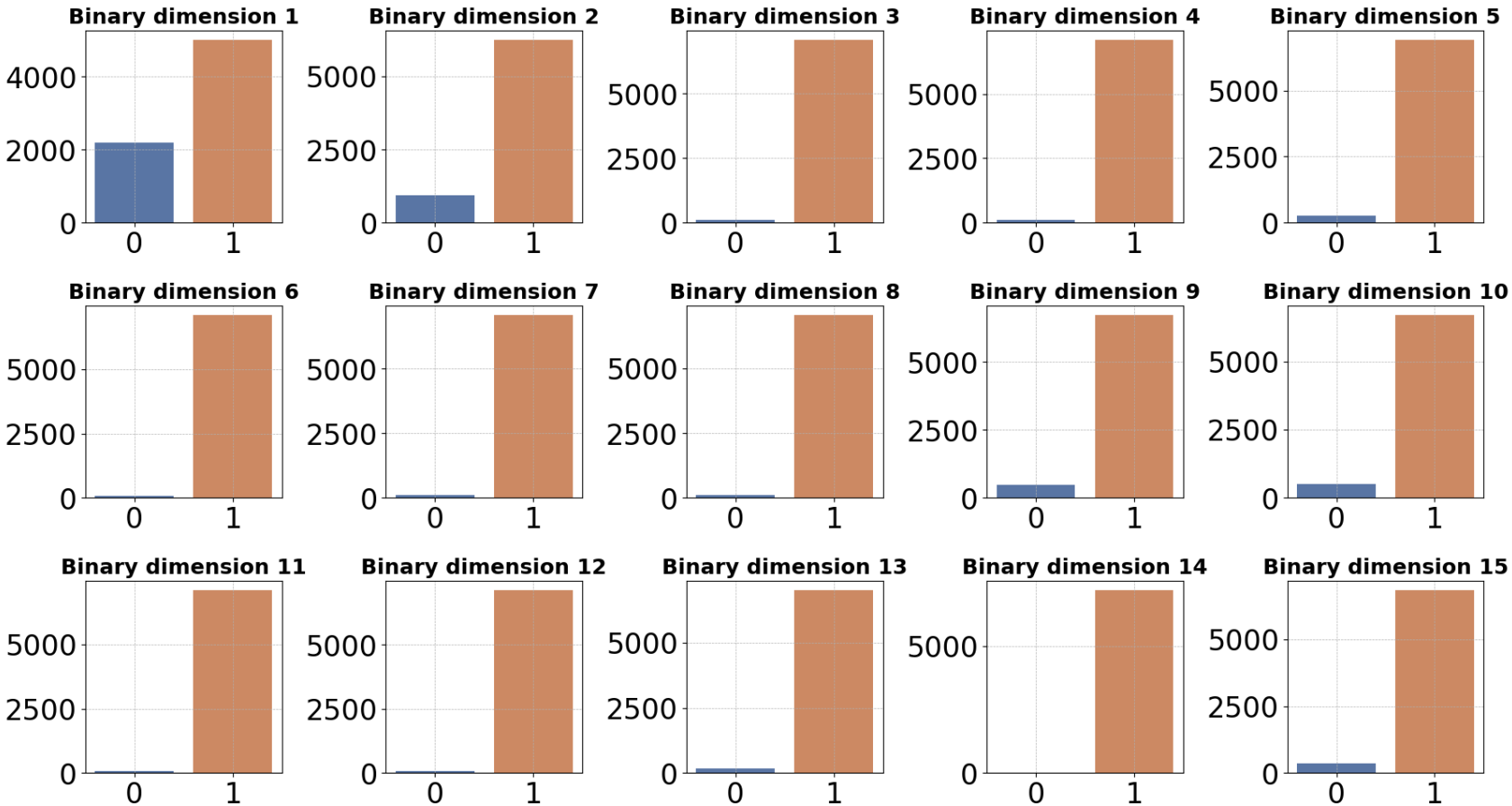


Figure: Distribution of binary values.

Difficulties

- Binary values are heavily imbalanced.
- For example, the binary dimension 14 has only one object whose value is 1 and 7199 zeros.
- Considering only the binary columns, there are only 135 unique objects.

Overall, the anomaly detection task on this dataset presents a significant challenge due to:

- The imbalanced binary values.
- The skewed distribution of continuous values.
- High dimensionality.
- Mixed data types.

Artificial Dataset for Model Testing

- 7200 objects with 6 continuous and 15 binary attributes.
- Continuous attributes follow various statistical distributions and are normalized to [0, 1].
- 5% of the objects (360) are turned into outliers.
  - Binary values are flipped.
  - Gaussian noise is added to continuous attributes.

Continuous Attributes (generated)

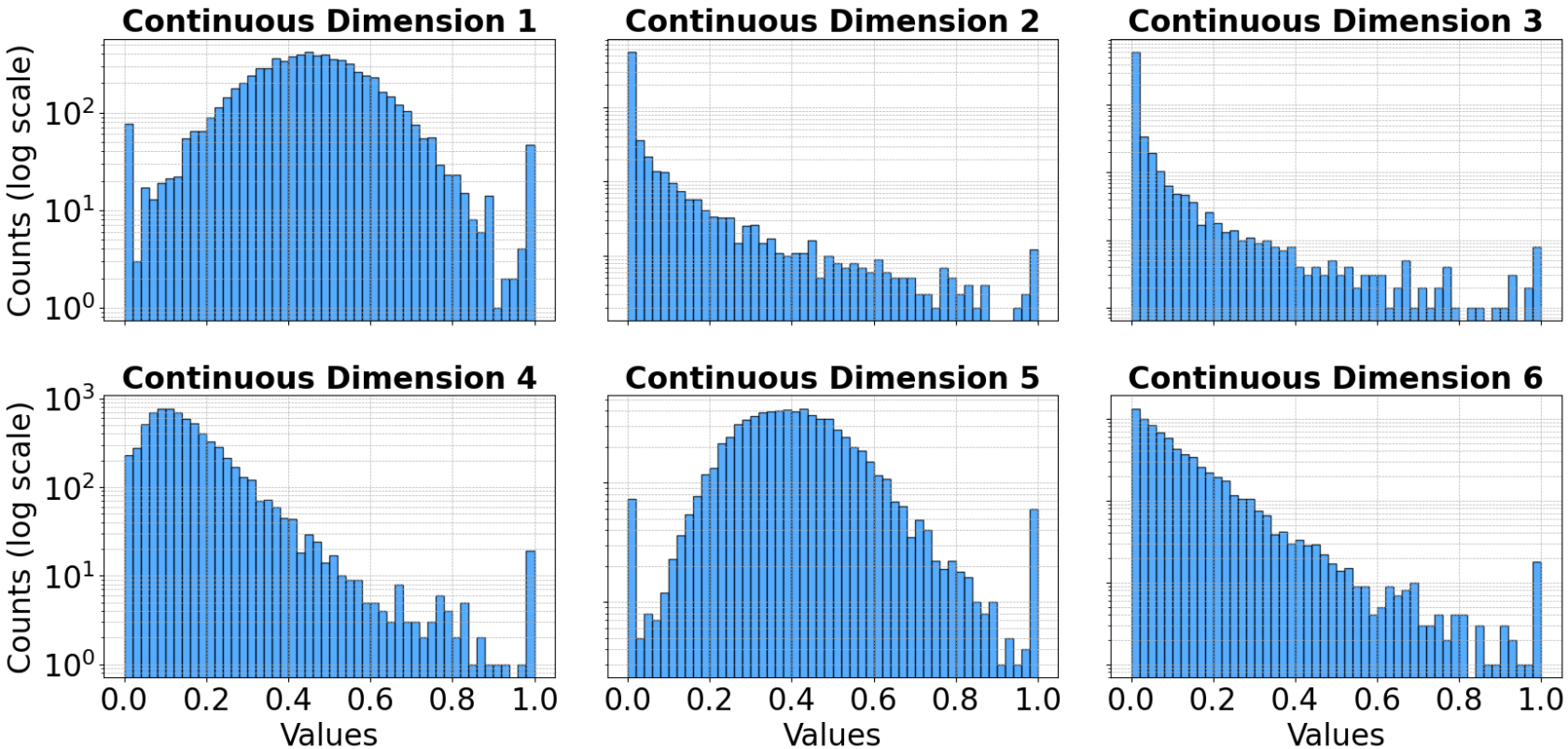


Figure: Distribution of Continuous values on a log scale.

Binary Attributes (generated)

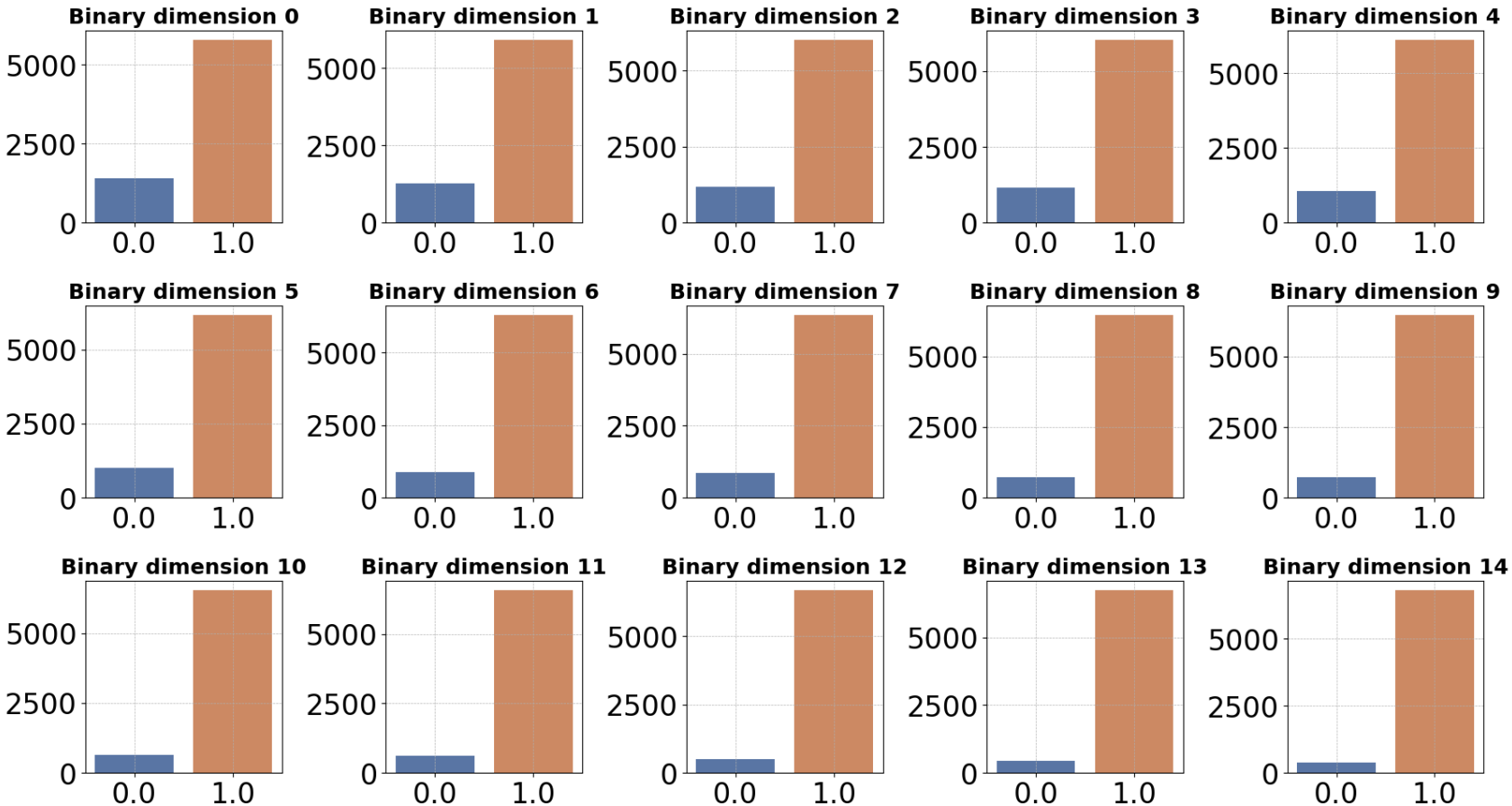
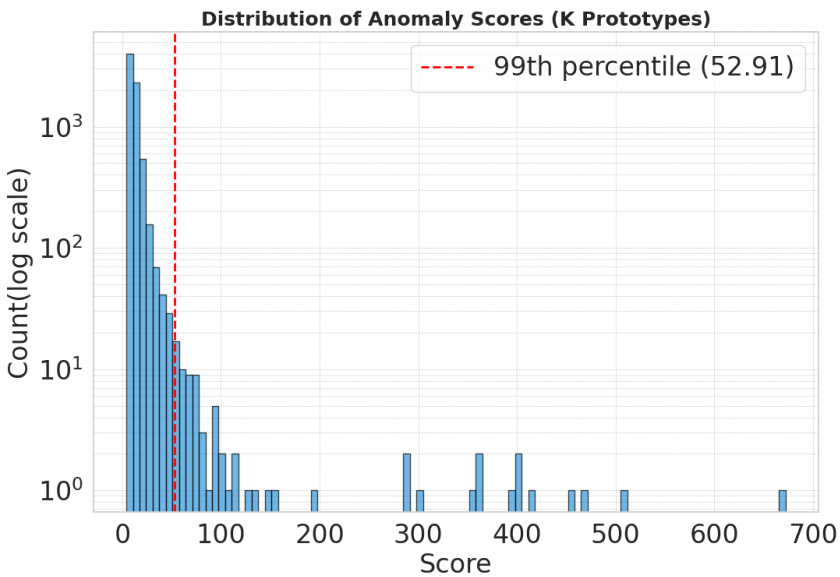


Figure: Distribution of binary values.

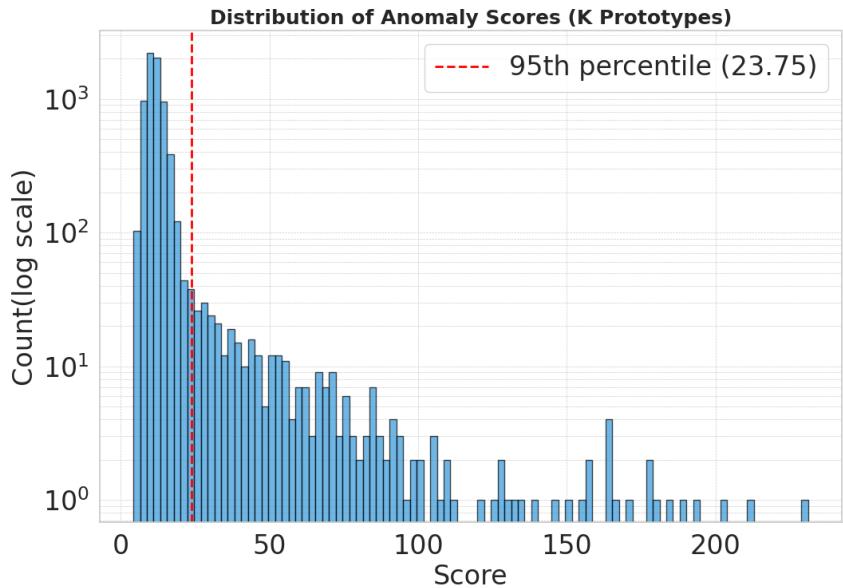
K-Prototypes

- K-Prototypes: combining K-means and K-modes for mixed-type data.
  - Euclidean distance for continuous attributes.
  - Matching distance for binary attributes.
  - Combined using weighted average.
- Normalizes continuous attributes.
- Forms specified number of clusters.
- Anomaly Detection:
  - Calculates distance of each point from its cluster centroid.
  - Points with significant deviation are detected as anomalies.

Distance distribution



Distance distribution (original dataset)



Distance distribution (generated dataset)

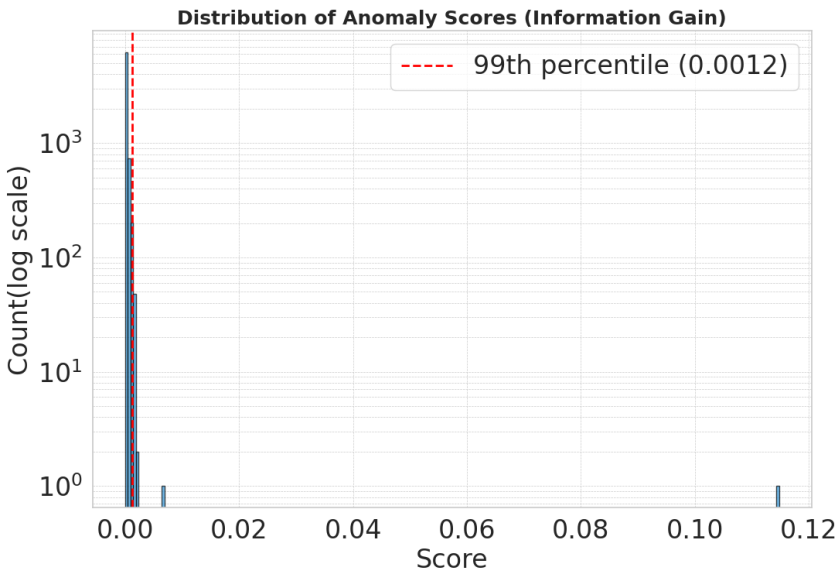
Information Gain Anomaly Detection

- Identifies anomalies based on information gain.
- Entropy Calculation:
  - Entropy is calculated for the entire dataset.
  - Separate computations for continuous and binary attributes.
  - Combined entropy is used for detection.
- Anomaly Detection:
  - Entropy of the dataset is recalculated after removing each data point.
  - Points causing high entropy changes are marked as anomalies.
  - The difference in entropy serves as the anomaly score.

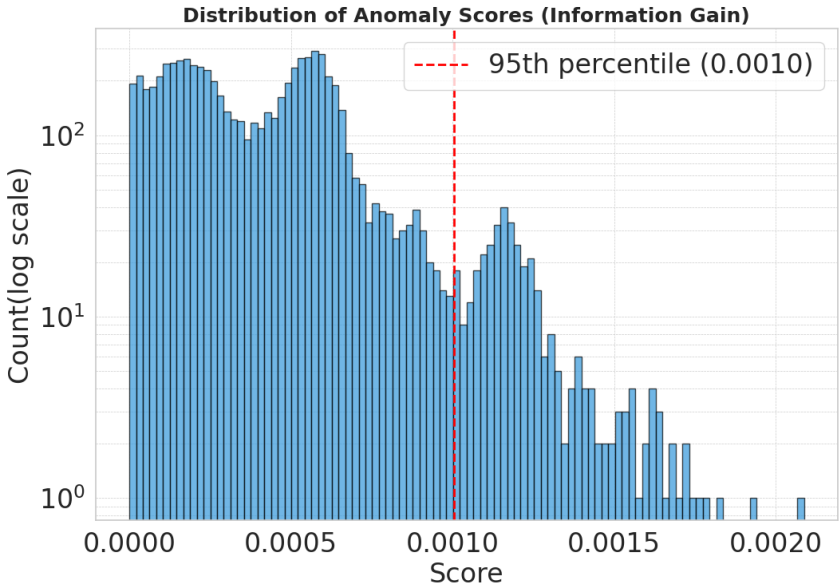
Drawback of Information Gain Anomaly Detection

- Main Drawback: Entropy is not well-defined for continuous variables.
- Approximation:
  - Continuous values are separated into bins to approximate differential entropy.
  - Differential entropy extends Shannon entropy to continuous variables by replacing the sum with an integral.
- Important Note:
  - Discrete and differential entropies are not equivalent and do not necessarily operate on the same scale.
  - This distinction must be considered when applying entropy-based methods to mixed data types.

Scores distribution



Anomaly scores distribution (original dataset)

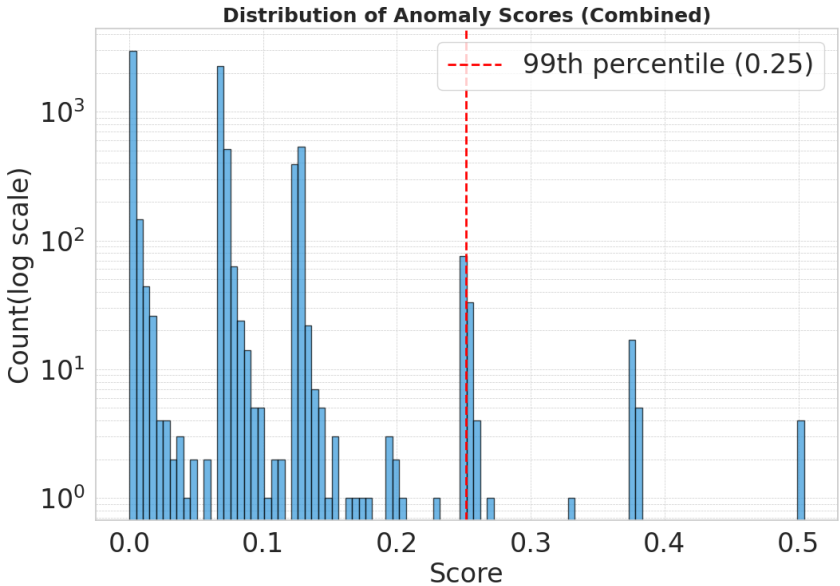


Anomaly scores distribution (generated dataset)

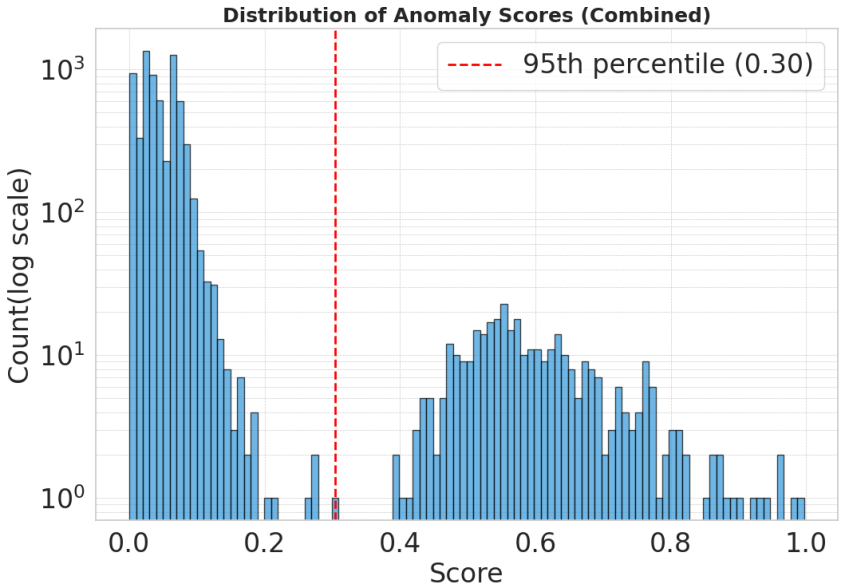
Hybrid Anomaly Detection Approach

- Motivation:
  - Clustering (K-Prototypes) is effective for continuous values.
  - Information gain approach is more natural for binary data.
- Hybrid Method:
  - Continuous Data:
    - Clustered using KMeans.
    - Distance from cluster centroid calculated as anomaly score.
  - Binary Data:
    - Information gain computed using entropy calculation.
    - Same idea as previous approach.
  - Combining Scores:
    - Scores normalized using Min-Max normalization.
    - Continuous and binary anomaly scores combined using weighted average.

Scores distribution



Anomaly scores distribution (original dataset)

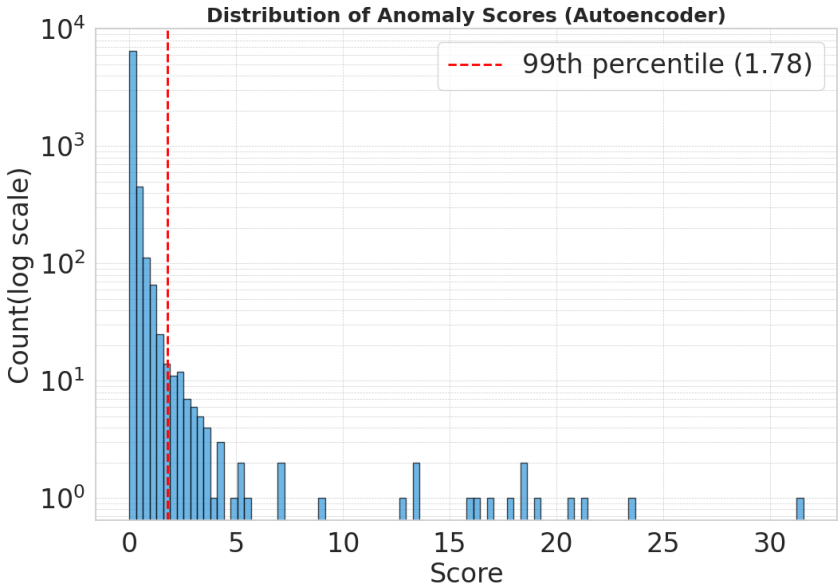


Anomaly scores distribution (generated dataset)

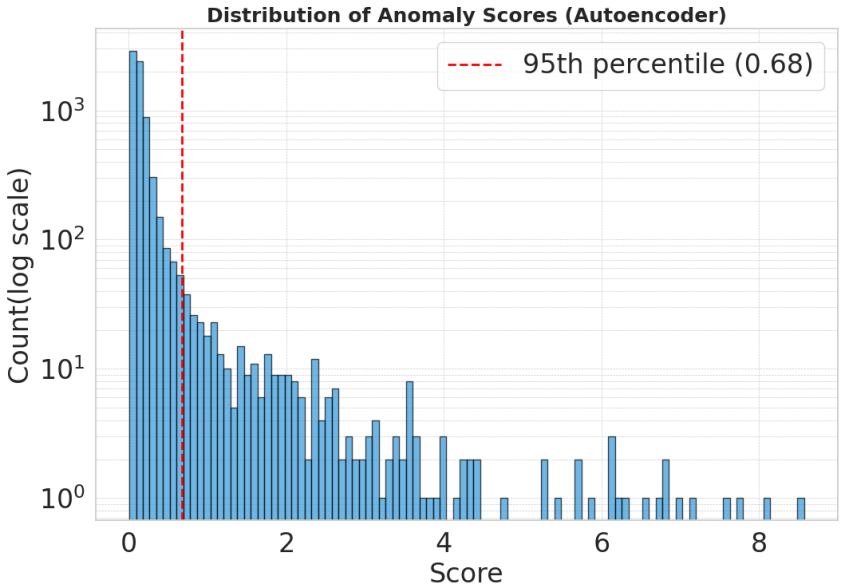
Autoencoder-Based Anomaly Detection

- Core Idea:
  - Autoencoders consist of two components: encoder and decoder.
  - Network is trained to minimize the difference between input data and reconstructed output.
  - Anomalies are detected by evaluating the reconstruction error, with normal data points having lower error.
- Encoder:
  - Compresses input data into a lower-dimensional representation.
  - Consists of three linear layers, ReLU activation functions, and dropout layers to prevent overfitting.
- Decoder:
  - Reconstructs data using a neural network with three linear layers and ReLU activations.
  - Includes dropout layers and applies a sigmoid activation function at the end to normalize values between 0 and 1.

Scores distribution



Anomaly scores distribution (original dataset)



Anomaly scores distribution (generated dataset)

Evaluation Results

- K Prototypes and Autoencoder are the most promising based on their distributions.
- For the generated dataset, we know which data points are outliers. All models were evaluated with the following results:

Approach	**Correctly Detected Outliers**
K Prototypes	295 / 360
Information Gain	140 / 360
Hybrid	360 / 360
Autoencoder	261 / 360

- Surprisingly, all outliers were detected using the Hybrid model. This may be a consequence of the way we built the outliers in the generated dataset.
  - Looking at the scores distribution, all the outliers are distinctly separated from the normal data with a noticeable gap in scores between these two groups.

Combination	Common Indices (max 360)
All Methods	107
KPrototypes and InfoBased	121
KPrototypes and Hybrid	295
KPrototypes and Autoencoder	319
InfoBased and Hybrid	140
InfoBased and Autoencoder	118
Hybrid and Autoencoder	261
KPrototypes, InfoBased, Hybrid	116
KPrototypes, InfoBased, Autoencoder	112
KPrototypes, Hybrid, Autoencoder	256
InfoBased, Hybrid, Autoencoder	111

Table: Common Indices Among Different Approaches for the Generated Dataset.

- Overall, all approaches except the Information-Based work well on the generated dataset.

- Now let's look at the results for the original dataset. All models defined 72 outliers (99% threshold). Below, we can see the similarity of the predictions.

Combination	**Common Indices (max 72)**
All Methods	0
KPrototypes and InfoBased	3
KPrototypes and Hybrid	2
KPrototypes and Autoencoder	61
InfoBased and Hybrid	29
InfoBased and Autoencoder	4
Hybrid and Autoencoder	3
KPrototypes, InfoBased, Hybrid	0
KPrototypes, InfoBased, Autoencoder	3
KPrototypes, Hybrid, Autoencoder	2
InfoBased, Hybrid, Autoencoder	1

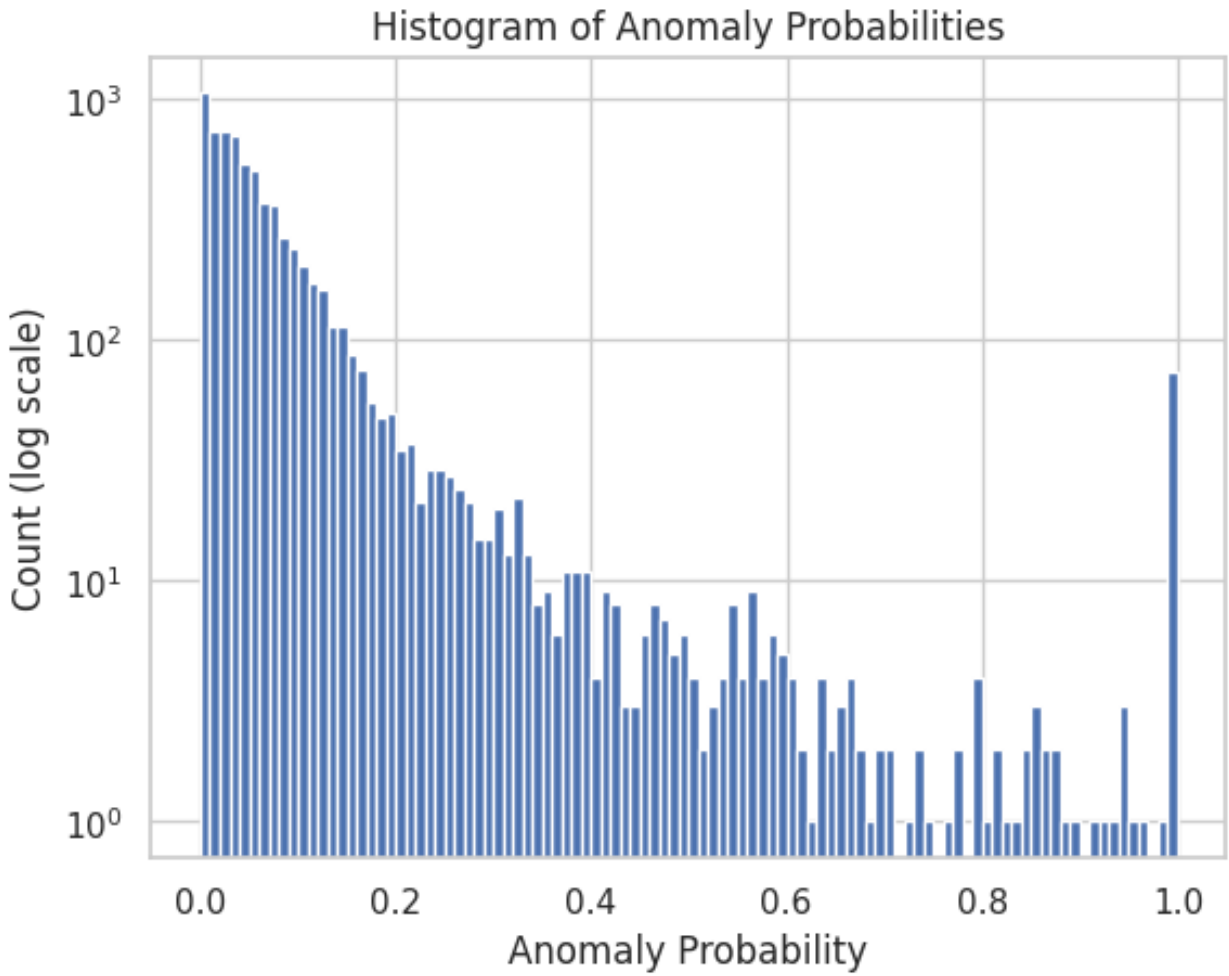
Table: Common Indices Among Different Approaches for the Original Dataset.

Final Observations and Model Selection

- K Prototypes and Autoencoder have the highest number of common detected anomalies.
- These two models were the most promising based on the distribution of anomaly scores.
- Given their significant differences, the common predictions are most likely the true anomalies.
- The Hybrid and InfoBased approaches performed very poorly.
- We need to select one model for the task.
- **Selected Model: Autoencoder**

Probability Estimation Based on Reconstruction Error

- Based on reconstruction error, we estimated probabilities of each object to be an anomaly.
- If the normalized error is above the threshold, we set the probability to 1; otherwise, we scale it.



Ideas for Improvement

- **K Prototypes**:
  - Optimize the number of clusters.
  - Experiment with different scaling methods.
- **InfoBased**:
  - Advanced binning or smooth approximations.
  - Experiment with entropy normalization.
- **Hybrid**:
  - Calculate weights for binary and continuous contributions.
  - Implement ideas from K Prototypes ('gamma' parameter).
- **Autoencoder**:
  - Tune hyperparameters.
  - Switch to variational autoencoder (VAE).
  - Explore different activation functions, learning rates, and batch sizes.

Thank You!