DOI: 10.12700/AIS.2025.029

# Accuracy Assessment of CatBoost for Storm Damage

Kuan Madiyarov
*Scientific and organizational department*
*Novosibirsk State University of Economics and Management*
Novosibirsk, Russia
kuan.mad@mail.ru

Alexey Letov
*Faculty of information technologies*
*Novosibirsk State University*
Novosibirsk, Russia
rorer232323@gmail.com

***Abstract*—This paper explores the use of gradient boosting to quantify the economic damage caused by storms and classify storm events as severe or mild. The study draws on a historical disaster open database. These figures confirm gradient boosting's high accuracy in analysing extreme meteorological events for insurance.**

***Keywords*— *CatBoost, storm-damage regression, storm-intensity classification, machine learning.***

## I. INTRODUCTION

Global warming is intensifying the frequency and economic impact of extreme weather phenomena—hurricanes, typhoons, and storms. Accurate forecasting of damage magnitude andearly identification of high-risk events are becoming critical. Modern AI and machine-learning techniques, especially boosting algorithms, have proved highly effectiveon tabular data[2, 4]. CatBoost, developedby Yandex and optimised for categorical features through ordered boosting, consistently outperforms other gradient-boosting implementations and has been successfully applied to climate-risk problems. In this study we deploy CatBoost to predict monetary storm losses and to flag events whose damage exceeds US$ 100 million. The aim is to demonstrate CatBoost's practical value for climate-disaster analytics and to benchmark its predictive performance [3, 13].

In this study, we leverage CatBoost to tackle two complementary disaster-analytics tasks using a historical storm disaster database: (i) a regression task to predict the monetary damage caused by individual storm events, and (ii) a classifcation task to fag "severe" storms, defned here as those causing losses above 100 million US. The goal is to assess CatBoost's accuracy and practical value in predicting storm impacts, while developing a robust modeling pipeline that spans data preprocessing, model training, and evaluation for both continuous and binary outcomes. By addressing both the magnitude of losses and a threshold-based intensity categorization, the approach provides a more holistic analysis of storm risk than either task alone. This dual focus is particularly important for translating predictive analytics into decision-making: precise loss estimates inform resource allocation and insurance/recovery planning, whereas early classifcation of high-damage events can trigger timely alerts and disaster response escalation. Moreover, our work flls a gap in the literature – whereas many prior studies apply CatBoost (and similar ML models) to hazard susceptibility mapping or event outcome classifcation, few have addressed direct economic loss regression for storms, especially in conjunction with an intensity classifcation. By benchmarking CatBoost on these tasks, we contribute evidence of its efectiveness inclimate-disaster contexts and identifystrengths and limitations of this approach.

The remainder of the paper is organized as follows. Section 2 describes the data and methodology, including preprocessing steps and CatBoost model setup for both regression and classifcation. Section 3 presents the results, with performance metrics evaluating the accuracy of damage predictionsandseverityclassifcation. In Section 3, we provide an in-depth discussion of these results in the context of related work, comparingourapproach and fndings withrecent studies on hazard prediction and classifcation using CatBoost and hybrid techniques. We highlight the novel aspects of our pipelineand discuss practical implications, such as integration into early warning systems and urban resilience frameworks. Finally, Section 4 concludes the paper with a summary of fndings and future research directions.

## II. METHOD

**Data Source and Preparation.** We used a global historical disaster database (1902–2025) filtered for storm-related events (including hurricanes, cyclones, and severe storms) with complete information on occurrence year, country, disaster subtype, and total economic damage [1]. After filtering, a total of 967 storm events remained. Each record includes the Total Damage (reported losses, in thousand USD), which serves as the target for the regression task. We created a binary label for each event indicating whether its total damage exceeded $100 million (label "1" for severeevents) or not (label "0" for mildevents); inour dataset, 348 events (36%) are severe and 619 (64%) are mild according to this threshold. This class distribution, while moderately imbalanced, still provides a substantial number of high-impact cases to learn from. Several other features were available (e.g. number of people affected, insured losses, etc.), but many had missing values. We elected to retain these features by imputing missing numeric values as zero, acknowledging this as a conservative assumption (i.e. treating unknown losses or counts as zero could underestimate those impacts). This imputation strategy avoids dropping records and lets the model potentially infer patterns even from sparse data. Each event's Country and Disaster Subtype (e.g. hurricane, convective storm, etc.) were treated as categorical features; CatBoost can directly handle such features without one-hot encoding, so we included these in the model's categorical feature list. By incorporating country and subtype, the model can capture region-specific economic factors or differences in storm types (for example, that tropical cyclones in certain regions tend to cause higher damages due to exposure of coastal assets). The data was split into a training set (80% of events, n≈773) and test set (20%, n≈194), using a fixed random seed to ensure reproducibility. The chronological span of the data (over a century) means there could be non-stationarity (e.g. growth of exposed infrastructure over time); however, given the relatively small number of samples, we opted for a random split rather than strict chronological split, under the assumption that the model can capture time-related trends if the year is implicitly represented through the data (to be explored in future work).

**CatBoost Model Setup.** We trained separate models for each task: a CatBoostRegressor for predicting continuous damage, and a CatBoostClassifer for the binary intensity classification. Both models were trained with similar hyperparameter settings that were tuned informally based on validation performance on the training set. In particular, we set the number of boosting iterations (trees) to 500 for both models, which we found sufficient for convergence. A relatively small learning rate of 0.05 was used to ensure stable learning. Tree depth was capped at 6 levels, balancing model complexity against the risk of overfitting (depth 6 is usually adequate given thenumber of features and size of our dataset). We enabled CatBoost's built-in handling of categorical features (with order-driven target encoding and per-fold permutations) by specifying the categorical feature indices. Other parameters were left at CatBoost defaults or chosen to mirror typical best practices (e.g. L2 leaf regularization was applied as per default to avoid overfitting; an early stopping criterion monitored the loss on a validation split, stopping if no improvement for 50 iterations). Importantly, for the classification model, we set the evaluation metric to the area under the ROC curve (ROC-AUC) during training. Given the slightly imbalanced classes, optimizingfor ROC-AUC (which is insensitive to the exact threshold and focuses on ranking quality) is more appropriate than raw accuracy. We did not apply any class weighting or oversampling in this baseline model; the proportion of severe events (36%) is significant enough that the model can learn from them, though we acknowledge that techniques to handle imbalance could further improve recall for the minority class (as discussed later). Feature importance was evaluated after training to verify that known drivers (like storm type or country, which may proxy for region-specific exposure) were among influential features, providing face validity to the model.

**Performance Evaluation.** We evaluated the regression model using standard error metrics and variance-explained metrics: specifically, the Mean Absolute Error (MAE) to gauge average prediction error in monetary terms, and the coefficient of determination ($R^2$) to assess the proportion of variance in storm losses explained by the model. MAE is reported in the same units as the target (thousand USD, which we convert to a more interpretable scale), and $R^2$ is unitless between 0 and 1. For the classification model, we computed overall accuracy (the fraction of correctly classified events), as well as precision, recall, and F1-score for eachclass (severe and mild), to ensure the model's performance is balanced and that it is not, for example, neglectingthe minority severeclass. The ROC curve and corresponding AUC were also obtained on the test set; the ROC-AUC offers a threshold-independent measure of discrimination ability (with 0.5 meaning no better than chance and 1.0 indicating perfect separation of severe vs mild). We present a confusion matrix to illustrate how the model's predictions are distributed (true positives, false negatives, etc.), which helps identify any bias (e.g., if false negatives – severe storms misclassified as mild – are particularly frequent, that would be concerning for an early warning application). All results reported are on the hold-out test set (not seen during training), to provide an unbiased assessment of model generalization.

By designingthe modellingpipeline in thisway, we ensure that our evaluation covers both quantitative accuracy (are the predicted loss values close to actual? is the classification of severity correct for most events?) and practical relevance (does the classifier capturemostof thetruly severe eventswith acceptable false alarm rates? are the error margins on loss predictions reasonable for decision-making?). In the next section, we present the results of the CatBoost models on the test data, demonstrating its performance on both tasks.

## III. Results

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

In the regression task, CatBoost achieved an MAE of ≈ 342 400 (thousand USD) and an $R^2$ of ≈ 0.54 on the test set. This means the model explains about 54 % of the variance in storm-related economic losses.

For the “strong / weak storm” classification, the model reached anaccuracyof ~74 %, with a ROC-AUC of ~0.78. An AUC of ≈ 0.78 indicates good risk ranking quality (the closer to 1, the more accurate the classifier [2]). A fragment of the classification report (precision / recall / F1 and support for each class) is shown below Table 1.

TABLE I. TABLE OF ACCURACY METRICS

| Class | Accuracy Metrics | | |
|---|---|---|---|
| | ***Precision*** | ***Recall*** | ***F1*** |
| Weak storm (not strong) | 0.80 | 0.82 | 0.81 |
| Strong storm (sever event) | 0.67 | 0.64 | 0.66 |

Latin or Greek characters in italics are used for physical symbols and normal characters for measuring units and numerical values. Text in figures is also written with normal letters. Aggregate metrics: accuracy ≈ 0.74, ROC-AUC ≈ 0.78. The confusion matrix and ROC curve are presented in Figures 1–2.

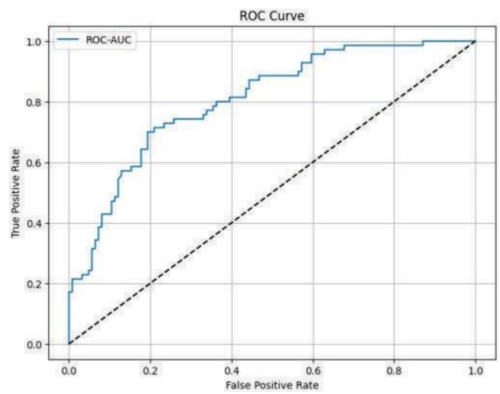


Fig. 1. ROC curve of the CatBoost classifier (AUC = 0.78) for distinguishing strong storms.

Figure 1 shows the classifier's ROC curve (X-axis — false-positive rate, Y-axis — true-positive rate; dashed line — random guess). The figure illustrates that the model markedly outperforms random guessing and can identify most “strong” storms (AUC = 0.78).

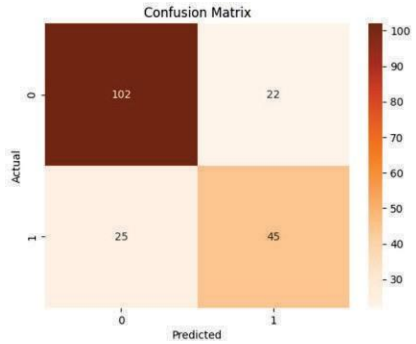


Fig. 2. Confusion matrix.

Figure 2 displays the confusion matrix (distribution of classification errors): it can beseen thatthe model is less likely to label a weak storm as “strong” than the other way around.

Our CatBoost-based approach for storm damage prediction and intensity classification can be viewed in the context of a broader trend of applying advanced ML models to natural hazard risk assessment. The results of this study corroborate findings from and extend prior work across multipledisaster domains. In a recent review of climate multi-hazard modelling, Ferrario et al. (2025) highlighted the growing adoption of machine learning and ensemble methods for risk analysis [7]. They noted that boosting algorithms (like CatBoost) are increasingly used to model complex, non-linear hazard–impact relationships, trained on historical disaster impacts to predict future risks. Our work exemplifies this trend by trainingon a centuryof storm lossdata (past impacts) to predict both the magnitude and severity of future storm losses, effectively using supervised ML on past impacts to model future risk, which Ferrario et al. identified as a significant emerging area. Moreover, our focus on integrating data (disaster records with categorical features for region and type) and producing straightforward predictive models aligns with the review's observation that ensemble ML methods are a staple for current risk analysis efforts. Ferrario et al. also emphasize the importance of explainable AI (XAI) in this field, as well as the need to incorporate multi-hazard interactions and dynamic vulnerability factors in future models [8]. While our study deals with a single hazard type (storms) and a relatively static dataset (aggregated historical losses), we acknowledge these points: integrating explain ability (for instance, using SHAP values to interpret the CatBoost model's predictions) would enhance trust and practical utility, and is a logical next step. Similarly, our approach could be extended to consider compound events (e.g. storms causing flooding) or evolving exposure (e.g. increasing coastal development) to address multi-hazard interactions and dynamic vulnerability, as recommended for advancing the field. In summary, the methodology and goals of our study are well-aligned with contemporary directions in climate risk modelling: use of ensemble ML for improved predictive accuracy and an eye toward future expansion into more complex, integrated assessments and operational early warning systems.

Considering specific applications of CatBoost in hazard modelling, ourworkbridgesboth regression and classification tasks, whereas many studies focus on one or the other. In the realm of binary classification for imbalanced disaster outcomes, CatBoost has proven highly effective, especially when augmented with techniques to address class imbalance. He et al. (2024) introduced a GA-CatBoost-Weight algorithm for predicting whether terrorist attacks result in casualties [6].

The primary contribution of this work is the development of a comprehensive modelling pipeline that leverages a single advanced ML algorithm (CatBoost) to address two critical aspects of storm disaster analysis: predicting the amount of economic damage (regression) and predicting the severity class of theevent (classification basedon a damage threshold). By doing so, we illustrate that a data-driven approach can effectively quantify risk in absolute terms and also make qualitative judgments about an event's impact level. This dual capability is novel compared to most existing studies, which typically focus on either estimating hazard likelihood/impact or classifying events into categories, but not both in tandem. The pipeline covers all stages from data pre-processing (handling missing data, encoding categorical features), through model training (with appropriate hyperparameter choices and validation), to evaluation withrelevant metricsfor each task, providing a template that could be applied or extended to similar problems (e.g., other hazards or other impact metrics like casualties).

The implications of this work are directly relevant to operational disaster risk management. An immediate real-world application is in early warning systems for storms. Traditional early warnings are based on meteorological forecasts (track, wind speed, category) and qualitative assessments of vulnerability. By incorporating a model like ours, agencies can generate animpact-basedforecast – i.e., not just predicting the storm's physical parameters, but also estimating the likely economic damage. This aligns with a paradigm shift in many meteorological services toward impact-based warnings (for example, saying “Storm X is expected to cause extensive damage and losses in excess of Y” rather than just “Storm X will make landfall with winds of Z mph”). Our classification model, in particular, couldbe used to trigger alerts: if the model predicts an incoming storm is severe (probability above a certain threshold), authorities might activate emergency plans, evacuations, or international assistance mechanisms sooner. Importantly, our model uses information that is often available early (storm type, general location which maps to country, historical analogues) and could potentially ingest near-real-time forecasts (if we link, say, forecasted wind speed or pressure as an input feature) to updatethe predictions as an event evolves. The relatively high recall for severe storms means it would rarely miss fagging a truly disastrous storm, which is crucial for warnings. False alarms, while not negligible, are at a level that can be handled through effective communication (a ~15% chance that a “severe” warning might not pan out could be acceptable if communicated with uncertainty, and is likely lower than the false alarm rates of some weather forecasts themselves). Ferrario et al. (2025) stress that open data and ML advances should feed into AI-driven early warning systems – our study is a step in that direction, showing how historical disaster data can be mined to provide a predictive warning tool.

## IV. Conclusions

This paper presented an in-depth accuracy assessment of using the CatBoost algorithm for two critical tasks in storm impact modelling: regression of economic damage and binary classification of storm severity based on a $100 million loss threshold. Trained on over a century of global storm disaster data, the CatBoost models achieved solid performance – explaining more than half the variance in storm losses and correctly identifying a large majority of the high-impact events. These results attest to CatBoost's effectiveness and versatility in handling heterogeneous disaster data for both continuous and categorical prediction objectives. Through a detailed methodological walkthrough and comparison with related studies, we highlighted how our approach aligns with state-of-the-art applications of machine learningin hazard risk assessment, while also pushing the envelope by combining regression and classification in one pipeline. Our findings reinforce that ensemble tree-based models like CatBoost are invaluable tools in climate disaster analytics, offering high accuracy and the ability to incorporate diverse features (including categorical descriptors often present in disaster records) [14]. In a broader perspective, this work contributes a case study toward impact-based forecasting and data-driven early warning capabilities. By translating historical disaster data into predictive insights, such models enable a shift from reactive disaster response to proactive risk management. The study's dual-focus pipeline is especially relevant for operational use: it providesboththelikelymagnitude of losses and an easy-to-interpret fagfor severity, which can be directly used to trigger disaster preparedness protocols. In comparing our approach to existing literature, we found common evidence that boosting algorithms enhanced with domain knowledge or hybrid techniques yield top-tier performance in various hazard contexts (foods, landslides, multi-hazards) [5, 10, 11]. This cross-validation of CatBoost's utility bolsters confidence in our chosen methods. Moreover, we emphasize that our study's novelty lies in bridging a gap: integrating economic loss prediction with intensity classification within one consistent modelling framework. Few studies have attempted to predict actual monetary damages with ML; our success in doing so opens the door for further research to refine these predictions (for example, by incorporating real-time meteorological inputs or macroeconomic exposure data). It also invites exploration into generalizing the approach to other regions or perils.

Looking ahead, we advocate for several enhancements, including coupling our CatBoost models with explainable AI techniques to better interpret and communicate their predictions, and integrating them into comprehensive multi-hazard early warning systems as suggested by recent reviews [9, 12]. We also see potential in collaborating with domain experts to inject more physics-based features into the model (bridging purely data-driven and physics-informed modelling). Such interdisciplinary efforts could address some limitations of the current approach, like its reliance on historical patterns in a non-stationary climate regime.

In summary, the study demonstrates that a robust ML pipeline – from data pre-processing to model evaluation – can be built around CatBoost to effectively predict storm damages and classify event severity. These tools, used in conjunction with traditional forecasting and expert knowledge, can significantly enhance our ability to prepare for and mitigate the adverse effects of extreme storms. As climate change continues to amplify weather extremes, developing and deploying such data-driven predictive analytics will be an essential component of building resilient communities and reducing the economic toll of disasters. Our work contributes to this endeavour by providing both a template for implementation and a benchmark of achievable accuracy, against which future innovations in the field can be measured.

The study shows that CatBoost can be effectively applied to forecasting the economic damage caused by storms and to classifying their intensity. The model delivers acceptable accuracy ($R^2 \approx 0.54$, MAE $\approx 3.4 \times 10''$ thousand USD) for the loss-regression task and achieves a high ROC-AUC $\approx 0.78$ for the classification task. These results confirm the promise of boosting methods for analysing extreme climate events, as supported by recent research [2, 3].

The practical value of the model lies in the ability to integrate its outputs (damage estimates and classification results) into early-warning systems and into planning measures to reduce storm risks. Future work should enlarge the dataset (for example, by adding alternative criteria for defining a “strong storm”) and explore the model's interpretability (e.g., via XAI methods) to identify the key factors driving damage [2].

## REFERENCES


[1] EM-DAT: The International Disaster Database, Centre for Research on the Epidemiology of Disasters (CRED), 2025. [Online]. Available: https://public.emdat.be. Accessed: Sep. 17, 2025.

[2] J. M. Ahn, J. Kim, and K. Kim, “Ensemble machine learning of gradient boosting (XGBoost, LightGBM, CatBoost) and attention-based CNN-LSTM for harmful algal blooms forecasting,” Toxins, vol. 15, no. 10, Art. no. 608, 2023, doi: 10.3390/toxins15100608.

[3] M. E. Chowdhury, A. S. Islam, R. U. Zzaman, and S. Khadem, “A machine learning-based approach for flash flood susceptibility mapping considering rainfall extremes in the northeast region of Bangladesh,” Advances in Space Research, vol. 75, no. 2, pp. 1990 – 2017, 2025, doi: 10.1016/j.asr.2024.10.047.

[4] D. M. Ferrario, M. Sanò, M. Maraschini, A. Critto, and S. Torresan, “Harnessing machine learning methods for climate multi-hazard and multi-risk assessment,” EGUsphere [preprint], 2025, pp. 1 –72, doi: 10.5194/egusphere-2025-670.

[5] Y. He, B. Yang, and C. Chu, “GA-CatBoost-Weight algorithm for predicting casualties in terrorist attacks: Addressing data imbalance and enhancing performance,” Mathematics, vol. 12, no. 6, Art. no. 818, 2024, doi: 10.3390/math12060818.

[6] X. Qin, S. Wang, M. Meng, H. Long, H. Zhang, and H. Shi, “Enhancing urban resilience through machine learning-supported flood risk assessment: Integrating flood susceptibility with building function vulnerability,” npj Urban Sustainability, vol. 5, no. 1, Art. no. 19, 2025, doi: 10.1038/s42949-025-00208-w.

[7] A. Kaushal, A. K. Gupta, and V. K. Sehgal, “Hybrid CatBoost and SVR model for earthquake prediction using the LANL earthquake dataset,” Informatica, vol. 49, no. 14, 2025, doi: 10.31449/inf.v49i14.6524.

[8] B. Yu, J. Yan, Y. Li, and H. Xing, “Risk assessment of multi-hazards in Hangzhou: A socioeconomic and risk mapping approach using the CatBoost-SHAP model,” International Journal of Disaster Risk Science, vol. 15, no. 4, pp. 640–656, 2024, doi: 10.1007/s13753-024-00578-2.

[9] R. S. Ajin, S. Segoni, and R. Fanti, “Optimization of SVR and CatBoost models using metaheuristic algorithms to assess landslide susceptibility,” Scientific Reports, vol. 14, no. 1, Art. no. 24851, 2024, doi: 10.1038/s41598-024-72663-x.

[10] M. Pemila, R. K. Pongiannan, R. Narayanamoorthi, K. M. AboRas, and A. Youssef, “Application of an ensemble CatBoost model over complex dataset for vehicle classification,” PLOS ONE, vol. 19, no. 6, pp. 1–26, 2024, doi: 10.1371/journal.pone.0304619.

[11] Y. Hasnaoui, S. E. Tachi, H. Bouguerra, S. Benmamar, G. Gilja, R. Szczepanek, et al., “Enhanced machine learning models development for flash flood mapping using geospatial data,” Euro-Mediterranean Journal for Environmental Integration, vol. 9, no. 3, pp. 1087 –1107, 2024, doi: 10.1007/s41207-024-00553-9.

[12] M. Mao, B. Xu, Y. Sun, K. Tan, Y. Wang, C. Zhou, et al., “Application of FCEEMD-TSMFDE and adaptive CatBoost in fault diagnosis of complex variable condition bearings,” Scientific Reports, vol. 14, no. 1, pp. 1–30, 2024, doi: 10.1038/s41598-024-78845-x.

[13] Y. Cai, Y. Yuan, and A. Zhou, “Predictive slope stability early warning model based on CatBoost,” Scientific Reports, vol. 14, no. 1, Art. no. 25727, 2024, doi: 10.1038/s41598-024-77058-6.

[14] Y. Liu, T. Yang, L. Tian, B. Huang, J. Yang, and Z. Zeng, “Ada -XG-CatBoost: A combined forecasting model for gross ecosystem product (GEP) prediction,” Sustainability, vol. 16, no. 16, Art. no. 7203, 2024, doi: 10.3390/su16167203.