Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 2, April 2024
https://doi.org/10.47709/cnahpc.v6i2.3790

Submitted : April 10, 2024
Accepted : April 27, 2024
Published : Apr 29, 2024

Analysis of Gradient Boosting, XGBoost, and CatBoost on Mobile Phone Classification

Agus Fahmi Limas Ptr[1)*], Muhammad Mizan Siregar[2)], Irwan Daniel[3)]
[1,2,3)]Universitas Deli Sumatera, Indonesia
[1)]agusfahmilimasptr@gmail.com, [2)]mizan.siregar1@gmail.com, [3)]irwandaniel@gmail.com

ABSTRACT

In the ever-evolving landscape of mobile phone technology, accurately classifying device specifications is paramount for market analysis and consumer decision-making. This research conducts a comprehensive analysis of mobile phone specification classification using three prominent machine learning algorithms: Gradient Boosting, XGBoost, and CatBoost. Through meticulous dataset acquisition and preprocessing steps, including resolution normalization and price categorization, features essential for classification analysis were standardized. Robust cross-validation techniques were employed to assess model performance effectively. The study demonstrates the significant impact of normalization techniques on improving model performance across all algorithms and fold variations. CatBoost consistently emerges as the top-performing algorithm, followed closely by XGBoost, with Gradient Boosting displaying respectable performance. Notably, CatBoost consistently achieves the highest AUC values and accuracy scores, demonstrating superior performance in accurately classifying mobile phone specifications. These findings underscore the importance of preprocessing methods and algorithm selection in achieving optimal classification results. For mobile phone manufacturers, leveraging machine learning algorithms for effective classification can inform product development strategies, optimizing offerings based on consumer preferences. Similarly, for data analysts, employing appropriate preprocessing techniques and algorithmic approaches can lead to more accurate predictions and informed decision-making. Future research avenues include exploring advanced preprocessing methods, investigating alternative algorithms, and incorporating additional features or datasets to enrich the classification process. Overall, this research contributes to understanding mobile phone specification classification through machine learning methodologies, offering actionable insights for industry practitioners and researchers to address evolving market dynamics and consumer preferences.

Keywords: Mobile phone classification; Gradient Boosting; XGBoost; CatBoost; Normalization

INTRODUCTION

In an era defined by technological advancements, the mobile phone industry stands as a cornerstone of innovation, continually introducing new devices with diverse specifications to cater to the varying needs of consumers (Singh & More, 2022). Accurately classifying these specifications is of paramount importance, serving as a fundamental step in market analysis, product development, and consumer decision-making processes (Sun, Luh, Zhao, & Sun, 2022). Understanding the distinct features and capabilities of mobile phones, categorized into different tiers such as entry-level, medium-level, and flagship devices, enables manufacturers to tailor their offerings to specific target demographics effectively (Kabeyi, 2018). Moreover, for consumers, precise classification facilitates informed purchasing decisions by providing clarity on performance expectations and value propositions (Chen, Samaranayake, Cen, Qi, & Lan, 2022). Consequently, the ability to classify mobile phone specifications accurately holds immense significance, shaping the dynamics of the industry and driving advancements in technology and consumer experiences (J. C. Wang, Hsieh, & Kung, 2023).

Gradient boosting, a powerful ensemble learning technique, has emerged as a cornerstone in the realm of machine learning, renowned for its exceptional predictive accuracy and versatility across various domains (Adler & Painsky, 2022). At its core, gradient boosting constructs a strong predictive model by iteratively combining an ensemble of weak learners, each of which seeks to minimize the errors of its predecessors (Callens, Morichon, Abadie, Delpey, & Liquet, 2020). By continuously refining the model through gradient descent, gradient boosting effectively captures complex patterns in the data, making it particularly adept at handling structured datasets with high-dimensional features (Jinan, Situmorang, & Rosnelly, 2023). This method has found widespread application in classification tasks, where it excels in distinguishing between multiple classes with remarkable precision (Aravind, Shyry, & Felix, 2019;

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 2, April 2024
https://doi.org/10.47709/cnahpc.v6i2.3790

Submitted : April 10, 2024
Accepted : April 27, 2024
Published : Apr 29, 2024

Fayaz et al., 2020; Suryana, Warsito, & Suparti, 2021). In this research, we delve into the nuances of gradient boosting and its efficacy in classifying mobile phone specifications across different tiers, shedding light on its potential to enhance decision-making processes in the mobile phone industry.

XGBoost, an optimized implementation of gradient boosting, has garnered considerable attention in recent years for its exceptional performance and scalability in handling large-scale datasets (Abdurohman & Putrada, 2023). Originating from the need to address the limitations of traditional gradient boosting algorithms, XGBoost introduces several innovative techniques, including regularization and parallel processing, to enhance model efficiency and generalization capabilities (Kumar, Kedam, Sharma, Mehta, & Caloiero, 2023). Leveraging a combination of tree-based ensemble models and advanced optimization strategies, XGBoost consistently achieves state-of-the-art results in various machine learning competitions and real-world applications (Ampomah, Qin, & Nyame, 2020). Its ability to balance model complexity and predictive accuracy makes it particularly well-suited for classification tasks, where it excels in accurately classifying complex datasets with high-dimensional features (Dwinanda, Satyahadewi, & Andani, 2023; Herni Yulianti, Oni Soesanto, & Yuana Sukmawaty, 2022; Siringoringo, Perangin Angin, & Rumahorbo, 2022). In this study, we explore the capabilities of XGBoost as a machine learning classifier for classifying mobile phone specifications across different tiers, offering insights into its performance and potential contributions to the mobile phone industry.

CatBoost, a relatively recent addition to the ensemble learning paradigm, has quickly gained prominence for its unique approach to handling categorical features and robustness to noisy data (Safaei et al., 2022). Developed by Yandex researchers, CatBoost introduces novel techniques, such as ordered boosting and feature combination trees, to effectively address challenges associated with categorical variables in machine learning tasks (Hancock & Khoshgoftaar, 2020). Unlike traditional gradient boosting algorithms, CatBoost requires minimal preprocessing of categorical features, making it particularly appealing for datasets with heterogeneous feature types (Breskuvien & Dzemyda, 2023). Additionally, CatBoost's adaptive regularization strategy mitigates overfitting concerns, ensuring reliable model performance in diverse real-world scenarios and enabling competitive performance with minimal hyperparameter tuning in classification tasks across various domains (Chang, Wang, Yang, & Qin, 2023; Ibrahim, Ridwan, Muhammed, Abdulaziz, & Saheed, 2020; Jasman, Fadhlullah, Pratama, & Rismayani, 2022). In this investigation, we delve into the capabilities of CatBoost as a machine learning classifier for classifying mobile phone specifications, elucidating its strengths and potential implications for the mobile phone industry.

The primary objective of this research is to conduct a comprehensive comparative analysis of three prominent machine learning classifiers—gradient boosting, XGBoost, and CatBoost—on the classification of mobile phone specifications across various tiers. With the rapid evolution of mobile technology, manufacturers continuously introduce a plethora of devices tailored to diverse consumer preferences and requirements. Understanding the distinct characteristics and performance attributes of these devices, categorized into different levels such as entry-level, medium-level, and flagship categories, is crucial for both industry stakeholders and consumers alike. By comparing the performance of these classifiers across different levels of mobile phone specifications, this study aims to elucidate their efficacy in accurately categorizing devices based on their features and capabilities. Through rigorous experimentation and evaluation, insights gained from this comparative analysis will not only contribute to advancing the field of machine learning but also provide valuable guidance for decision-making processes within the mobile phone industry, ultimately facilitating informed product development strategies and consumer choices.

LITERATURE REVIEW

Gradient Boosting Algorithm

Gradient boosting is a powerful ensemble learning technique that has gained significant traction in the field of machine learning due to its ability to produce highly accurate predictions across various domains. At its core, gradient boosting constructs an ensemble of weak learners, typically decision trees, in a sequential manner, where each subsequent learner focuses on reducing the errors made by the previous ones (Devos, Meert, & Davis, 2020). By iteratively optimizing a predefined loss function through gradient descent, gradient boosting effectively learns complex relationships within the data and produces a strong predictive model (Heydarizad, Pumijumnong, Sorí, Salari, & Gimeno, 2022). In classification tasks, gradient boosting algorithms, such as Gradient Boosting Classifier (GBC) and its variants, have demonstrated remarkable performance in distinguishing between multiple classes with high precision and recall (Boldini, Grisoni, Kuhn, Friedrich, & Sieber, 2023).

A plethora of studies in the literature have showcased the effectiveness of gradient boosting algorithms in various classification tasks, including but not limited to sentiment analysis (Neelakandan & Paulraj, 2020), disease diagnosis (Suhendra et al., 2023), and fraud detection (Sopiyan, Fauziah, & Wijaya, 2022). For instance, in the domain of

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 2, April 2024
https://doi.org/10.47709/cnahpc.v6i2.3790

Submitted : April 10, 2024
Accepted : April 27, 2024
Published : Apr 29, 2024

sentiment analysis, gradient boosting models have been employed to accurately classify text data into positive, negative, or neutral sentiments, leveraging features extracted from textual content. Similarly, in medical diagnosis, gradient boosting algorithms have been utilized to distinguish between different disease states based on patient symptoms and clinical data, aiding healthcare professionals in making accurate diagnoses and treatment decisions. Moreover, in financial fraud detection, gradient boosting techniques have been applied to identify fraudulent transactions by analyzing patterns and anomalies in transactional data.

Overall, the literature underscores the versatility and efficacy of gradient boosting algorithms in tackling complex classification tasks across diverse domains. The adaptability of these algorithms to handle high-dimensional data and capture intricate relationships within the data make them indispensable tools for researchers and practitioners seeking robust solutions for classification problems. As such, in the context of this research, the application of gradient boosting algorithms to classify mobile phone specifications holds promise for providing accurate and reliable insights into the features and capabilities of mobile devices across different tiers.

XGboost Algorithm

XGBoost, short for eXtreme Gradient Boosting, has emerged as a leading ensemble learning algorithm that extends traditional gradient boosting with several innovative features. Developed by Tianqi Chen, XGBoost introduces enhancements in regularization, parallel computing, and tree optimization techniques, resulting in superior performance and scalability compared to conventional gradient boosting methods (Tusher, Rahman, Islam, & Hossain, 2024). One of the key advantages of XGBoost lies in its ability to handle sparse data efficiently, making it particularly well-suited for datasets with high-dimensional features or missing values (Zhang, 2022). Moreover, XGBoost incorporates both first-order and second-order gradient information during the optimization process, enabling it to capture complex relationships within the data more effectively (Qinghe, Wen, Boyan, Jong, & Junlong, 2022). These advancements contribute to enhanced model generalization and reduced overfitting, resulting in improved predictive accuracy and robustness.

A wealth of literature exists demonstrating the efficacy of XGBoost across various classification tasks, ranging from text classification (Putri, Dwifebri, & Adiwijaya, 2023) and image recognition (Liew, Hameed, & Clos, 2021) to financial forecasting (Jabeur, Mefteh-Wali, & Viviani, 2024) and bioinformatics (N. Wang, Zeng, Li, Wu, & Li, 2021). In the domain of text classification, researchers have utilized XGBoost to achieve state-of-the-art results in sentiment analysis, topic modeling, and spam detection by leveraging its ability to handle high-dimensional textual features and imbalanced datasets. Similarly, in image recognition tasks, XGBoost has been employed to classify images into different categories based on extracted features, demonstrating its versatility in handling diverse data modalities. Furthermore, in financial applications, XGBoost has been utilized for credit risk assessment, stock price prediction, and algorithmic trading, where its robustness to noisy data and interpretability are highly valued.

Overall, the literature underscores the effectiveness of XGBoost as a versatile and powerful algorithm for classification tasks across various domains. Its ability to leverage advanced optimization techniques and handle complex data structures makes it a preferred choice for researchers and practitioners seeking high-performance solutions. In the context of this research, the application of XGBoost to classify mobile phone specifications offers the potential to deliver accurate and reliable insights into the characteristics of mobile devices across different tiers, thereby informing decision-making processes within the mobile phone industry.

CatBoost Algorithm

CatBoost, a relatively recent addition to the ensemble learning paradigm, distinguishes itself from traditional boosting algorithms through its innovative approach to handling categorical features and robustness to noisy data (Hussain et al., 2021). Developed by Yandex researchers, CatBoost introduces several novel techniques that set it apart from its counterparts, such as gradient boosting and XGBoost. One of the key features of CatBoost is its ability to handle categorical variables seamlessly without the need for extensive preprocessing or one-hot encoding (Fedorov & Petrichenko, 2020). By employing an efficient algorithm for gradient boosting with decision trees, CatBoost incorporates a specialized treatment for categorical features during the training process, effectively capturing the interactions between categories and numerical variables (Dutta & Wahab Sait, 2024). Additionally, CatBoost implements an ordered boosting scheme and utilizes feature combination trees to enhance model performance and reduce overfitting (Jabeur, Gharib, Mefteh-Wali, & Arfi, 2021).

A growing body of literature showcases the effectiveness of CatBoost in addressing a wide range of classification problems across diverse domains. In the realm of natural language processing (NLP), researchers have utilized CatBoost to achieve competitive results in sentiment analysis (Karanikola, Davrazos, Liapis, & Kotsiantis, 2023), text

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 2, April 2024
https://doi.org/10.47709/cnahpc.v6i2.3790

Submitted : April 10, 2024
Accepted : April 27, 2024
Published : Apr 29, 2024

classification (Gonçalves Freitas, Edokawa, Carvalho Valadares Rodrigues, Thomé de Farias, & Rodrigues de Alencar, 2023), and document categorization tasks (Avelino, Felizmenio, & Naval, 2022) by exploiting its ability to handle categorical features inherent in textual data. Moreover, in healthcare applications, CatBoost has been employed for disease diagnosis (Zheng et al., 2023), patient risk stratification (Lo, Liao, Chen, Chang, & Li, 2021), and medical image analysis (Lin et al., 2023), where its robustness to noisy and heterogeneous data has proven beneficial. Overall, the literature underscores the utility of CatBoost as a versatile and powerful algorithm for classification tasks, offering significant advantages in handling complex datasets and delivering robust predictive models.

In the context of this research, the application of CatBoost to classify mobile phone specifications presents an opportunity to leverage its unique capabilities in handling categorical features and noisy data. By comparing the performance of CatBoost with other boosting algorithms, such as gradient boosting and XGBoost, insights can be gained into the strengths and limitations of each approach, informing decision-making processes within the mobile phone industry.

METHOD

Dataset Acquisition

The dataset utilized in this research comprises mobile phone specifications collected from the kaggle.com website. This dataset contains 980 mobile phone data from a diverse range of brands available in the market. Each mobile phone entry in the dataset includes a variety of features essential for classification analysis. These features encompass processor brand (Snapdragon, Exynos, Dimensity, Bionic, Helio, Unisoc, Tiger, Google, SC9863A, Spreadtrum, Fusion, Kirin or Mediatek), CPU Cores (4, 6, or 8), CPU speed (from 1.2 to 3.22 GHz), RAM capacity (from 1 to 18 GB), internal storage capacity (from 8 to 1024 GB), screen size (from 3.54 to 8.03 inch), screen refresh rate (from 60 to 240 mHz), screen resolutions (from480 X 640 to 2460 X 1080 pixels), front and rear camera resolution (in megapixels), battery capacity (from 1821 to 2200 mAh), and price.

Data collection was performed meticulously to ensure accuracy and consistency across entries. Additionally, efforts were made to cover a wide spectrum of mobile phone models, ranging from entry-level to flagship devices, to provide a comprehensive representation of the market landscape. The dataset serves as a valuable resource for conducting comparative analysis and evaluating the performance of machine learning classifiers in classifying mobile phone specifications across different tiers.

Preprocessing

In this section, we describe the preprocessing steps applied to the features used in our research on mobile phone classification using Gradient Boosting, XGBoost, and CatBoost algorithms.

1. Resolution Normalization
   The resolution feature, which originally consisted of 957 categories, was normalized using a category labeling method to reduce the dimensionality and improve model performance. The following rules were applied to categorize resolutions into broader classes:
   Low-end: Resolutions ranging from 480 X 640 to 854 X 480.
   Mid-range: Resolutions ranging from 1080 X 1920 to 1080 X 2640.
   High-end: Resolutions ranging from 1116 X 2480 to 1560 X 720.
   Premium: Resolutions ranging from 1600 X 720 to 2460 X 1080.
2. Price Categorization
   The price feature was categorized into three levels to capture the price range of mobile phones:
   Entry Level: Prices below 15,000 units.
   Medium Level: Prices equal to or greater than 15,000 units and less than 35,000 units.
   Flagship: Prices equal to or greater than 35,000 units.

These preprocessing steps were essential to standardize the features and make them compatible with the machine learning algorithms utilized in our classification task. By reducing the resolution categories and categorizing prices into meaningful levels, we aimed to enhance the efficiency and accuracy of our models in predicting mobile phone classifications.

Model Configuration

In this section, we delineate the setup of the Gradient Boosting, XGBoost, and CatBoost algorithms utilized to classify mobile phones in our study. The parameters employed for each model include:

1. Number of Trees: 100

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 2, April 2024
https://doi.org/10.47709/cnahpc.v6i2.3790

Submitted : April 10, 2024
Accepted : April 27, 2024
Published : Apr 29, 2024

2. Learning Rate: 0.1
3. Maximum Depth of Each Tree: 5

Each algorithm adheres to its standard procedures, as described below:

1. Gradient Boosting
   - Step 1: Initialize the model with an initial guess (using the mean of the target values).
   - Step 2: Fit a decision tree to the residuals (errors) of the previous model.
   - Step 3: Update the model by adding a fraction of the new tree to the previous model.
   - Step 4: Repeat Steps 2 and 3 for the specified number of trees.
   - Step 5: The final model is an ensemble of decision trees, where predictions are made by summing the predictions from each tree.
2. XGBoost (Extreme Gradient Boosting)
   - Step 1: Initialize the model with an initial guess.
   - Step 2: Fit a decision tree to the residuals of the previous model.
   - Step 3: Update the model by adding the prediction of the new tree multiplied by a shrinkage factor (learning rate) to the previous model.
   - Step 4: Apply regularization techniques to control overfitting.
   - Step 5: Repeat Steps 2-4 for the specified number of trees.
   - Step 6: The final model is an ensemble of decision trees, where predictions are made by summing the predictions from each tree.
3. CatBoost (Categorical Boosting)
   - Step 1: Handle categorical features internally.
   - Step 2: Initialize the model with an initial guess.
   - Step 3: Fit a decision tree to the residuals of the previous model.
   - Step 4: Apply special handling for categorical features to avoid overfitting.
   - Step 5: Update the model by adding the prediction of the new tree to the previous model.
   - Step 6: Repeat Steps 3-5 for the specified number of trees.
   - Step 7: The final model is an ensemble of decision trees, where predictions are made by summing the predictions from each tree.

By configuring each algorithm with 100 trees, a learning rate of 0.1, and a maximum depth of 5 for each tree, we aimed to strike a balance between model complexity and predictive performance. These algorithms were selected for their effectiveness in handling classification tasks and their ability to capture complex relationships within the data.

Evaluation Metrics

In this section, we present the evaluation metrics used to assess the performance of our models—Gradient Boosting, XGBoost, and CatBoost—employed for classifying mobile phones. We utilize cross-validation techniques with variations of K = 5, K = 10, and K = 20 to robustly evaluate the models' performance, measuring metrics including the Area Under the Receiver Operating Characteristic Curve (AUC), accuracy, F1 score, precision, and recall.

In the evaluation procedure, each model is assessed using the mentioned metrics to determine its efficacy in classification tasks. We compute the AUC, accuracy, F1 score, precision, and recall by employing the confusion matrix results alongside the respective formulas for each metric, as shown in formula (1) to (4) (Pardede, Firmansyah, Handayani, Riandini, & Rosnelly, 2022; Pardede & Hayadi, 2023).

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \quad (1)$$

$$Precision = \frac{True\ Positives}{True\ Positives+False\ Positives} \quad (2)$$

$$Recall = \frac{True\ Positives}{True\ Positives+False\ Negatives} \quad (3)$$

$$F1 = \frac{2\ X\ Precision\ X\ Recall}{Precision+Recall} \quad (4)$$

Additionally, we conduct a comparative analysis to evaluate the impact of normalization on model performance, contrasting results obtained with and without the application of the category labeling method. This comparison offers insights into the normalization technique's effectiveness in enhancing model performance and generalizability. Regarding the interpretation of metrics, the AUC measures the model's ability to differentiate between positive and

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 2, April 2024
https://doi.org/10.47709/cnahpc.v6i2.3790

Submitted : April 10, 2024
Accepted : April 27, 2024
Published : Apr 29, 2024

negative classes, with a higher AUC indicating better discrimination. Accuracy reflects the proportion of correctly classified instances, while the F1 score provides a balance between precision and recall, especially beneficial for imbalanced datasets. Precision signifies the proportion of true positive predictions among all positive predictions, whereas recall measures the proportion of true positive predictions relative to all actual positive instances in the dataset.

Through comparing the performance of models with and without normalization, we aim to demonstrate the efficacy of the category labeling method in enhancing model performance, thus highlighting the significance of preprocessing techniques in bolstering the robustness and effectiveness of machine learning models in real-world applications. This comparison, coupled with comprehensive evaluation using multiple metrics and cross-validation techniques, enables us to gain insights into the strengths and weaknesses of our models, thereby facilitating informed decision-making in mobile phone classification tasks.

RESULT

Performance Without Normalization

Below, we present the performance results of the models without the application of normalization, obtained through 5-fold, 10-fold, and 20-fold cross-validation evaluations for each Gradient Boosting, XGBoost, and CatBoost model, as shown in Table 1.

Table 1. Performance Results Without Normalization

K-Fold	Model	AUC	Accuracy	F1	Precision	Recall
5	Gradient Boosting	0,946	0,821	0,822	0,822	0,821
	XGboost	0,950	0,832	0,832	0,832	0,832
	Catboost	0,955	0,843	0,843	0,843	0,843
10	Gradient Boosting	0,949	0,819	0,820	0,820	0,819
	XGboost	0,949	0,831	0,831	0,831	0,831
	Catboost	0,955	0,837	0,837	0,837	0,837
20	Gradient Boosting	0,948	0,818	0,819	0,819	0,818
	XGboost	0,951	0,824	0,825	0,825	0,824
	Catboost	0,953	0,831	0,831	0,832	0,831

In Table 1, we present the performance results obtained from the models without the application of normalization, as assessed through 5-fold, 10-fold, and 20-fold cross-validation evaluations. The models evaluated include Gradient Boosting, XGBoost, and CatBoost, with each model assessed based on various performance metrics such as AUC (Area Under the Curve), accuracy, F1 score, precision, and recall. Across all fold variations, CatBoost consistently demonstrated the highest performance, with AUC values ranging from 0.955 to 0.953 and accuracy scores ranging from 0.837 to 0.831. XGBoost also exhibited competitive performance, with AUC values ranging from 0.950 to 0.951 and accuracy scores ranging from 0.832 to 0.824. Gradient Boosting, while slightly trailing behind CatBoost and XGBoost, still displayed respectable performance, with AUC values ranging from 0.946 to 0.949 and accuracy scores ranging from 0.821 to 0.818. These results provide valuable insights into the effectiveness of each model in classifying mobile phones without the utilization of normalization techniques.

Performance With Normalization

Table 2 presents the performance results obtained from 5-fold, 10-fold, and 20-fold cross-validation evaluations for each model—Gradient Boosting, XGBoost, and CatBoost—after applying normalization techniques.

Table 2. Performance Results With Normalization

K-Fold	Model	AUC	Accuracy	F1	Precision	Recall
5	Gradient Boosting	0,949	0,831	0,831	0,832	0,831
	XGboost	0,951	0,835	0,835	0,835	0,835
	Catboost	0,955	0,841	0,841	0,842	0,841
10	Gradient Boosting	0,951	0,822	0,823	0,823	0,822

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 2, April 2024
https://doi.org/10.47709/cnahpc.v6i2.3790

Submitted : April 10, 2024
Accepted : April 27, 2024
Published : Apr 29, 2024

	XGboost	0,952	0,834	0,834	0,835	0,834
	Catboost	0,953	0,837	0,837	0,838	0,837
	Gradient Boosting	0,950	0,820	0,821	0,821	0,820
20	XGboost	0,952	0,824	0,825	0,825	0,824
	Catboost	0,954	0,836	0,836	0,837	0,836

Table 2 illustrates the performance results achieved with normalization techniques applied across various fold configurations (5-fold, 10-fold, and 20-fold) for each model—Gradient Boosting, XGBoost, and CatBoost. With normalization in place, we observe notable improvements in performance metrics such as AUC, accuracy, F1 score, precision, and recall. CatBoost consistently outperforms the other models, demonstrating the highest AUC values ranging from 0.955 to 0.954 and accuracy scores ranging from 0.841 to 0.836 across different fold variations. XGBoost also exhibits competitive performance, with AUC values ranging from 0.951 to 0.952 and accuracy scores ranging from 0.835 to 0.824. Gradient Boosting, while slightly trailing behind CatBoost and XGBoost, still shows improved performance with AUC values ranging from 0.949 to 0.950 and accuracy scores ranging from 0.831 to 0.820. These findings underscore the effectiveness of normalization techniques in enhancing the models' classification performance in the context of mobile phone classification tasks.

DISCUSSIONS

The interpretation of the results obtained from both the models without normalization and those with normalization provides valuable insights into the classification of mobile phone specifications. Notably, the application of normalization techniques has resulted in notable improvements across various performance metrics, including AUC, accuracy, F1 score, precision, and recall. This suggests that preprocessing methods play a crucial role in enhancing the robustness and effectiveness of machine learning models in classifying mobile phones.

Furthermore, a comparison of the strengths and weaknesses of gradient boosting, XGBoost, and CatBoost in this context reveals interesting patterns. CatBoost consistently outperformed the other models across different fold variations, demonstrating superior performance in terms of AUC values and accuracy scores. This indicates that CatBoost is particularly effective in capturing complex relationships within the data and making accurate predictions for mobile phone classification tasks. On the other hand, XGBoost also exhibited competitive performance, showcasing its effectiveness as a versatile algorithm for classification tasks. While Gradient Boosting slightly trailed behind CatBoost and XGBoost in performance, it still demonstrated respectable performance, highlighting its potential utility in certain scenarios.

Insights into which classifier performs best for different levels of mobile phone specifications suggest that CatBoost generally excels across various levels, followed closely by XGBoost. However, the choice of classifier may depend on specific requirements and constraints of the classification task. For instance, Gradient Boosting may be suitable for scenarios where computational resources are limited, while CatBoost may be preferred for tasks where maximizing predictive accuracy is paramount.

Overall, these findings underscore the importance of employing appropriate preprocessing techniques and selecting suitable classification algorithms for effectively classifying mobile phone specifications, thus facilitating informed decision-making in the realm of mobile technology.

CONCLUSION

This research presents a comprehensive analysis of mobile phone specification classification using machine learning algorithms, specifically Gradient Boosting, XGBoost, and CatBoost. Through meticulous dataset acquisition and preprocessing steps, including resolution normalization and price categorization, we standardized features essential for classification analysis. Our evaluation utilized robust cross-validation techniques across varying fold configurations to assess model performance effectively. The study demonstrates the significant impact of normalization techniques on improving model performance, as evidenced by notable enhancements in AUC, accuracy, F1 score, precision, and recall metrics across all algorithms and fold variations. CatBoost consistently emerges as the top-performing algorithm, followed closely by XGBoost, with Gradient Boosting displaying respectable performance. These findings underscore the importance of preprocessing methods and algorithm selection in achieving optimal classification results. For mobile phone manufacturers, our research highlights the relevance of leveraging machine learning algorithms to classify mobile phone specifications effectively. Insights derived from the study can inform product development strategies, enabling manufacturers to optimize product offerings and market positioning based

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 2, April 2024
https://doi.org/10.47709/cnahpc.v6i2.3790

Submitted : April 10, 2024
Accepted : April 27, 2024
Published : Apr 29, 2024

on consumer preferences and competitive landscapes. Similarly, for data analysts, the study emphasizes the significance of employing appropriate preprocessing techniques and algorithmic approaches to enhance model performance. This can lead to more accurate predictions and informed decision-making in the realm of mobile technology. While this study provides valuable insights into mobile phone specification classification, there are several avenues for future research. Further exploration of advanced preprocessing methods, such as feature engineering and dimensionality reduction techniques, could offer additional improvements in model performance. Additionally, investigating the applicability of other machine learning algorithms or ensemble methods may yield novel insights and enhanced classification capabilities. Furthermore, exploring the impact of additional features or incorporating external datasets could enrich the classification process and broaden the scope of analysis. In conclusion, this research contributes to the understanding of mobile phone specification classification through machine learning methodologies. By emphasizing the importance of preprocessing techniques, algorithm selection, and cross-validation evaluation, the study provides actionable insights for both industry practitioners and researchers. Moving forward, continued exploration and refinement of classification approaches in the context of mobile technology hold promise for addressing evolving market dynamics and consumer preferences.

REFERENCES

Abdurohman, M., & Putrada, A. G. (2023). Forecasting Model for Lighting Electricity Load with a Limited Dataset using XGBoost. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, *8*(2), 571–580. https://doi.org/10.22219/kinetik.v8i2.1687

Adler, A. I., & Painsky, A. (2022). Feature Importance in Gradient Boosting Trees with Cross-Validation Feature Selection. *Entropy*, *24*(5). https://doi.org/10.3390/e24050687

Ampomah, E. K., Qin, Z., & Nyame, G. (2020). Evaluation of Tree-Based Ensemble Machine Learning Models in Predicting Stock Price Direction of Movement. *Information*, *11*(6), 332. https://doi.org/10.3390/info11060332

Aravind, K. R. N. V. V. D., Shyry, S. P., & Felix, Y. (2019). Classification of Healthy and Rot Leaves of Apple Using Gradient Boosting and Support Vector Classifier. *International Journal of Innovative Technology and Exploring Engineering*, *8*(12), 2868–2872. https://doi.org/10.35940/ijitee.L3049.1081219

Avelino, J. N. M., Felizmenio, E. P., & Naval, P. C. (2022). Unraveling COVID-19 Misinformation with Latent Dirichlet Allocation and CatBoost. *Communications in Computer and Information Science*, *1653 CCIS*(July), 16–28. https://doi.org/10.1007/978-3-031-16210-7_2

Boldini, D., Grisoni, F., Kuhn, D., Friedrich, L., & Sieber, S. A. (2023). Practical guidelines for the use of gradient boosting for molecular property prediction. *Journal of Cheminformatics*, *15*(1), 1–13. https://doi.org/10.1186/s13321-023-00743-7

Breskuvien, D., & Dzemyda, G. (2023). Categorical Feature Encoding Techniques for Improved Classifier Performance when Dealing with Imbalanced Data of Fraudulent Transactions. *International Journal of Computers, Communications and Control*, *18*(3), 1–17. https://doi.org/10.15837/ijccc.2023.3.5433

Callens, A., Morichon, D., Abadie, S., Delpey, M., & Liquet, B. (2020). Using Random forest and Gradient boosting trees to improve wave forecast at a specific location. *Applied Ocean Research*, *104*(July), 102339. https://doi.org/10.1016/j.apor.2020.102339

Chang, W., Wang, X., Yang, J., & Qin, T. (2023). An Improved CatBoost-Based Classification Model for Ecological Suitability of Blueberries. *Sensors*, *23*(4). https://doi.org/10.3390/s23041811

Chen, T., Samaranayake, P., Cen, X. Y., Qi, M., & Lan, Y. C. (2022). The Impact of Online Reviews on Consumers' Purchasing Decisions: Evidence From an Eye-Tracking Study. *Frontiers in Psychology*, *13*(June). https://doi.org/10.3389/fpsyg.2022.865702

Devos, L., Meert, W., & Davis, J. (2020). Fast Gradient Boosting Decision Trees with Bit-Level Data Structures. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11906 LNAI*, 590–606. https://doi.org/10.1007/978-3-030-46150-8_35

Dutta, A. K., & Wahab Sait, A. R. (2024). A Fine-Tuned CatBoost-Based Speech Disorder Detection Model. *Journal of Disability Research*, *3*(3), 1–8. https://doi.org/10.57197/JDR-2024-0027

Dwinanda, M. W., Satyahadewi, N., & Andani, W. (2023). Classification of Student Graduation Status Using Xgboost Algorithm. *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, *17*(3), 1785–1794. https://doi.org/10.30598/barekengvol17iss3pp1785-1794

Fayaz, M., Khan, A., Rahman, J. U., Alharbi, A., Uddin, M. I., & Alouffi, B. (2020). Ensemble machine learning model for classification of spam product reviews. *Complexity*, *2020*. https://doi.org/10.1155/2020/8857570

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 2, April 2024
https://doi.org/10.47709/cnahpc.v6i2.3790

Submitted : April 10, 2024
Accepted : April 27, 2024
Published : Apr 29, 2024

Fedorov, N., & Petrichenko, Y. (2020). Gradient Boosting–Based Machine Learning Methods in Real Estate Market Forecasting. *Proceedings of the 8th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2020)*, *174*(Itids), 203–208. Paris, France: Atlantis Press. https://doi.org/10.2991/aisr.k.201029.039

Gonçalves Freitas, L. J., Edokawa, P. S. D., Carvalho Valadares Rodrigues, T., Thomé de Farias, A. H., & Rodrigues de Alencar, E. (2023). Catboost algorithm application in legal texts and UN 2030 Agenda. *Revista de Informatica Teorica e Aplicada*, *30*(2), 51–58. https://doi.org/10.22456/2175-2745.128836

Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: an interdisciplinary review. *Journal of Big Data*, *7*(1). https://doi.org/10.1186/s40537-020-00369-8

Herni Yulianti, S. E., Oni Soesanto, & Yuana Sukmawaty. (2022). Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit. *Journal of Mathematics: Theory and Applications*, *4*(1), 21–26. https://doi.org/10.31605/jomta.v4i1.1792

Heydarizad, M., Pumijumnong, N., Sorí, R., Salari, P., & Gimeno, L. (2022). Fractional Importance of Various Moisture Sources Influencing Precipitation in Iran Using a Comparative Analysis of Analytical Hierarchy Processes and Machine Learning Techniques. *Atmosphere*, *13*(12), 6–8. https://doi.org/10.3390/atmos13122019

Hussain, S., Mustafa, M. W., Jumani, T. A., Baloch, S. K., Alotaibi, H., Khan, I., & Khan, A. (2021). A novel feature engineered-CatBoost-based supervised machine learning framework for electricity theft detection. *Energy Reports*, *7*, 4425–4436. https://doi.org/10.1016/j.egyr.2021.07.008

Ibrahim, A. A., Ridwan, R. L., Muhammed, M. M., Abdulaziz, R. O., & Saheed, G. A. (2020). Comparison of the CatBoost Classifier with other Machine Learning Methods. *International Journal of Advanced Computer Science and Applications*, *11*(11), 738–748. https://doi.org/10.14569/IJACSA.2020.0111190

Jabeur, S. Ben, Gharib, C., Mefteh-Wali, S., & Arfi, W. Ben. (2021). CatBoost model and artificial intelligence techniques for corporate failure prediction. *Technological Forecasting and Social Change*, *166*(February), 120658. https://doi.org/10.1016/j.techfore.2021.120658

Jabeur, S. Ben, Mefteh-Wali, S., & Viviani, J. L. (2024). Forecasting gold price with the XGBoost algorithm and SHAP interaction values. *Annals of Operations Research*, *334*(1–3), 679–699. https://doi.org/10.1007/s10479-021-04187-w

Jasman, T. Z., Fadhlullah, M. A., Pratama, A. L., & Rismayani, R. (2022). Analisis Algoritma Gradient Boosting, Adaboost dan Catboost dalam Klasifikasi Kualitas Air. *Jurnal Teknik Informatika Dan Sistem Informasi*, *8*(2), 392–402. https://doi.org/10.28932/jutisi.v8i2.4906

Jinan, A., Situmorang, Z., & Rosnelly, R. (2023). Bulldog Breed Classification Using VGG-19 and Ensemble Learning. *International Conference on Information Science and Technology Innovation (ICoSTEC)*, *2*(1), 29–33. https://doi.org/10.35842/icostec.v2i1.29

Kabeyi, M. J. B. (2018). Michael porter's five competitive forces and generetic strategies, market segmentation strategy and case study of competition in global smartphone manufacturing industry. *International Journal of Applied Research*, *4*(10), 39–45. https://doi.org/10.22271/allresearch.2018.v4.i10a.5275

Karanikola, A., Davrazos, G., Liapis, C. M., & Kotsiantis, S. (2023). Financial sentiment analysis: Classic methods vs. deep learning models. *Intelligent Decision Technologies*, *17*(4), 893–915. https://doi.org/10.3233/IDT-230478

Kumar, V., Kedam, N., Sharma, K. V., Mehta, D. J., & Caloiero, T. (2023). Advanced Machine Learning Techniques to Improve Hydrological Prediction: A Comparative Analysis of Streamflow Prediction Models. *Water (Switzerland)*, *15*(14). https://doi.org/10.3390/w15142572

Liew, X. Y., Hameed, N., & Clos, J. (2021). An investigation of XGBoost-based algorithm for breast cancer classification. *Machine Learning with Applications*, *6*(September), 100154. https://doi.org/10.1016/j.mlwa.2021.100154

Lin, C. H., Hsu, P. I., Tseng, C. D., Chao, P. J., Wu, I. T., Ghose, S., … Lee, T. F. (2023). Application of artificial intelligence in endoscopic image analysis for the diagnosis of a gastric cancer pathogen-Helicobacter pylori infection. *Scientific Reports*, *13*(1), 1–12. https://doi.org/10.1038/s41598-023-40179-5

Lo, Y. T., Liao, J. C. hen, Chen, M. H., Chang, C. M., & Li, C. Te. (2021). Predictive modeling for 14-day unplanned hospital readmission risk by using machine learning algorithms. *BMC Medical Informatics and Decision Making*, *21*(1), 1–11. https://doi.org/10.1186/s12911-021-01639-y

Neelakandan, S., & Paulraj, D. (2020). A gradient boosted decision tree-based sentiment classification of twitter data. *International Journal of Wavelets, Multiresolution and Information Processing*, *18*(4).

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 2, April 2024
https://doi.org/10.47709/cnahpc.v6i2.3790

Submitted : April 10, 2024
Accepted : April 27, 2024
Published : Apr 29, 2024

https://doi.org/10.1142/S0219691320500277

Pardede, D., Firmansyah, I., Handayani, M., Riandini, M., & Rosnelly, R. (2022). Comparison Of Multilayer Perceptron's Activation And Optimization Functions In Classification Of Covid-19 Patients. *JURTEKSI (Jurnal Teknologi Dan Sistem Informasi)*, *8*(3), 271–278. https://doi.org/10.33330/jurteksi.v8i3.1482

Pardede, D., & Hayadi, B. H. (2023). Klasifikasi Sentimen Terhadap Gelaran MotoGP Mandalika 2022 Menggunakan Machine Learning. *Jurnal TRANSFORMATIKA*, *20*(2), 42–50.

Putri, D. J., Dwifebri, M., & Adiwijaya, A. (2023). Text Classification of Indonesian Translated Hadith Using XGBoost Model and Chi-Square Feature Selection. *Building of Informatics, Technology and Science (BITS)*, *4*(4), 1732–1738. https://doi.org/10.47065/bits.v4i4.2944

Qinghe, Z., Wen, X., Boyan, H., Jong, W., & Junlong, F. (2022). Optimised extreme gradient boosting model for short term electric load demand forecasting of regional grid system. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-22024-3

Safaei, N., Safaei, B., Seyedekrami, S., Talafidaryani, M., Masoud, A., Wang, S., … Moqri, M. (2022). E-CatBoost: An efficient machine learning framework for predicting ICU mortality using the eICU Collaborative Research Database. In *PLoS ONE* (Vol. 17). https://doi.org/10.1371/journal.pone.0262895

Singh, S., & More, R. (2022). Mobile Phone Companies Increasing Market Share through Innovations, R&D Spending and Patents. *EMAJ: Emerging Markets Journal*, *12*(1), 76–85. https://doi.org/10.5195/emaj.2022.251

Siringoringo, R., Perangin Angin, R., & Rumahorbo, B. (2022). Model Klasifikasi Genetic-XGBoost Dengan T-Distributed Stochastic Neighbor Embedding Pada Peramalan Pasar. *Jurnal Times*, *XI*(1), 30–36. Retrieved from https://archive.ics.uci.edu/ml/datasets/online+retail

Sopiyan, M., Fauziah, F., & Wijaya, Y. F. (2022). Fraud Detection Using Random Forest Classifier, Logistic Regression, and Gradient Boosting Classifier Algorithms on Credit Cards. *JUITA: Jurnal Informatika*, *10*(1), 77. https://doi.org/10.30595/juita.v10i1.12050

Suhendra, R., Husdayanti, N., Suryadi, S., Juliwardi, I., Sanusi, S., Ridho, A., … Ikhsan, I. (2023). Cardiovascular Disease Prediction Using Gradient Boosting Classifier. *Infolitika Journal of Data Science*, *1*(2), 56–62. https://doi.org/10.60084/ijds.v1i2.131

Sun, F., Luh, D.-B., Zhao, Y., & Sun, Y. (2022). Product Classification With the Motivation of Target Consumers by Deep Learning. *IEEE Access*, *10*, 62258–62267. https://doi.org/10.1109/ACCESS.2022.3181624

Suryana, S. E., Warsito, B., & Suparti, S. (2021). PENERAPAN GRADIENT BOOSTING DENGAN HYPEROPT UNTUK MEMPREDIKSI KEBERHASILAN TELEMARKETING BANK. *Jurnal Gaussian*, *10*(4), 617–623. https://doi.org/10.14710/j.gauss.v10i4.31335

Tusher, A. S., Rahman, M. A., Islam, M. R., & Hossain, M. J. (2024). Adversarial training-based robust lifetime prediction system for power transformers. *Electric Power Systems Research*, *231*, 110351. https://doi.org/10.1016/j.epsr.2024.110351

Wang, J. C., Hsieh, C. Y., & Kung, S. H. (2023). The impact of smartphone use on learning effectiveness: A case study of primary school students. In *Education and Information Technologies* (Vol. 28). Springer US. https://doi.org/10.1007/s10639-022-11430-9

Wang, N., Zeng, M., Li, Y., Wu, F. X., & Li, M. (2021). Essential Protein Prediction Based on node2vec and XGBoost. *Journal of Computational Biology*, *28*(7), 687–700. https://doi.org/10.1089/cmb.2020.0543

Zhang, X. (2022). A Model Combining LightGBM and Neural Network for High-frequency Realized Volatility Forecasting. *Proceedings of the 2022 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022)*, *648*(Icfied), 2906–2912. https://doi.org/10.2991/aebmr.k.220307.473

Zheng, Y., Guo, X., Yang, Y., Wang, H., Liao, K., & Qin, J. (2023). Phonocardiogram transfer learning-based CatBoost model for diastolic dysfunction identification using multiple domain-specific deep feature fusion. *Computers in Biology and Medicine*, *156*(1), 106707. https://doi.org/10.1016/j.compbiomed.2023.106707

* Corresponding author