Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)
Vol. 11, No. 1, March 2025, pp. 124-138
ISSN: 2338-3070, DOI: 10.26555/jiteki.v11i1.30610

Prediction of Purchase Volume Coffee Shops in Surabaya Using Catboost with Leave-One-Out Cross Validation

Calvien Danny Nariyana, Mohammad Idhom, Trimono
Department of Computer Science,UPN Veteran Jawa Timur, Surabaya 60294 , Indonesia

ARTICLE INFO

Article history:

Received January 21, 2025
Revised March 17, 2025
Published March 21, 2025

Keywords:

Catboost;
LightGBM;
Coffee Shops;
Gradient Boosting

ABSTRACT

Indonesia's coffee consumption grew from 265,000 tons in 2015 to 294,000 tons in 2020. Averaging 2% annual growth with a projected 368,000 tons by 2024. One of the coffee businesses is coffee shops, Coffee shop businesses often struggle to attract customers quickly, risking low purchase volume within their first five years. In their first year, challenges include management, company size, service quality, and customer preferences. This study adopts a quantitative approach and new solutions to develop a purchase prediction application based on machine learning and strategy to enhance purchase volumes for three coffee shops in Surabaya. It utilizes CatBoost, with LightGBM as a comparison, across multiple coffee shop locations. *LOOCV (Leave-One-Out Cross-Validation*) is used in this model to address research limitations, such as data overfitting and biases, while enhancing evaluation accuracy. As a result, the study established CatBoost as the superior model for purchase prediction, providing insights and practical applications in business forecasting. The Catboost model achieved an MAE of 0.91 and MAPE of 15%, outperforming LightGBM's MAE of 1.13 and MAPE of 18%. These results confirmed CatBoost's effectiveness for the coffee shop industry with good accuracy. This research also contributes to helping coffee shop owners in Surabaya understand market characteristics, such as the most profitable coffee types and high-customer-density locations. Additionally, it aids in optimizing purchase volume to leverage profit by developing new strategies based on prediction result. In conclusion, CatBoost accurately predicts purchase volume, helping coffee shops identify target markets and refine strategies based on customer preferences.

Corresponding Author:

Mohammad Idhom, Department of Computer Science, UPN Veteran Jawa Timur, Surabaya, 60294, Indonesia
Email: idhom@upnjatim.ac.id

1. INTRODUCTION

MSMEs are vital to Indonesia's economy, making up 99.99% of businesses (56.54 million units) and significantly contributing to GDP. Their adaptability allows them to sustain economic growth and withstand financial crises [1]. MSMEs contribute 61% to Indonesia's non-oil and gas GDP (2023), driving growth and reducing disparities. With proper support, they can ensure sustainable development, though innovation is often overlooked [2]. Effective commercial strategies are key to MSME success. This study explores increasing coffee shops as MSME's profits through purchase volume prediction using machine learning regression with CatBoost [3].

Surabaya has experienced rapid growth in MSMEs, with the number increasing annually. Data from the Surabaya Cooperative and Micro Enterprises Office shows that the number of small and medium-sized businesses has exceeded 4,800. To thrive, MSMEs must adopt entrepreneurial orientation and innovation, which are critical for crafting competitive business strategies [4].

One of the rapidly growing MSME sectors is the coffee shop business, which has evolved with diverse concepts. Coffee shops have expanded beyond roadside locations to malls and offices, offering more than coffee and snacks. With a cozy ambiance and emotional appeal like prestige and warmth, they have become a lifestyle choice for urban communities [11].

Coffee shops bussiness often struggle to quickly attract consumers, putting their survival at risk within the first five years. Among businesses established in 2011, the survival rates are 75% after 2 years, 60% after 3 years, 49% after 4 years, and 44% after 5 years. However, in the accommodation and food services sector, only 34.6% of businesses manage to survive for 5 years [5]. The first year is the most critical period for new businesses, with the highest risk of closure. It indicates the factors such as customer preferences, and purchase volume play a significant role [5].

Saidin Nainggolan et al. (2022) state that internal and external factors in consumers preferences such as age, occupation, education level, marital status, culture, family, environment, and social media also play a role [2]. Businesses leverage business intelligence, machine learning such as Catboost, and data analysis to gain insights into consumer needs and preferences by analyzing sales data like purchase volumes, purchasing behavior, and market trends. This allows business owners to develop data-driven strategies, such as pricing optimization, enhancing efficiency and decision-making accuracy [1].

A previous study using the FP-Growth Algorithm found it effective for identifying rules to create shopping coffee shop packages based on customer preferences, enhancing service quality and increasing revenue. However, it has limitations, such as potentially overlooking subtle purchasing patterns and requiring regular data updates to match changing customer preferences [6]. A related study on coffee forecasting produced an LSTM method with MSE loss function graph and an MAE measure, both approaching zero, indicating the high accuracy of LSTM forecasting in predicting coffee prices [7].

Since the prediction method uses LSTM and evaluates results based on MAE, the outcomes may appear favorable but sensitive to categorical data, with a value below 0.6. A related study on LSTM forecasting for coffee prices showed strong performance, as indicated by the MSE loss function graph and MAE approaching zero [7]. Another study related for increasing purchase volume is using LightGBM for predicting coffee quality to get increasing purchasing volumes, the research result is getting accuracy pretty good with 72% accuracy, The study found that psychological factors and the marketing mix significantly influence purchasing decisions, while price affects satisfaction and location is less crucial. To attract customers, coffee shops should focus on competitive pricing and enhancing the buying experience [8].

This research provides solutions for business owners to develop product strategies, such as market trends and the most purchased locations and a newer approach like Catboost compared to existing gradient boosting machine methods like LightGBM [9], [10]. Research by Prokhorenkova et al. (2018) highlights that CatBoost (short for "Categorical Boosting") surpasses the leading implementations of gradient-boosted decision trees, specifically XGBoost and LightGBM, across a variety of popular machine learning tasks [9]. For instance, Shengquan et al. (2023) notes that LightGBM has drawbacks, such as a heightened risk of overfitting with small sample sizes and increased sensitivity to noisy datasets [10].

CatBoost is preferred over other gradient boosting methods due to a statistical issue identified by Prokhorenkova et al. (2018) prediction shift, a specific form of target leakage affecting all gradient boosting models. Standard preprocessing techniques for categorical features also encounter similar challenges. A common approach in gradient boosting is converting categories into target statistics, yet this can result in target leakage and prediction bias. To mitigate these issues, CatBoost introduces a new algorithm for handling categorical features. This open-source library, designed specifically for "Categorical Boosting," has shown superior performance compared to state-of-the-art models like XGBoost and LightGBM across various machine learning tasks [9].

This research examines factors from previous studies, such as name, occupation, age, marital status, coffee variety, location, price, purchase time, taste, and serving preferences. Data is collected from three Surabaya coffee shops through interviews with owners, employees, and customers, supported by a personal dataset. The chosen coffee shops are having criteria such as third wave coffee type, highlighting consumer involvement in the coffee-making process, quality of beans, and a front bar as a hallmark [2].

The location of the coffee shops is near from university or high school with high rating in Google Map, because average 16 – 25 years old with average 22.58 show high demand for third generation or third wave coffee shops [11], [12]. The population of the customers coffee shops aged 16–25, as this age group represents a significant portion of coffee enthusiasts.

The uniqueness of this research or research gaps from the previous study lies in its multi-location approach, analyzing data from various coffee shops to provide a broader perspective on the industry. Unlike

earlier studies that focused on a single coffee shop, this study incorporates machine learning techniques for more accurate predictions [13]. The research took by August 2024 – September 2024, and the place took 3 coffee shops that are third wave era coffee shops in Surabaya. The participants in this research include all coffee shop patrons, owners, and staff members.

The sampling technique that will be used is purposive snowball sampling. Purposive snowball sampling is useful for reaching hard-to-access populations by utilizing social networks to identify and recruit participants who are challenging to find through conventional methods [14]. While this method effectively reaches hard-to-access populations through social networks, it may still introduce bias and lack representativeness. To address these challenges, the researcher incorporates local coffee shop data and gathers insights through in-depth interviews with industry experts, including coffee shop owners to reduce the bias and could affect the generalization [14].

This research contributes by introducing the use of CatBoost, a modern machine learning approach, and compares its performance with other boosting algorithms like LightGBM and other method like FP-Growth and LSTM. Selecting a high-accuracy algorithm is crucial for effective target prediction. Proper management minimizes losses, enhancing profitability and sustainability. Predictive modeling helps prioritize products, boosting customer satisfaction and repeat purchases [15]. Another contribution of this research is providing coffee businesses with a cost-effective way to gain insights into future profits by predicting purchase volume.

By predicting purchase volume, they can develop pricing strategies or identify profitable locations for sales. For example, if Genteng District has the highest sales, owners can optimize sales efforts there. If latte is the most purchased coffee, they can focus on latte production with an optimal price, such as 15,000 Rupiah. Conversely, if the predicted sales are below average, they can adjust coffee types and locations to increase future purchase volume.

2. METHODS

This research encompasses several steps, including data collection, visualization, cleaning, encoding, splitting, modeling, and model evaluation. The study utilizes 200 rows of data across 11 columns, employing two regression algorithms: CatBoost Regression as the primary algorithm and LightGBM Regression for comparison. Model evaluation is conducted using parameters such as MAE, RMSE, MAPE, and MSE. A detailed explanation of each step is provided below.

2.1. Machine Learning

Machine learning analyzes patterns to interpret task-related data. With advanced algorithms and computational power, it enhances large-scale data processing and fraud detection, offering fast, real-time solutions. Machine learning has the capability to learn autonomously from data and adjust to evolving conditions. It excels at managing complex, non-linear relationships between variables and can integrate data from various sources to deliver precise predictions [16], [17], [18].

Machine learning, especially ensemble techniques, enhances financial early warning systems but faces challenges with imbalanced risk data and interpretability. Gradient Boosting Decision Tree (GBDT) improves predictions by sequentially refining decision trees, focusing on misclassified samples [19], [16]. To improve business failure prediction, we use a tree ensemble method within a boosting framework. Ensemble learning combines multiple models to minimize errors. Based on this approach, prediction algorithms fall into Bagging and Boosting techniques [16].

2.2. Catboost

CatBoost is a Gradient Boosting algorithm using decision trees, optimized for categorical and ordered features while reducing overfitting with Bayesian estimators. Unlike many models, it requires minimal preprocessing and supports both CPU and GPU, with the GPU version offering faster training. It enhances performance by permuting data, utilizing the full dataset, and addressing imbalance through class weight hyperparameters [20].

Adjusting hyperparameters in machine learning algorithms positively influences the final outcomes. However, the extent of this impact varies based on the algorithm used. Research has shown that the Gradient Boosted Decision Tree algorithm benefits the most from hyperparameter tuning, achieving an average performance improvement of 8-11% [21].

A low value may reduce its performance compared to other GBDT models. Model can adjust the maximum depth of decision trees and the number of categorical feature combinations to balance resource usage and performance [22]. The study demonstrate higher objective values were associated with a low bagging temperature, moderate tree depth, a high number of iterations, low L2 leaf regularization, a high learning rate,

a short od wait, and a high random strength. Bagging temperature contributed most to optimizing the CatBoost model (30%), followed by iterations (26%) and learning rate (22%), while other hyperparameters had less than 10% impact [23].

CatBoost, developed by Yandex researchers, is utilized for various tasks such as search, recommendation systems, virtual assistants, self-driving cars, and weather forecasting, both at Yandex and other companies. It supports both CPU and GPU implementations, enabling faster training. CatBoost is particularly effective for small datasets and excels in handling categorical features [21]. The overall process of the CatBoost algorithm can be summarized in the following steps [24]:

1. Data Permutation: The algorithm begins with a training dataset 'D' containing 'n' instances. To introduce diversity, it randomly shuffles the dataset d times, creating d different training sets denoted as Dr, where r ranges from 1 to d.
2. Matrix Initialization: The algorithm constructs a matrix M where each element $M(r,i)$ represents the initial prediction value for an instance i in the training set Dr. Initially, these values are set to zero.
3. CatBoost Training on a Random Set: The algorithm randomly selects one of the permutation sets, Dr, for further processing.
   - Categorical Feature Encoding: Categorical features are encoded using Ordered Target Statistics (TS) Encoding.
   - Tree Construction: A new Ordered Boosting tree (T) is built, which approximates the gradient or residual for each instance in Dr, utilizing the $M(r,)$ matrix during gradient calculations.
   - Gradient Boosting Update: The algorithm uses the newly created tree T to predict outcomes for all permutation datasets, then updates the matrix M based on these predictions through a gradient boosting method.
4. Ensemble Prediction: The algorithm repeats Step 3 N times to construct N trees. Finally, predictions for any instance are made by averaging the predictions from all N trees. This ensemble method is similar to traditional Gradient Boosting.

The expression formula of this algorithm is [24]:

$$x_{i,k} = \frac{\sum_{j=1}^{p-1} \left[x_{\sigma_{j,k}} = x_{\sigma_{p,k}}\right] . \ Y_j + \alpha \ . \ p}{\sum_{j=1}^{p-1} \left[x_{\sigma_{j,k}} = x_{\sigma_{p,k}}\right] + \alpha} \quad (1)$$

In the formula, σ_j indicates the model's output for the jth data point,while $x_{\sigma_{i,k}}$refers to the discrete feature in the kth column of the ith row in the training dataset,The variable α represents a prior weight, and p stands for the prior distribution term.

CatBoost stands out from other gradient boosting methods through several key features. It employs ordered boosting to prevent target leakage, performs well on smaller datasets, and natively handles categorical features without extensive preprocessing. Instead of numerical encoding, it creates binary features for each category and uses random permutations to estimate leaf values, reducing overfitting. Finally, CatBoost utilizes binary decision trees as its base learners.

2.3. LightGBM

LightGBM (Light Gradient Boosting Machine) is a framework designed by Microsoft for implementing the GBDT algorithm [20]. It is known for its efficient parallel training capabilities, offering faster training speeds, lower memory usage, improved accuracy, and support for distributed processing, making it well-suited for handling large datasets. The process of building a LightGBM model includes the following steps [25]:

1. Train the LightGBM model using the training dataset.
2. Determine feature importance based on the trained LightGBM model.
3. Evaluate the model's performance by applying it to both training and testing datasets.

LightGBM is a fast and efficient gradient-boosting framework for classification, regression, and ranking tasks. It excels in handling large datasets with high accuracy, thanks to tree pruning, histogram-based splitting, and efficient memory use. Its distributed training capability further enhances performance on extensive data [26].

Parameter optimization is a crucial step in both training and forecasting [27]. Key hyperparameters in LightGBM include learning_rate, which controls the learning speed (lower values require more trees but enhance generalization), max_depth, which limits tree depth to prevent overfitting, and n_estimators, which determines the number of weak learners (higher values boost accuracy but increase computation time) [28].

LightGBM also achieves high accuracy across a variety of machine learning tasks due to its reliance on decision tree and gradient boosting algorithms.The base learner in LightGBM is a decision tree, and the gradient boosting process iteratively adds new decision trees to enhance prediction accuracy. The mathematical representation of LightGBM can be expressed as [26]:

$$g(x) = f(x) + \beta * (y - f(x)) \quad (2)$$

In this context, **y** represents the true value associated with **x**. The equation demonstrates that the predicted value for **x** is a linear combination of the base learner and the gradient boosting step. The learning rate, denoted as β , plays a crucial role in regulating the contribution of the gradient boosting step to the overall prediction, ensuring an optimal balance between learning speed and model accuracy.

2.4. Model Performance Metrics

RMSE (Root Mean Squared Error), Mean Squarred Error (MSE), MAE (Mean Absolute Error), and MAPE (Mean Absolute Percentage Error) are commonly used evaluation metrics in regression to assess prediction accuracy. These three metrics differ in terms of sensitivity to outliers and the ease of interpreting the results [29], [30]. In this study, the determination Mean Squarred Error (MSE) , Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) , and Mean Absolute Error (MAE), were used to evaluate the model accuracy [24], [25].

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (3)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (4)$$

$$MAPE = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{y_i - y'_i}{y_i}\right| \quad (5)$$

$$MAE: \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (6)$$

In the given context, y_i denotes the true value of the ith sample, $\hat{y}_i$ refers to the predicted value for the ith sample, and N represents the total number of samples, MAPE measures the percentage of mistakes [24].

In a regression problem, the Root Mean Squared Error (RMSE) is a frequently used statistic that quantifies the discrepancy between expected and actual values. It is the square root of the mean of the squared deviations between the values that were expected and those that were observed. The model fits the data better when the RMSE value is smaller. It is frequently used in regression situations to assess a model's performance [31]. Mean Absolute Percentage Error (MAPE) are statistical measure expresses how much two continuous variables differ from one another.

Comparing the predicted values to the dataset's actual values is a popular method of assessing a predictive model's accuracy. The average of the absolute discrepancies between the expected and actual values is used to compute it. Performance Metric Table *o*f MAPE shown in Table 1.

Table 1. Performance Metric Table of MAPE [31], [32]

Percentage (MAPE)	Meaning
0% – 10%	Accurate Result Forecasting
10% – 20%	Good Result Forecasting
20% - 50%	Reasonable Result Forecasting
> 50%	Bad Result Forecasting

The MAPE number can be interpreted as follows: less than 10% indicates highly accurate predicting; 10–20% indicates good forecasting; 20–50% indicates reasonable forecasting; and more than 50% indicates poor

forecasting [33], [34]. Since forecasting errors are squared, the Mean Squared Error (MSE) is a forecasting method that effectively handles large prediction errors.

MSE assigns greater weight to larger errors compared to smaller ones, making it particularly useful in cases where significant deviations from actual values are more critical or undesirable. This characteristic makes MSE a preferred metric in applications where penalizing substantial errors helps enhance the model's accuracy and reliability. Given that MSE and RMSE are closely related, a lower MSE value indicates a better model fit to the data [35].

2.5. Leave-One-Out Cross Validation

After evaluating the model, the researchers used optimization techniques such as Leave-One-Out Cross Validation The LOOCV method treats each observation in the dataset as the validation set while using the remaining N-1 observations as the training set. This process is repeated for every observation in the dataset, covering the entire sample size (N) [36].

As noted by Angelika Geroldinger et al. (2023), a key advantage of Leave-One-Out Cross-Validation (LOOCV) is its suitability for small sample sizes, where methods like ten-fold or five-fold cross-validation may perform poorly, particularly if some subsets of the data include only one category of a binary outcome. Nevertheless, LOOCV remains widely used for "cross-validating" c-statistics and calculated probabilities due to its capacity to handle specific challenges in model evaluation effectively [37].

For internal validation, resampling methods such as leave-one-out cross-validation (LOOCV) are frequently employed to address optimism in the model's performance metrics. This approach is especially relevant in scenarios involving small sample sizes or rare events, where traditional validation methods might be less effective. Several performance metrics can be utilized to evaluate the model's effectiveness for binary outcomes [36].

Examples include the concordance statistic (c-statistic or area under the curve), which measures the model's ability to distinguish between outcome categories, the discrimination slope, which quantifies the separation between predicted probabilities for different outcome,; and the Brier score, which assesses the accuracy of probability predictions. Together, these metrics provide a comprehensive assessment of the model's predictive capability and robustness [37]. Although Leave-One-Out Cross-Validation (LOOCV) offers an unbiased assessment of generalization error, it is computationally demanding and prone to high variance, especially when applied to large datasets [36].

Although LOOCV is well-suited for small datasets, it can sometimes be computationally expensive and exhibit high variance in certain cases [38]. The study found that Ridge regularization mitigates high variance in regression coefficient estimates by applying a penalty on their magnitude, effectively addressing multicollinearity and stabilizing the estimation process, resulting in more reliable and interpretable model outcomes [39].

Ridge regression is a method designed to address multicollinearity by adjusting Ordinary Least Squares (OLS) to generate more stable coefficient estimates. It minimizes coefficient variability, leading to slightly biased estimates that still closely approximate the true parameter values [40]. Ridge regularization in regression incorporates a penalty into the loss function, effectively compressing the eigenvalues of the covariance matrix.

This approach enhances estimator performance in high-dimensional contexts by mitigating over-dispersion and improving the stability of parameter estimates [41]. Ridge regression is favored over other methods as it effectively handles multicollinearity, delivers more reliable coefficient estimates, minimizes prediction variance, and lowers the mean square error, leading to a more stable and precise model than least squares regression [42].

The *l2_leaf_reg* hyperparameter in CatBoost regulates L2 regularization on leaf weights to mitigate overfitting and high variance model. Its value plays a crucial role in model performance, and since CatBoost is highly sensitive to the parameter, careful tuning is essential for achieving optimal results [22]. To prevent the computationally expensive such as a long time process, the study profound early stop parameter allows for stopping the observation process as soon as a decision rule is satisfied, effectively shortening the time [43].

2.6. Research Flowchart

The flowchart outlines a systematic process for data analysis and predictive modeling (Fig. 1). It starts with data acquisition, in this phase data was collected by purposive snowball sampling, Even though this method is effective for reaching hard-to-access populations through social networks, there is still a possibility of bias and non-representativeness. To mitigate these issues, the researcher utilizes local coffee shop data and conducts in-depth interviews with industry experts, such as coffee shop owners [14]. After data collection

phase, followed by checking the quality of the data. If the data is clean, it proceeds to exploratory data analysis (EDA) and visualization to uncover insights and patterns. If the data is dirty, it undergoes a cleaning process, including handling outlier with IQR Method, the method adjusts data distribution, it is extensively utilized in multiple disciplines to enhance the reliability and performance of analytical methods [44]. And deleting certain columns that contain personal information and fields less relevant from a business perspective, such as occupation, timestamp, name, and coffee shop name [45], [46], [47].

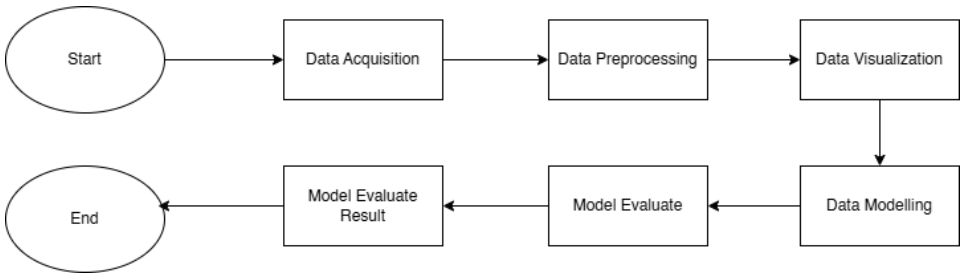

Fig. 1. Research Flows

Once prepared, The data categorizes categorical features using Target Statistic Encoding and One-Hot Encoding, and dependent and independent variables are identified. Two models, LightGBM and CatBoost, are trained and evaluated for accuracy. The model with the highest accuracy is selected, and its performance is visualized by comparing predicted and actual data. Finally, a prediction function is created for future use, marking the end of the workflow. This structured approach ensures reliable data preparation, analysis, and modeling for accurate predictions.

3. RESULTS AND DISCUSSION

This section provides a detailed explanation of the research process, outlining each step from data collection to model evaluation. Every stage is described comprehensively to ensure a clear understanding of the methods and outcomes.

3.1. Dataset Acquisition

Third-wave coffee defines high-quality beans based on factors such as their origin (where they are grown and harvested), roasting level, grinding precision, and the barista's skill during the brewing process, ensuring quality remains a priority throughout [20], [48]. A total of 31 subdistricts were identified, and data collection was conducted through in-depth interviews with coffee shop business owners, their employees, and customers. The dataset consists of 14 columns and 200 rows, collected from third-wave coffee shops in Surabaya, categorized by subdistrict.

Since purposive snowball sampling may not always be representative of the broader coffee shop customer population, additional steps were taken to minimize bias and enhance representativeness. To address this, the research validates and strengthens the dataset's reliability by incorporating in-depth interviews with industry experts, including coffee shop owners, and cross-referencing the findings with local coffee shop sales data.

In-depth interviews with structured questions and standardized evaluation criteria help reduce bias and improve data validity by ensuring consistency and minimizing subjective interpretation. Avoiding sensitive personal questions prevents social desirability bias, as participants may fear judgment. Building trust and familiarity with interviewers enhances comfort, encouraging honest responses and reducing response bias, ultimately improving data reliability [49], [50].

The dataset includes a timestamp, respondent demographics such as name, age, gender, marital status, occupation, and subdistrict in Surabaya (Table 2). It documents coffee preferences, including type (e.g., Cappuccino, Latte, Espresso), price, taste preference (e.g., Sour, Sweet, Bitter), and serving style (Hot or Cold). Additionally, it records the typical purchase time (Morning, Afternoon, Evening, or Night), purchase frequency, and the coffee shop name. The dependent variable of dataset is purchase volume, the independent variable is consumer preferences and location detail like lokasi,waktu_pembelian,umur,etc. And the machine learning method using Catboost Regression for purchase volume on 3 third wave era coffee shops in Surabaya, The effectiveness of CatBoost in forecasting purchase volume was demonstrated by comparing its results with several benchmark algorithms, including Light Gradient Boosting Machine (LightGBM).

Table 2. Data Example

Timestamp	2024	2024	2024	2024
Name	Consumen 1	Consumen 2	Consumen 3	Consumen 4
Age	23	22	23	20
Gender	Pria	Pria	Pria	Pria
Location	Wonokromo	Dukuh Pakis	Karangpilang	Karangpilang
Marital_status	Belum menikah	Belum menikah	Belum menikah	Belum menikah
Occupation	Mahasiswa	Mahasiswa	Mahasiswa	Mahasiswa
Coffee_purchased	Espresso	Americano	Americano	Americano
Coffee_price	23000	24000	23000	17000
Coffee_preferences	Pahit	Pahit	Pahit	Pahit
Serving_preferences	Panas	Panas	Panas	Panas
Purchase_time	Siang	Siang	Siang	Siang
Frequency_purchase	5	5	6	6
Coffee_shop_name	brain coffee	Garasi Kopi 75	Kopi Imaji	Kopi Imaji

3.2. Dataset Preprocessing

After dataset collected the process is check the missing value and outlier. The dataset profound 9 outlier with missing value, because importance to the dataset, missing value can be erased or deleted from dataset, To cleaning outlier the researchers use IQR method. Since the IQR approach normalizes data, it is widely used in many different domains to enhance the quality of applicable methods [44]. The resulting value, the IQR, which is the 75th percentile of all data values, will be used to cover the feature outlier values. In order to calculate it, the data values are first sorted in ascending order before being divided into four equal sections, or quartiles.

The IQR method detects and removes outliers by calculating Q1 and Q3, then determining the IQR (Q3 - Q1). Outliers fall beyond 1.5 times the IQR from these percentiles and are removed to improve model accuracy and reliability [51], [52]. To provide better dataset, privacy and unrelevant data like timestamp,name,occupation and coffee shops name are erased from dataset [46], [45]. The result of data cleaning are 190 rows and 10 coloumns in total.

After determining the first and third quartiles, the researchers may subtract them to determine the IQR value, which is an inner range of values. the researchers cap all values and use these inner values for the entire dataset [44]. The result of data cleaning can be shown in Fig. 2. As can be seen in Fig. 2, The image shows purchase frequency data before and after outlier handling. Initially (left, red), the data is highly skewed with extreme values above 30,000, distorting the distribution. After handling outliers (right, green), the distribution becomes more balanced, with values ranging from 2 to 12. This adjustment makes the data more representative and improves analysis reliability.

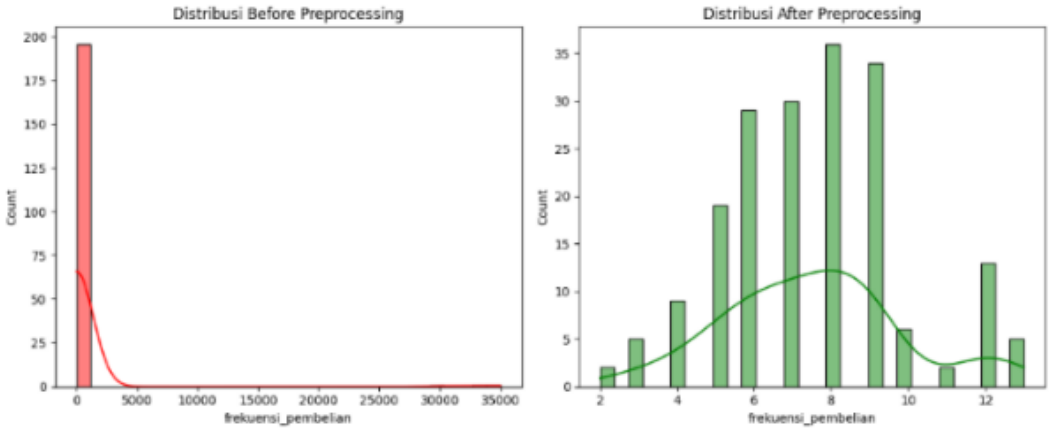

Fig. 2. Data Preprocessing Result

3.3. Data Visualization

After data cleaning process, the dataset can be visualitated to get better insight for upcoming process. The dynamic and varied subject of data visualization integrates knowledge from many domains, adapts to a range of audiences and situations, and commonly uses tacit knowledge [53]. The visualization use Piechart and Barplot. Some the visualization result can be seen as Fig. 3.

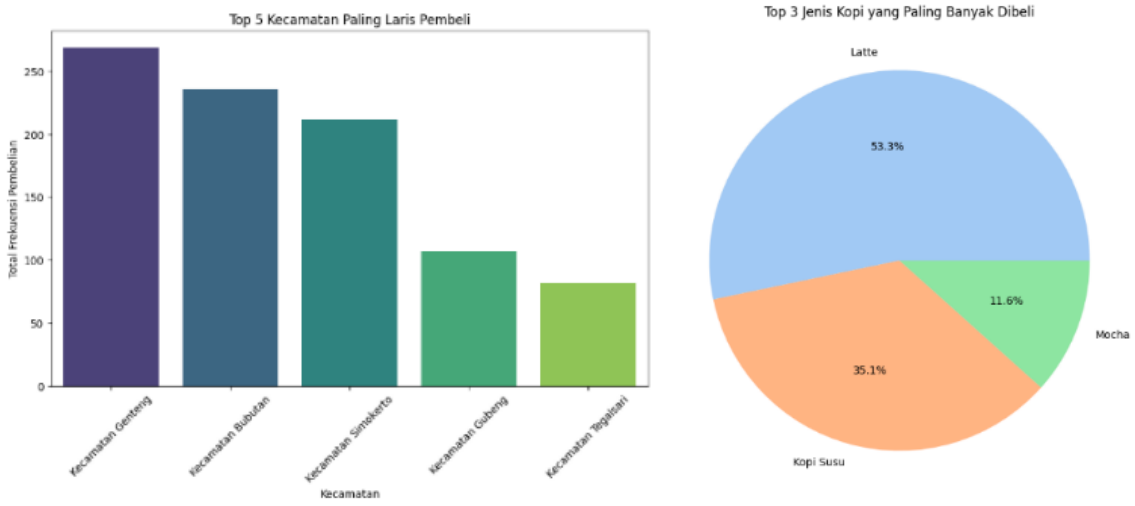

Fig. 3. Top Subdistrict Sales and Most Buyed Coffee

Fig. 3 presents the five subdistricts with the highest coffee purchase frequency, with Kecamatan Genteng leading, followed by Bubutan, Simokerto, Gubeng, and Tegalsari. This information is valuable for identifying prime locations for marketing efforts or establishing new coffee shops. Additionally, the Fig. 4 highlights the three most popular coffee types. Latte holds the highest share, contributing 53.3% of total purchases, followed by Kopi Susu (Milk Coffee) at 35.1%. Mocha ranks third, accounting for 11.6% of purchases.

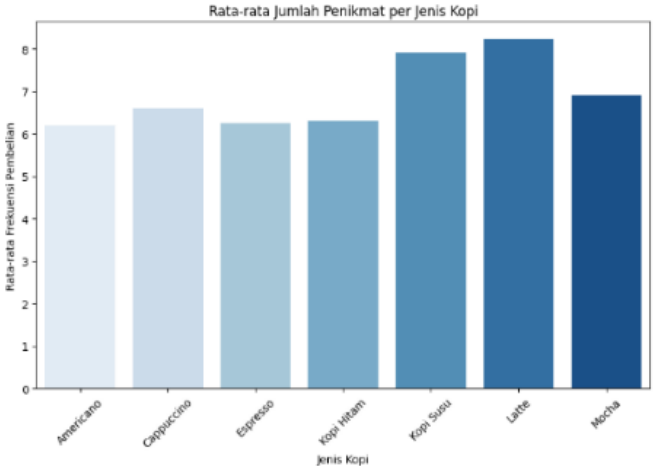

Fig. 4. Average Coffee Buyer

Fig. 4 shows the average number of coffee enthusiasts per coffee type. The visualization shows that Latte and Kopi Susu have the highest average purchase frequency compared to other coffee types. Meanwhile, Americano, Espresso, and Kopi Hitam have lower averages. And if all the buyers are counted, the total is 48.

3.4. Data Modelling

In data modeling, it was found that CatBoost outperformed LightGBM when parameters were tuned using Leave-One-Out Cross-Validation optimization. Before Tuned, the categorical features like location,marital status,and coffee type are encoded using Target Statistic Encoding if Catboost, and One Hot Encoding if LightGBM. Target statistic encoding uses the target variable's statistics to encode categories, capturing relationships with the target, while one-hot encoding creates orthogonal binary vectors for each category, failing to capture any relationship or similarity between categories, especially in high-cardinality settings [54].

The parameters applied for both CatBoost and LightGBM are presented in Table 3. The selected parameters were chosen and supported by previous studies, which suggest that higher objective values are associated with a low bagging temperature, a moderate tree depth (commonly 6), a large number of iterations (normally 500), low L2 leaf regularization (normally 6), a high learning rate (commonly 0.5), a short overfitting

detection wait time, and high random strength (normally 1) [55], [23], [56]. The functions of the parameters used in the two methods are similar, as illustrated in Table 3.

Table 3. Model Parameters

Method	Parameter
Catboost	Iteration = 500, learning_rate=0.5, depth=6, l2_leaf_reg=6, random_strength=1, verbose=100, od_type='IncToDec', od_wait=50
LightGBM	n_estimators=500, learning_rate=0.5, max_depth=6, lambda_l2=6, min_split_gain=1, verbosity=1, eval_metric='rmse' early_stopping_rounds=50

To mitigate potential overfitting and high variance also ehance the robustness of the model, the l2_leaf_reg parameter in CatBoost is increased from its default value of 3 to 6. Additionally, the Overfitting Detector (od_type and od_wait) is employed to prevent overfitting and reduce the computational time of LOOCV, stopping training when overfitting is detected. Furthermore, the model depth is set to 6 to balance bias and variance, ensuring stable performance with LOOCV [22].

3.5. Model Evaluate

The model is evaluated using LOOCV for an unbiased generalization error assessment, where each data point serves as a validation set while the rest are used for training, repeated iteratively for all observations. K-Fold Cross-Validation is also used for comparison, dividing the dataset into multiple folds [36], [38]. Table 4 presents key metrics (accuracy, MAE, RMSE, MSE, MAPE) to assess performance, identify improvements, and ensure model reliability and optimization.The fold used for comparison is 5,10,15,20 fold and other parameter used in K-Fold is set to default to get same comparison and insight [57], [58].

Table 4. Model Result

Method	Evaluation Model	LOOCV Value	K-fold Value (5 fold)	K-fold Value(10 Fold)	K-fold Value(15 Fold)	K-fold Value(20 Fold)
Catboost	MAE	0.91	1.51	1.48	1.37	1.40
	MSE	2.31	3.95	3.70	3.46	3.53
	RMSE	0.91	1.97	1.89	1.82	1.81
	MAPE	15%	24%	23%	22%	23%
LightGBM	MAE	1.1	1.50	1.48	1.46	1.46
	MSE	2.9	4.01	4.00	3.95	3.79
	RMSE	1.1	1.98	1.96	1.93	1.88
	MAPE	18%	24%	24%	24%	24%

Based on the results, the CatBoost model with LOOCV achieved the with an MAE of 0.91, an MSE of 2.31, an RMSE of 0.91, and a MAPE of 15%. Meanwhile, the LightGBM model had a slightly higher MAE of 1.1, an MSE of 2.9, an RMSE of 1.1, and a MAPE of 18%. This indicates that CatBoost outperforms LightGBM in the LOOCV and K-Fold scenario with good accuracy [31].

In K-fold validation with 5, 10, 15, and 20 folds, the evaluation error values tend to decrease as the number of folds increases, suggesting that the model becomes more stable with a higher number of folds. For CatBoost, the MAE ranges from 1.51 to 1.37, MSE from 3.95 to 3.53, RMSE from 1.97 to 1.81, and MAPE from 24% to 22%. Meanwhile, LightGBM exhibits a relatively higher MAE than CatBoost, ranging from 1.50 to 1.46, with MSE, RMSE, and MAPE values that are also generally higher in most scenarios [59], [60]. The LOOCV model demonstrates strong generalization, supported by effective cross-validation and hyperparameter tuning, which help prevent overfitting and enhance performance, especially on small datasets [61].

Fig. 5 shows a strong correlation between actual and predicted values using the CatBoost model with the Leave-One-Out Cross Validation (LOOCV) method. The x-axis represents actual values, while the y-axis represents predictions. Green dots indicate model predictions, and the dashed red line serves as a reference where predictions equal actual values (y = x). Predictions closer to this line indicate better performance, while deviations suggest errors.

Although LOOCV has limitation such as high computational complexity and may lead to high variance in some situations, this research addresses these issues by using the Ridge regularization parameter to prevent high variance results and the Overfitting Detector to prevent prolonged processing time and overfitting when applying the LOOCV method [41], [23].

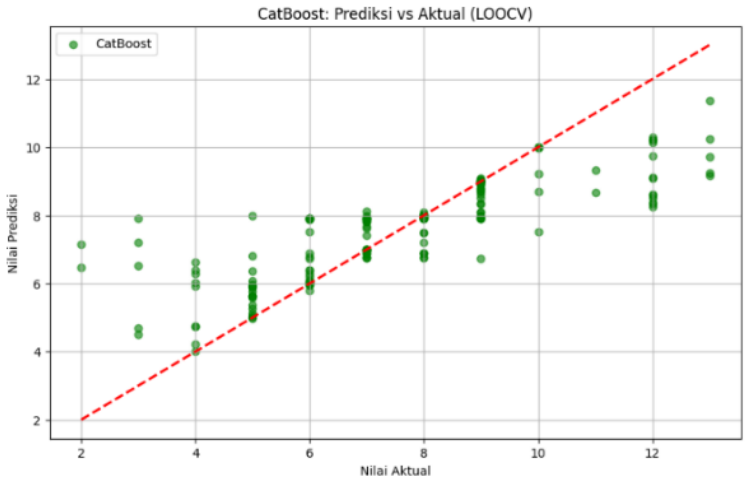

Fig. 5. Predicted Result and Actual Data Comparison

The Fig. 6 illustrates the Feature Importance of the CatBoost model, showing the contribution of each feature to the prediction results. Coffee price is the most influential factor, significantly affecting the model's predictions. Location, coffee type, and preferred coffee taste also play important roles. Meanwhile, features like age, gender, marital status, and serving preference have lower importance scores, indicating a lesser impact on the model's predictions. This result is also strengthened by previous studies, indicating that price and location greatly affect buying interest or purchase volume in MSMEs, particularly in the coffee business. This reinforces the crucial role of price in purchase frequency, consistent with earlier research on its positive influence on consumer interest [62], [63].

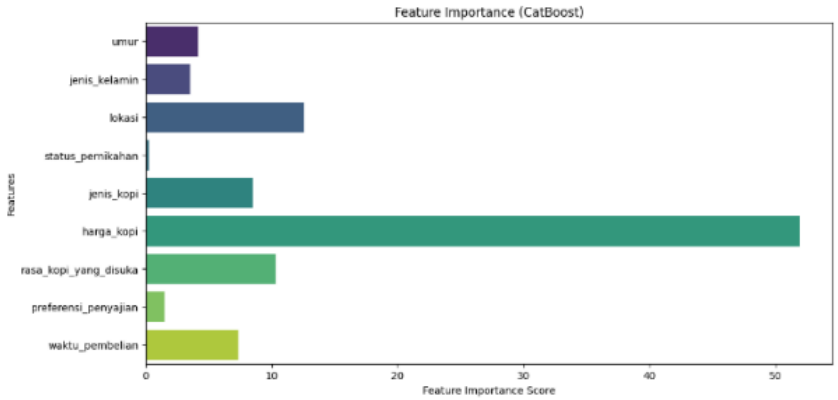

Fig. 6. Feature Importance Analysis

To optimize and implement model performance, achieving a MAPE of 15% indicates that CatBoost has an accuracy of 85%. Business owners can leverage these predictions for strategic decision-making. If the business owner sets the age parameter to 28, selects espresso as the coffee type, sets the selling time to night, and chooses "not married" as the marital status, with a predicted average frequency of buyers of 48 or lower, it is advisable to focus on Latte coffee and target the Genteng subdistrict to boost sales. However, if the predicted frequency exceeds 48, the focus should shift to the next best-selling coffee, such as Mocha or milk coffee, catering to customers who prefer sweeter flavors.

Comparing the results of this study with previous research reveals certain differences. Previous studies conducted by Adya *et. al.* implemented prediction of coffee quality by LightGBM in classification scope with train test split threshold method that set to 70% , that indicates pretty good accuracy for prediction that acquire 72% of accucary [8], It can be concluded that the Light Gradient Boosting algorithm in this study successfully predicted coffee quality, although there remains a 28% chance that the predicted results may differ from the actual outcomes. Other studies conducted by Wahyuningsih *et. al* implemented in enhancing coffee shop sales in association scope using FP-growth method, the confidence value of the method is 0.692 indicates a strong likelihood of consumers purchasing these items together [6].

A comparison of these studies shows that various methods and approaches can enhance business profits and reduce future losses. This research focuses on multiple coffee shop locations and consumer preferences using regression, whereas previous studies examined a single branded coffee shop, emphasizing coffee quality or production through classification and association methods. Despite different approaches, all studies achieved satisfactory accuracy in supporting coffee business growth.

4. CONCLUSION

This study was motivated by previous findings indicating that many coffee business owners experience losses due to a lack of understanding of market demands and target segments. As a response, the research aims to assist business owners and new entrants in minimizing potential future losses. The analysis results demonstrate that the CatBoost machine learning model outperforms LightGBM in predicting coffee shop sales. Specifically, model evaluation shows that CatBoost achieves a MAPE of 15%, which is 3% lower than LightGBM's 18%.

In addition, CatBoost performs better in terms of MAE and RMSE, both scoring 0.91, with a 0.2-point advantage over LightGBM. However, for MSE, CatBoost records a slightly higher value of 2.31 compared to LightGBM. Although CatBoost with LOOCV outperforms other models, it has some limitations, including high computational costs and result variance. To mitigate these issues, applying strong regularization and utilizing an overfitting detector can help minimize these challenges.

This study employs the Purposive Snowball Sampling method, which effectively reaches hard-to-access populations through social networks. However, this approach has limitations, including potential bias and lack of representativeness. To address these issues, the method is reinforced with in-depth structured interviews and supported by local coffee shop sales data, ensuring more reliable and representative results across a broader area. The main contribution of this research lies in offering data-driven solutions powered by machine learning, enabling business owners to mitigate the risk of losses.

With a MAPE of 15%, the model's predictions possibility deviate from actual results by 15%, meaning CatBoost achieves 85% accuracy, supporting business decisions. If a 28-year-old, unmarried customer prefers espresso at night and the predicted buyer frequency is 48 or lower, it is recommended to focus on Latte in Genteng. If it exceeds 48, shifting to Mocha or milk coffee to cater to sweeter preferences is advisable.

For future research, efforts should be directed toward developing more accurate machine learning models, such as time series or advanced regression techniques. Implementing ARIMA, SARIMA, or deep learning can improve predictive accuracy. Additionally, optimizing model performance through other cross-validation methods, including K-Fold, Stratified K-Fold, or other techniques, is strongly recommended.

REFERENCES

[1] F. Wang and J. Aviles, "Enhancing Operational Efficiency: Integrating Machine Learning Predictive Capabilities in Business Intellgence for Informed Decision-Making," *Front. Business, Econ. Manag.*, vol. 9, no. 1, pp. 282–286, May 2023, https://doi.org/10.54097/fbem.v9i1.8694.

[2] S. Nainggolan, E. Kernalis, and D. Z. Carolin, "Analysis of Factors Affecting the Behavior of Coffee Shop Consumers in Jambi City," *Randwick Int. Soc. Sci. J.*, vol. 3, no. 1, pp. 53–60, 2022, https://doi.org/10.47175/rissj.v3i1.369.

[3] A. Mukhlis, A. Moeins, and W. Sunaryo, "Development Strategies for Micro, Small, and Medium Enterprises (Msme) By Improving the Quality of Human Resources," *Int. J. Econ. Educ. Entrep.*, vol. 2, no. 2, pp. 525–536, 2022, https://doi.org/10.53067/ije3.v2i2.91.

[4] A. Daengs, GS, B. Pramono, A. I. Soemantri, and R. B. Kusumo Negoro, "Orientation Entrepreneurial Effects on MSME Performance Facilitated by Surabaya Commerce Department through Marketing Strategy as a Moderating Variable," *Int. J. Adv. Eng. Manag. Res.*, vol. 08, no. 05, pp. 30–41, 2023, https://doi.org/10.51505/ijaemr.2023.8503.

[5] D. A. N. Menengah, "Factor Affecting Business Sustainability of Small and Medium Coffee Shop," *J. Teknol. Ind. Pertan.*, vol. 30, no. 3, pp. 308–318, 2020, https://doi.org/10.24961/j.tek.ind.pert.2020.30.3.308.

[6] W. Wahyuningsih and P. T. Prasetyaningrum, "Enhancing Sales Determination for Coffee Shop Packages through Associated Data Mining: Leveraging the FP-Growth Algorithm," *J. Inf. Syst. Informatics*, vol. 5, no. 2, pp. 758–770, 2023, https://doi.org/10.51519/journalisi.v5i2.500.

[7] L. Setiyani and W. H. Utomo, "Arabica Coffee Price Prediction Using the Long Short Term Memory Network (LSTM) Algorithm," *Sci. J. Informatics*, vol. 10, no. 3, pp. 287–296, 2023, https://doi.org/10.15294/sji.v10i3.44162.

[8] A. Z. Putra, C. Chalvin, A. Nurhadi, A. E. Tambun, and S. Defha, "Coffee Quality Prediction with Light Gradient Boosting Machine Algorithm Through Data Science Approach," *Sinkron*, vol. 8, no. 1, pp. 563–573, 2023, https://doi.org/10.33395/sinkron.v8i1.12169.

[9] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: Unbiased boosting with categorical features," *Adv. Neural Inf. Process. Syst.*, pp. 6638–6648, 2018, https://doi.org/10.48550/arXiv.1706.09516.

[10] S. Chen, H. Jin, and L. Li, "Analysis and Comparison of House Price Prediction Based on XGboost and LightGBM," *Adv. Econ. Manag. Polit. Sci.*, vol. 46, no. 1, pp. 55–61, 2023, https://doi.org/10.54254/2754-1169/46/20230317.
[11] A. N. Karabulut, "Comparing the Young People's Coffee Shop Perceptions with Their Senses of Taste," *Yönetim ve Ekon. Derg.*, vol. 30, no. 1, pp. 1–19, 2023, https://doi.org/10.18657/yonveek.1244119.
[12] A. M. B. Wicaksana, S. Suharno, and W. Supartono, "The Impact of Consumer Behavior and Marketing Mix on the Decision to Buy Coffee at Coffee Shops in the Sleman Region During the Covid-19 Pandemic," *Agroindustrial J.*, vol. 8, no. 1, p. 520, 2022, https://doi.org/10.22146/aij.v8i1.73543.
[13] A. S. R. M. Sinaga, R. E. Putra, and A. S. Girsang, "Prediction measuring local coffee production and marketing relationships coffee with big data analysis support," *Bull. Electr. Eng. Informatics*, vol. 11, no. 5, pp. 2764–2772, Oct. 2022, https://doi.org/10.11591/eei.v11i5.4082.
[14] W. A. Limont, J. T. Łukasiewicz-Wieleba, A. Demianowska, and M. Jabłonowska, "The Snowball Sampling Strategy in the Field of Social Sciences. Contexts and Considerations.," *Przegląd Badań Eduk. (Educational Stud. Rev.*, vol. 2, no. 43, pp. 87–104, Sep. 2024, https://doi.org/10.12775/PBE.2022.001.
[15] J. Sayyad, K. Attarde, and N. Saadouli, "Optimizing e-commerce Supply Chains with Categorical Boosting: A Predictive Modeling Framework," *IEEE Access*, 2024, https://doi.org/10.1109/ACCESS.2024.3447756.
[16] Y. Zou, C. Gao, and H. Gao, "Business Failure Prediction Based on a Cost-Sensitive Extreme Gradient Boosting Machine," *IEEE Access*, vol. 10, pp. 42623–42639, 2022, https://doi.org/10.1109/ACCESS.2022.3168857.
[17] M. Idhom, A. Fauzi, T. Trimono, and P. Riyantoko, "Time Series Regression: Prediction of Electricity Consumption Based on Number of Consumers at National Electricity Supply Company," *TEM J.*, vol. 12, no. 3, pp. 1575–1581, 2023, https://doi.org/10.18421/TEM123-39.
[18] M. Idhom, I. G. P. A. Buditjahjanto, Munoto, Trimono, and P. A. Riyantoko, "Antithesis of Human Rater: Psychometric Responding to Shifts Competency Test Assessment Using Automation (AES System)," *Stud. Learn. Teach.*, vol. 4, no. 2, pp. 329–340, 2023, https://doi.org/10.46627/silet.v4i2.291.
[19] W. Liang, S. Luo, G. Zhao, and H. Wu, "Predicting hard rock pillar stability using GBDT, XGBoost, and LightGBM algorithms," *Mathematics*, vol. 8, no. 5, pp. 1–17, 2020, https://doi.org/10.3390/MATH8050765.
[20] A. Odeh, Q. A. Al-Haija, A. Aref, and A. A. Taleb, "Comparative Study of CatBoost, XGBoost, and LightGBM for Enhanced URL Phishing Detection: A Performance Assessment," *J. Internet Serv. Inf. Secur.*, vol. 13, no. 4, pp. 1–11, 2023, https://doi.org/10.58346/JISIS.2023.I4.001.
[21] Y. F. Zamzam, T. H. Saragih, R. Herteno, Muliadi, D. T. Nugrahadi, and P. H. Huynh, "Comparison of CatBoost and Random Forest Methods for Lung Cancer Classification using Hyperparameter Tuning Bayesian Optimization-based," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 6, no. 2, pp. 125–136, 2024, https://doi.org/10.35882/jeeemi.v6i2.382.
[22] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," *J. Big Data*, vol. 7, no. 1, 2020, https://doi.org/10.1186/s40537-020-00369-8.
[23] M. Nagassou, R. W. Mwangi, and E. Nyarige, "A Hybrid Ensemble Learning Approach Utilizing Light Gradient Boosting Machine and Category Boosting Model for Lifestyle-Based Prediction of Type-II Diabetes Mellitus," *J. Data Anal. Inf. Process.*, vol. 11, no. 04, pp. 480–511, 2023, https://doi.org/10.4236/jdaip.2023.114025.
[24] X. Lv, D. Gu, X. Liu, J. Dong, and Y. li, "Momentum prediction models of tennis match based on CatBoost regression and random forest algorithms," *Sci. Rep.*, vol. 14, no. 1, pp. 1–17, 2024, https://doi.org/10.1038/s41598-024-69876-5.
[25] Z. Lu, "Study of Mother-infant Behavioural Relationships based on Structural Equation Modelling and LightGBM Regression Models," *Sci. J. Intell. Syst. Res.*, vol. 6, no. 7, pp. 1–9, 2024, https://doi.org/10.54691/m7eqms74.
[26] A. Alsubayhin, M. S. Ramzan, and B. Alzahrani, "Crime Prediction Model using Three Classification Techniques: Random Forest, Logistic Regression, and LightGBM," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 1, pp. 240–251, 2024, https://doi.org/10.14569/IJACSA.2024.0150123.
[27] Y. Zhang, C. Zhu, and Q. Wang, "Lightgbm-based model for metro passenger volume forecasting," *IET Intell. Transp. Syst.*, vol. 14, no. 13, pp. 1815–1823, 2020, https://doi.org/10.1049/iet-its.2020.0396.
[28] L. Lin, J. Zhang, N. Zhang, J. Shi, and C. Chen, "Optimized LightGBM Power Fingerprint Identification Based on Entropy Features," *Entropy*, vol. 24, no. 11, 2022, https://doi.org/10.3390/e24111558.
[29] A. Botchkarev, "A New Typology Design of Performance Metrics to Measure Errors in Machine Learning Regression Algorithms," *Interdiscip. J. Information, Knowledge, Manag.*, vol. 14, no. 113, pp. 45–79, 2019, https://doi.org/10.28945/4184.
[30] A. T. Damaliana and S. Hidayati, "Implementation of Quantile Regression Neural Network Model for Forecasting Electricity Demand in East Java," *Proceeding - IEEE 8th Inf. Technol. Int. Semin. ITIS*, pp. 229–234, 2022, https://doi.org/10.1109/ITIS57155.2022.10009045.
[31] A. Uribeetxebarria, A. Castellón, and A. Aizpurua, "Optimizing Wheat Yield Prediction Integrating Data from Sentinel-1 and Sentinel-2 with CatBoost Algorithm," *Remote Sens.*, vol. 15, no. 6, 2023, https://doi.org/10.3390/rs15061640.
[32] A. M. Aviolla Terza Damaliana and D. A. Prasetya, "Forecasting The Occupancy Rate Of Star Hotels In Bali," *J. Stat.*, vol. 12, no. 1, pp. 24–33, 2024, https://doi.org/10.14710/JSUNIMUS. 12.1.2024.24-33.
[33] N. Putu, V. Ginanti, C. Wiedyaningsih, and E. Yuniarti, "Comparison Of Forecasting Drug Needs Using Time Series Methods In Healthcare Facilities : A Systematic Review.," *J. Farm. Sains dan Prakt.*, vol. 10, no. 2, pp. 156–165, 2024, https://doi.org/10.31603/pharmacy.v10i2.11145.

[34] E. Vivas, H. Allende-Cid, and R. Salas, "A Systematic Review of Statistical and Machine Learning Methods for Electrical Power Forecasting with Reported MAPE Score," *Entropy*, vol. 22, no. 12, p. 1412, Dec. 2020, https://doi.org/10.3390/e22121412.
[35] Y. L. Sukestiyarno, D. T. Wiyanti, L. Azizah, and W. Widada, "Algorithm Optimizer in GA-LSTM for Stock Price Forecasting," *Contemp. Math.*, vol. 5, no. 1, pp. 1–12, Jan. 2024, https://doi.org/10.37256/cm.5120243367.
[36] V. Lumumba, D. Kiprotich, M. Mpaine, N. Makena, and M. Kavita, "Comparative Analysis of Cross-Validation Techniques: LOOCV, K-folds Cross-Validation, and Repeated K-folds Cross-Validation in Machine Learning Models," *Am. J. Theor. Appl. Stat.*, vol. 13, no. 5, pp. 127–137, Oct. 2024, https://doi.org/10.11648/j.ajtas.20241305.13.
[37] A. Geroldinger, L. Lusa, M. Nold, and G. Heinze, "Leave-one-out cross-validation, penalization, and differential bias of some prediction model performance measures—a simulation study," *Diagnostic Progn. Res.*, vol. 7, no. 1, 2023, https://doi.org/10.1186/s41512-023-00146-0.
[38] I. Tougui, A. Jilbab, and J. El Mhamdi, "Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications," *Healthc. Inform. Res.*, vol. 27, no. 3, pp. 189–199, 2021, https://doi.org/10.4258/HIR.2021.27.3.189.
[39] C. O. Chavez-Chong, C. Hardouin, and A.-K. Fermin, "Ridge regularization for spatial autoregressive models with multicollinearity issues," *AStA Adv. Stat. Anal.*, vol. 109, no. 1, pp. 25–52, Mar. 2025, https://doi.org/10.1007/s10182-024-00496-0.
[40] A. R. Nur, A. K. Jaya, and S. Siswanto, "Comparative Analysis of Ridge, LASSO, and Elastic Net Regularization Approaches in Handling Multicollinearity for Infant Mortality Data in South Sulawesi," *J. Mat. Stat. dan Komputasi*, vol. 20, no. 2, pp. 311–319, 2023, https://doi.org/10.20956/j.v20i2.31632.
[41] C. M. Le, K. Levin, P. J. Bickel, and E. Levina, "Comment: Ridge Regression and Regularization of Large Matrices," *Technometrics*, vol. 62, no. 4, pp. 443–446, Oct. 2020, https://doi.org/10.1080/00401706.2020.1796815.
[42] C. Tirink, S. H. Abaci, and H. Onder, "Comparison of Ridge Regression and Least Squares Methods in the Presence of Multicollinearity for Body Measurements in Saanen Kids," *Iğdır Üniversitesi Fen Bilim. Enstitüsü Derg.*, vol. 10, no. 2, pp. 1429–1437, 2020, https://doi.org/10.21597/jist.671662.
[43] D. Barragán-Guerrero, M. Au, G. Gagnon, F. Gagnon, and P. Giard, "Early-detection scheme based on sequential tests for low-latency communications," *Eurasip J. Wirel. Commun. Netw.*, vol. 2023, no. 1, 2023, https://doi.org/10.1186/s13638-023-02240-9.
[44] A. Alabrah, "An Improved CCF Detector to Handle the Problem of Class Imbalance with Outlier Normalization Using IQR Method," *Sensors*, vol. 23, no. 9, 2023, https://doi.org/10.3390/s23094406.
[45] B. Dym and C. Fiesler, "Ethical and privacy considerations for research using online fandom data," *Transform. Work. Cult.*, vol. 33, pp. 1–19, 2020, https://doi.org/10.3983/twc.2020.1733.
[46] A. C. Haber, U. Sax, and F. Prasser, "Open tools for quantitative anonymization of tabular phenotype data: literature review," *Brief. Bioinform.*, vol. 23, no. 6, pp. 1–10, 2022, https://doi.org/10.1093/bib/bbac440.
[47] S. Sardjono, R. Y. R. Alamsyah, M. Marwondo, and E. Setiana, "Data Cleansing Strategies on Data Sets Become Data Science," *Int. J. Quant. Res. Model.*, vol. 1, no. 3, pp. 145–156, 2020, https://doi.org/10.46336/ijqrm.v1i3.71.
[48] D. A. Prasetya, A. P. Sari, P. A. Riyantoko, and T. M. Fahrudin, "The Effect of Information Quality and Service Quality on User Satisfaction of the Government of Kabupaten Malang," *TIERS Inf. Technol. J.*, vol. 4, no. 1, pp. 32–42, 2023, https://doi.org/10.38043/tiers.v4i1.4328.
[49] I. Bergelson, C. Tracy, and E. Takacs, "Best Practices for Reducing Bias in the Interview Process," *Curr. Urol. Rep.*, vol. 23, no. 11, pp. 319–325, Nov. 2022, https://doi.org/10.1007/s11934-022-01116-7.
[50] K. Isaksen *et al.*, "Interviewing adolescent girls about sexual and reproductive health: a qualitative study exploring how best to ask questions in structured follow-up interviews in a randomized controlled trial in Zambia," *Reprod. Health*, vol. 19, no. 1, pp. 1–11, 2022, https://doi.org/10.1186/s12978-021-01318-1.
[51] Q. Zheng, C. Yu, J. Cao, Y. Xu, Q. Xing, and Y. Jin, "Advanced Payment Security System:XGBoost, CatBoost and SMOTE Integrated," *arXiv e-prints, arXiv-2406*, 2024, https://doi.org/10.1109/MetaCom62920.2024.00063.
[52] C. S. K. Dash, A. K. Behera, S. Dehuri, and A. Ghosh, "An outliers detection and elimination framework in classification task of data mining," *Decis. Anal. J.*, vol. 6, p. 100164, 2023, https://doi.org/10.1016/j.dajour.2023.100164.
[53] B. Bach *et al.*, "Challenges and Opportunities in Data Visualization Education: A Call to Action," *IEEE Trans. Vis. Comput. Graph.*, vol. 30, no. 1, pp. 649–660, 2024, https://doi.org/10.1109/TVCG.2023.3327378.
[54] P. Cerda and G. Varoquaux, "Encoding High-Cardinality String Categorical Variables," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 3, pp. 1164–1176, Mar. 2022, https://doi.org/10.1109/TKDE.2020.2992529.
[55] M. Fan, K. Xiao, L. Sun, S. Zhang, and Y. Xu, "Automated Hyperparameter Optimization of Gradient Boosting Decision Tree Approach for Gold Mineral Prospectivity Mapping in the Xiong'ershan Area," *Minerals*, vol. 12, no. 12, 2022, https://doi.org/10.3390/min12121621.
[56] A. Maulana, R. P. F. Afidh, N. B. Maulydia, G. M. Idroes, and S. Rahimah, "Predicting Obesity Levels with High Accuracy: Insights from a CatBoost Machine Learning Model," *Infolitika J. Data Sci.*, vol. 2, no. 1, pp. 17–27, 2024, https://doi.org/10.60084/ijds.v2i1.195.
[57] K. M. Hindrayani, T. M. Fahrudin, R. Prismahardi Aji, and E. M. Safitri, "Indonesian Stock Price Prediction including Covid19 Era Using Decision Tree Regression," *2020 3rd Int. Semin. Res. Inf. Technol. Intell. Syst. ISRITI*, pp. 344–347, 2020, https://doi.org/10.1109/ISRITI51436.2020.9315484.

[58] I. G. S. M. Diayasa, M. Idhom, A. Fauzi, and A. T. Damaliana, "Stacking Ensemble Methods to Predict Obesity Levels in Adults," *Proceeding - IEEE 8th Inf. Technol. Int. Semin. ITIS*, pp. 339–344, 2022, https://doi.org/10.1109/ITIS57155.2022.10010260.
[59] P. Bagus, P. Putra Budiartha, C. Wiedyaningsih, E. Yuniarti, A. Agung, and A. Prithadewi, "Forecasting Drug Demand Using The Single Moving Average At Prof. dr. I.G.N.G. Ngoerah Hospital," *Maj. Farm.*, vol. 19, no. 3, pp. 394–402, 2023, https://doi.org/10.22146/farmaseutik.v19i3.86207.
[60] M. Hani'ah, M. Z. Abdullah, W. I. Sabilla, S. Akbar, and D. R. Shafara, "Google Trends and Technical Indicator based Machine Learning for Stock Market Prediction," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 22, no. 2, pp. 271–284, Mar. 2023, https://doi.org/10.30812/matrik.v22i2.2287.
[61] P. Charilaou and R. Battat, "Machine learning models and over-fitting considerations," *World J. Gastroenterol.*, vol. 28, no. 5, pp. 605–607, Feb. 2022, https://doi.org/10.3748/wjg.v28.i5.605.
[62] E. Efendi, M. Butarbutar, R. M. Girsang, E. Chandra, and V. Candra, "Purchase Interest Reviewed Based On Price And Location At Danu Jaya Birdshop Pematang Siantar," *Mak. J. Manaj.*, vol. 9, no. 1, pp. 119–126, Jun. 2023, https://doi.org/10.37403/mjm.v9i1.579.
[63] A. Syaidah, M. Munawaroh, and L. Susilowati, "Influence of Price And Location on Belikopi Ploso Consumer Purchasing Decisions," *J. Bus. Manag. Econ. Dev.*, vol. 1, no. 03, pp. 556–564, 2023, https://doi.org/10.59653/jbmed.v1i03.297.

BIOGRAPHY OF AUTHORS

Calvien Danny Nariyana, is an undergraduate student in Computer Science at Universitas Pembangunan Nasional "Veteran" Jawa Timur, Surabaya. He has a keen interest in machine learning and artificial intelligence, with his primary research focus on leveraging data science to tackle a variety of challenges in science and technology. Calvien is deeply passionate about utilizing advanced methods in data analysis to create innovative and practical solutions. Email: 21083010040@student.upnjatim.ac.id Orchid: https://orcid.org/0009-0009-6990-1759

Mohammad Idhom is a lecturer in Informatics Engineering at UPN Veteran Jawa Timur, Indonesia. He holds a Bachelor's in Computer Science from UPN Veteran Jawa Timur, a Master's in Computer Engineering from Atma Jaya University Yogyakarta, and a Doctorate from Universitas Negeri Surabaya. Since 2022, he has served as Head of Career Development and Entrepreneurship and Editor-in-Chief of the *International Journal of Computer, Network Security, and Information System*. He has obtained multiple patents and is a patent compiler at UPN Veteran East Java. His research interests include Machine Learning, AI, robotics, network security, intelligent control, and IoT. Email: idhom@upnjatim.ac.id. Orchid: https://orcid.org/0000-0002-6460-9507.

Trimono, is a lecturer in the Computer Science Study Program, UPN Veteran Jawa Timur, Indonesia. He earned a Bachelor's degree in Statistics from Diponegoro State University and a Master's degree in Mathematics from Bandung Institute of Technology. He has a strong passion for Risk Management, Time Series Analysis, and Financial Statistics. As a lecturer, Trimono actively teaches and mentors Mathematics students, integrating multiple scientific disciplines through the application of Data Science. Email: trimono.stat@upnjatim.ac.id. Orchid: https://orcid.org/0000-0003-4380-57840.