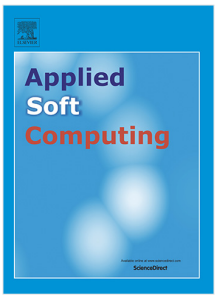
Journal Pre-proof



Green AI in the finance industry: Exploring the impact of feature engineering on the accuracy and computational time of Machine Learning models

Marcos R. Machado, Amin Asadi, Renato William R. de Souza, Wallace C. Ugulino

PII: S1568-4946(24)01117-7
DOI: https://doi.org/10.1016/j.asoc.2024.112343
Reference: ASOC 112343

To appear in: *Applied Soft Computing*

Received date : 30 March 2024
Revised date : 11 October 2024
Accepted date : 12 October 2024

Please cite this article as: M.R. Machado, A. Asadi, R.W.R. de Souza et al., 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* (2024), doi: https://doi.org/10.1016/j.asoc.2024.112343.

Green AI in the Finance Industry: Exploring the Impact of Feature Engineering on the Accuracy and Computational Time of Machine Learning Models

Abstract

As research and practice in Artificial Intelligence (AI) applications rapidly expand, the support for AI deployment is also increasing. While the abundance of data allows for sophisticated feature engineering techniques that can enhance accuracy, it is crucial to highlight both the computational costs and the efficiency with which these models operate. This paper compares the processing time and accuracy of individual and hybrid Machine Learning (ML) models in predicting customer loyalty within financial contexts. Frameworks that incorporate feature engineering and green AI principles are used separately in both individual and hybrid approaches. The individual models are commonly used regressor-based algorithms applied to business problems. The hybrid models first use k-Means to cluster customers, followed by the application of individual regressor-based models (e.g., decision trees, gradient boosting, and LightGBM). The present results show that using fewer features results in only a marginally lower accuracy compared to models with more features (a difference of ≈ 0.01 in MAE when comparing the use of 18 versus 85 features). Additionally, this article clearly demonstrate the trade-off between higher accuracy and longer computational time in hybrid ML models versus lower accuracy and shorter computational time in individual models when predicting customer loyalty. Hybrid models exhibit a lower MSE (≈ 0.88) compared to individual models (≈ 0.91). These findings provide managers with insights on selecting the most appropriate model based on their organization's specific needs.

Keywords: `Feature Engineering`, Green AI, Machine Learning, Hybrid Machine Learning, Customer Loyalty, Finance Industry

1. Introduction

Many Marketing scholars have recognized that customers are the significant assets of a firm, and as so, they require individual and portfolio management [1], with the implementation of campaigns based on models that account for their behaviour and other indicators (e.g., loyalty, risk, and engagement). However, [2] points out that other variables should be considered when computing the firms' liquid profits (e.g., personnel, equipment, and energy expenses). Thus, there are companies aiming to increase the accuracy of their models and become more profitable in the face of their competitors; however, researchers are concerned about the methods adopted to reach higher accuracy on different Artificial Intelligence (AI) models (i.e., red AI). Companies not only aim to avoid losses (i.e., electricity consumption) but also need

to decide whether it is desirable and reasonable to run an algorithm for many hours. They need to be informed about the existence and possibility of applying green AI methods, which are presented in the recent literature as alternative frameworks that provide similar accuracy using a lower number of features (resulting in lower processing time) [3, 4, 5, 6]. A solution to this trade-off, increasing the number of features and processing time versus lowering the number of independent variables reducing processing time, depends on the management decisions. Further, no formal or scientific methods can support the one to own to his decision in this matter.

Adding to this context, hundreds of different studies exploring applications of Machine Learning (ML) models have demonstrated that ML algorithms can outperform traditional statistical tools in processing time and accuracy [7, 8, 9, 10, 11, 12]. This "new research stream" has aimed to empirically prove that ML has higher predictive power than traditional methods. More recently, the use of hybrid ML methods, which integrate different algorithms (e.g., k-Means, Density-based Spatial Clustering of Applications with Noise (DBSCAN), Decision Tree (DT), Gradient Boosting (GB), Random Forest (RF), Light Gradient Boosting Machine (LightGBM) and Adaboost (AB)). Specifically distinct ML methods has demonstrated across various domains, including medical diagnostics, bio-informatics, recommender systems, and specifically, financial forecasting or consumer segmentation [7, 13] that these models can provide more accurate predictions by effectively handling datasets with a large number of dimensions and combining the strengths of different models while minimizing their limitations [14, 15]. Consequently, by clustering examples before performing predictions, hybrid ML approaches group the number of instances, thereby enhancing the predictive performance of individual models [16, 17, 18, 19, 20]. The effectiveness of hybrid approaches remains a topic of debate in the academic literature. Nevertheless, some researchers recommend that they be employed to investigate the optimal attributes of each method [14, 15, 18]. Therefore, one can assume that the combination of unsupervised clustering algorithms with supervised regressors or classifiers in a prediction task will yield superior outcomes compared to hybrid models that just use supervised learning approaches.

In contrast to the research stream that aims to boost accuracy at any cost (e.g., using ensemble models or hybrid approaches), the green AI strategy supports the idea of adopting simpler models or avoiding enlarged data dimensionality [21]. It claims that reasonable predictive power can be reached at a reasonable processing time, minimizing the use of electricity (i.e., lower carbon footprint emissions). Studies adopting the green AI principles [3, 4, 5, 6] have demonstrated that in some cases, the amount of carbon emission that can be avoided is comparable to the amount generated by driving a car for a day.

ML models are increasingly used to estimate, predict and monitor Customer Relationship Management (CRM) metrics. For instance, individual ML pipelines that aim to forecast credit risk, customer satisfaction, customer value, customer churn, and many other metrics are daily trained and deployed in different industries [22, 18, 19]. While ML has been proven to improve predictive accuracy and maximize profitability for firms, no research to date has explored the costs of these benefits in models that aim to assess CRM

metrics. On the other hand, companies face several challenges when deploying AI models, the key among them being resource constraints and the escalating demand for computational efficiency. Resource constraints often manifest in limited access to high-quality data, insufficient computational power, and the lack of skilled personnel required to develop, manage, and interpret AI systems. As AI technologies advance, the complexity of models increases, demanding more powerful hardware and greater energy consumption. This escalation poses significant issues, particularly for smaller organizations with limited budgets. Furthermore, the increasing emphasis on sustainability necessitates the development of more computationally efficient AI models that minimize energy usage and environmental impact. This drive towards Green AI not only aims to make AI more accessible across varying resource availabilities but also aligns with broader environmental sustainability goals. Consequently, research by academics and practitioners must navigate these constraints while striving to harness the full potential of AI technologies, ensuring they remain competitive and innovative.

This paper aims to provide insights and support management decisions related to adopting Feature Engineering (FE) techniques, Green AI (GAI) principles, or combining both when implementing individual or hybrid ML frameworks to forecast CRM metrics in the financial industry. To accomplish this objective, different experiments are implemented to assess the loyalty of customers from a credit card company [23]. These experiments mainly differentiate how variables are selected to be part of the independent features used to train the different ML models. In the first experiment, Data Preprocessing Treatment (DPT) is performed considering only some steps of data cleaning and preprocessing without applying feature engineering or green AI principles. In the second experiment, both DPT and FE are considered. Here, existing features are explored and many others are created aiming only to increase accuracy. In the third and fourth experiments, DPT+GAI(5) and DPT+GAI(10), the green AI principles are considered by dropping five and ten features from the main dataset (DPT), respectively. Extensive pipelines are explored, frameworks trained through 5 and 10-fold cross-validations are implemented, and the individual (AB, GB, DT, RF, LightGBM) and hybrid ML models are used. For the latter, data is clustered using unsupervised learning (k-Means), and then customer loyalty is predicted using different supervised learning methods applied to the obtained clusters (AB, GB, DT, RF, LightGBM) similar to [19]. However, different from this paper, which only considered maximizing the number of features in the DPT, this article experimented with the reduction of dimensionality, focusing on the green AI aspects of the predictive frameworks. Finally, each experiment's processing time, errors, and feature importance are extracted.

We contribute to the literature in different ways:

- **First and most importantly**, many studies have used ML models to forecast customer loyalty, and some have applied hybrid ML algorithms for this purpose. In this research, not only individual and hybrid ML models are implemented but also results and performance comparisons of these strategies with frameworks that adopt green AI principles are provided.
- **Second**, while no study has contrasted the processing time concerning the number of features used

to assess CRM metrics, the article experiments with other frameworks that use a lower and higher number of independent features than the base dataset. Performance in terms of processing time, accuracy, and most essential features are compared.

- **Third and final**, This article compares the vast amount of results obtained by testing for significant differences and identifying the most critical factors to consider when assessing customers' loyalty across the different regressor-based ML algorithms, deriving insights for managers.

The rest of the paper is organized as follows. Section 2 presents the literature review related to this study. Section 3 presents the ML algorithms used in this study and includes an overview of the experimental setup. Section 4 presents the results and discusses the findings. Finally, Section 6 concludes the study, listing its limitations and presenting some venues for future work.

2. Background

2.1. Machine Learning (ML) Algorithms: Supervised and Unsupervised Methods

Supervised learning algorithms are mathematical methods that use labelled datasets to predict outcomes [24]. There are two types of supervised learning algorithms: classifiers, where the target is a categorical feature, and regressors, where the target is a continuous value. This research aims to forecast customer loyalty, a continuous label (y), based on a set of independent features (X_i). Some of the most commonly used regressor-based ML algorithms are selected to solve business problems, namely, AB, GB, DT, RF, and LighGBM [7, 8, 9, 10, 11, 12, 25]. In fact, Machado et al. [22] have applied LightGBM to forecast customer loyalty on the same dataset that is used in this study. However, besides the comparison of processing time and prediction power between Machado et al. [22] and the present frameworks, this work is built based on their research and different frameworks are implemented to explore the trade-off between feature engineering and green AI.

Unsupervised learning algorithms learn patterns or information not identified from unlabeled datasets [26]. There are different unsupervised learning models that can be grouped based on their primary task. For instance, clustering, dimension reduction, anomaly detection, and topic modelling are some examples of unsupervised learning model categories. Similar to other studies in the financial settings [7, 8, 9, 10, 11, 12], this research applies clustering methods (i.e., k-Means) to group customers before implementing regressor-based models that aim to assess customers loyalty. Also, it is important to highlight that k-Means is selected in this setting, as it is, by far, the most commonly used algorithm to cluster customers in business problems [16].

Given the absence of formal literature that explicitly classifies and certifies an ML model as Green AI (or a green model), this study selects models for experimentation based on their prevalent use in the application area (customer loyalty assessments and CRM metrics predictions). It is important to emphasize that the objective of this research is not to claim or define a specific ML algorithm as green or not. Rather, the aim

is to evaluate existing models to determine which ones perform more favourably in this context, according to the Green AI criteria previously outlined.

The order of complexity and computational costs for these ML models starts with DT, which has the lowest complexity and computational cost due to their relatively simple structure and fast training process. AB follows, with a moderate computational cost, as it sequentially combines multiple weak learners, usually decision trees, to improve model accuracy. RF are more complex and computationally expensive than AB, as they involve building numerous decision trees independently and aggregating their results, which increases both the time and resources required for training. GB is at the higher end of the complexity spectrum, with a significantly higher computational cost due to its iterative process of sequentially optimizing DTs, each correcting the errors of the previous one. Lastly, LightGBM, while also a boosting algorithm, is optimized for speed and efficiency, reducing its computational cost compared to traditional GB, yet still remaining more complex than RFs and AB due to its advanced techniques for handling large datasets and more complex models [7, 8, 9, 10, 11, 12, 25].

2.2. Hybrid ML models

Hybrid ML models integrate multiple simple algorithms to complement and augment each other, specifically entirely different ML techniques [13, 27, 28, 29]. In the financial setting, to boost predictive performance, various studies have clustered customers before implementing regressor or classifiers-based methods [13, 7, 16, 18, 30]. For instance, Machado and Karray [18] compare the predictive performance of individual models (AB, GB, DT, RF, Support Vector Machines (SVM), and Neural Networks) with hybrid frameworks composed by k-Means and all aforementioned individual algorithms when assessing commercial customers' credit scores in financial settings. They find that the combination of k-Means with either DT or RF presents the best validation metrics. Hence, in this study, a set of hybrid frameworks are implemented to compare results with base methods and frameworks that include more or less independent features. The predictive power of all implemented approaches are also compared.

2.3. ML Applications to Assess Customer Loyalty

Due to the increasing market competition that can lead to customer churn, businesses have been investigating efficient methods to assess their clients' CRM metrics, especially identifying practices that can better predict customer loyalty levels.

The literature recommends different methods to measure customer loyalty or to at least provide insights on how to assess this metric. These include the net promoter score [31], repurchase ratio [32], customer engagement score [33], and customer lifetime value [2]. Recently, many studies have used supervised ML algorithms to predict customer loyalty scores, considered as a pre-defined label [34, 35, 22, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47]. These studies are applied to different industries and use different frameworks. Some examples are presented by Buckinx et al. [35], who uses regressor-based algorithms, random forest, and neural networks over behavioural and demographics datasets from customers of a grocery store to

assess their loyalty level. Lee and Jiang [36] apply a hybrid ML approach to classify whether customers of a fashion retailer from China, Taiwan, and Hong Kong were loyal. In their study, they define a loyal customer as a client that remains a top buyer for two years. The hybrid approach consists of using k-means to cluster customers (based on Recency, Frequency, and Monetary (RFM) variables), then classification-based methods (DT, LR, RF, GB) to predict loyalty. Finally, Wassouf et al. [39] assesses customer loyalty in the telecommunication industry. They use more than 125 million examples containing demographic and behavioural features. In their predictive frameworks, they segment customers based on their RFM variables and then assess their loyalty using GB, DT, RF, and Neural Networks (NN).

Table 1 presents a summary of the recent studies using ML algorithms to assess customer loyalty. We can see that there are no studies that have experimented with green AI principles. Also, barely any study highlights feature engineering, the process of converting unprocessed data into significant and predictive characteristics [36], steps applied before implementing the ML algorithms, nor experiments with any hybrid ML approach. Feature engineering is crucial for the effectiveness of ML models. Hence, addressing the gap in studies and exploring the use of feature engineering methods prior to ML (or hybrid ML) implementation in CRM predictive frameworks is of high importance. Moreover, effective feature engineering is crucial in the context of CRM because organizations depend on customer data to make relevant decisions and extract pertinent insights. By looking more closely at this important preprocessing step, practitioners might be able to make CRM models more accurate and easy to understand, find hidden patterns in how customers behave, and ultimately find the best ways to manage their relationships with customers. By bridging this gap, firms gain the ability to make well-informed decisions, allocate resources more efficiently, and provide improved customer experiences, eventually leading to the maximization of their profits and boosting customer satisfaction. The exceptions of studies addressing both gaps are Lee and Jiang [36], Wassouf et al. [39]; however, they cluster customers based solely on their RFM variables and ultimately use classifiers to predict whether a customer is loyal.

Table 1: Summary of recent studies using DPT, FE, or Hybrid ML models in CRM predictive models.

Author	**Settings**	**Green AI**	**Feature Eng.**	**Hybrid Model**	**Main predictive methods[1]**	**Metrics for validation[1]**
Machado et al. [22]	Retail: Credit card	-	✓	-	LightGBM and Xboosting	RMSE
Lee and Jiang [36]	Retail: Fashion	-	-	✓	k-Means, DT, RF, GB	Confusion matrix, lift chart, and ROC/AUC
Sulistiani et al. [38]	Retail: Consumer goods	-	✓	-	SVM and NB	Accuracy, precision, recall, and F1
Wassouf et al. [39]	Retail: Telecommunications	-	✓	✓	GB, DT, RF, NN	Accuracy, recall, precision, and ROC/AUC
Zaki et al. [41]	Commercial customers: B2B	-	✓	-	ANN, Bayesian Network	Accuracy, Lift, and ROC/AUC
Zhang et al. [43]	Retail: Automotive (Repair)	-	✓	-	RF, DT, ANN, and SVM	Accuracy
Latheef and Vineetha [44]	Retail: Banking	-	✓	-	LightGBM, ANN, GB, and CNN	Accuracy and ROC/AUC

Hadiyat [45]	Retail: Tourism (hotel)	-	✓	-	ANN	Accuracy
Surya and Anitha [46]	Retail: Telecom-munication	-	✓	-	RF, LR, KNN, SVM	Accuracy, F1/F2, and ROC/AUC
Khodabandehlou and NikNafs [47]	Retail: Grocery stores	-	✓	-	ANN, and SVM	Accuracy
This paper	Retail: Credit card	✓	✓	✓	k-Means, AB, GB, DT, RF, LightGBM	MAE, MSE, and MedAE

[1] Note: DT refers to Decision Trees, RF is Random Forest, GB is Gradient Boosting, SVM is Support Vector Machines, NB is Naive Bayes, ANN is Artificial Neural Networks, KNN refers to K-Nearest Neighbour, CNN is Convolution Neural Networks, AB is AdaBoost, RMSE is the Root Mean Squared Error, ROC/AUC refers to the Area Under the ROC Curve, MAE is the Mean Absolute Error, MSE is Mean Squared Error, and MedAE refers to the Median Absolute Error.

2.4. Feature Engineering versus Green AI

To effectively discuss the tradeoff between feature engineering and Green AI, it is essential to clarify these concepts. In this study, we rely on established definitions from the literature. Feature engineering involves the process of selecting, transforming, and creating input variables (features) to enhance the performance of machine learning models. This process requires domain knowledge and creativity to design features that improve the model's learning capabilities [22, 39]. Green AI, on the other hand, focuses on developing and using machine learning models that prioritize energy efficiency and environmental sustainability, aiming to reduce the carbon footprint of AI research and applications [21, 48]. In contrast, Red AI emphasizes achieving state-of-the-art performance, often without considering computational costs, which leads to higher energy consumption and a greater environmental impact [21, 48].

Schwartz et al. [21] has demonstrated how much the increasingly large and computationally costly ML algorithms have impacted carbon footprint emission. They define this trend as red AI and claim that the AI community has focused on reaching higher accuracy without considering efficiency. As opposed to this trend, many studies have been focusing on exploring efficient frameworks, for instance, by indicating metrics that enable comparison of processing time [3, 4, 5, 6].

Although the computational cost of Deep Learning (DL) models is not comparable to individual ML algorithms (e.g., regressors and classifiers), the extensive use of feature engineering can lead to an increase in processing time without highly impacting model accuracy [21].

In addition to the ongoing debate about the trade-off between computational power and accuracy, recent studies on the use of AI cover a wide range of research topics that explore the significant effects of AI on society and the environment. One particular aspect, as emphasized by Verdecchia et al. [48], acknowledges the increasing carbon emissions caused by AI, which has given rise to the field of Green AI. The authors provide a comprehensive analysis of the existing body of research on Green AI, which demonstrates a growing level of interest, particularly in the period following 2020. The review primarily highlights tactics such as monitoring the environmental impact of AI models, optimizing hyperparameters, and establishing benchmarks. Vasiliu et al. [49] tackles the difficulties related to insufficient and flawed financial market data in the finance industry. They suggest generating artificial datasets to evaluate AI-powered financial

algorithms, with a particular focus on their use in green investing and trading. Their findings show that synthetic data provides an impartial, comprehensive, and cost-efficient solution. König et al. [50] examine the public's attitudes towards AI, revealing that although individuals prioritize AI transparency, they prefer to disregard concerns regarding energy efficiency (e.g., Green AI). These findings have substantial ramifications for AI policy and regulation. It is critical because recognizing the public's prioritization of AI transparency over energy efficiency underscores important factors to be taken into account when formulating AI policies and regulations that are in line with society's preferences and values. Furthermore, Yigitcanlar et al. [51] emphasizes the necessity of adopting a Green AI strategy in smart cities while also criticizing previous AI approaches that were overly focused on technology. This viewpoint paper argues in favour of AI systems that promote efficiency, sustainability, and equity. It does so by relying on a comprehensive analysis of AI and smart city literature to provide guidance to urban planners and authorities. This recent literature [48, 49, 50, 51] offers a thorough perspective on the complex issues and opportunities related to Green AI. It covers several aspects, such as environmental sustainability, constraints in financial data, public preferences, and the impact of AI on creating smarter and more sustainable urban futures.

In the financial setting, the Bank for International Settlements (BIS) has already recommended that banks and financial institutions cooperate with each other and with their central banks to mitigate the environmental costs of big data [52]. It also mentions that feature engineering, which can involve merging datasets from different sources to include more features into predictive frameworks to boost their accuracy, can increase processing time and result in non-efficient models, leading to red AI.

Similar to Machado et al. [22], Lee and Jiang [36], and even more closely to Wassouf et al. [39], the loyalty of credit card company (Elo) customers is assessed. However, in the present experiments, the value of using feature engineering in predictive frameworks is investigated. Models to reduce the number of features to verify the impact on accuracy are implemented. The proposed models are implemented using individual ML regressor-based algorithms (AB, GB, DT, RF, LightGBM) and a hybrid approach, where at first customers are clustered (based on a set of variables) using k-Means, then the regressor-based algorithms (AB, GB, DT, RF, LightGBM) to predict loyalty are implemented. To evaluate the current frameworks, 5 and 10-fold cross-validation are implemented, validation metrics (MSE, MAE, and Median Absolute Error (MedAE)) are extracted, and paired-wise statistical tests to verify the equality of these approaches are performed.

2.4.1. Balancing Computational Efficiency and Sustainability in AI: Insights from Recent Research

The increasing emphasis on environmental sustainability in artificial intelligence has sparked a growing interest in Green AI, which focuses on optimizing energy efficiency and reducing the computational resources required by AI models. Dehghani and Larijani [53] developed an efficient algorithm for predicting stock market indices by integrating the MID algorithm with a neural network, using only seven features. This approach not only enhances prediction accuracy but also significantly reduces computational overhead, aligning with the principles of Green AI. By minimizing feature sets and optimizing computational efficiency, the study demonstrates how sustainable AI models can be developed for financial forecasting

without compromising performance.

In a related context, Qiu and Wang [54] explored the application of machine learning for credit card customer segmentation, highlighting the importance of efficient clustering algorithms in managing large datasets typical of customer behaviour analysis. The study systematically evaluates various clustering models to determine the most effective approach, ultimately contributing to economic stability by optimizing resource usage and model performance. This work emphasizes the potential of Green AI in the financial sector, where computational efficiency can directly impact economic outcomes.

Similarly, the work by Tehranian [55] on predicting economic recessions leverages machine learning techniques, such as RFs, GB, and ANN, alongside traditional econometric models. The study's use of feature-reduction techniques highlights the necessity of managing computational resources, especially when dealing with extensive and high-dimensional data. This reflects a conscious effort to balance model performance with the environmental and economic costs associated with AI, a critical aspect of the ongoing debate between Green and Red AI.

Moreover, in the domain of cloud computing, Alahdadi et al. [56] proposed a resource allocation mechanism designed to improve truthfulness and efficiency in double auctions. Given the significant computational demands of cloud-based AI models, this study is particularly relevant as it addresses the environmental and economic sustainability of AI deployments. By encouraging honest bidding and improving budget balance, Alahdadi et al. [56] 's work contributes to reducing the carbon footprint of cloud computing, further advancing the goals of Green AI.

Collectively, these studies illustrate the growing importance of developing AI algorithms and models that prioritize both computational efficiency and environmental sustainability. Whether through optimized feature selection, efficient clustering, or resource-conscious algorithm design, these efforts are paving the way for more sustainable and responsible AI practices across various sectors.

3. Materials & Methods

3.1. Experimental Setup

An overview of the designed framework, presenting the tasks performed to accomplish this study's objectives, is shown in Figure 1. At first, data prepossessing treatment is applied to the entire dataset, in which individual regressor-based algorithms and hybrid ML models are directly applied to assess customer loyalty (the baseline ML application). Then, using the pre-processed data, two other different frameworks are explored. First, feature engineering to increase the number of independent variables aiming to maximize accuracy is examined before the individual and hybrid ML implementations. Second, an experiment is explored to analyze the relationship between the number of independent features used, processing time, and accuracy of the predictive models. In this scenario, features are dropped based on their importance in the baseline application. Two experiments are considered; first, the five most minor essential features are dropped and then the ten least important features. This experiment follows the main concepts from the

green AI practices [21] and can support the decision on model selection and implementations given their explainability and accuracy. Finally, the last task is to extract validation metrics and implement statistical tests to evaluate the different algorithms' performance. Each of these steps are described in the following subsections.

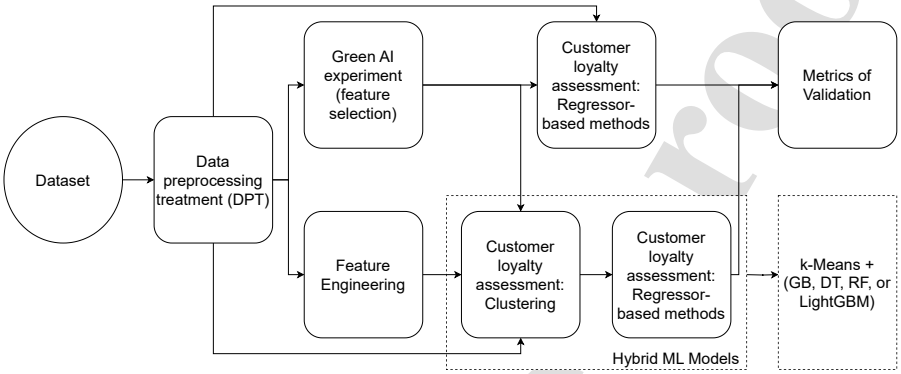


Figure 1: Schematic framework for this study.

3.1.1. Data, Experimental Design, and Parameter Setting

Experiments are conducted using data from EloGroup, one of the largest payment brands in South America. EloGroup was formed in 2010 based on a partnership of three of the largest banks in Brazil: Bradesco, Caixa Economica Federal, and Banco do Brasil, and by the end of 2021, they had issued more than 140 million credit cards. Besides their portfolio of products, they also offer multiple services, such as instalment payments for customers [23].

To expand and build on their partnership with merchants and strengthen their relationship with customers, they have implemented ML models to understand customers' preferences and behaviour. Thus, assessing customer loyalty at an individual level can drive different policies and support decision-making when implementing various marketing-related campaigns. The Kaggle competition they proposed in 2018 required data scientists' predictive models with high accuracy for their label customer loyalty score, a target given in their datasets [23]. Although most features were anonymous, they made 28 features available with more than 325 thousand examples. These described historical transactions (e.g., purchase date and the amount and the number of instalments), merchant characteristics (e.g., the business where cards were used, average sales, and identification group), and more general information within their train dataset (e.g., the first month of activation, and target). Descriptive statistics for some of these features are presented in Table 2.

The experiments in this research aim to assess the customer loyalty of each credit card owner in the dataset and to compare the predictive performance of the different implemented frameworks. Namely, individual and hybrid ML models are applied in the (i) preprocessed dataset, (ii) preprocessed and augmented datasets using feature engineering methods, and (iii) preprocessed and low-dimensional dataset, following

Table 2: Descriptive statistics of relevant features among different samples.

Descriptive Indicator	Variables							
	Target	New Merchant	New Purchase	New Merchant Category	Hist Merchant	Hist Merchant Category	Hist Subsector	Elapsed Time
Mean	-0.0308	54.0000	2281214.00	18.0000	33.7163	18.2867	11.4624	380.3197
St. Dev.	0.0000	0.0000	1825757.00	0.0000	30.7445	11.4826	5.1853	292.9694
Minimum	-17.6093	0.0000	12326.2500	1.0000	1.0000	1.0000	1.0000	0.0000
Q1	-0.8447	2.0000	2518292.00	2.0000	13.0000	10.0000	8.0000	153.0000
Q2	-0.0077	4.0000	3961418.75	4.0000	24.0000	16.0000	11.0000	306.0000
Q3	0.7773	8.0000	1.0000	7.0000	44.0000	24.0000	15.0000	488.0000
Maximum	17.9687	109.0000	5343723.00	40.0000	398.0000	95.0000	34.0000	2284.00

main green AI principles. The ML algorithms used in this study are k-Means (in the hybrid models) and AB, GB, DT, RF, and LightGBM to forecast the loyalty level. The k-means clustering algorithm was selected as a component of the hybrid ML experiments for many reasons, including Firstly, due to its ability to be interpreted and its capacity for segmentation; k-means clustering is an algorithm that can categorize examples into separated segments or clusters based on their similarities [57]. In the present example, by dividing customers into different clusters, valuable insights can be obtained into their behaviours and characteristics, which can aid in understanding why certain customers are more loyal than others. These clusters can function as significant attributes in the hybrid models, enabling them to capture fluctuations in consumer behaviour that impact loyalty [18, 16]. Furthermore, through the process of clustering, the noise can be reduced in the data and make the model more robust [57]. This can result in better predictions of customer loyalty scores, as the model is better equipped to capture underlying trends and patterns in the data. It can also help prevent overfitting, which is a common issue in predictive modelling. Finally, among the clustering algorithms commonly employed in corporate settings, k-Means has been the subject of a greater number of earlier studies. Also, it is important to highlight that the selection of regressor-based algorithms (AB, GB, DT, RF, and LightGBM) is motivated by two factors. Firstly, these methods have been widely used by academics and practitioners to forecast CRM-related metrics [7, 8, 9, 10, 11, 12, 25]. Secondly, the aim is to compare the predictive power with benchmark studies that utilized LightGBM to forecast customer loyalty using the same dataset [22]. Some of the hyperparameters tuned were the learning rates (0.001, 0.01, and 0.1), the maximum number of iterations (2,500, 5,000, and 10,000), the minimum number of observations per leaf (10, 50, 100, and 150), the number of trees (10, 100, 200, 500, and 1000), the minimum number of leaves (5, 10, 20, 30, 40, 50, and 100), and the number of clusters (3, 4, and 5). To support the validation of the present methods, cross-validation ($5x2$, and $10x2$) are implemented in all supervised learning algorithms [7]. This means that the dataset is randomly divided into two parts: one is retained for validation, and the other is used as training data. The cross-validation method is repeated five (or ten) times, and results are averaged to obtain one individual assessment. Also, validation metrics are extracted to compare the results from individual and hybrid ML models. Then, the feature importance of the different frameworks are extracted, which is computed by assigning a score to the features that are most

used to split data when growing trees in the other supervised learning models implemented (GB, DT, RF, and LightGBM). Finally, a pairwise t-test is implemented to assess the equality of the various frameworks [58] and analyze the trade-off between feature importance and green AI and its impact on this application's management and decision-making process.

3.1.2. Data Preprocessing Treatment (DPT)

At the data preprocessing treatment stage, the following tasks are implemented:

1. The dataset is cleaned, and duplicated features are removed;
2. Categorical features were treated by implementing an ordinal or one hot encoder depending on the number of possible categories in each feature;
3. The outliers are treated as follows. Critical cases, where examples have more than 1% of outliers, were evaluated by creating dummies to analyze their behaviour and included in the predictive frameworks. If outliers represented less than 1% of examples for a given feature, they were dropped from the dataset;
4. The missing values are handled depending on their frequency across the different features. When less than 5% of examples were missing in a given feature, they were replaced by the mean, mode, or median. Otherwise, dummies were created to analyze the existence of a hidden and unknown reason for the missing pattern for that given feature;
5. Correlation was tested by measuring the variance inflation factor. In cases where the correlation is over 90%, features are dropped from the final dataset, and
6. Different methods of scaling were tested and compared, and a standardization process, where the mean is removed, and data is scaled to unit variance, was selected based on the predictive performance of baseline models when this method was used.

Two other steps in the DPT tasks relate to feature engineering and green AI experiments implemented in this study. Figure 2 represents the experiments' main idea, showing their major outputs (reducing or increasing the number of features). For instance, the DPT tasks are used in the primary dataset (Section 3.1.2), then separately feature engineering and green AI are implemented:

1. When implementing feature engineering in the dataset, the aim is to identify and create new variables based on the existent information and test their use or relevance when assessing the given target (i.e., customer loyalty). For example, the present primary dataset shows the monthly purchase and sales amount. Therefore, one can create the average, median, minimum, maximum purchase, and sales amount by month features. In the present study, more than fifty features are created, and thus, the frameworks using feature engineering (FE) contain a training set of 85×199.710 dimension. Mathematically, the feature engineering process can be represented as:
   (a) **Dataset Description:** The primary dataset D contains monthly purchase and sales data. Assume N records in the dataset.

(b) **Features from the Dataset:** Multiple statistical features such as average, median, minimum, and maximum are derived for multiple features. Assume:

- $f_1, f_2, \ldots, f_k$ for purchase-related features.
- $g_1, g_2, \ldots, g_k$ for sales-related features.

(c) **Total Features Created:** Denoted as K where $K > 50$.

(d) **Framework for Feature Engineering:** The resulting dataset D' after feature engineering has dimensions of 85 rows × 199,710 columns.

(e) **Mathematical Representation of the Transformation:**

- Let x_{ij} represent the data for the j-th month in the i-th record of the original dataset D, where $i \in \{1, 2, \ldots, N\}$ and j represents the months.
- Define transformation functions $T_1, T_2, \ldots, T_K$ where each function T_k takes monthly data x_{ij} and computes a new feature. For example, $T_1(x_{ij})$ might compute the average purchase amount for month j in record i.

(f) **Constructed Feature Set:** Each record i in D transforms into a set of K features per month in D'. Hence, the transformed record i can be represented as:

$$\text{record}'_i = [T_1(x_{i1}), T_2(x_{i1}), \ldots, T_K(x_{i1}), \ldots, T_1(x_{iM}), \ldots, T_K(x_{iM})]$$

where M is the number of months recorded in D.

Final Form of the Dataset D':

$$D' = \{\text{record}'_1, \text{record}'_2, \ldots, \text{record}'_{85}\}$$

This mathematical representation highlights how the original dataset D is transformed into D' through the application of multiple feature transformation functions $T_1, T_2, \ldots, T_K$, significantly expanding the dataset's dimensions and potentially enhancing its utility for assessing, in this case, customer loyalty.

2. When implementing green AI, the number of features is gradually reduced. The variables are dropped based on their relative importance when assessing customer loyalty using the primary dataset (DPT). First, the worst five features are dropped, and then ten features are not included in a second scenario. As a result of these two scenarios exploring green AI aspects, the training datasets will have 23 and 18 for the Green AI minus 5 (GAI5) and minus 10 (GAI10) feature cases respectively.

In summary, we use the DPT as a baseline to assess the number of features. To increase the number of features, we apply feature engineering as described above. For the Green AI experiments, we reduce the number of features based on their importance in predicting the target within the baseline dataset.

The primary motivation to explore the use (or not) of feature engineering and green AI philosophy in financial settings relates to the trade-off of processing time versus predictive performance. Feature engineering methods are well-known techniques that significantly impact the accuracy of ML models [59, 60, 61]. Green AI principles explore how the number of features impacts predictions. This exploration is done by assessing

the value of a model concerning the time (and energy) that can be saved by reducing the number of independent features [21, 51]. In this study, these two cases are implemented and a set of benchmark models (that only implement DPT methods, as shown in Section 3.1.2) considering using individual and hybrid[1] ML models. All the results referring to predictive performance, processing time, and feature importance are collected, contrasted, and discussed.

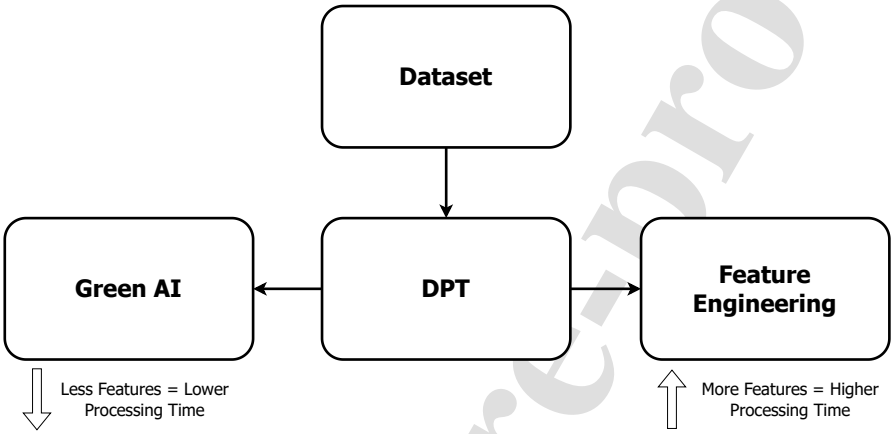


Figure 2: Experiments conducted in this research.

3.1.3. Evaluation Criteria

The customer loyalty assessment is evaluated through the extraction of metrics of validation commonly used in regression-based algorithms, described in Section 2.1 [7, 8, 9, 10, 11, 12, 22]. Also, cross-validation (10 and 5-Fold) is implemented to evaluate individual and hybrid models [62] and pairwise t-tests [58] is performed to compare the statistical difference between the different explored frameworks.

4. Results and Discussion

This section is organized as follows:

1. The preliminary results related to the study are presented (e.g., correlation analysis and target frequency distribution);
2. Predictive performance of the different frameworks is explored, and feature importance analysis is assessed;
3. Processing time of the various models are compared;
4. Results are discussed and compared.

[1] *The hybrid ML models follow an integrated approach that combines unsupervised learning, k-Means clustering, and supervised ML, several individual regressor-based algorithms in series. No further elaboration on these algorithms' main models is provided.*

4.1. Preliminary Results

Table 2 shows the mean, standard deviation, minimum, maximum, first, second, and third quartile values for some of the most important features used in the experiments. The information is presented before performing the respective transformations as shown in Section 3.1.2. For instance, the average loyalty level (target) is -0.0308, while its maximum and minimum values are 17.9687 and -17.6093, respectively. Also, the standard elapsed time is 380.31s while the minimum elapsed time is zero, and its maximum value is 2284s.

Figure 3 presents the number of features, or independent variables, used in each experiment. Originally, 28 features are considered (DPT case). For the feature engineering experiment, 85 features are taken into account (DPT+FE case). In the green AI experiments, 5 (DPT+GAI(5) case) and 10 (DPT+GAI(10) case) features are dropped from the original dataset, reaching 23 and 18 features, respectively.

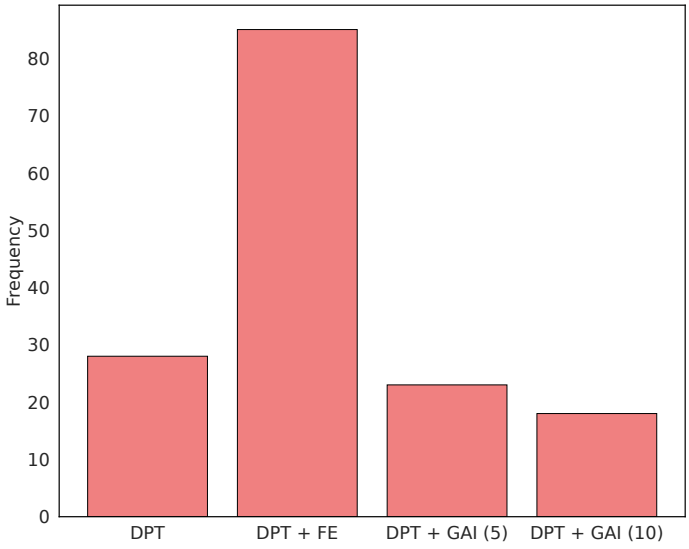


Figure 3: Number of features in each experiment.

Figure 4 shows the frequency distribution for the target variable (customer loyalty score) after outliers treatments (as shown in Section 3.1.2). It is clear that loyalty levels average around zero, with customers having positive and negative scores in a symmetrical distribution. A normality test considering 90% confidence shows that this distribution is normal with an average of zero. The company does not explain the meaning of the target variable in its dataset. Still, it is implicit that higher positive values represent higher loyalty levels. For instance, it is easier to see in Table 2 that the highest customer loyalty score is 17.96 and the lowest is -17.60.

Figure 5 presents the correlation among the features used in this study (DPT). It is clear that the target is not well correlated to any other feature, and only a few variables are strongly correlated. For instance, the

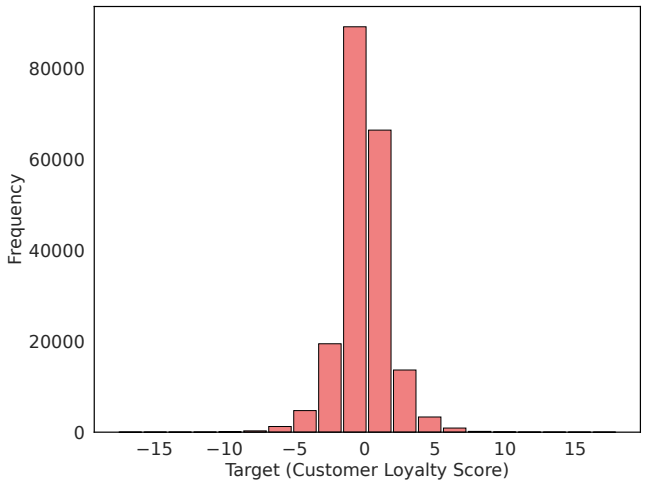


Figure 4: Frequency distribution of the target variable.

elapsed time has a correlation of -0.9 with the year. Also, the history of the subsector has a correlation of 0.9 with the history merchant category, and the merchant category is also 90% correlated with the history merchant. The reasons for the above results are illustrated in Section 3.1.2.

4.2. Predictive Performance

According to the literature, the most commonly used ML algorithms applied in business [7, 8, 9, 10, 11, 12] are implemented. To facilitate visualization and analysis and to enable comparison with other ML studies that explored the same dataset in the past, the different error measures (MSE, MAE, and MedAE) is presented for the individual and hybrid models across the different ML algorithms (AB, GB, DT, RF, LightGBM). The results presented in this section focus on the experiments that used a 10-Fold cross-validation method[2]. Table 3 presents the errors for each experimental framework, contrasting the performance of the individual and hybrid approaches.

4.2.1. Individual versus Hybrid Models Throughout the Different Experiments

It is clear in Table 3 that GB, DT, and LightGBM metrics improve when using the hybrid models compared to individual ML models. The error range across the different experiments averages 0.1. For instance, MSE for the hybrid models using LightGBM ranges from 0.86 to 1.17, and the error range for the hybrid models using GB reaches a minimum of 0.84 and a maximum of 0.91. Moreover, the excessive use of feature engineering does not imply reaching a significantly lower error. For example, the individual use of GB

[2]Tables A1 and A2 in the appendix contain the results for the experiments using a 5-Fold cross-validation for both individual and hybrid models, respectively. Table A3, also in the appendix, presents the results for the hybrid models (10-Fold) open by clusters.

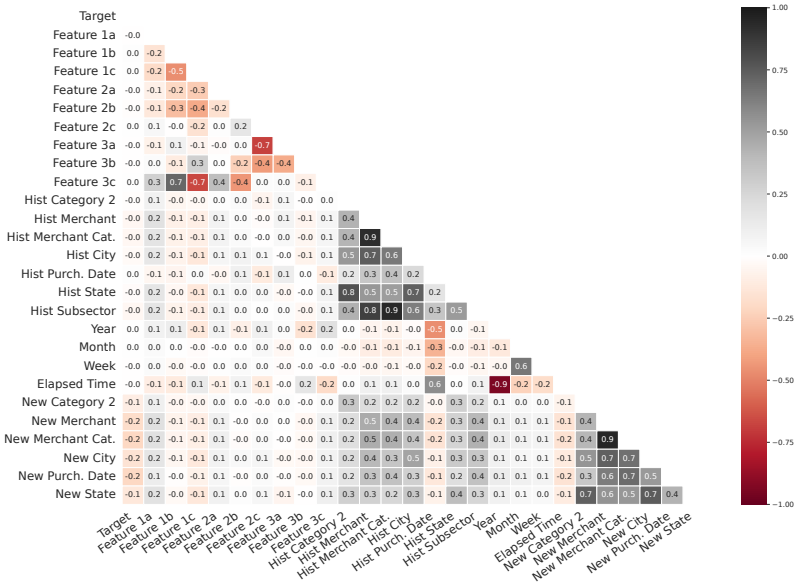


Figure 5: Correlation among the different independent features.

provides an MAE of 0.66 in the DPT+FE case and an MSE of 0.67 in the case of DPT+GAI(10), meaning that using 85 or 18 independent features to assess the customer loyalty score will result in an error difference of 0.01.

Specifically, by contrasting the results in the first experiment (DPT), one can notice that the error is reduced for all cases when using the hybrid approach except when AB is applied. For instance, the maximum error reduction occurs for DT, from 1.02 when using individual models to 0.79 when hybrid models are used. The minimum reduction in error is observed for RF and LightGBM, with a 0.01 difference between the individual and hybrid models. In the other experiments (DPT+FE, DPT+GAI(5), and DPT+GAI(10)), a similar

Table 3: Metrics of validation extracted for all individual ML models.

Metric	Experimental Framework	Adaboost		Gradient Boosting		Decision Tree		Random Forest		LightGBM	
		Individual	Hybrid	Individual	Hybrid	Individual	Hybrid	Individual	Hybrid	Individual	Hybrid
MAE	DPT	0.6862	0.7840	0.6758	0.6399	1.0283	0.7974	0.6928	0.6860	0.6733	0.6681
	DPT+FE	0.6758	0.7746	0.6615	0.6323	0.9690	0.6699	0.6559	0.7680	0.6447	0.6585
	DPT+GAI(5)	0.6861	0.7839	0.6757	0.6400	1.0245	0.7783	0.6949	0.6963	0.6735	0.6687
	DPT+GAI(10)	0.6862	0.7842	0.6758	0.6400	1.0266	0.7777	0.6970	0.6974	0.6740	0.6686
MSE	DPT	0.9253	1.1859	0.9246	0.9191	1.9867	1.2267	0.9419	0.9328	0.9069	0.9371
	DPT+FE	0.8968	1.1601	0.8872	0.8415	1.7779	1.1767	0.8478	0.8110	0.8284	0.8633
	DPT+GAI(5)	0.9253	1.1860	0.9247	0.8886	1.9619	1.2030	0.9474	0.9119	0.9073	1.1717
	DPT+GAI(10)	0.9254	1.1862	0.9248	0.8890	1.9727	1.2013	0.9517	0.9092	0.9083	1.1713
MedAE	DPT	0.4903	0.6550	0.4684	0.6499	0.7505	0.6586	0.4957	0.6343	0.4742	0.6256
	DPT+FE	0.4909	0.6412	0.4565	0.6418	0.7050	0.6107	0.4692	0.6226	0.4566	0.6115
	DPT+GAI(5)	0.4902	0.6549	0.4684	0.6496	0.7519	0.6279	0.4983	0.6498	0.4742	0.6269
	DPT+GAI(10)	0.4902	0.6554	0.4684	0.6497	0.7518	0.6278	0.4990	0.6486	0.4744	0.6266

pattern is observed: the error decreases for frameworks that use GB, DT, and RF, and it increases when AB and LightGBM are used. For instance, when using DT in the DPT+FE experiment, error reduction reaches an absolute difference of 0.6 (regarding MSE metric). In contrast, the maximum increase in error occurs for the use of LightGBM in DPT+GAI(5) and DPT+GAI(5), with an absolute error difference regarding MSE metric (between individual and hybrid results) of 0.27.

Regarding the MedAE, one can observe across the experiments' outputs (Table 3) that using hybrid ML models increases the error. This happens for all four experiments (DPT, DPT+FE, DPT+GAI(5), and DPT+GAI(10)) and all ML models used, except for the DT algorithm. For instance, the GB algorithm in the DPT experiment provides a MedAE of 0.46 and 0.64 in the individual and hybrid frameworks, respectively. Similarly, using AB in the DPT+GAI(5) experiment results in errors of 0.49 (individual) and 0.65 (hybrid models). These results are related to the MedAE intrinsic characteristics, as the median is unaffected by values at the tails.

Further comparisons are presented in Figure 6 and A1. These figures present the MSE of the different predictive models (AB, GB, DT, RF, and LGBM) for each experiment (individual and hybrid). First, for the DPT and DPT+GAI(5) experiments, the MSEs are statistically equal with 95% confidence. However, for the other experiments, one can see that the individual models present a higher error in the individual approaches for GB, DT, and LightGBM. Performing a statistical test, with 95% confidence is ensured that they are, on average, greater for individual models on those algorithms than when using their hybrid versions.

Because they can capture more complex data relationships, hybrid models outperform individual models in most cases [19, 20]. For example, a hybrid model that combines k-Means with DT might be taking advantage of the k-Means Euclidean distances computations between the examples and the DT's interpretability to capture non-linear relationships. However, hybrid models may not outperform individual models in other cases [63, 64]. This can happen if models are highly accurate and there is little room for improvement, causing overfitting or underfitting to lower accuracy if the hybrid model is not designed correctly or trained, or it might be related to the characteristics of the problem being addressed (dataset).

4.2.2. Feature Importance

Feature importance is used to provide insights into models' explainability [7, 17] and list the important factors is listed when assessing customer loyalty. For that, the top features for each of the five algorithms implemented (AB, GB, DT, RF, and LightGBM) is listed when extracting their errors through 10-fold cross-validation, as they provide, on average, lower errors than the ones obtained in the 5-Fold cross-validation strategy (Table 3). Figures 7 and 8 present the top five features per experimental framework (DPT, DPT+FE, DPT+GAI(5), and DPT+GAI(10)) and per individual ML models (AB, GB, DT, RF, and LGBM). Similar plots for hybrid ML models (per model: AB, GB, DT, RF, and LGBM) are presented in the Appendix A2. The scores were scaled, with the most important variable having a score of 100% and computing the other features' relative importance from the most relevant factor. Table 4 presents the top 3 features in each experiment and every model when assessing customer loyalty.

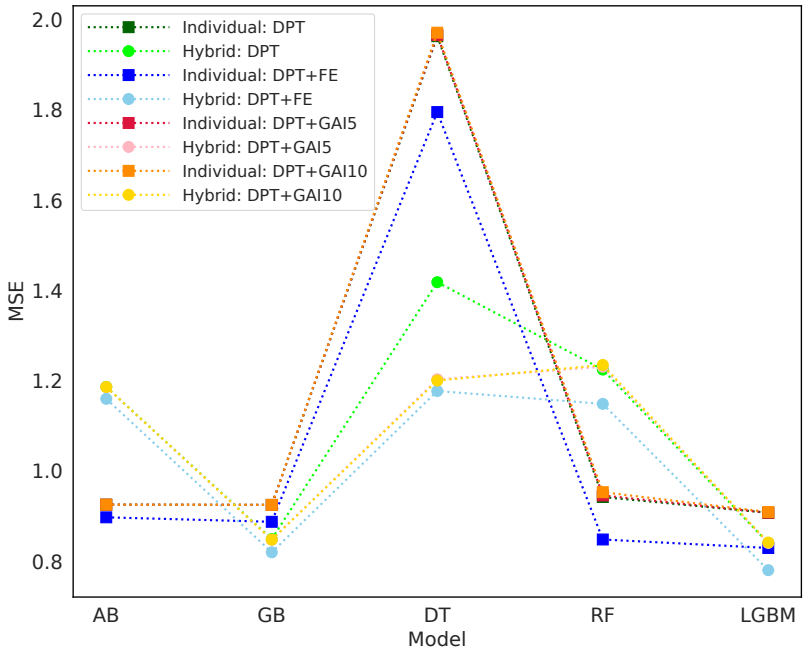
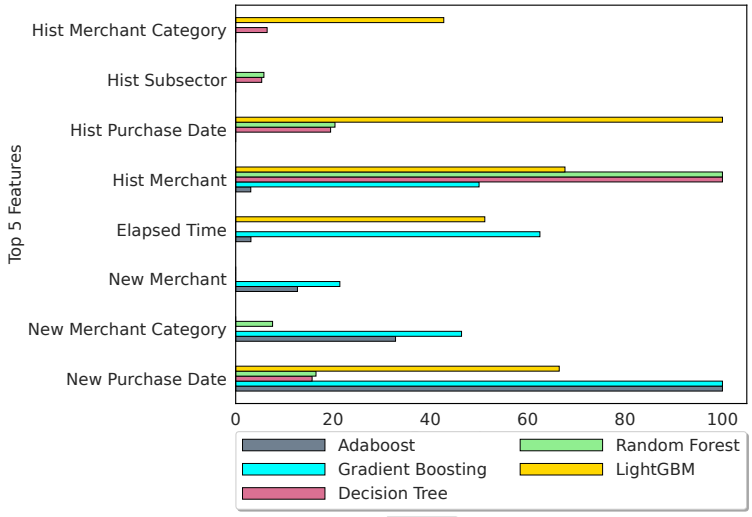


Figure 6: MSE for the different frameworks: 10-Fold.

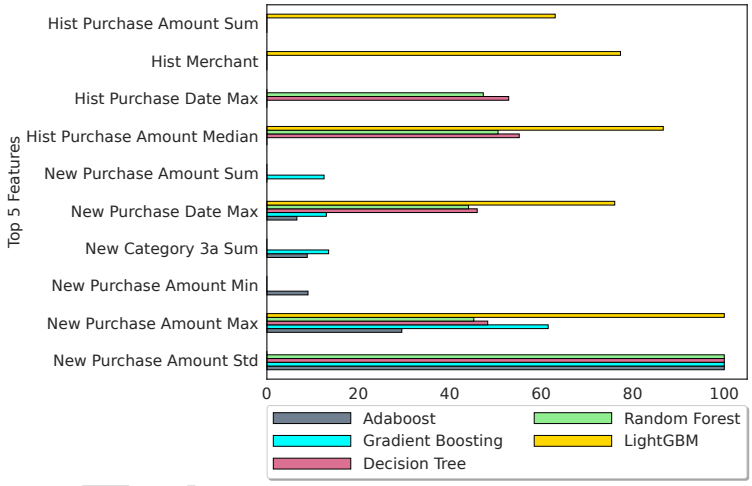
The results demonstrate that the most critical feature for forecasting customer loyalty of credit card firms for all these frameworks and tests is either Hist Purchase, New Purchase Date, or Hist Merchant. Similar results could not be found in the literature, as this is the first study experimenting with the number of features used to assess customer loyalty. However, one can compare the current analysis with the results from the competition where the company enquires data scientists and Kaggle users to propose predictive models to assess their customer loyalty levels [23]. The teams that ended the competition in the first three positions found that the essential features are Purchase Data and History Merchant. Therefore, these findings are consistent with work done using the same dataset [23]. However, some novelty can be observed. For example, one can notice that in the DPT+FE experiment, the top features listed were created during the data exploration analysis phase (e.g., Hist Purchase Amount Sum, New Purchase Amount Maximum, Hist Purchase Amount Maximum). These features are the most critical variables in the present study; however, they are not mentioned in the top 5 features listed by the practitioners in the competition [23].

4.3. Processing Time

Figure 9 presents the processing time of individual and hybrid ML models for all experiments (DPT, DPT+FE, DPT+GAI(5), and DPT+GAI(10)), for 5- and 10-fold cross-validation and all ML models ex-

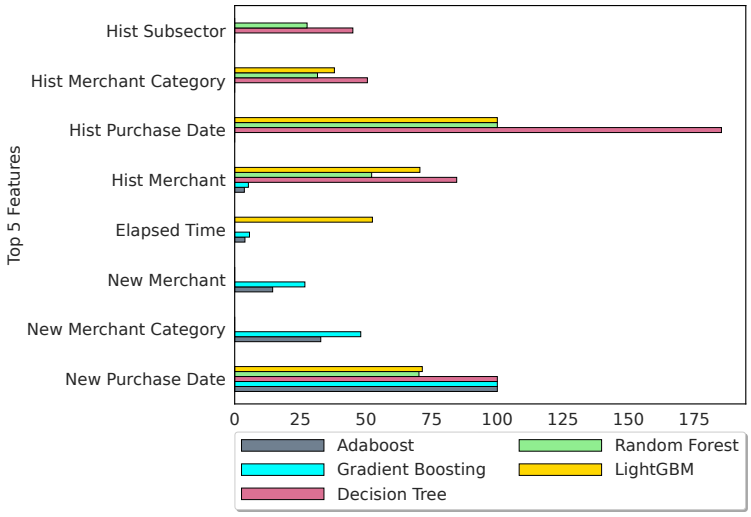


(a) Feature Importance - Individual Models: DPT Framework.

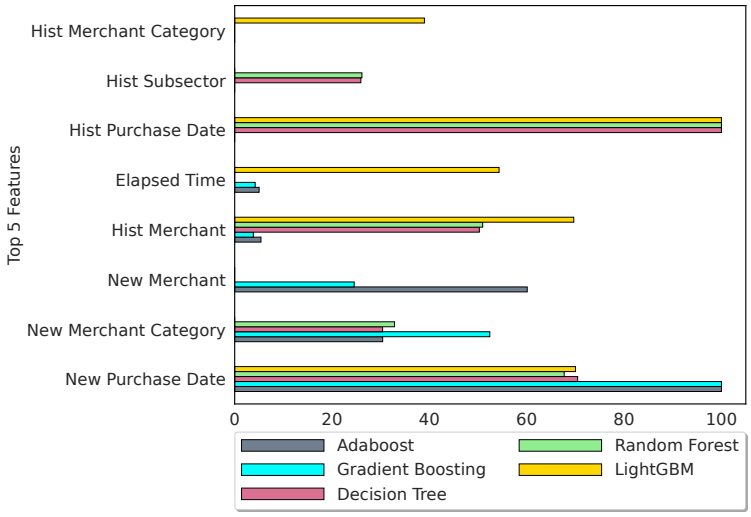


(b) Feature Importance - Individual Models: DPT+FE Framework.

Figure 7: Feature Importance for the Different Experiments (DPT & DPT+FE) in the Individual Frameworks, using 5-Fold cross-validation strategy.



(a) Feature Importance - Individual Models: DPT+GAI(5) Framework.



(b) Feature Importance - Individual Models: DPT+GAI(10) Framework.

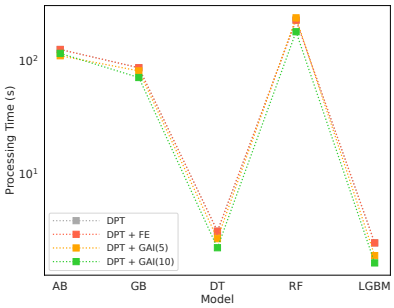
Figure 8: Feature importance for the different experiments (DPT+GAI(5) & DPTGAI(10)) in the individual frameworks, using a 5-Fold cross-validation strategy.

Table 4: Feature Importance (Top 3 Features) per Experiment and Model used (red (and italic) and black fonts refer to individual and hybrid models, respectively).

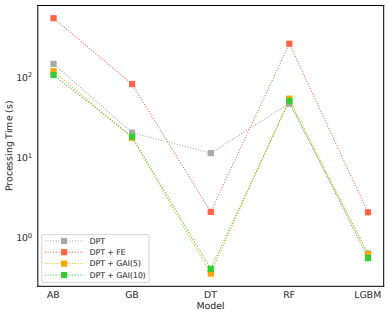
Experiment	Model	Top 3 Features in Each Experiment and Framework					
		1st Feature (Importance 100%)	2nd Feature (Name and Importance)		3rd Feature (Name and Importance)		
			Name	(%)	Name	(%)	
DPT	AB	*New purchase date*	*New merchant category*	*32.8*	*New merchant*	*12.7*	
		New purchase date	History merchant category	41.3	New merchant	18.4	
	GB	*New purchase date*	*Elapsed time*	*62.5*	*History merchant*	*50*	
		New purchase date	History merchant category	40.5	History merchant	18.1	
	DT	*History merchant*	*History purchase date*	*19.6*	*New purchase date*	*15.7*	
		History merchant	History purchase date	24.9	History purchase date	13.4	
	RF	*History merchant*	*History purchase date*	*20.4*	*New purchase date*	*16.5*	
		History merchant	New purchase date	19.4	New merchant	12.1	
	LightGBM	*History purchase date*	*History merchant*	*67.6*	*New purchase date*	*66.4*	
		History purchase date	New purchase date	71.5	New merchant	62.6	
DPT + FE	AB	*New purchase amount std*	*New purchase amount max*	*29.5*	*New purchase amount min*	*8.9*	
		New purchase amount max	New purchase amount std	60.3	New purchase amount min	49.8	
	GB	*New purchase amount std*	*New purchase amount max*	*55.2*	*New category 3a*	*52.8*	
		New purchase amount max	New purchase amount std	49.8	New purchase amount min	46.6	
	DT	*New purchase amount std*	*History purchase amt median*	*55.1*	*History purchase date max*	*52.8*	
		History authorization flag mean	New purchase amount std	96.2	History purchase amount min	91.5	
	RF	*New purchase amount std*	*History purchase amt median*	*50.5*	*History purchase date max*	*47.3*	
		History authorization flag mean	New purchase amount std	97.1	History purchase amount max	87.7	
	LightGBM	*New purchase amount max*	*History purchase amt median*	*88.6*	*History merchant*	*77.3*	
		New purchase amount max	History purchase amount min	88.4	History purchase amount max	71.2	
DPT + GAI(5)	AB	*New purchase date*	*New merchant category*	*32.7*	*New merchant*	*14.4*	
		History purchase date	New purchase date	56.3	New merchant category	39.2	
	GB	*New purchase date*	*New merchant category*	*48*	*New merchant*	*26.7*	
		History purchase date	New purchase date	47.7	New merchant category	33.3	
	DT	*New purchase date*	*History merchant category*	*85*	*History merchant*	*84.5*	
		History purchase date	New purchase date	93.6	History merchant category	82.9	
	RF	*History purchase date*	*New purchase date*	*70.1*	*History merchant*	*52.1*	
		History purchase date	New purchase date	95.1	History merchant category	87.6	
	LightGBM	*History purchase date*	*New purchase date*	*71.4*	*History merchant*	*70.5*	
		History purchase date	History merchant	68.4	New purchase date	62.2	
DPT + GAI(10)	AB	*New purchase date*	*New merchant*	*60.1*	*New merchant category*	*30.4*	
		History purchase date	New merchant	35.3	Elapsed time	33.5	
	GB	*New purchase date*	*New merchant category*	*52.4*	*New merchant*	*24.5*	
		History purchase date	New merchant	37.8	Elapsed time	31.9	
	DT	*History purchase date*	*New purchase date*	*70.4*	*History merchant*	*50.2*	
		History purchase date	New purchase date	93.5	History merchant	91.1	
	RF	*History purchase date*	*New purchase date*	*67.9*	*History merchant*	*50.9*	
		History purchase date	New purchase date	88	History merchant	86.2	
	LightGBM	*History purchase date*	*New purchase date*	*70*	*History merchant*	*69.6*	
		History purchase date	History merchant	62.7	New purchase date	62.7	

plored (AB, GB, DT, RF, and LGBM). Also, as for analyzing the processing time of the different frameworks, it is relevant to contrast the performance (related to processing time) regarding the use of a different number of folders in the cross-validation methods across all different ML algorithms used (AB, GB, DT, RF, and LightGBM). As expected and demonstrated in the literature [65, 66], it shows that the 5-fold cross-validation has a lower processing time than the 10-fold for individual and hybrid ML models. Also, it demonstrates that in the green AI experiments (DPT+GAI(5) and DPT+GAI(10)), the processing time is lower than when implementing feature engineering strategies (DPT and DPT+FE). For instance, for the 10-fold cross-validation and when using the hybrid ML models, DT models finalize training in approximately one second (DPT+GAI(10)). At the same time, it takes more than ten seconds for the DPT+FE strategy; as the dimension of the dataset used as input in the different ML frameworks increases, the order of complexity of these models grows at the same levels. Which model is selected for deployment would depend

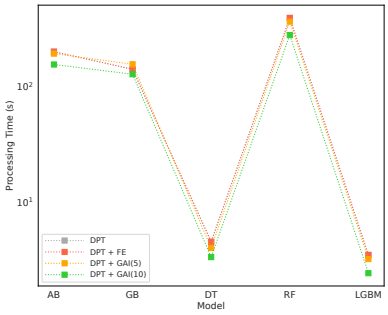
on management decisions in the trade-off of accuracy versus processing time.



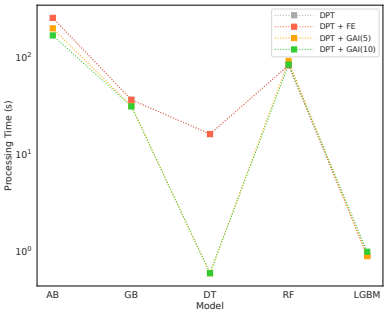
(a) Processing time for the implementations done with a 5-fold cross-validation strategy through the use of individual ML models.



(b) Processing time for the implementations done with a 5-fold cross-validation strategy through the use of hybrid ML models.



(c) Processing time for the implementations done with a 10-fold cross-validation strategy through the use of individual ML models



(d) Processing time for the implementations done with a 10-fold cross-validation strategy through the use of hybrid ML models.

Figure 9: Training process time for different customer loyalty assessment strategies (individual and hybrid ML models) across the different ML regressor-based algorithms used (AB, GB, DT, RF, and LightGBM), and multiple experiments (DPT, DPT+FE, DPT+GAI(5), and DPT+GAI(10)).

4.4. Comparison with State-of-the-Art Studies

It is important to underscore the absence of studies investigating the prediction of customer loyalty scores as a regression problem while simultaneously evaluating the efficacy of hybrid machine learning (ML) models in comparison to individual ML models. This includes an analysis of the trade-off between processing time (pertinent to green AI principles) and accuracy. With this context, we compare our results with existing studies that have designed and implemented prescriptive or predictive frameworks for customer loyalty, extending also to the concept of customer churn due to their related nature.

Firstly, Buckinx et al. [67] predicts customer loyalty using individual ML models such as regression and decision trees, achieving an average Mean Squared Error (MSE) of 0.6 and a Root Mean Squared Error

(RMSE) of 0.77. Our predictions yield a slightly higher MSE (0.8); however, it is crucial to note that their study utilizes Customer Lifetime Value (CLV) to assess loyalty, a feature absent in our dataset, which may explain the discrepancy in accuracy. Machado et al. [22] further explores the use of tree-based models to assess customer loyalty, reporting an RMSE on the order of 3, significantly higher than our findings. Notably, their study also reports processing times—often omitted in other studies—wherein their XGBoost and LightGBM models require approximately 3 seconds for training, almost thrice the duration required in our frameworks.

Additional studies, such as those by Kim and Lee [68] and Latheef and Vineetha [69], propose predictive frameworks for customer churn. While not directly comparable, these are mentioned here due to the close connection between churn and loyalty concepts. These studies, employing hybrid ML models, address a classification-based problem (churn vs. non-churn), achieving accuracies between 0.8 and 0.9. Finally, Wassouf et al. [70] utilizes ML to evaluate whether customer loyalty increases following a series of marketing actions. Although their focus is not directly on assessing customer loyalty, the demographic and behavioural features they consider are similar to ours, and they report an accuracy of approximately 60%.

5. Management Implications

The management implications of the experiments conducted in this study can be understood from two distinct perspectives. First, the ability to accurately predict CRM metrics (i.e., customer loyalty) and the importance of different features in both individual and hybrid ML models that extensively utilize feature engineering and dimensionality reduction approaches. Second, the analysis of processing time (and, therefore, carbon emissions) is crucial for selecting the most suitable model.

The results of the present studies, which examine the efficiency (predictive power and processing time) of individual and hybrid ML models in predicting CRM metrics (i.e., customer loyalty), have significant implications for management in different industries. The current research shows that the hybrid models consistently demonstrate slightly higher levels of prediction accuracy when compared to their individual counterparts. It suggests that employing the synergistic collaboration of multiple ML algorithms could lead to more accurate and reliable predictions of critical CRM metrics, such as customer loyalty, retention, credit risk, and lifetime value. Finance industry experts can enhance their decision-making processes by adopting hybrid ML models. This can result in improved customer segmentation, focused marketing strategies, and ultimately, enhanced financial success. The current analysis of both individual and hybrid ML models has yielded valuable insights that can guide management decisions regarding the prediction of CRM indicators (i.e., customer loyalty). The analysis of feature importance reveals that certain variables have a greater impact on predictions. Finance industry leaders can utilize this knowledge to ascertain the sequence of resource allocation and focus on maximizing these crucial components. Moreover, the hybrid models show a higher ability to understand nuanced feature interactions, highlighting their suitability for addressing the intricate dynamics of financial markets and customer behaviours.

The present study entails evaluating the processing time of extensive feature engineering and dimensionality reduction techniques leading to assessing carbon emissions, which provides a holistic perspective on model deployment in the banking industry. With that, management can thoroughly evaluate the trade-offs between computational efficiency and environmental sustainability when selecting models. For instance, a precise assessment is provided regarding how hybrid ML models provide lower processing time needed for improved generalization and performance when using extensive feature engineering. As a result, the current analysis provides quantitative insights about the trade-off regarding the additional computing cost that may be justified for certain critical financial applications or in case new data or information has to be considered in the models. This observation empowers financial professionals to make more informed decisions that align with both operational efficiency and corporate social responsibility objectives.

Hence, the balance between processing time and accuracy poses a significant challenge when choosing ML models for CRM metric forecasts (i.e., customer loyalty), as it directly affects operational efficiency and decision-making efficacy in the banking sector. In this particular situation, the capacity to efficiently handle and evaluate extensive quantities of consumer data is essential for prompt and well-informed decision-making, allowing firms to adapt to market fluctuations and meet customer demands quickly. Nevertheless, compromising accuracy in favour of swiftness can result in less-than-ideal outcomes, thereby compromising crucial business judgments. Achieving an optimal equilibrium between processing time and accuracy is crucial for finance professionals. This enables them to effectively utilize sophisticated predictive models that are both practical and responsive to real-world requirements. Consequently, this leads to improved results in CRM.

5.1. Carbon Emission Reduction versus Decision-Making Quality

The trade-off between "Green AI and Carbon Emission" and "Decision-Making Quality" is a critical consideration where precision and processing time are essential. Green AI emphasizes energy efficiency to reduce carbon emissions, which is crucial for sustainability [21, 71]. However, optimizing for energy efficiency may sometimes involve model simplifications or reduced computational resources, potentially compromising decision-making quality.

To navigate this trade-off effectively, businesses must carefully evaluate the specific use cases and their environmental commitments [72]. They may opt for a balanced approach, employing energy-efficient AI for tasks where high precision isn't critical while relying on more resource-intensive models for high-stakes, time-sensitive decisions [73]. Striking the right balance is essential to align with environmental goals without compromising the quality of the decision-making [72, 73, 74].

In the finance industry, achieving the right balance between "Green AI and Carbon Emission Reduction" and "Decision-Making Quality" is pivotal. Financial institutions must weigh the environmental impact of their AI operations against the necessity for precise and timely decision-making [75]. In the financial market, for example, specifically in algorithmic trading, using energy-efficient models with lower computational demands may lead to slightly slower execution times, which could impact the ability to capitalize on rapid

market fluctuations. The carbon footprint of ML algorithms is closely tied to their processing time, as longer computation durations consume more energy and result in higher greenhouse gas emissions. Optimizing processing time not only enhances the efficiency of these algorithms but also significantly reduces their environmental impact [72, 75]. On the other hand, more resource-intensive AI models can provide more accurate predictions but consume more energy, contributing to a higher carbon footprint. Other examples in the finance industry might relate to training ML algorithms for credit risk, customer loyalty, customer value, and any other customer metric, and the generalizations adhere to the same set of guidelines.

Another set of examples related to, for instance, routine tasks such as customer service chatbots or basic data analysis demonstrates the suitability of using energy-efficient AI models [74]. These tasks don't demand the highest level of decision-making quality, and the energy savings and reduced carbon emissions from efficient models align well with sustainability goals. However, precision and speed are relevant for high-frequency trading algorithms or risk assessment models. In these scenarios, optimizing solely for energy efficiency might lead to suboptimal results. Financial institutions may find it more prudent to allocate greater computational resources, accepting a higher carbon footprint in exchange for superior decision-making quality and the potential for increased profitability.

In summary, the strategy depends on the specific use case, environmental commitments, and regulatory requirements. Some financial organizations may opt for a mixed approach, using energy-efficient AI where feasible and relying on more resource-intensive models when precision is paramount. This way, they can contribute to carbon emission reduction while preserving the quality of their financial decision-making processes, ensuring a well-rounded and sustainable approach in the dynamic landscape of the finance industry.

5.2. Green AI and CRM

The significance of the relationship between Green AI and Customer Relationship Management (CRM) in the banking business is growing due to various compelling reasons [76, 77]. Primarily, the financial sector functions within a wider ecosystem that is increasingly aware of environmental issues and focusing more on sustainability. Financial institutions are facing increasing pressure to match their operations with the expectations of customers, shareholders, and regulatory bodies, who are demanding more environmentally friendly activities [78, 79, 80]. CRM is crucial in this scenario as it directly influences consumer interactions, which might result in environmental implications.

Banks and financial institutions face different environmental challenges, such as the need to comply with regulations that mandate the disclosure of carbon emissions, the influence of shareholder activism promoting environmentally sustainable investments, and the increased public consciousness regarding climate change concerns [78, 79]. The aforementioned challenges provide a compelling motivation for financial institutions to minimize their carbon emissions. This dedication also encompasses their utilization of AI and ML algorithms, which have the potential to be demanding in terms of resources and energy consumption. In order to achieve significant reductions in emissions, financial institutions may need to make decisions that require some compromise in algorithm efficiency while yet maintaining a balance between profitability and

sustainability. Essentially, it is a calculated reaction to evolve market dynamics and cultural expectations, demonstrate the financial industry's dedication to solve environmental issues while still prioritizing its core business goals.

6. Conclusion

Accurately assessing customers' loyalty levels is crucial to any business's success since it provides evidence that might support managers when designing marketing initiatives to boost retention, acquisition, or loyalty indicators. The use of ML models has provided high accuracy levels when assessing such CRM metrics in the financial industry; however, what are the costs of these approaches? This paper explores different strategies to assess the customer loyalty levels of a credit card company in Brazil. The objective is to verify the value of using hybrid ML models, analyzing their accuracy in contrast with the individual model's performance. Also, this paper aims to evaluate how the depreciation of the number of features and their impact on processing time used in the different predictive frameworks would affect the accuracy and processing time of the different pipelines.

This paper forecasts customer loyalty levels in credit card settings. A real dataset is used from the Elo Group, a company operating in South America (their data is available at Kaggle [23]). Using hybrid ML models is proposed in the current prediction task. These models combine k-Means with regressor-based algorithms (AB, GB, DT, RF, and LightGBM) and have been shown to speed up processing time and increase accuracy in previous studies exploring similar datasets [19, 18]. Frameworks using 5 and 10-fold cross-validation and four different experiments, namely 1. DPT, 2. DPT+FE, 3. DPT+GAI(5), and DPT+GAI(10) is also considered. Then, metrics of error (MSE, MAE, and MedAE) are extracted and the processing time of each model are listed. These vast number of frameworks (individual and hybrid) test the other models' performance, reducing the number of independent variables used to forecast customer loyalty. This is different from recent literature in this domain, which has focused on increasing the number of features to boost accuracy only [81, 82, 83, 84].

The experiments explored in this study demonstrate that the hybrid ML models perform better than their individual versions. Also, for most cases, a lower error is reached when using 10-fold cross-validation, as well as lower processing time and a lower number of features without significant impact on accuracy. In fact, when using 10-Fold cross-validation, the individual use of LightGBM provides an MAE of 0.6733 in the DPT experiment, while the hybrid model (k-Means + LightGBM) in the same experiment has an MAE of 0.6681. Also, comparing the DPT with DPT+GAI(10) in the individual models for the 5-fold cross-validation strategy provides an MSE of 0.9422 and 0.9419, respectively. This can be explained by the dimensionality reduction achieved in the hybrid models and the DPT+GAI(5) and DPT+GAI(10) experiments. Further, explaining the models by extracting and analyzing the importance of each feature is taken into account.

While the scope of the case study in this research is somewhat limited due to the dimensionality of the

dataset and the nature of the predictions (specifically, CRM metrics forecasting), it is possible to extrapolate the findings and discuss them more broadly. This study highlights the intricate balance between processing time, which contributes to energy consumption and carbon emissions, and the precision of the predictions, which can significantly influence the direction of decision-making processes. It also delves into the complexities involved in these decisions across various frameworks and the availability of multiple models. The main findings of the paper are particularly valuable for analyzing decision-making processes, allowing considerations to extend beyond mere accuracy to include the efficiency of the models as well.

This work has a few limitations that can be addressed in future research. First, the explainability of the models can be further explored through the implementation of different tools (e.g., LIME and SHAP) to provide more insights into the main features used to assess customer loyalty with extensive supporting evidence. Second, the reduction in the number of features for the green AI experiments should be automatized, and its financial impact should be measured in terms of saved time and other economic aspects. Third, hybrid ML models should be considered using other ML models to cluster customers prior to implementing the regressor-based models. This might impact the accuracy and, consequently, increase the firms' profits. Fourth, the applicability of green AI principles in individual and hybrid ML models should be investigated within other domains of the finance industry, such as fraud detection and investment forecasting, also exploring big data in these settings and industries beyond finance. Fifth and final, ensuring that ethical principles are considered while using AI algorithms can guarantee that customers have equal access to products and services, that the algorithms fairly assess customers' different indicators, and that institutions can be held accountable in case of bias and discrimination resulting from AI-based models.

References

[1] Kumar V, Reinartz W. Customer relationship management. Springer; 2018.

[2] Gupta S, Hanssens D, Hardie B, Kahn W, Kumar V, Lin N, et al. Modeling customer lifetime value. Journal of service research 2006;9(2):139–55.

[3] König PD, Wurster S, Siewert MB. Consumers are willing to pay a price for explainable, but not for green ai. evidence from a choice-based conjoint analysis. Big Data & Society 2022;9(1):20539517211069632.

[4] Strubell E, Ganesh A, McCallum A. Energy and policy considerations for modern deep learning research. In: Proceedings of the AAAI Conference on Artificial Intelligence; vol. 34. 2020, p. 13693–6.

[5] Lana I, Sanchez-Medina JJ, Vlahogianni EI, Del Ser J. From data to actions in intelligent transportation systems: A prescription of functional requirements for model actionability. Sensors 2021;21(4). URL: https://www.mdpi.com/1424-8220/21/4/1121. doi:10.3390/s21041121.

[6] Castanyer RC, Martínez-Fernández S, Franch X. Which design decisions in ai-enabled mobile applications contribute to greener ai? arXiv preprint arXiv:210915284 2021;.

[7] De Caigny A, Coussement K, De Bock KW. A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operational Research 2018;269(2):760–72. doi:10.1016/j.ejor.2018.02.009.

[8] Jamalian E, Foukerdi R. A hybrid data mining method for customer churn prediction. Engineering, Technology & Applied Science Research 2018;8(3):2991–7.

[9] Verbeke W, Martens D, Baesens B. Social network analysis for customer churn prediction. Applied Soft Computing

2014;14:431–46. URL: http://www.sciencedirect.com/science/article/pii/S1568494613003116. doi:https://doi.org/10.1016/j.asoc.2013.09.017.

[10] Amin A, Al-Obeidat F, Shah B, Adnan A, Loo J, Anwar S. Customer churn prediction in telecommunication industry using data certainty. Journal of Business Research 2018;:1–12URL: https://doi.org/10.1016/j.jbusres.2018.03.003. doi:10.1016/j.jbusres.2018.03.003.

[11] Bishop CM. Pattern recognition and machine learning. Information Science and Statistics. Berlin, Germany: Springer-Verlag; 2006. ISBN 0387310738.

[12] Negahban A. Simulation-based estimation of the real demand in bike-sharing systems in the presence of censoring. European Journal of Operational Research 2019;277(1):317–32. URL: https://ideas.repec.org/a/eee/ejores/v277y2019i1p317-332.html. doi:10.1016/j.ejor.2019.02.01.

[13] Castillo O, Melin P, Pedrycz W. Hybrid intelligent systems: Analysis and design. Studies in fuzziness and soft computing; Berlin, Germany: Springer; 2010. ISBN 9783642072239. URL: https://books.google.com.gi/books?id=UbOacQAACAAJ.

[14] Brazdil P, Carrier CG, Soares C, Vilalta R. Metalearning: Applications to data mining. Springer Science & Business Media; 2008.

[15] Cios KJ, Kurgan ŁA. Hybrid inductive machine learning: An overview of clip algorithms. New learning paradigms in soft computing 2002;:276–322.

[16] Bao W, Lianju N, Yue K. Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. Expert Systems with Applications 2019;128:301–15. URL: http://www.sciencedirect.com/science/article/pii/S0957417419301472. doi:https://doi.org/10.1016/j.eswa.2019.02.033.

[17] Bequé A, Lessmann S. Extreme learning machines for credit scoring: An empirical evaluation. Expert Systems with Applications 2017;86:42–53.

[18] Machado MR, Karray S. Assessing credit risk of commercial customers using hybrid machine learning algorithms. Expert Systems with Applications 2022;:116889URL: https://www.sciencedirect.com/science/article/pii/S0957417422003311. doi:https://doi.org/10.1016/j.eswa.2022.116889.

[19] Machado MR, Karray S. Applying hybrid machine learning algorithms to assess customer risk-adjusted revenue in the financial industry. Electronic Commerce Research and Applications 2022;:101202URL: https://www.sciencedirect.com/science/article/pii/S1567422322000850. doi:https://doi.org/10.1016/j.elerap.2022.101202.

[20] Machado MR, Karray S. Assessing credit risk of commercial customers using hybrid machine learning algorithms. Expert Systems with Applications 2022;200:116889. URL: https://www.sciencedirect.com/science/article/pii/S0957417422003311. doi:https://doi.org/10.1016/j.eswa.2022.116889.

[21] Schwartz R, Dodge J, Smith NA, Etzioni O. Green ai. Commun ACM 2020;63(12):54–63. URL: https://doi.org/10.1145/3381831. doi:10.1145/3381831.

[22] Machado MR, Karray S, de Sousa IT. Lightgbm: An effective decision tree gradient boosting method to predict customer loyalty in the finance industry. In: 2019 14th International Conference on Computer Science and Education (ICCSE). IEEE; 2019, p. 1111–6.

[23] Kaggle Competition EMCR. Elo merchant category recommendation competition: Help understand customer loyalty. 2018. URL: https://www.kaggle.com/c/elo-merchant-category-recommendation.

[24] Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd international conference on Machine learning. 2006, p. 161–8.

[25] Al-Mashraie M, Chung SH, Jeon HW. Customer switching behavior analysis in the telecommunication industry via push-pull-mooring framework: A machine learning approach. Computers & Industrial Engineering 2020;144:106476. URL: https://www.sciencedirect.com/science/article/pii/S0360835220302102. doi:https://doi.org/10.1016/j.cie.2020.106476.

[26] Ghahramani Z. Unsupervised learning. In: Summer school on machine learning. Springer; 2003, p. 72–112.

[27] Corchado E, Abraham A, de Carvalho A. Hybrid intelligent algorithms and applications. Information Sciences 2010;180(14):2633–4. URL: http://www.sciencedirect.com/science/article/pii/S0020025510000848. doi:https://doi.org/10.1016/j.ins.2010.02.019; including Special Section on Hybrid Intelligent Algorithms and Ap-

plications.

[28] Kajdanowicz T, Kazienko P, Kraszewski J. Boosting algorithm with sequence-loss cost function for structured prediction. In: Graña Romay M, Corchado E, Garcia Sebastian MT, editors. Hybrid artificial intelligence systems. Berlin, Germany: Springer. ISBN 978-3-642-13769-3; 2010, p. 573–80.

[29] Gao X, Lee GM. Moment-based rental prediction for bicycle-sharing transportation systems using a hybrid genetic algorithm and machine learning. Computers & Industrial Engineering 2019;128:60–9. URL: https://www.sciencedirect.com/science/article/pii/S0360835218306260. doi:https://doi.org/10.1016/j.cie.2018.12.023.

[30] De Bock KW, De Caigny A. Spline-rule ensemble classifiers with structured sparsity regularization for interpretable customer churn modeling. Decision Support Systems 2021;150:113523. URL: https://www.sciencedirect.com/science/article/pii/S0167923621000336. doi:https://doi.org/10.1016/j.dss.2021.113523; interpretable Data Science For Decision Making.

[31] Kristensen K, Eskildsen J. Is the net promoter score a reliable performance measure? In: 2011 IEEE International Conference on Quality and Reliability. IEEE; 2011, p. 249–53.

[32] Yi Y, La S. What influences the relationship between customer satisfaction and repurchase intention? investigating the effects of adjusted expectations and customer loyalty. Psychology & Marketing 2004;21(5):351–73.

[33] Bielski L. Guided by feedback: measuring customer engagement. American Bankers Association ABA Banking Journal 2008;100(8):44.

[34] Patel N, Trivedi S. Leveraging predictive modeling, machine learning personalization, nlp customer support, and ai chatbots to increase customer loyalty. Empirical Quests for Management Essences 2020;3(3):1–24.

[35] Buckinx W, Verstraeten G, Van den Poel D. Predicting customer loyalty using the internal transactional database. Expert systems with applications 2007;32(1):125–34.

[36] Lee HF, Jiang M. A hybrid machine learning approach for customer loyalty prediction. In: International Conference on Neural Computing for Advanced Applications. Springer; 2021, p. 211–26.

[37] Ballestar MT, Grau-Carles P, Sainz J. Predicting customer quality in e-commerce social networks: a machine learning approach. Review of Managerial Science 2019;13(3):589–603.

[38] Sulistiani H, Muludi K, Syarif A. Implementation of dynamic mutual information and support vector machine for customer loyalty classification. In: Journal of Physics: Conference Series; vol. 1338. IOP Publishing; 2019, p. 012050.

[39] Wassouf WN, Alkhatib R, Salloum K, Balloul S. Predictive analytics using big data for increased customer loyalty: Syriatel telecom company case study. Journal of Big Data 2020;7(1):1–24.

[40] Ulkhaq MM, Adyatama A, Fidiyanti F, Rozaq R, Raharjo MFM. An artificial neural network approach for predicting customer loyalty: a case study in an online travel agency. International Journal of Machine Learning and Computing 2020;10(2):283–9.

[41] Zaki M, Kandeil D, Neely A, McColl-Kennedy JR. The fallacy of the net promoter score: Customer loyalty predictive model. Cambridge Service Alliance 2016;10:1–25.

[42] Wijaya A, Girsang AS. Use of data mining for prediction of customer loyalty. CommIT (Communication and Information Technology) Journal 2016;10(1):41–7.

[43] Zhang S, Tan X, Wang J, Chen J, Lai X. Modeling customers' loyalty using ten years' automobile repair and maintenance data: Machine learning approaches. In: Proceedings of the 2019 2nd International Conference on Data Science and Information Technology. 2019, p. 242–8.

[44] Latheef J, Vineetha S. Predicting customer loyalty in banking sector with mixed ensemble model and hybrid model. In: Smart Computing Techniques and Applications. Springer; 2021, p. 363–71.

[45] Hadiyat MA. Combined structural equation modelling–artificial neural networks model for predicting customer loyalty. In: IOP Conference Series: Materials Science and Engineering; vol. 703. IOP Publishing; 2019, p. 012024.

[46] Surya PSS, Anitha K. Comparative analysis of accuracy and prediction of customer loyalty in the telecom industry using novel diverse algorithm. In: 2022 International Conference on Business Analytics for Technology and Security (ICBATS). IEEE; 2022, p. 1–7.

[47] Khodabandehlou S, NikNafs AA. Improving customer loyalty evaluation methods in the grocery retail industry: a data mining approach. International Journal of Electronic Customer Relationship Management 2016;10(2-4):158–78.

[48] Verdecchia R, Sallou J, Cruz L. A systematic review of green ai. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2023;:e1507.

[49] Vasiliu LA, Roman D, Prodan R. Extreme and sustainable graph processing for green finance investment and trading. In: Companion of the 2023 ACM/SPEC International Conference on Performance Engineering. 2023, p. 249–50.

[50] König PD, Wurster S, Siewert MB. Consumers are willing to pay a price for explainable, but not for green ai. evidence from a choice-based conjoint analysis. Big Data & Society 2022;9(1):20539517211069632.

[51] Yigitcanlar T, Mehmood R, Corchado JM. Green artificial intelligence: Towards an efficient, sustainable and equitable technology for smart cities and futures. Sustainability 2021;13(16):8952.

[52] for International Settlements (BIS) B. Working paper (no 930) - big data and machine learning in central banking. 2021. URL: https://www.bis.org/publ/work930.pdf.

[53] Dehghani F, Larijani A. An algorithm for predicting stock market's index based on mid algorithm and neural network. Available at SSRN 4448033 2023;.

[54] Qiu Y, Wang J. A machine learning approach to credit card customer segmentation for economic stability. In: Proceedings of the 4th International Conference on Economic Management and Big Data Applications, ICEMBDA 2023, October 27–29, 2023, Tianjin, China. 2024, p. 19.

[55] Tehranian K. Can machine learning catch economic recessions using economic and market sentiments? arXiv preprint arXiv:230816200 2023;.

[56] Alahdadi A, Safaei AA, Ebadi MJ. A truthful and budget-balanced double auction model for resource allocation in cloud computing. Soft Computing 2023;27(23):18263–84.

[57] Hamerly G, Elkan C. Learning the k in k-means. Advances in neural information processing systems 2003;16.

[58] Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput 1998;10(7):1895–923. URL: https://doi.org/10.1162/089976698300017197. doi:10.1162/089976698300017197.

[59] Turner CR, Fuggetta A, Lavazza L, Wolf AL. A conceptual basis for feature engineering. Journal of Systems and Software 1999;49(1):3–15.

[60] Nargesian F, Samulowitz H, Khurana U, Khalil EB, Turaga DS. Learning feature engineering for classification. In: Ijcai. 2017, p. 2529–35.

[61] Verdonck T, Baesens B, Óskarsdóttir M, et al. Special issue on feature engineering editorial. Machine Learning 2021;:1–12.

[62] Refaeilzadeh P, Tang L, Liu H. Cross-validation. Encyclopedia of database systems 2009;5:532–8.

[63] Finlay S. Multiple classifier architectures and their application to credit risk assessment. European Journal of Operational Research 2011;210(2):368–78.

[64] Bijak K, Thomas LC. Does segmentation always improve model performance in credit scoring? Expert Systems with Applications 2012;39(3):2433–42.

[65] Barstugan M, Ozkaya U, Ozturk S. Coronavirus (covid-19) classification using ct images by machine learning methods. arXiv preprint arXiv:200309424 2020;.

[66] Shrivastava VK, Londhe ND, Sonawane RS, Suri JS. Reliable and accurate psoriasis disease classification in dermatology images using comprehensive feature space in machine learning paradigm. Expert Systems with Applications 2015;42(15-16):6184–95.

[67] Buckinx W, Verstraeten G, Van den Poel D. Predicting customer loyalty using the internal transactional database. Expert systems with applications 2007;32(1):125–34.

[68] Kim S, Lee H. Customer churn prediction in influencer commerce: An application of decision trees. Procedia Computer Science 2022;199:1332–9.

[69] Latheef J, Vineetha S. Predicting customer loyalty in banking sector with mixed ensemble model and hybrid model. In: Smart Computing Techniques and Applications: Proceedings of the Fourth International Conference on Smart Computing and Informatics, Volume 2. Springer; 2021, p. 363–71.

[70] Wassouf WN, Alkhatib R, Salloum K, Balloul S. Predictive analytics using big data for increased customer loyalty: Syriatel telecom company case study. Journal of Big Data 2020;7(1):29.

[71] Ying Y, Jin S. Digital transformation and corporate sustainability: The moderating effect of ambidextrous innovation. Systems 2023;11(7):344.

[72] Delanoë P, Tchuente D, Colin G. Method and evaluations of the effective gain of artificial intelligence models for reducing co2 emissions. Journal of Environmental Management 2023;331:117261. URL: https://www.sciencedirect.com/science/article/pii/S030147972300049X. doi:https://doi.org/10.1016/j.jenvman.2023.117261.

[73] Zhen K, Kun B. Optimizing carbon emission reduction strategies through ai-enabled modeling: paving the way to a sustainable future. In: International Conference on Mechatronics and Intelligent Control (ICMIC 2023); vol. 12793. SPIE; 2023, p. 168–74.

[74] Alpan K, Tuncal K, Ozkan C, Sekeroglu B, Ever YK. Design and simulation of global model for carbon emission reduction using iot and artificial intelligence. Procedia Computer Science 2022;204:627–34.

[75] Tang Z. Research on digital technology empowering green finance: Evidence from financial enterprises. In: 3rd International Conference on Internet Finance and Digital Economy (ICIFDE 2023). Atlantis Press; 2023, p. 168–75.

[76] Ferrer-Estévez M, Chalmeta R. Sustainable customer relationship management. Marketing Intelligence & Planning 2023;41(2):244–62.

[77] Gil-Gomez H, Guerola-Navarro V, Oltra-Badenes R, Lozano-Quilis JA. Customer relationship management: digital transformation and sustainable business model innovation. Economic research-Ekonomska istraživanja 2020;33(1):2733–50.

[78] Paris Agreement . Bluebook citation: Paris agreement to the united nations framework convention on climate change, dec. 12, 2015, t.i.a.s. no. 16-1104. 2015. URL: https://treaties.un.org/Pages/showDetails.aspx?objid=0800000280458f37&clang=_en.

[79] UN . For a livable climate: Net-zero commitments must be backed by credible action. 2023. URL: https://www.un.org/en/climatechange/net-zero-coalition.

[80] Carolina Rezende de Carvalho Ferreira M, Amorim Sobreiro V, Kimura H, Luiz de Moraes Barboza F. A systematic review of literature about finance and sustainability. Journal of Sustainable Finance & Investment 2016;6(2):112–47.

[81] Gosiewska A, Kozak A, Biecek P. Simpler is better: Lifting interpretability-performance trade-off via automated feature engineering. Decision Support Systems 2021;150:113556.

[82] Yun KK, Yoon SW, Won D. Prediction of stock price direction using a hybrid ga-xgboost algorithm with a three-stage feature engineering process. Expert Systems with Applications 2021;186:115716.

[83] Chen Y, Yang W, Wang K, Qin Y, Huang R, Zheng Q. A neuralized feature engineering method for entity relation extraction. Neural Networks 2021;141:249–60.

[84] Gibert D, Planes J, Mateu C, Le Q. Fusing feature engineering and deep learning: A case study for malware classification. Expert Systems with Applications 2022;207:117957.

Appendix

Table A1: Metrics of validation extracted when using 5-Fold cross-validation.

Metric	Experimental Framework	Adaboost		Gradient Boosting		Decision Tree		Random Forest		LightGBM	
		Individual	Hybrid	Individual	Hybrid	Individual	Hybrid	Individual	Hybrid	Individual	Hybrid
MAE	DPT	0.6862	0.7848	0.6758	0.6660	1.0258	0.8545	0.6929	0.7908	0.6735	0.6690
	DPT+FE	0.6758	0.7746	0.6615	0.6555	0.9748	0.7708	0.6560	0.7689	0.6450	0.6414
	DPT+GAI(5)	0.6861	0.7839	0.6758	0.6733	1.0242	0.7791	0.6946	0.7968	0.6735	0.6711
	DPT+GAI(10)	0.6861	0.7840	0.6758	0.6921	1.0266	0.7780	0.6971	0.7977	0.6743	0.7352
MSE	DPT	0.9253	1.1863	0.9247	0.8491	1.9634	1.4183	0.9422	1.2246	0.9072	0.8400
	DPT+FE	0.8972	1.1598	0.8873	0.8197	1.7952	1.1771	0.8481	1.1486	0.8293	0.7801
	DPT+GAI(5)	0.9252	1.1859	0.9247	0.8477	1.9662	1.2031	0.9457	1.2309	0.9074	0.8404
	DPT+GAI(10)	0.9252	1.1861	0.9248	0.8477	1.9711	1.2002	0.9529	1.2349	0.9089	0.8414
MedAE	DPT	0.4903	0.6561	0.4683	0.6536	0.7508	0.7143	0.4964	0.6393	0.4745	0.6265
	DPT+FE	0.4906	0.6410	0.4566	0.6416	0.7100	0.6117	0.4703	0.6233	0.4573	0.6007
	DPT+GAI(5)	0.4901	0.6548	0.4684	0.6494	0.7488	0.6291	0.4979	0.6492	0.4749	0.6283
	DPT+GAI(10)	0.4904	0.6550	0.4684	0.6493	0.7524	0.6285	0.5005	0.6484	0.4749	0.6274

Table A2: Metrics of validation (MAE, MSE, and MedAE) for the different experiments (DPT, DPT+FE, DPT+GAI(5), and DPT+GAI(5)) for all ML models implemented (AB, GB, DT, RF, LighGBM) using 5-Fold cross-validation training strategy.

Error Metric	Experiment	Clusters	Adaboost	Gradient Boosting	Decision Tree	Random Forest	LightGBM
MAE	DPT	Cluster 1	0.9528	0.6157	1.1738	0.9611	0.6327
		Cluster 2	0.5181	0.5181	0.5240	0.5228	0.5162
		Cluster 3	0.8834	0.8642	0.8656	0.8885	0.8581
		Average	0.7848	0.6660	0.8545	0.7908	0.6690
		SD	0.2335	0.1785	0.3251	0.2349	0.1738
	DPT+FE	Cluster 1	0.9435	0.6026	0.9450	0.9528	0.6190
		Cluster 2	0.5142	0.5139	0.5205	0.5102	0.5066
		Cluster 3	0.8661	0.8501	0.8469	0.8437	0.7987
		Average	0.7746	0.6555	0.7708	0.7689	0.6414
		SD	0.2288	0.1742	0.2223	0.2306	0.1474
	DPT+GAI(5)	Cluster 1	0.9503	0.6375	0.9460	0.9734	0.8307
		Cluster 2	0.5181	0.5181	0.5241	0.5238	0.5163
		Cluster 3	0.8831	0.8644	0.8670	0.8933	0.8587
		Average	0.7839	0.6733	0.7791	0.7968	0.7352
		SD	0.2326	0.1759	0.2243	0.2398	0.1901
	DPT+GAI(10)	Cluster 1	0.9505	0.6937	0.9457	0.9746	0.6934
		Cluster 2	0.5181	0.5181	0.5234	0.5244	0.5016
		Cluster 3	0.8833	0.8644	0.8651	0.8940	0.8184
		Average	0.7840	0.6921	0.7780	0.7977	0.6711
		SD	0.2327	0.1732	0.2242	0.2400	0.1595
MSE	DPT	Cluster 1	1.7022	0.6973	2.3773	1.7528	0.6788
		Cluster 2	0.3896	0.3896	0.4012	0.3993	0.3878
		Cluster 3	1.4670	1.4605	1.4764	1.5217	1.4534
		Average	1.1863	0.8491	1.4183	1.2246	0.8400
		SD	0.6999	0.5514	0.9893	0.7240	0.5508
	DPT+FE	Cluster 1	1.6863	0.6758	1.7259	1.6951	0.6444
		Cluster 2	0.3842	0.3838	0.3992	0.3815	0.3757
		Cluster 3	1.4090	1.3997	1.4061	1.3693	1.3202
		Average	1.1598	0.8197	1.1771	1.1486	0.7801
		SD	0.6859	0.5230	0.6924	0.6840	0.4867
	DPT+GAI(5)	Cluster 1	1.7011	0.6917	1.7273	1.7558	0.6798
		Cluster 2	0.3896	0.3896	0.4012	0.4009	0.3878
		Cluster 3	1.4670	1.4618	1.4808	1.5361	1.4536
		Average	1.1859	0.8477	1.2031	1.2309	0.8404
		SD	0.6995	0.5529	0.7053	0.7272	0.5507
	DPT+GAI(10)	Cluster 1	1.7016	0.6916	1.7249	1.7580	0.6822
		Cluster 2	0.3897	0.3896	0.4004	0.4020	0.3880
		Cluster 3	1.4671	1.4618	1.4755	1.5447	1.4539
		Average	1.1861	0.8477	1.2002	1.2349	0.8414
		SD	0.6997	0.5529	0.7038	0.7292	0.5505
MedAE	DPT	Cluster 1	0.8178	0.8117	1.0198	0.7737	0.7653
		Cluster 2	0.4729	0.4734	0.4741	0.4735	0.4700
		Cluster 3	0.6775	0.6755	0.6489	0.6707	0.6442
		Average	0.6561	0.6536	0.7143	0.6393	0.6265
		SD	0.1735	0.1702	0.2787	0.1526	0.1484
	DPT+FE	Cluster 1	0.7975	0.7879	0.7496	0.7848	0.7436
		Cluster 2	0.4688	0.4693	0.4664	0.4591	0.4587
		Cluster 3	0.6567	0.6675	0.6191	0.6259	0.5997
		Average	0.6410	0.6416	0.6117	0.6233	0.6007
		SD	0.1649	0.1609	0.1417	0.1628	0.1424
	DPT+GAI(5)	Cluster 1	0.8139	0.7990	0.7615	0.8031	0.7681
		Cluster 2	0.4732	0.4735	0.4742	0.4738	0.4702
		Cluster 3	0.6773	0.6757	0.6516	0.6708	0.6465
		Average	0.6548	0.6494	0.6291	0.6492	0.6283
		SD	0.1715	0.1643	0.1450	0.1657	0.1498
	DPT+GAI(10)	Cluster 1	0.8145	0.7984	0.7629	0.8039	0.7681
		Cluster 2	0.4731	0.4738	0.4731	0.4747	0.4701
		Cluster 3	0.6773	0.6758	0.6496	0.6667	0.6442
		Average	0.6550	0.6493	0.6285	0.6484	0.6274
		SD	0.1718	0.1639	0.1461	0.1654	0.1497

Table A3: Metrics of validation (MAE, MSE, and MedAE) for the different experiments (DPT, DPT+FE, DPT+GAI(5), and DPT+GAI(5)) for all Hybrid ML models (10-Fold cross-validation).

Error Metric	Experiment	Clusters	Adaboost	Gradient Boosting	Decision Tree	Random Forest	LightGBM
MAE	DPT	Cluster 1	0.9499	0.5374	1.0022	0.6464	0.6319
		Cluster 2	0.5181	0.5181	0.5244	0.5227	0.5161
		Cluster 3	0.8840	0.8642	0.8655	0.8890	0.8564
		Average	0.7840	0.6399	0.7974	0.6860	0.6681
		SD	0.2326	0.1945	0.2461	0.1864	0.1730
	DPT+FE	Cluster 1	0.9430	0.5327	0.6452	0.9520	0.6521
		Cluster 2	0.5142	0.5140	0.5206	0.5103	0.5064
		Cluster 3	0.8665	0.8503	0.8438	0.8418	0.8170
		Average	0.7746	0.6323	0.6699	0.7680	0.6585
		SD	0.2287	0.1890	0.1630	0.2299	0.1554
	DPT+GAI(5)	Cluster 1	0.9499	0.5375	0.9454	0.6716	0.6321
		Cluster 2	0.5181	0.5181	0.5239	0.5237	0.5162
		Cluster 3	0.8838	0.8643	0.8657	0.8938	0.8578
		Average	0.7839	0.6400	0.7783	0.6963	0.6687
		SD	0.2326	0.1945	0.2239	0.1863	0.1737
	DPT+GAI(10)	Cluster 1	0.9506	0.5375	0.9445	0.6740	0.6322
		Cluster 2	0.5181	0.5181	0.5236	0.5247	0.5163
		Cluster 3	0.8839	0.8643	0.8649	0.8935	0.8573
		Average	0.7842	0.6400	0.7777	0.6974	0.6686
		SD	0.2328	0.1945	0.2236	0.1855	0.1734
MSE	DPT	Cluster 1	1.7011	0.9080	1.7990	0.9097	0.9776
		Cluster 2	0.3896	0.3896	0.4015	0.3988	0.3877
		Cluster 3	1.4669	1.4597	1.4795	1.4900	1.4461
		Average	1.1859	0.9191	1.2267	0.9328	0.9371
		SD	0.6995	0.5352	0.7322	0.5460	0.5304
	DPT+FE	Cluster 1	1.6865	0.9607	1.7327	0.8865	0.9006
		Cluster 2	0.3843	0.3838	0.3990	0.3813	0.3755
		Cluster 3	1.4094	1.1800	1.3984	1.1652	1.3138
		Average	1.1601	0.8415	1.1767	0.8110	0.8633
		SD	0.6860	0.4113	0.6939	0.3974	0.4703
	DPT+GAI(5)	Cluster 1	1.7007	0.8157	1.7293	0.9484	1.6773
		Cluster 2	0.3896	0.3896	0.4008	0.4006	0.3877
		Cluster 3	1.4678	1.4606	1.4789	1.3868	1.4500
		Average	1.1860	0.8886	1.2030	0.9119	1.1717
		SD	0.6995	0.5392	0.7059	0.4942	0.6884
	DPT+GAI(10)	Cluster 1	1.7010	0.8167	1.7269	0.7800	1.6783
		Cluster 2	0.3896	0.3896	0.4004	0.4020	0.3878
		Cluster 3	1.4678	1.4606	1.4766	1.5455	1.4476
		Average	1.1862	0.8890	1.2013	0.9092	1.1713
		SD	0.6996	0.5391	0.7048	0.5826	0.6882
MedAE	DPT	Cluster 1	0.8140	0.8003	0.8494	0.7582	0.7642
		Cluster 2	0.4726	0.4737	0.4753	0.4740	0.4691
		Cluster 3	0.6784	0.6757	0.6511	0.6707	0.6436
		Average	0.6550	0.6499	0.6586	0.6343	0.6256
		SD	0.1719	0.1648	0.1872	0.1455	0.1484
	DPT+FE	Cluster 1	0.7974	0.7890	0.7463	0.7821	0.7809
		Cluster 2	0.4685	0.4690	0.4679	0.4613	0.4570
		Cluster 3	0.6576	0.6673	0.6179	0.6243	0.5965
		Average	0.6412	0.6418	0.6107	0.6226	0.6115
		SD	0.1651	0.1615	0.1394	0.1604	0.1624
	DPT+GAI(5)	Cluster 1	0.8143	0.7990	0.7598	0.8019	0.7664
		Cluster 2	0.4729	0.4736	0.4744	0.4726	0.4692
		Cluster 3	0.6775	0.6763	0.6494	0.6749	0.6452
		Average	0.6549	0.6496	0.6279	0.6498	0.6269
		SD	0.1718	0.1643	0.1439	0.1661	0.1494
	DPT+GAI(10)	Cluster 1	0.8155	0.7991	0.7607	0.8027	0.7648
		Cluster 2	0.4731	0.4737	0.4749	0.4748	0.4702
		Cluster 3	0.6775	0.6763	0.6478	0.6684	0.6448
		Average	0.6554	0.6497	0.6278	0.6486	0.6266
		SD	0.1722	0.1643	0.1439	0.1648	0.1482

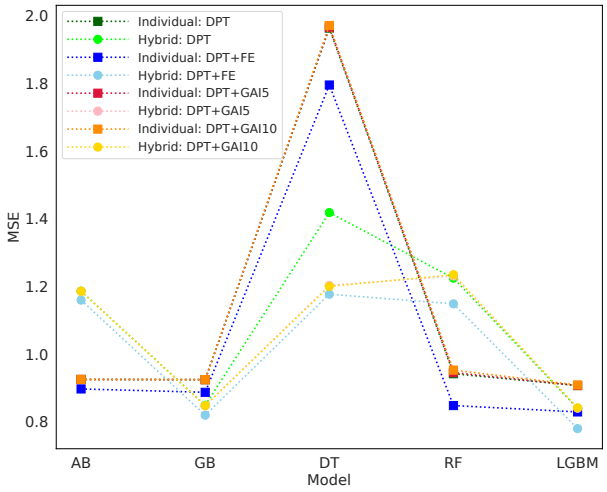
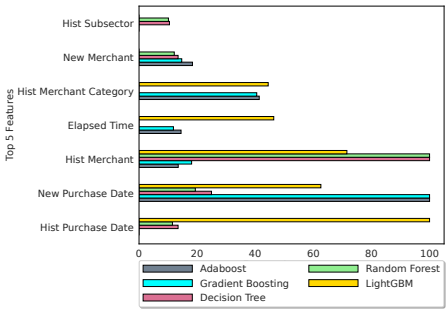
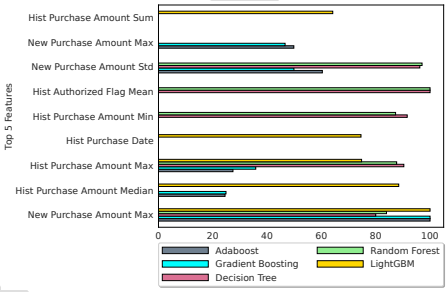


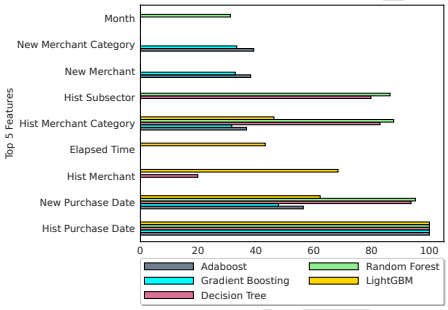
Figure A1: MSE for the different frameworks: 5-Fold.



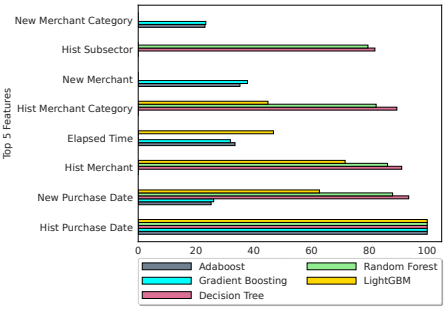
(a) Feature importance - Hybrid Models: DPT Framework.



(b) Feature importance - Hybrid Models: DPT+FE Framework.



(c) Feature importance - Hybrid Models: DPT+GAI(5) Framework.



(d) Feature importance - Hybrid Models: DPT+GAI(10) Framework.

Figure A2: Feature importance for the different experiments in the Hybrid ML frameworks (10-Fold cross-validation).

Green AI in the Finance Industry: Exploring the Impact of Feature Engineering on the Accuracy and Computational Time of Machine Learning Models

Marcos R. Machado[a,*], Amin Asadi[a], Renato William R. de Souza[b], Wallace C. Ugulino[c]

[a]*Department of Industrial Engineering and Business Information Systems, University of Twente, 7500 AE Enschede, The Netherlands*

[b]*Federal Institute of Education Science and Technology of Ceara Alameda Jose Quintino Cedro 63.300-000 Ceara Brazil*

[c]*Department of Semantics, Cybersecurity & Services (SCS), University of Twente, 7500 AE Enschede, The Netherlands*

Abstract

As research and practice in Artificial Intelligence (AI) applications rapidly expand, the support for AI deployment is also increasing. While the abundance of data allows for sophisticated feature engineering techniques that can enhance accuracy, it is crucial to highlight both the computational costs and the efficiency with which these models operate. This paper compares the processing time and accuracy of indi- vidual and hybrid Machine Learning (ML) models in predicting customer loyalty within financial contexts. Frameworks that incorporate feature engineering and green AI principles are used separately in both individ- ual and hybrid approaches. The individual models are commonly used regressor-based algorithms applied to business problems. The hybrid models first use k-Means to cluster customers, followed by the application of individual regressor-based models (e.g., decision trees, gradient boosting, and LightGBM). The present results show that using fewer features results in only a marginally lower accuracy compared to models with more features (a difference of $\approx$ 0.01 in MAE when comparing the use of 18 versus 85 features). Addition- ally, this article clearly demonstrate the trade-off between higher accuracy and longer computational time in hybrid ML models versus lower accuracy and shorter computational time in individual models when predicting customer loyalty. Hybrid models exhibit a lower MSE ($\approx$ 0.88) compared to individual models ($\approx$ 0.91). These findings provide managers with insights on selecting the most appropriate model based on their organization's specific needs.

Keywords: Feature Engineering, Green AI, Machine Learning, Hybrid Machine Learning, Customer Loyalty, Finance Industry

*Corresponding author: Marcos R. Machado (m.r.machado@utwente.nl, Tel. +31 534899045).

Email addresses: m.r.machado@utwente.nl (Marcos R. Machado), amin.asadi@utwente.nl (Amin Asadi), renato.william@ifce.edu.br (Renato William R. de Souza), w.corbougulino@utwente.nl (Wallace C. Ugulino)

Highlights

- Green AI and Feature Engineering are used in predictive frameworks in the finance industry.
- The trade-off processing time and accuracy is analyzed from different perspectives.
- Results can be generalized and support managers decision-making.
- Experiments with Hybrid ML models are conducted.
- Real customer data from the finance industry is used.

Declaration of interests

☐ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☒ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Marcos Machado reports financial support was provided by EU Framework Programme for Research and Innovation Euratom. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.