Optimized Stochastic Multi-Dimensional Scaling Sampling Heuristic

Vourvachakis Georgios

6/11/2024

1 Introduction

The strategy implemented in the `optimized_stratified_mds_analysis` function is designed to ensure data distribution integrity and achieve reliable performance insights through efficient dimensionality reduction techniques. The approach leverages Multi-Dimensional Scaling (MDS) to project high-dimensional data into 2D and 3D spaces, while maintaining computational efficiency and robustness in model performance assessment.

2 Objective

The key objectives of this strategy are:

1. **Distribution-Aware Sampling**: Ensuring that the sampled training set accurately reflects the distribution of the full dataset. This is critical for achieving meaningful dimensionality reduction analysis and for obtaining performance metrics that generalize well.

2. **Efficient Dimensionality Reduction**: Utilizing MDS for 2D and 3D projections to reduce computational complexity while preserving the essential structure of high-dimensional data.

3. **Robust Model Performance Assessment**: Minimizing the impact of sampling variability by conducting multiple runs and aggregating performance metrics.

4. **Baseline Comparison**: Establishing a performance benchmark without dimensionality reduction to understand the impact of MDS transformations clearly.

3 Methodology

The strategy is broken down into several key components:

3.1 Distribution Fitting

The function employs a `DistributionAwareStratifiedSampler` to segment the target variable into bins. This enables stratified sampling, ensuring proportional representation of each segment in the sampled dataset. The distribution summary and bin-specific details are logged to validate the sampling's effectiveness.

3.2 Sampling Validation

Sampling validation involves checking the distribution summary and ensuring that the binned segments are accurately represented in the sample. This step is crucial for confirming that the stratified sampling approach preserves the original dataset's characteristics.

3.3 Dimensionality Reduction

MDS transformations are applied to project the data into both 2D and 3D spaces. Regression models are then trained and tested on the reduced-dimensional data. Performance metrics, such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2, are calculated and averaged over multiple runs to enhance robustness. Timing metrics for MDS transformations and overall analysis execution are recorded to evaluate computational efficiency.

4 Performance Reporting

The results are presented in a comprehensive manner, comparing the performance of models trained on the reduced-dimensional data with the baseline model. This comparison provides insights into the trade-offs involved in applying MDS, including potential gains or losses in model performance.

5 Conclusion

This strategy provides a robust and efficient method for dimensionality reduction while maintaining data distribution integrity. The use of stratified sampling, robust metric aggregation, and detailed performance reporting ensures a reliable assessment of the impact of MDS on predictive modeling.

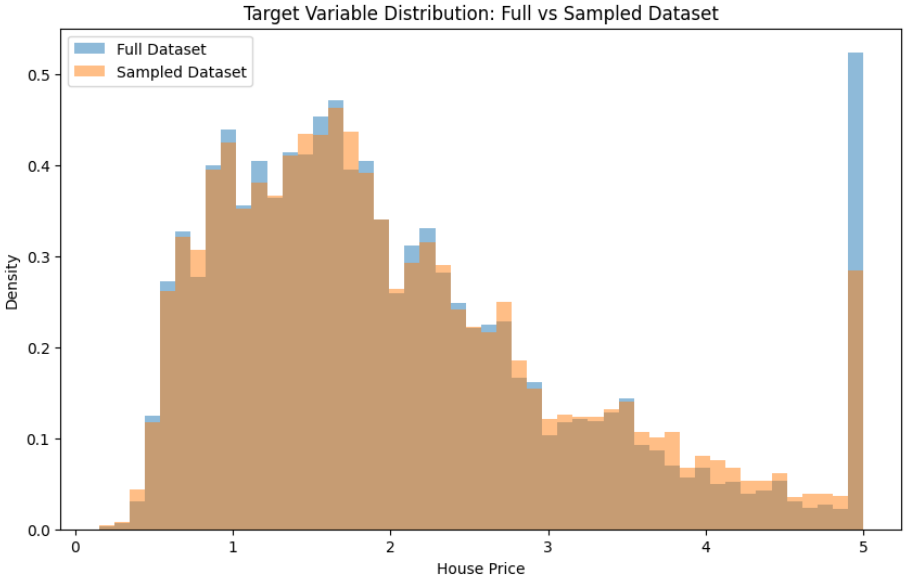

Figure 1: Comparison of Full vs. Sampled Dataset Distribution