Journal of Computing and Communication Vol 3 , No.2 , PP. 23-30 , 2024

Evaluation of the Performance of Categorical Boosting Algorithm for Flood Prediction in Osun River Basin

Oluwatosin I. Ogundolie*[a], Stephen O. Olabiyisi[a], Rafiu A. Ganiyu[b], Yetomiwa S. Jeremiah[a] Frank A. Ogundolie[c]

[a] Computer Science Department, Ladoke Akintola University of Technology, Ogbomoso, Nigeriat.

[b] Computer Engineering Department, Ladoke Akintola University of Technology, Ogbomoso, Nigeria.

[c] Department of Biotechnology, Baze University, Abuja, Nigeria

[b]Department of Computer Science, Misr International University, Cairo, Egypt

*Corresponding Author: Oluwatosin I. Ogundolie [***mdtosin@gmail.com***]

ARTICLE DATA

Article history:
Received 01 May 2024
Revised 04 June 2024
Accepted 24 June 2024
Available online

Keywords:
Categorical Boosting;
Algorithm;
Flooding;
Machine learning;
Evaluation.

ABSTRACT

Flooding is the third biggest disaster in the world according to the World Meteorological Organization. Several methods like numerical models, physical models, and Machine Learning (ML) models have been engaged in flood prediction to minimize the impact of flooding. Despite the improvements experienced in the use of some ML methods, there are still drawbacks due to accuracy. Hence, this study evaluated Categorical Boosting Algorithm (CatBoost) for flood prediction based on some evaluation metrics. Relevant flood-predictive factors were identified from the Osun River basin. The data was split into 70% for the training and 30% for the testing of the algorithm. The flood dataset was imported into the CatBoost Algorithm using Python programming language with the default parameters of the algorithm. The algorithm was evaluated using accuracy, precision, sensitivity, and multiclass loss function. The results showed that the accuracy, precision, and sensitivity of the CatBoost Algorithm were 92.48%, 63.82%, and 85.86% respectively. The result of the multiclass loss function during validation was 0.165874, which was significantly lower than the result during training, which was 0.925104. This indicates that the algorithm is overfitting the training data and is not generalizing well to new data. This can be a prospect for further study.

1. Introduction

The Categorical Boosting Algorithm (CatBoost) algorithm has gained prominence in developing predictive models by leveraging historical data [4]. CatBoost, a Gradient Boosting Decision Trees (GBDT) method, excels in handling categorical features during training [2], making it suitable for diverse applications like ranking, forecasting, and recommendations [11]. Floods, exacerbated by climate change, pose severe threats globally, necessitating accurate prediction. While numerous factors contribute to flooding, including climate change and human activities, developing precise flood prediction models remains challenging [13], [14]. This study focuses on applying the CatBoost algorithm to flood prediction, an area where it has not been previously implemented. The Osun River basin in Nigeria serves as the study area to acquire satellite remote sensing data for flood resource variables. Floods are recurrent and impactful in Nigeria due to climate change, necessitating reliable prediction systems for mitigation and emergency response [10],[5]. The study aims to assess the CatBoost algorithm's performance using accuracy, precision, sensitivity, and multiclass loss function as metrics, to enhance flood spatial prediction accuracy. Even though machine learning has demonstrated promise in predicting floods from remotely sensed Earth

Observation data [12], achieving high prediction accuracy remains a challenge [1]. This research seeks to address this gap by evaluating the suitability of CatBoost for flood prediction, potentially offering a valuable tool for improving flood control, emergency response, and water resources management. As floods continue to pose significant threats, accurate flood prediction models are crucial for minimizing damage and ensuring effective disaster management [15],[3]. This study intends to contribute to advancing flood prediction capabilities, thereby enhancing societal resilience and response strategies.

2. Background

In this study, ten flood conditioning factors, including rainfall, topographic water index (TWI), slope, drainage density, digital elevation model (DEM), soil data, distances, NDVI, and land use, were analyzed for the 2018 flood event. The Ọṣun River Basin is the study area, the Osun River flows southwards through Yoruba land in southwestern Nigeria, passing six states. The 10 conditioning flood factors were weighed based on their importance using the Analytic Hierarchy Process (AHP). AHP is a decision-making method used in remote sensing and geographic information science to organize and analyze complex decisions. It combines mathematics and psychology to compare several options and assign each criterion an importance weight based on pairwise comparisons. Rainfall data from PERSIANN-CDR and DEM from Copernicus were used to assess flooding. ArcGIS processed DEM, slope, Landsat imagery, TWI, and drainage density. Soil data from FAO's portal determined water-holding capacity. Land use and wetlands were mapped in ArcMap, drainage lines calculated density, and factors like distance from rivers and LULC were evaluated. Geomorphic elements were derived from DEM, while NDVI indicated vegetation health. TWI identified potential floodplains and the AHP prioritized flood factors. The created maps were converted into numerical data which were extracted at every 100-meter interval into an Excel sheet in a CSV file. A sample of the extracted data is shown in Table 1.

TABLE 1: A Cross Section of the dataset for the year 2018

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Longitude	Latitude	TWI	Soil Type	Slope	Road	River	Rainfall	NDVI	LULC	Drainage Density	DEM	Flood Susceptibility
2	3.817450	6.462866	4	2	5	5	5	5	1	4	1	5	4
3	3.818354	6.462864	4	2	5	5	5	5	1	2	1	5	4
4	3.819259	6.462863	4	2	5	5	5	5	1	3	1	5	4
5	3.820163	6.462862	4	2	5	5	5	5	1	2	1	5	4
6	3.821067	6.462860	4	2	5	5	5	5	1	2	1	5	4
7	3.821972	6.462859	4	2	5	5	5	5	1	4	1	5	4
8	3.822876	6.462857	3	2	5	4	5	5	1	2	1	5	4
9	3.823780	6.462856	3	2	4	4	5	5	1	1	1	5	4
10	3.824684	6.462854	3	2	4	4	5	5	1	2	1	5	4
11	3.825589	6.462853	3	2	4	4	5	5	1	2	1	5	4
12	3.826493	6.462851	3	2	5	4	5	5	1	2	1	5	4
13	3.827397	6.462850	3	2	4	4	5	5	1	2	1	5	4
14	3.828302	6.462848	3	2	4	5	5	5	1	2	1	5	4
15	3.829206	6.462847	3	2	3	5	5	5	1	2	1	5	4
16	3.830110	6.462845	3	2	4	5	5	5	1	2	1	5	4
17	3.831014	6.462844	3	2	4	5	5	5	1	2	1	5	4

It shows the longitude, latitude, the 10 conditioning factors and the flood susceptibility based on the ten factors. The data was classified based on flood susceptibility as; 1 representing Very low, 2 representing Low, 3 representing Moderate, 4 representing High and 5 representing Very High. A total of 930,789 locations represented by longitude and latitude were identified in the study area. The key metrics occupy columns of the Excel sheet which form the dataset in the raw format. Data preprocessing libraries such as pandas and machine learning libraries like spark ml were used while the datasets were converted into a data frame. This data frame serves as the input to the machine learning pipeline. With the data frame, all the columns of the data were checked for abnormalities such as missing values, null values, outliers and others.

3. Evaluation of the Performance of the CatBoost Algorithm

From the extract of the dataset, the flood susceptibility is the dependent feature while the conditioning factors are the independent features which are rainfall data, topographic water index (TWI), slope, drainage density, digital elevation model, soil data, distances from roads, normalized difference vegetation index (NDVI), distance from rivers, and land use land cover (LULC) data. This makes this prediction a multiclass prediction using the CatBoost Algorithm. The study utilized the CatBoost algorithm for flood prediction, employing the geospatial database with a 70:30 data split for training and validation. Metrics including accuracy, precision, sensitivity, and the multiclass loss function assess the algorithm's efficacy. CatBoost Algorithm's distinctive features like ordered boosting and categorical handling are highlighted, addressing target leakage via random permutations. It operated in ordered and plain modes, influencing tree structure choice. Hyperparameters, such as learning rate, tree count, regularization, and categorical feature combinations, were examined, guided by default settings. These key hyperparameters are analyzed within constraints. Tree count ranges from 100 to 1000 to control model complexity. The learning rate, affecting adaptation, is restricted between 0.001 and 0.01. Maximum tree depth, influencing complexity, ranges from 1 to 10. Regularization (L1 and L2) is constrained between 0 and 3 to control complexity. These constraints impact the multiclass equation in categorical boosting, contributing to improved flood prediction.

4. Results from Flood Prediction Using CatBoost Algorithm

The dataset includes flood susceptibility as the dependent feature and various conditioning factors as independent features, making it a multiclass prediction problem using the CatBoost Algorithm. After importing the data into Jupyter Notebook, 930,789 locations were analyzed for the year 2018. To ensure accurate predictions, the dataset was split into training and test sets, with a test size of 0.3 and a corresponding training set size of 0.7. This split helps validate the model's performance on new data. The CatBoost classifier was applied to the dataset with initial settings of 100 iterations and a learning rate of 0.5, while later optimizing the learning rate to 0.1658737041 and the model tree count to 399. The results of training and validating the model are presented in Table 2, showing accuracy. To assess the model's performance, the loss function was used to determine feature importance for improved training. The confusion matrix was utilized to calculate metrics such as accuracy, precision, and sensitivity. The matrix contains five rows and columns, representing different flood risk classes (very low to very high). Table 3, Table 4, Table 5, Table 6 and Table 7 show the results of the analyzed specific class predictions: It was observed from class 1 that the model correctly predicted 474 instances and the false negative predictions are 209 instances. For class 2, the model correctly predicted 85 instances and falsely predicted 285 instances as negative. For class 3, the model correctly predicted 107,616 instances and falsely predicted 22,316 instances also as negative. For class 4, the model correctly predicted 133,829 instances and falsely predicted 11,294 instances as negatives. Lastly, for class 5, the model correctly predicted 1,854 instances and falsely predicted 268 instances as negative. This analysis demonstrates how the CatBoost Algorithm was employed for multiclass flood risk prediction, splitting the dataset for training and testing, optimizing model parameters, and evaluating predictions through various performance metrics, ultimately providing insights into different flood risk levels.

TABLE 2: Results of Flood Prediction Using 2018 Dataset

	Training	**validation**
Total Dataset (930,789)		
Accuracy	0.92708	0.925104
Multiclass	0.9251	0.165874

TABLE 3: Results for Very Low Flood Susceptibility

Actual Class		**Predicted Class**
Total Test Data (279,236)		
	Positive	**Negative**
Class 1	474 (TP)	278,586 (TN)
	134 (FP)	209 (FN)

TABLE 4: Results for Low Flood Susceptibility

Actual Class		**Predicted Class**
Total Test Data (279,236)		
	Positive	**Negative**
Class 2	85 (TP)	278,846 (TN)
	20 (FP)	285 (FN)

TABLE 5: Results for Moderate Flood Susceptibility

Actual Class		**Predicted Class**
Total Test Data (279,236)		
	Positive	**Negative**
Class 3	107,616 (TP)	138,009 (TN)
	11,295 (FP)	22,316 (FN)

TABLE 6: Results for High Flood Susceptibility

Actual Class		**Predicted Class**
Total Test Data (279,236**)**		
	Positive	**Negative**
Class 4	133,829 (TP)	110,444 (TN)
	23,669 (FP)	11,294 (FN)

TABLE 7: Results for Very High Flood Susceptibility

Actual Class		**Predicted Class**
Total Test Data (279,236**)**		
	Positive	**Negative**
Class 5	1,854 (TP)	273,700 (TN)
	334 (FP)	2,268 (FN)

5. Error Analysis

The evaluation of the algorithm for flood prediction focused on accuracy, employing performance metrics such as precision, sensitivity, and accuracy in order to examine the misclassifications in the prediction. Accuracy is determined by the correct utilization of the error between predicted and actual values. The model's performance is assessed using a test dataset representing 30% of the total dataset. Precision, sensitivity (or recall rate), and accuracy are vital metrics in this evaluation. Accuracy measures the proportion of true outcomes (true positives and true negatives) in the system, indicating the system's exactness. Sensitivity, or recall rate, gauges the model's ability to correctly identify positive cases of flooding relative to the total actual positive cases. High sensitivity signifies effective flood detection, while low sensitivity indicates the model's failure to detect some positive cases, potentially leading to incomplete flood predictions. Precision, a performance metric, assesses the proportion of accurate flood predictions among all predictions made by the model. High precision indicates accurate predictions, enhancing the model's trustworthiness for informed flood response and mitigation. A low precision, however, leads to more false positive predictions, possibly resulting in unnecessary emergency measures. Thus, a flood prediction model with low precision might require algorithm refinements to reduce false alarms and improve its overall performance.

The values of the performance metrics are shown in Table 8. The bar chart of the performance metrics is shown in Figure 1. The model has high accuracy for all classes, ranging from 89.5% to 93.2%. The values of the sensitivity for the five classes range from 29.4% to 96.9%, indicating that the model has a varying degree of confidence in its positive predictions for each class. The precision values for the five classes range from 66.6% to 95.0%, indicating that the model is better at detecting some classes than others. The precision and sensitivity values of the model indicate that the model's performance may vary depending on the class being predicted. Decision tree-based machine learning models particularly Extreme Gradient Boosting (XGBoost) ML has been adopted for flood prediction and performed best with an accuracy of 0.84 [6]. Another study suggested employing a random forest algorithm for flash-flood forecasting [9]. The findings show that, given ideal inputs made up solely of features that account for 80% of the model's outcome variance, the generated parsimonious models can achieve validation efficiencies [9]. Although still quite high, the precision can be raised. Using an artificial neural network (ANN) with rainfall and temperature as inputs, a similar study created a flood prediction model in Nigeria that predicted the Standard Precipitation Index (SPI) [7]. Limitations in flood prediction include problems with data management, network architecture, and model interpretation, even with a 76% accuracy rate. ANNs result in a relatively low accuracy in flood prediction which is a serious drawback [8].

TABLE 8: Performance Metrics of the Classes

Class predicted	**Accuracy (%)**	**Sensitivity (%)**	**Precision (%)**
Class 1	92.3	29.4	87.7
Class 2	89.5	34.4	66.6
Class 3	93.2	87.1	95
Class 4	92.6	96.9	91.1
Class 5	92.9	71.3	88.9

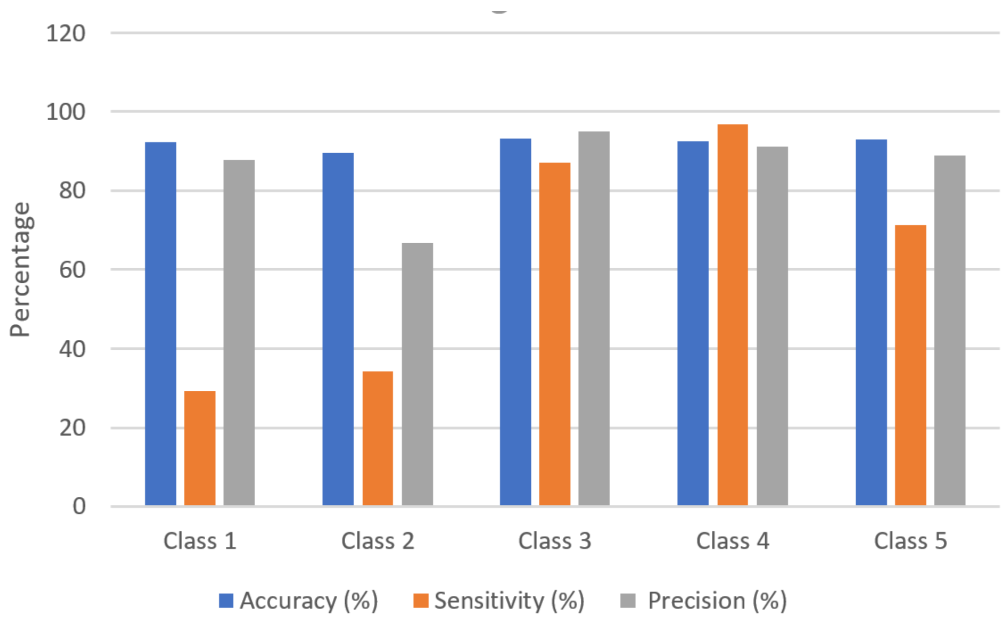

Figure 1 Bar chart Performance Metrics for the Existing CatBoost Algorithm

6. Conclusions and Future work

The multiclass loss function during validation is 0.165874, significantly lower than the training result of 0.925104, suggesting overfitting to training data. A low loss indicates accurate classification, while a high value suggests errors. Overfitting arises when a complex model learns training noise, hurting generalization to new data. It excels in training but fails in validation. To counter this, there is a need to optimize categorical boosting by tuning hyperparameters like learning rate, tree depth, and regularization strength. L1/L2 regularization penalizes large weights, promoting simpler models for better generalization. Early stopping closes training on degraded validation performance. Optimal hyperparameters enhance generalization, reducing loss on both sets.

The future work that needs to be done is to design and implement an optimized CatBoost Algorithmic model for flood prediction. There is a need to explore the integration of additional data sources and to test the model in different geographic regions in future works to be done. Ensemble learning techniques can also be adopted to further enhance prediction accuracy. Lastly, to compare the performance with the existing CatBoost Algorithm.

Acknowledgement

The research was partly funded by L'Oreal UNESCO For Women in Science Foundation. Thank you very much for your support towards this research.

References

[1] Deng, G., Chen, H., & Wang, S. Risk assessment and prediction of rainstorm and flood disaster based on henan province, China. Mathematical Problems in Engineering, (2022).

[2] Dorogush, A. V., Ershov, V., and Gulin, A. CatBoost: gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363, 2018.

[3] Gebrehiwot, A., Hashemi-Beni, L., Thompson, G., Kordjamshidi, P., and Langan, T. E.. Deep convolutional neural network for flood extent mapping using unmanned aerial vehicles data. Sensors, (2019), 19(7): 1486.

[4] Huntingford, C., Jeffers, E. S., Bonsall, M. B., Christensen, H. M., Lees, T. and Yang, H.. Machine learning and artificial intelligence to aid climate change research and preparedness. Environmental Research Letters, (2019), 14(12), 124007.

[5] Ighile, E. H., Shirakawa, H., and Tanikawa, H.. Application of GIS and Machine Learning to Predict Flood Areas in Nigeria. Sustainability, (2022), 14(9), 5039.

[6] Ma, M., Zhao, G., He, B., Li, Q., Dong, H., Wang, S., and Wang, Z.. XGBoost-based method for flash flood risk assessment. Journal of Hydrology, (2021), 598, 126382.

[7] Michael, E. B. and Patience O. E.. Flood prediction in Nigeria using artificial neural network. American Journal of Engineering Research, (2018), 7(9): 15-21.

[8] Mosavi, A., Ozturk, P., and Chau, K. W.. Flood prediction using machine learning models: Literature review. Water, (2018), 10(11): 1536.

[9] Muñoz, P., Orellana-Alvear, J., Willems, P., and Célleri, R.. Flash-flood forecasting in an Andean mountain catchment—Development of a step-wise methodology based on the random forest algorithm. Water, (2018), 10(11): 1519.

[10] Obeta, M. C.. Institutional Approach to Flood Disaster Management in Nigeria: Need for a Preparedness Plan. British Journal of Applied Science & Technology, (2014), 4(33): 4575-4590.

[11] Peretz T.. Mastering The New Generation of Gradient Boosting. https://towardsdatascience.com/https-medium-com-talperetz24-mastering-the-new-generation-of-gradient-boosting-db04062a7ea2 . (2018) Accessed on 29th December 2021

[12] Shaharkar, Y., Sonar, A., Sonar, C. and Pawar D. Flood Damage Estimation using Machine Learning in GIS. International Research Journal of Engineering and Technology. (2020), 7(6) :5756 – 5760

[13] Syifa, M., Park, S., Achmad, A., Lee, C., and Eom, J.. Flood Mapping Using Remote Sensing Imagery and Artificial

Intelligence Techniques: A Case Study in Brumadinho, Brazil. Journal of Coastal Research, 2019, 197-204.

[14] Umar, N., and Gray, A.. Flooding in Nigeria: a review of its occurrence and impacts and approaches to modelling flood data. International Journal of Environmental Studies, 2022, 1-22.

[15] Yin, J., and Li, N.. Ensemble learning models with a Bayesian optimization algorithm for mineral prospectivity mapping. Ore Geology Reviews, 2022(145), 104916