ML application for power consumption prediction

Samral Tahirli (samral.tahirli@studio.unibo.it)

June 2023

Content

1 Introduction

Supercomputers play a crucial role in various domains, serving as powerful tools for solving complex problems. However, their energy consumption and the lack of comprehensive data on their behavior pose significant challenges. This paper presents the outcomes of a decade-long project where they deployed the EXAMON monitoring framework at the CINECA data center in Italy, housing top-tier supercomputers. Additionally, I utilized various regression algorithms to preprocess the data and make accurate predictions, gaining valuable insights into the behavior and performance of the Marconi100 system. These findings contribute to the improvement of supercomputer efficiency and sustainability.

2 Problem

2.1 Description

Supercomputers are incredibly powerful machines that play a important role in various aspects of society, such as economics, industry, and overall development. They are extensively used by scientists, engineers, decision-makers, and data analysts to solve complex problems through computational means. However, these supercomputers and their associated data centers are complex systems that consume a significant amount of power. Enhancing their efficiency, availability, and resilience is of utmost importance, and many researchers and engineers are working towards achieving these goals.

In this paper, I present the outcomes of a decade-long project aimed at creating a monitoring framework called EXAMON, which was deployed at the CINECA data center in Italy. They unveil the first comprehensive dataset of a tier-0 Top10 supercomputer, specifically the Marconi100, which covers two and a half years of operation. To facilitate access to the data and its utilization, they have also developed open-source software modules, accompanied by practical examples.

In this research, I diligently prepared the collected data before employing various predictive models. To facilitate accurate predictions, I utilized a range of regression algorithms, including Linear Regression, Lasso, Ridge, ElasticNet, RandomForestRegressor, GradientBoostingRegressor, and DecisionTreeRegressor.
The integration of diverse algorithms provided us with a comprehensive understanding of the behavior and performance of the system, enabling us to extract valuable insights.

2.1 Dataset

High-performance computing (HPC) systems are intricate machines consisting of numerous diverse components. These components include computing nodes with thousands of parts, cooling infrastructure, network connectors, and various software elements. They have made available the largest dataset to the public, approximately 49.9TB in size before compression, and it is now accessible through Zenodo. The Marconi100 dataset is distributed as 12 different datasets, stored as a partitioned Parquet dataset. The partitioning hierarchy is based on the year and month ("YY-MM"), plugin, and metric categories.In this research, the focus will be on computing the average power consumption for each job.

3 Data processing

Data processing plays a significant role in preparing the dataset before feeding it into the models. In this research, the dataset was voluminous, necessitating certain steps to facilitate model computation on a laptop. Firstly, it was essential to filter the dataset based on a specific time period, selecting data points only within that range. This reduced the dataset size, making it more manageable for analysis.

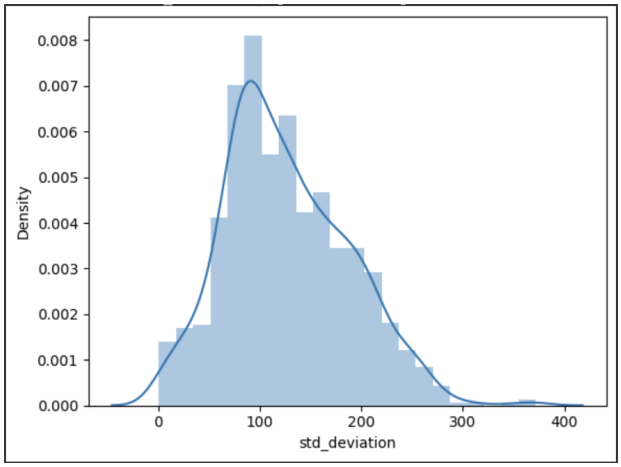
Furthermore, certain data type adjustments were required for effective modeling. Particularly, the "nodes" column initially had a string data type that contained

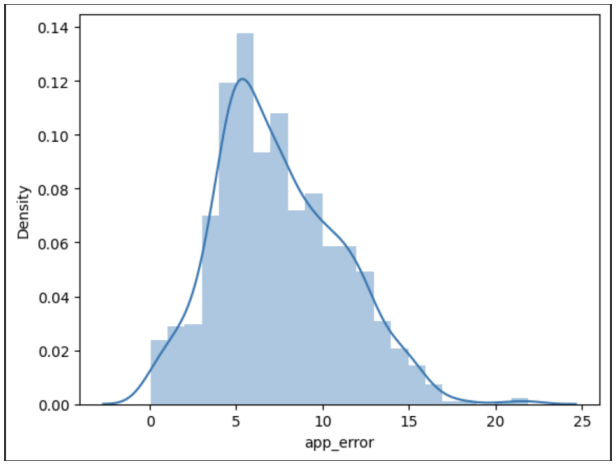
lists, with each list containing multiple items. To ensure compatibility for merging and further analysis, the "nodes" column was transformed using the "explode" function, resulting in separate individual items.

Additionally, the dataset contained null values that needed to be addressed. In this research, the approach involved either removing rows with null values or filling them using appropriate techniques, ensuring the dataset's integrity and completeness.

By employing these data preprocessing steps, the research was able to create a refined and structured dataset suitable for subsequent modeling and analysis.

Data preprocessing steps were applied to prepare the dataset, followed by the computation of key metrics including mean average power, standard deviation, and approximation error. Density plots were generated to visually represent the distribution and variability of these metrics across different nodes. The density plot for standard deviation revealed patterns and variations in power consumption, while the plot for approximation error assessed the accuracy of predictive models. These visualizations provided valuable insights into the dataset, aiding in informed decision-making and facilitating a better understanding of power consumption behavior and model performance.





4 Evaluation and Discussion and Results

In the modeling phase, a range of regression algorithms, including Linear Regression, Lasso, Ridge, ElasticNet, RandomForestRegressor, GradientBoostingRegressor, and DecisionTreeRegressor, were employed to predict power consumption. To assess the impact of normalization on model performance, both normalized and non-normalized datasets were utilized. Standard scaling was applied as the normalization technique.

The objective was to compare the train and test errors between the normalized and non-normalized datasets. It was acknowledged that the choice of scaling could significantly influence the prediction results, as scaling plays a crucial role in ensuring fair and accurate comparisons.

The results, as depicted in the accompanying images, clearly indicated the advantages of normalization. The train and test errors were observed to be consistently lower in the normalized dataset compared to the non-normalized dataset. This highlights the significance of scaling in achieving more reliable and robust predictions.

Furthermore, when examining the performance of individual models, it is evident that both Random Forest and Decision Tree models exhibited the lowest train errors, with negligible differences between them. However, in terms of test errors, Random Forest outperformed Decision Trees, demonstrating superior generalization capability and better prediction accuracy.

These findings underscore the importance of normalization and highlight the effectiveness of Random Forest as a powerful algorithm for modeling power consumption. The comprehensive analysis of train and test errors provides valuable insights into the performance and suitability of different regression models, aiding in the selection of optimal models for future power consumption prediction tasks.

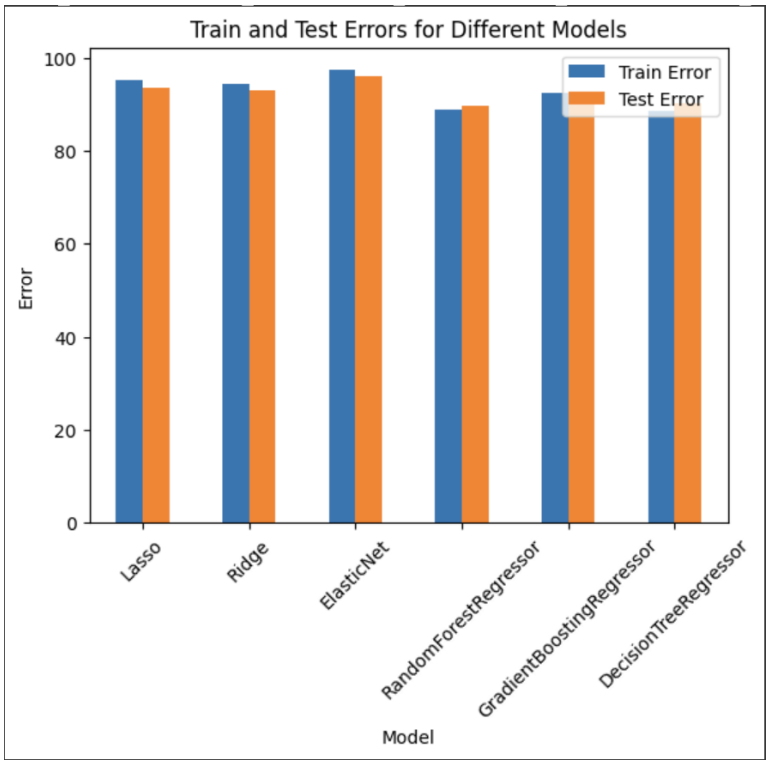
Non-normalization

	model	train_error	test_error
0	LinearRegression	94.463700	92.984726
1	Lasso	95.093348	93.493405
2	Ridge	94.464510	92.961984
3	ElasticNet	97.320132	95.958911
4	RandomForestRegressor	88.793226	89.578991
5	GradientBoostingRegressor	92.380943	91.130623
6	DecisionTreeRegressor	88.659427	90.134174

Normalization

	model	train_error	test_error
0	LinearRegression	0.935650	0.921008
1	Lasso	0.969997	0.957027
2	Ridge	0.935650	0.920996
3	ElasticNet	0.957574	0.944356
4	RandomForestRegressor	0.879401	0.887135
5	GradientBoostingRegressor	0.915020	0.902636
6	DecisionTreeRegressor	0.878159	0.892181

Graph below shows prediction error with normalized data



5 Conclusion

In conclusion, supercomputers play a crucial role in society, but their power consumption and efficiency remain significant challenges. This research project addressed these challenges by presenting the EXAMON monitoring framework and releasing the largest-ever comprehensive dataset of a tier-0 Top10 supercomputer, the Marconi100. The dataset covers an extensive period of operation and is accompanied by open-source software modules for easier access and utilization.

Through diligent data preparation and the utilization of various regression algorithms, this research provided valuable insights into the behavior and performance of the supercomputer. The integration of diverse models allowed for a comprehensive understanding of the system, paving the way for improved efficiency, availability, and resilience.

These findings contribute to ongoing efforts in enhancing supercomputer performance and facilitating sustainable and optimized utilization.

6 References

1.M100 ExaData: a data collection campaign on the 2 CINECA's Marconi100 Tier-0 supercomputer.pdf