Evaluating Computational and Carbon Costs of Machine Learning: The Impact of Feature Dimensionality During Model Development

Antonio Everton C. Teixeira[1], Fernanda Scarcela[1], Lívian Torres Mariano[1], Maria Clara Pereira[1], Marcos Roberto Machado[2], Wellington Franco[1], Bruno R. Silva[1], Renato William R. de Souza[3]

[1] Campus of Crateús, Federal University of Ceará, Crateús, Brazil

[2]Department of Financial Engineering and Business Information Systems, University of Twente, Enschede, The Netherlands

[3]Federal Institute of Education, Science and Technology of Ceará, Boa Viagem, Brazil

tteverton2024@gmail.com, m.r.machado@utwente.nl, wellington@crateus.ufc.br

Abstract. *The expansion of AI increases energy demand, making Green AI crucial. This study evaluates computational and carbon trade-offs in financial credit risk across six architectures under feature variations ($\pm$10% to $\pm$50%). Using CodeCarbon, we propose ECAR to quantify emissions per unit of predictive performance. Results show that across several model–dataset configurations, reducing feature dimensionality by 20% to 50% lowered emissions without significant losses in F1-score. In contrast, the evaluated feature augmentation increased computational costs without proportional gains in predictive performance. These findings highlight feature dimensionality and model architecture as key factors in the environmental efficiency of financial ML systems.*

1. Introduction

Recent advances in Artificial Intelligence (AI) have expanded its use in important areas such as healthcare and finance, where it supports decision-making tasks such as medical diagnosis and credit risk assessment [Zavyalova et al. 2023]. However, this continued advancement has also increased the computational demands and environmental impact of AI systems. Prior work frequently highlights that training large-scale models can emit hundreds of tons of CO_2 [Delanoë et al. 2023]; nevertheless, the sustainability challenge is not limited to isolated training runs. In practical machine learning (ML) workflows, repeated experimentation, hyperparameter optimization, cross-validation, and large-scale inference can also accumulate substantial computational costs. This issue is particularly relevant in financial applications, where models often process large volumes of continuously updated data and are commonly optimized primarily for predictive performance. Therefore, evaluating predictive performance is no longer sufficient: model development must also consider energy consumption and carbon emissions.

In this scenario, Green AI has emerged as a framework for developing AI systems that optimize predictive performance while minimizing computational and environmental costs throughout the ML lifecycle [Haakman et al. 2021]. Energy consumption is affected by factors such as hardware utilization [Xu et al. 2021], hyperparameter optimization [de Chavannes et al. 2021, Morteza et al. 2023], framework efficiency

[Georgiou et al. 2022], and data characteristics. Because the structure, quality, and dimensionality of input data directly influence computational workload, recent studies have highlighted data-centric approaches as promising strategies for improving sustainability from the earliest stages of the ML pipeline [Salehi and Schmeink 2024].

In this context, feature engineering and feature selection play central roles. Feature engineering can expand the original feature space by deriving new variables that capture temporal, aggregated, or domain-specific patterns, often improving predictive performance. However, increasing feature dimensionality may also introduce redundancy, raise computational cost, and increase energy consumption. Conversely, feature selection and dimensionality reduction can simplify the learning problem, mitigate the curse of dimensionality, reduce memory usage, and decrease processing time, but may also remove information relevant to prediction. Thus, feature dimensionality represents a key trade-off between predictive effectiveness and environmental efficiency in ML pipelines [Machado et al. 2024, Li et al. 2024, Jiao et al. 2023].

However, it remains unclear how systematic feature expansion and reduction affect the cumulative carbon cost of ML development workflows, especially when hyperparameter tuning and cross-validation are included in the evaluation process. To address this limitation, this study investigates the trade-off between predictive performance and environmental cost in credit risk modeling. Specifically, we evaluate systematic variations in feature dimensionality, ranging from -50% to $+50\%$, across two financial datasets and six ML architectures. Carbon emissions are monitored with *CodeCarbon* throughout the model development process, including hyperparameter tuning and cross-validation.

The core contributions of this paper are threefold: (i) quantifying how controlled feature expansion and reduction affect predictive performance, computational cost, and carbon emissions in credit risk models; (ii) evaluating these effects across multiple model families, enabling the analysis of how lightweight and more computationally demanding architectures respond to changes in feature dimensionality; and (iii) introducing the Emission-to-Classification Assessment Ratio (ECAR), a development-cycle metric designed to express the cumulative carbon cost required to achieve a given level of predictive effectiveness. This metric supports the guiding research question: *What is the cumulative CO_2 cost required to maintain an acceptable baseline F1-score during model development?*

2. Related Work

Data-centric approaches, including feature engineering, feature selection, dimensionality reduction, and data pruning, have been increasingly investigated to improve the predictive performance and computational efficiency of machine learning (ML) models. In particular, feature engineering can expand the original feature space by incorporating temporal, aggregated, or domain-specific variables, thereby capturing more complex patterns and improving predictive accuracy [Medeiros et al. 2024, Machado et al. 2024]. For instance, one study showed that feature engineering and relevance-based feature selection can reduce overfitting and improve the scalability of XGBoost in large-scale sales forecasting [Shilong et al. 2021]. Similarly, another study used Pearson correlation to select relevant features in an improved LightGBM framework for predicting the Remaining Useful Life (RUL) of lithium-ion batteries under driving conditions, achieving higher predictive

accuracy and robustness than alternative models [Jiao et al. 2023].

Beyond predictive performance, recent surveys on Green AI and Data-Centric Green AI indicate that dataset-level interventions can also reduce the energy consumption and environmental impact of ML workflows [Verdecchia et al. 2023, Salehi and Schmeink 2024]. An exploratory study showed that feature selection can reduce training energy consumption by up to 76% with negligible or no accuracy degradation, also identifying a positive correlation between the number of features and the energy consumed during model training [Verdecchia et al. 2022]. However, its experimental design relies on a simple train-test split, making the results strongly dependent on a single data partition, and it does not provide a normalized metric that integrates predictive performance and environmental cost.

In the financial sector, a closely related study investigated Green AI strategies for Customer Relationship Management (CRM) models, comparing ensemble methods and k-Means-based hybrid approaches [Machado et al. 2024]. The study showed that reducing the feature space from 85 to 18 variables resulted in only a marginal MAE difference of approximately 0.01, while substantially reducing processing time. Conversely, expanding the feature space from 28 to 85 variables through feature engineering aimed to improve predictive performance but increased computational cost. Nevertheless, the study infers environmental impact from processing time rather than directly measuring energy consumption or carbon emissions, and it does not provide a normalized metric relating predictive performance to environmental cost.

Overall, the literature shows that feature-level interventions can improve predictive performance and computational efficiency. However, most studies either rely on indirect efficiency proxies or estimate energy consumption without integrating it with predictive performance in a unified metric. In particular, few works combine controlled variations in feature dimensionality, multiple ML model families, hyperparameter optimization, cross-validation, and software-based carbon monitoring within a single experimental framework. To address this gap, the present work evaluates the trade-off between predictive performance and carbon emissions in credit risk modeling and proposes ECAR as a development-cycle-oriented metric for comparing the environmental efficiency of different model and feature configurations.

3. Methodology

This section details the experimental protocol used to compare model performance, computational cost, and carbon emissions under controlled scenarios with varying feature dimensionality. As shown in Figure 1, the experimental workflow comprises four main stages: dataset selection, data preprocessing and feature dimensionality manipulation, model training, and carbon emissions monitoring. All source code is publicly available in our GitHub repository[1].

3.1. Datasets

Two real-world datasets from the credit risk domain were selected. These datasets were chosen for their relevance to financial analytics and their diversity in scale and structure.

[1]https://github.com/EvertonTeix/greenai-credit-risk-ecar

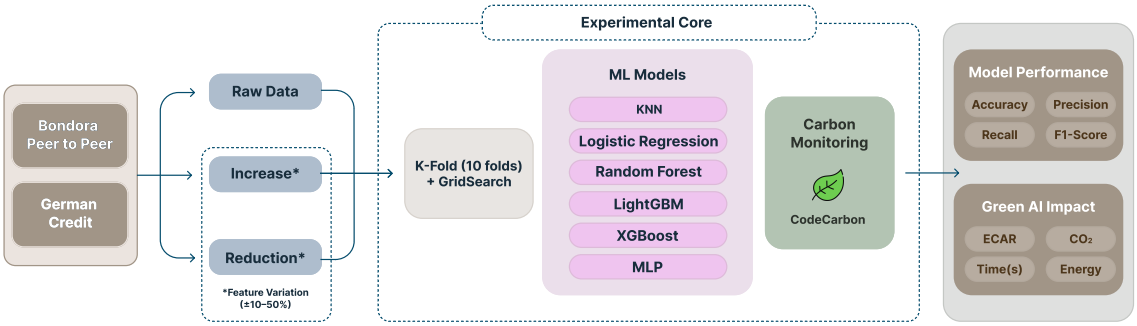


Figure 1. Methodological experimental flow: from feature dimensionality variation (±10–50%) to comparative carbon emission monitoring.

- **German Credit Dataset**[2]: This widely used benchmark dataset contains 13 features and 1,000 instances, where each record represents a bank loan applicant classified as either a "good" or "bad" credit risk, with 300 instances labeled as "bad" credit risk.
- **Bondora Peer-to-Peer (P2P)**[3]: This dataset contains historical records from the European lending platform Bondora (2009–2021) and was used to predict credit default. It includes 121,461 instances and 112 features, comprising demographic and financial variables that describe borrowers' risk behavior in an unsecured lending environment.

Before dimensionality manipulation, the datasets were preprocessed by handling missing values with imputation, using the mean for numerical variables and the mode for categorical ones. Next, categorical attributes were converted into numerical representations via One-Hot Encoding to ensure compatibility with the machine learning algorithms and subsequent statistical analyses. To eliminate data redundancy, a Pearson correlation matrix was then computed across all numerical and encoded features, and one feature was removed from each pair presenting a high correlation coefficient (≥ 0.7).

3.1.1. Feature Engineering and Feature Selection

After preprocessing, the workflow proceeded to the dimensionality manipulation steps used to evaluate the impact of feature-space changes:

- **Dimensionality Reduction (-10% to -50%):** Reduction scenarios were generated by ranking features according to their Mutual Information (MI) score with the target variable (*default*) and progressively removing the least informative attributes. MI was chosen because it captures both linear and non-linear statistical dependencies, making it suitable for credit risk datasets where complex relationships may exist between predictors and the target.
- **Dimensionality Augmentation (+10% to +50%):** Conversely, derived features were systematically engineered in increments of 10%. Although specific transformations were tailored to each dataset's schema, all scenarios followed a common

[2]https://www.kaggle.com/code/yiitcancokun/risk-prediction-with-tree-classifiers/input
[3]https://www.kaggle.com/datasets/sid321axn/bondora-peer-to-peer-lending-loan-data

feature-engineering rationale. Table 1 summarizes the engineered features used across the five expansion levels.

The complete dataset-specific list of engineered features is available in the anonymous repository.

Table 1. Summary of engineered features used in the dimensionality augmentation scenarios. The exact number of features added at each step is denoted for the Bondora and German Credit datasets.

Scenario	Features Added		Type of engineered features	Examples
	Bondora	German		
+10%	+11	+1	Financial ratios and payment-burden indicators	Credit per month; loan amount vs. income; loan amount vs. free cash; estimated monthly payment; payment-to-income ratio.
+20%	+10	+1	Risk segmentation and income/source composition indicators	Young applicant indicator; alternative income ratio; principal income ratio; pension ratio; manual/API/portfolio bid ratios.
+30%	+10	+2	Log transformations and exposure indicators	Safe logarithmic transformations for skewed financial variables; job–credit interaction; high loan amount indicator.
+40%	+11	+2	Polynomial terms and risk-normalized ratios	Credit-to-age risk ratio; duration-to-age ratio; squared terms for age, interest rate, income, and liabilities.
+50%	+10	+1	High-risk purpose groups, solvency indicators, and higher-order interactions	High-risk loan purpose indicator; debt-to-income interactions with age and interest rate; variables normalized by existing liabilities.
Total Added	**+52**	**+7**		

3.2. Experimental Design

The experimental phase was designed to assess how changes in feature dimensionality affect energy consumption across different feature workloads. Each model was evaluated using 10-fold cross-validation combined with grid search. For each fold, MI was computed on training data before optimizing hyperparameters via an internal 80/20 split; the best configuration was then retrained and evaluated on the independent test fold. This workflow was repeated across 11 scenarios per dataset: one baseline scenario using the original preprocessed dataset without dimensionality changes, five reduction scenarios that progressively removed 10%, 20%, 30%, 40%, and 50% of the features with the lowest MI scores, and five expansion scenarios that increased dimensionality by the same percentages through derived features generated via feature engineering.

3.2.1. Machine Learning Models

Six machine learning models were evaluated: K-Nearest Neighbors (KNN) [Guo et al. 2003], Logistic Regression (LR) [LaValley 2008], Random Forest (RF) [Rigatti 2017], LightGBM [Ke et al. 2017], XGBoost [Chen et al. 2015], and Multi-Layer Perceptron (MLP) [Riedmiller and Lernen 2014]. These models represent complementary families widely used for tabular classification: LR provides an interpretable linear baseline, KNN a distance-based learner, RF a bagging ensemble, LightGBM and XGBoost computationally distinct gradient-boosting implementations, and MLP a neural

baseline. This diversity enables comparisons across different learning mechanisms and computational profiles. The evaluated hyperparameters included the number of neighbors and distance metric for KNN, regularization strength for LR, the number of estimators and maximum depth for tree-based models, and hidden-layer size and regularization parameter for MLP. The complete hyperparameter grid is available in the repository.

3.2.2. Carbon Monitoring Framework and Experimental Setup

To quantify environmental impact, the experiments used CodeCarbon[4], an open-source Python library that estimates hardware energy consumption (CPU, GPU, RAM) and converts it into CO2-equivalent emissions using a carbon-intensity coefficient.

All experiments were executed on a local workstation located in Crateús, Ceará, Brazil, equipped with an Intel Xeon W-1350 CPU @ 3.30 GHz, 16 GB of RAM, and an NVIDIA RTX A4500 (20 GB VRAM), running Ubuntu Linux. To ensure reproducibility and avoid temporal grid fluctuations, we bypassed CodeCarbon's dynamic retrieval and applied a fixed carbon-intensity factor of 46.1 gCO_2/kWh (the 2025 average for the Brazilian National Interconnected System)[5]. Thus, total emissions become a deterministic function of hardware energy consumption, isolating model efficiency from external variables.

3.3. Evaluation Metrics

To evaluate the experimental trade-offs, predictive performance was measured using standard classification metrics commonly adopted in classification and credit risk tasks: Accuracy, Precision, Recall, and F1-score. The environmental dimension was assessed through three physical metrics: Time (s), Energy (kWh), and Carbon Footprint (g of CO_2). In addition, we formulated a composite indicator termed ECAR, which relates cumulative carbon emissions to predictive effectiveness, as defined in Equation 1:

$$ECAR = \frac{Total\ Cumulative\ Energy\ (kWh) \times CI}{Mean\ F1\text{-}score} \tag{1}$$

In this formulation, *Total Cumulative Energy* is the sum of the hardware energy consumption (kWh) recorded throughout the modelling process, including hyperparameter tuning via grid search and evaluation under the 10-fold cross-validation protocol. This consumption is multiplied by *CI*, a fixed national Carbon Intensity factor (46.1 gCO_2/kWh) that isolates the metric from temporal grid fluctuations. The denominator is the mean F1-score across the cross-validation folds. The F1-score was adopted because credit risk datasets are often class-imbalanced, making it a more appropriate measure of predictive effectiveness than accuracy alone.

The ECAR follows the broader Green AI perspective of reporting environmental impact alongside model performance. Existing approaches have proposed carbon-aware evaluation metrics and reporting frameworks that combine predictive performance, energy consumption, and carbon emissions, including benefit-centric indicators such as the Carbon Footprint Efficiency Ratio (CFER), which measures optimization benefit per

[4]https://github.com/mlco2/codecarbon

[5]https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/sirene/dados-e-ferramentas/fatores-de-emissao

unit of carbon emission [Yu et al. 2026, Kaur et al. 2026, Zhou and Bu 2026]. However, the objective of ECAR is different: rather than measuring performance return per unit of emission, it adopts a cost-centric perspective by quantifying how much carbon is emitted to achieve a given level of predictive performance.

Therefore, ECAR is not intended to replace traditional predictive or environmental metrics, but to complement them as a comparative indicator under the same experimental protocol.

4. Results and Discussion

To evaluate the trade-off between predictive performance and environmental cost, we first established a baseline scenario using 100% of the original features after preprocessing. Table 2 presents the predictive metrics, computational cost, and environmental footprint of the six models across both datasets. Results are reported as the mean and standard deviation across the 10 cross-validation folds, directly yielding 10 paired samples for the subsequent Wilcoxon signed-rank test.

Table 2. Baseline predictive performance and environmental cost evaluation across both credit risk datasets (Mean $\pm$ SD over 10 folds).

Dataset	Model	Accuracy	Precision	Recall	F1-score	Time (s)	Emissions (g)	Energy (kWh)	ECAR
Bondora	KNN	0.6822 ± 0.0032	0.6892 ± 0.0045	0.7962 ± 0.0050	0.7388 ± 0.0029	152.63	5.1147	0.0529	3.3012
	LR	0.6833 ± 0.0043	0.6869 ± 0.0043	0.8067 ± 0.0049	0.7420 ± 0.0040	32.66	**0.4683**	**0.0048**	**0.2980**
	LightGBM	0.7163 ± 0.0038	0.7081 ± 0.0045	0.8464 ± 0.0043	0.7711 ± 0.0025	**15.38**	0.6228	0.0063	0.3786
	MLP	0.6981 ± 0.0039	0.7025 ± 0.0062	0.8072 ± 0.0205	0.7511 ± 0.0069	663.76	29.5034	0.3019	18.5309
	RF	$\mathbf{0.7215 \pm 0.0038}$	$\mathbf{0.7129 \pm 0.0051}$	0.8484 ± 0.0028	$\mathbf{0.7748 \pm 0.0030}$	125.85	1.1057	0.0113	0.6735
	XGBoost	0.7135 ± 0.0122	0.7031 ± 0.0207	$\mathbf{0.8558 \pm 0.0279}$	0.7713 ± 0.0036	41.67	2.4162	0.0248	1.4815
German Credit	KNN	0.6960 ± 0.0414	0.7095 ± 0.0404	0.9549 ± 0.0364	0.8135 ± 0.0337	**1.29**	0.0016	**0.0000**	0.0009
	LR	0.7060 ± 0.0369	0.7167 ± 0.0396	0.9585 ± 0.0360	0.8193 ± 0.0293	1.47	**0.0008**	**0.0000**	**0.0005**
	LightGBM	$\mathbf{0.7130 \pm 0.0531}$	$\mathbf{0.7325 \pm 0.0578}$	0.9334 ± 0.0683	0.8184 ± 0.0424	1.80	0.0420	0.0004	0.0241
	MLP	0.7050 ± 0.0347	0.7316 ± 0.0396	0.9128 ± 0.0277	0.8116 ± 0.0285	3.96	0.0504	0.0005	0.0291
	RF	0.7090 ± 0.0441	0.7200 ± 0.0535	$\mathbf{0.9607 \pm 0.0253}$	$\mathbf{0.8216 \pm 0.0298}$	2.89	0.0351	0.0004	0.0200
	XGBoost	0.7110 ± 0.0484	0.7296 ± 0.0597	0.9405 ± 0.0425	0.8194 ± 0.0324	1.64	0.0860	0.0009	0.0492

The baseline results reveal distinct trade-offs between predictive performance and environmental cost across model architectures. In the Bondora dataset, tree-based models achieved the highest F1-score, with RF, XGBoost, and LightGBM reaching values close to 0.77. In the German Credit dataset, performance differences were narrower, and RF achieved the highest F1-score (0.821). These results indicate that predictive effectiveness depends not only on model complexity, but also on dataset scale and structure.

When the environmental dimension is considered through the proposed ECAR metric, additional differences emerge. For example, in the Bondora dataset, the MLP achieved a competitive F1-score (0.7511) but required over 660 seconds of processing time, resulting in an ECAR of 18.53. In contrast, LightGBM achieved a higher F1-score (0.7711) while maintaining a substantially lower ECAR of 0.38. This indicates that models with similar predictive performance may impose very different carbon costs.

ECAR also differentiates models that appear similar under traditional predictive metrics: although XGBoost and LightGBM achieve nearly identical F1-score on Bondora (0.7713 vs. 0.7711), XGBoost's ECAR (1.48) is nearly four times that of LightGBM (0.378). These baseline results motivate the subsequent experiments on dimensionality variation.

4.1. The Green AI Sweet Spot: Dimensionality Reduction

To compare how different learning mechanisms respond to feature reduction, all six architectures were evaluated across the dimensionality-reduction scenarios. Figure 2 illustrates the trade-off between the F1-score (bars, left vertical axis) and the proposed ECAR metric (lines, right vertical axis) across reduction scenarios ranging from 0% (baseline) to -50% for the Bondora and German Credit datasets.

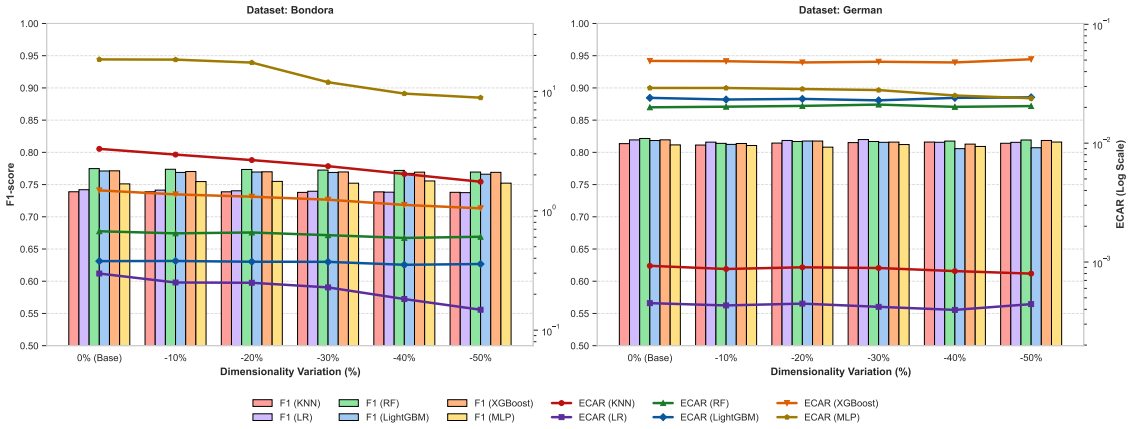


Figure 2. Trade-off between F1-score and ECAR across dimensionality reduction scenarios for all six architectures on the Bondora and German Credit datasets.

Evaluating all six architectures reveals distinct behavioral profiles. Tree-based models (RF, XGBoost, and LightGBM) achieve the highest F1-score while maintaining low, stable ECAR values below the 10^0 mark throughout the reduction range, and German Credit's ECAR remains low and stable across all scenarios. In contrast, the MLP and KNN start from disproportionately high carbon costs on Bondora and show the largest ECAR reductions as features are removed-the MLP drops from 18.53 at baseline to approximately 10 at -50%, while LR remains the most carbon-efficient model throughout. A 'Green AI sweet spot' therefore emerges at the maximum evaluated reduction level (50%), where cumulative carbon costs reach their lowest values while predictive performance remains competitive.

4.2. Environmental Cost of Dimensionality Augmentation

Figure 3 illustrates how feature expansion can translate into an environmental penalty. By adding engineered features to the datasets, we observe three behaviors that challenge the assumption that additional features necessarily improve predictive performance for the evaluated datasets and engineered variables.

First, the results show predictive stagnation. Across augmentation scenarios, the F1-score remains nearly unchanged, around 0.769 for the Bondora dataset and with only minor fluctuations in the German Credit dataset. This suggests that the additional engineered features did not provide discriminative information under the evaluated settings.

Second, while predictive performance tends to stagnate, computational and environmental costs continue to increase: for XGBoost on Bondora, ECAR rose from 1.482 at baseline to 2.590 at +50% (Figure 3). These results suggest that, beyond

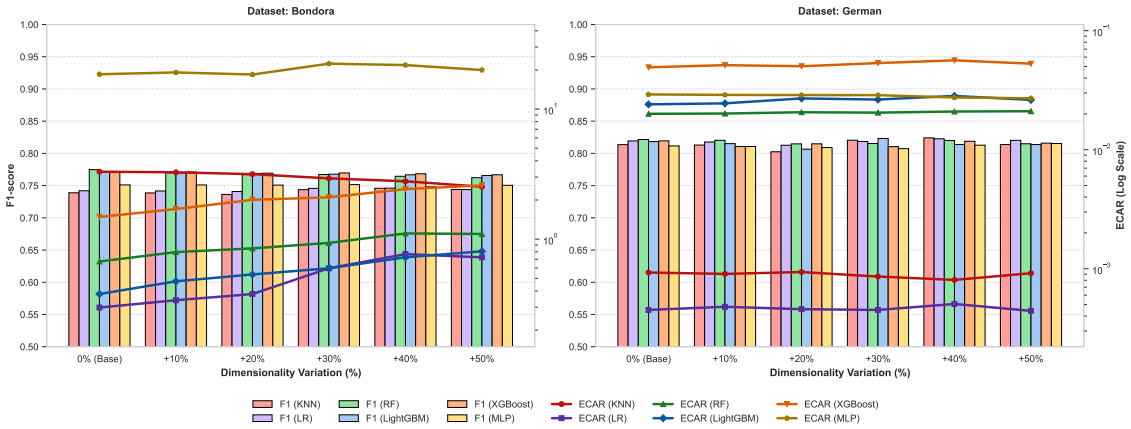


Figure 3. Trade-off between F1-score and ECAR across dimensionality augmentation scenarios for all six architectures on the Bondora and German Credit datasets.

an optimal information threshold, the applied feature augmentation strategies became environmentally inefficient without proportional predictive gains.

Finally, the augmentation scenarios expose architectural differences in high-dimensional settings: even with +50% augmentation on Bondora, LightGBM maintained an ECAR of 0.806, compared to XGBoost's 2.590, despite delivering a comparable F1-score (0.766 vs. 0.767).

4.3. Statistical Validation

We applied the Wilcoxon signed-rank test on the cross-validation folds to assess if differences in both F1-score and ECAR against the baseline (0%) were statistically significant. We used the Benjamini-Hochberg correction for all comparisons; adjusted p-values are summarized in Figure 4.

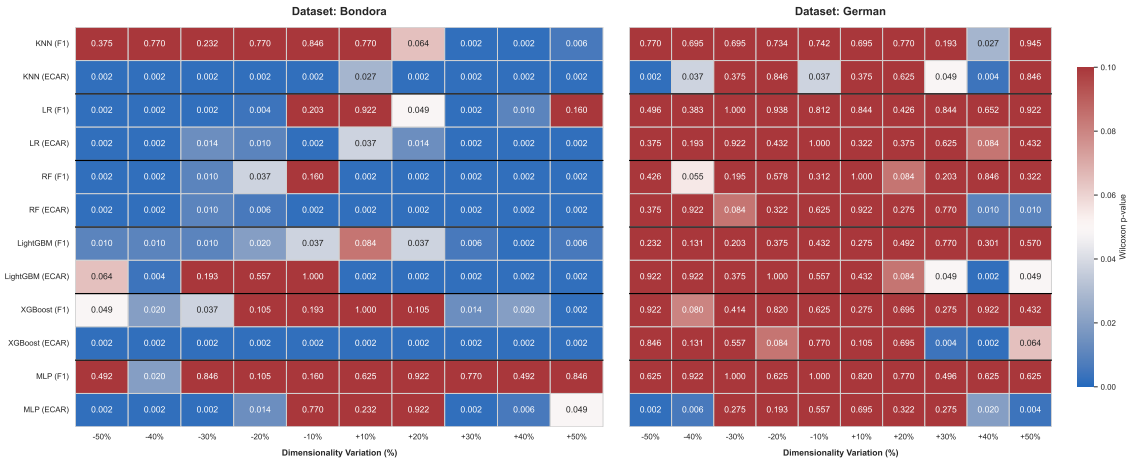


Figure 4. Heatmap of Wilcoxon signed-rank test adjusted p-values comparing F1-score and ECAR against the baseline. Blue cells ($p \leq 0.05$) indicate significant differences; red cells ($p > 0.05$) denote no statistically significant difference.

The heatmap highlights distinct behaviors: for F1-score, most scenarios (red

cells, $p > 0.05$) show no significant difference from baseline despite dimensionality changes, whereas ECAR rows are dominated by blue cells ($p \leq 0.05$), confirming substantial environmental cost variations. For instance, Bondora augmentation scenarios significantly increase ECAR across most models, without proportional increases in predictive performance.

5. Limitations and Future Work

This study has limitations related to the experimental setting and scope. Energy and carbon estimates depend on the hardware, software environment, and carbon-intensity configuration, so ECAR should be interpreted primarily as a relative indicator within the same experimental setting. The computational costs of feature reduction and augmentation were also not included in the carbon measurements. Furthermore, confidence intervals were not reported for environmental metrics, and the experiments were restricted to two credit risk datasets and fixed feature-engineering strategies.

Future work will extend the framework to other financial tasks and data modalities, including evaluating carbon-performance trade-offs in Large Language Models (LLMs) applied to unstructured financial data, such as aspect-based sentiment analysis. Future research will also investigate automated carbon-aware hyperparameter optimization to identify more efficient model configurations during development.

6. Conclusion

This study investigated the trade-off between predictive performance and environmental cost in credit risk modeling under controlled variations in feature dimensionality. The proposed ECAR metric complements traditional performance and environmental measures by expressing the carbon cost associated with predictive effectiveness.

Results show that dimensionality reduction can substantially decrease computational and carbon costs while preserving predictive performance, depending on the dataset and learning algorithm. In contrast, the evaluated feature-augmentation strategies increased computational costs without providing proportional predictive gains. Overall, the full experimental workflow consumed 4.14 kWh and produced 191.27 g of CO2.

These findings reinforce the importance of considering feature dimensionality and model architecture when developing financial ML systems. Lower-dimensional and lower-ECAR configurations should be preferred when additional complexity does not yield statistically supported predictive gains, supporting more computationally and environmentally efficient model-development practices.

Declaration on the Use of Generative AI

Generative AI tools were used only to support language revision, text organization, and clarity improvements. All scientific content, methodology, experiments, results, and conclusions were verified and validated by the authors, who take full responsibility for the integrity, originality, and final content of the paper.

References

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., et al. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4.

de Chavannes, L. H. P., Kongsbak, M. G. K., Rantzau, T., and Derczynski, L. (2021). Hyperparameter power impact in transformer language model training. In *Proceedings of the second workshop on simple and efficient natural language processing*, pages 96–118.

Delanoë, P., Tchuente, D., and Colin, G. (2023). Method and evaluations of the effective gain of artificial intelligence models for reducing co2 emissions. *Journal of Environmental Management*, 331:117261.

Georgiou, S., Kechagia, M., Sharma, T., Sarro, F., and Zou, Y. (2022). Green ai: Do deep learning frameworks have different costs? In *Proceedings of the 44th International Conference on Software Engineering*, pages 1082–1094.

Guo, G., Wang, H., Bell, D., Bi, Y., and Greer, K. (2003). Knn model-based approach in classification. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 986–996. Springer.

Haakman, M., Cruz, L., Huijgens, H., and Van Deursen, A. (2021). Ai lifecycle models need to be revised: An exploratory study in fintech. *Empirical Software Engineering*, 26(5):95.

Jiao, Z., Wang, H., Xing, J., Yang, Q., Yang, M., Zhou, Y., and Zhao, J. (2023). A lightgbm based framework for lithium-ion battery remaining useful life prediction under driving conditions. *IEEE Transactions on Industrial Informatics*.

Kaur, R., Kundu, T., Park, K. M., and Pinsky, E. (2026). The carbon cost of intelligence: A domain-specific framework for measuring ai energy and emissions. *Energies*, 19(3):642.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

LaValley, M. P. (2008). Logistic regression. *Circulation*, 117(18):2395–2399.

Li, G., Yu, Z., Yang, K., Lin, M., and Chen, C. P. (2024). Exploring feature selection with limited labels: A comprehensive survey of semi-supervised and unsupervised approaches. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6124–6144.

Machado, M. R., Asadi, A., de Souza, R. W. R., and Ugulino, W. C. (2024). Green ai in the finance industry: Exploring the impact of feature engineering on the accuracy and computational time of machine learning models. *Applied Soft Computing*, 167:112343.

Medeiros, E. P., Machado, M. R., de Freitas, E. D. G., da Silva, D. S., and de Souza, R. W. R. (2024). Applications of machine learning algorithms to support covid-19 diagnosis using x-rays data information. *Expert Systems With Applications*, 238:122029.

Morteza, A., Yahyaeian, A. A., Mirzaeibonehkhater, M., Sadeghi, S., Mohaimeni, A., and Taheri, S. (2023). Deep learning hyperparameter optimization: Application to electricity and heat demand prediction for buildings. *Energy and Buildings*, 289:113036.

Riedmiller, M. and Lernen, A. (2014). Multi layer perceptron. *Machine learning lab special lecture, University of Freiburg*, 24:11–60.

Rigatti, S. J. (2017). Random forest. *Journal of insurance medicine*, 47(1):31–39.

Salehi, S. and Schmeink, A. (2024). Data-centric green artificial intelligence: A survey. *IEEE Transactions on Artificial Intelligence*, 5(5):1973–1989.

Shilong, Z. et al. (2021). Machine learning model for sales forecasting by using xgboost. In *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, pages 480–483. IEEE.

Verdecchia, R., Cruz, L., Sallou, J., Lin, M., Wickenden, J., and Hotellier, E. (2022). Data-centric green ai an exploratory empirical study. In *2022 international conference on ICT for sustainability (ICT4S)*, pages 35–45. IEEE.

Verdecchia, R., Sallou, J., and Cruz, L. (2023). A systematic review of green ai. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(4):e1507.

Xu, J., Zhou, W., Fu, Z., Zhou, H., and Li, L. (2021). A survey on green deep learning. *arXiv preprint arXiv:2111.05193*.

Yu, M., Yang, H., Wei, X., Fu, S., Xu, J., and Zhang, J. (2026). Carbon footprint efficiency ratio: A unified indicator for green evaluation of evolutionary algorithms. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

Zavyalova, E. B., Volokhina, V. A., Troyanskaya, M. A., and Dubova, Y. I. (2023). A humanistic model of corporate social responsibility in e-commerce with high-tech support in the artificial intelligence economy. *Humanities and social sciences communications*, 10(1):274.

Zhou, Y. and Bu, W. (2026). Artificial intelligence adoption, energy management, and corporate energy transition: Evidence from energy consumption, energy intensity, and carbon emission intensity. *Energies*, 19(3):821.