Pavlo Mysak
12/16/2024

This project investigates the relationship between dataset metadata and the performance of various feature selection algorithms in predictive modelling. Synthetic datasets with controlled properties were generated to simulate diverse conditions, including varying distribution types, levels of dimensionality, noise ratios, interaction terms, and polynomial terms. Feature selection methods were assessed using two primary criteria: Root Mean Squared Error (RMSE) of a predictive model and a "Feature Correctness" score. The "Feature Correctness" metric, calculated as the harmonic mean of a pseudo -precision and -recall, captures each method's ability to identify informative features accurately.

The study identified the best-performing feature selection method for each dataset based on the highest "Feature Correctness" score, with RMSE used as a tiebreaker. Metadata properties were then analyzed using multinomial logistic regression to determine their influence on the likelihood of a specific method being optimal. Ongoing refinements to the methodology aim to enhance the robustness of the findings. Future work will include examining the relationship between RMSE reduction achieved through feature selection and dataset metadata, providing further insights into the role of data characteristics in predictive performance.

The findings will provide actionable insights for practitioners aiming to reduce dimensionality and improve feature interpretability in high-dimensional datasets. By understanding the relationship between dataset characteristics and algorithm performance, practitioners can make informed choices about feature selection methods, streamlining workflows while preserving model accuracy and robustness. This research highlights the critical role of dataset-specific metadata in guiding method selection and optimizing machine learning pipelines.