ReDI School of
Digital Integration

Economic Damage Prediction and Anomaly Detection in Global Natural Disaster Data Using Machine Learning

Machine Learning Online final Project

We use tech to connect human potential and opportunity with dignity & humility

by Prudensy Febreine Opit

Motivation Behind the Project 

With a background in humanitarian logistics, I wanted to combine my research experience with my new data science skills to build a useful, data-driven tool for estimating disaster impact and detecting anomalies, helping improve preparedness and response in real-world scenarios.

Project Workflow

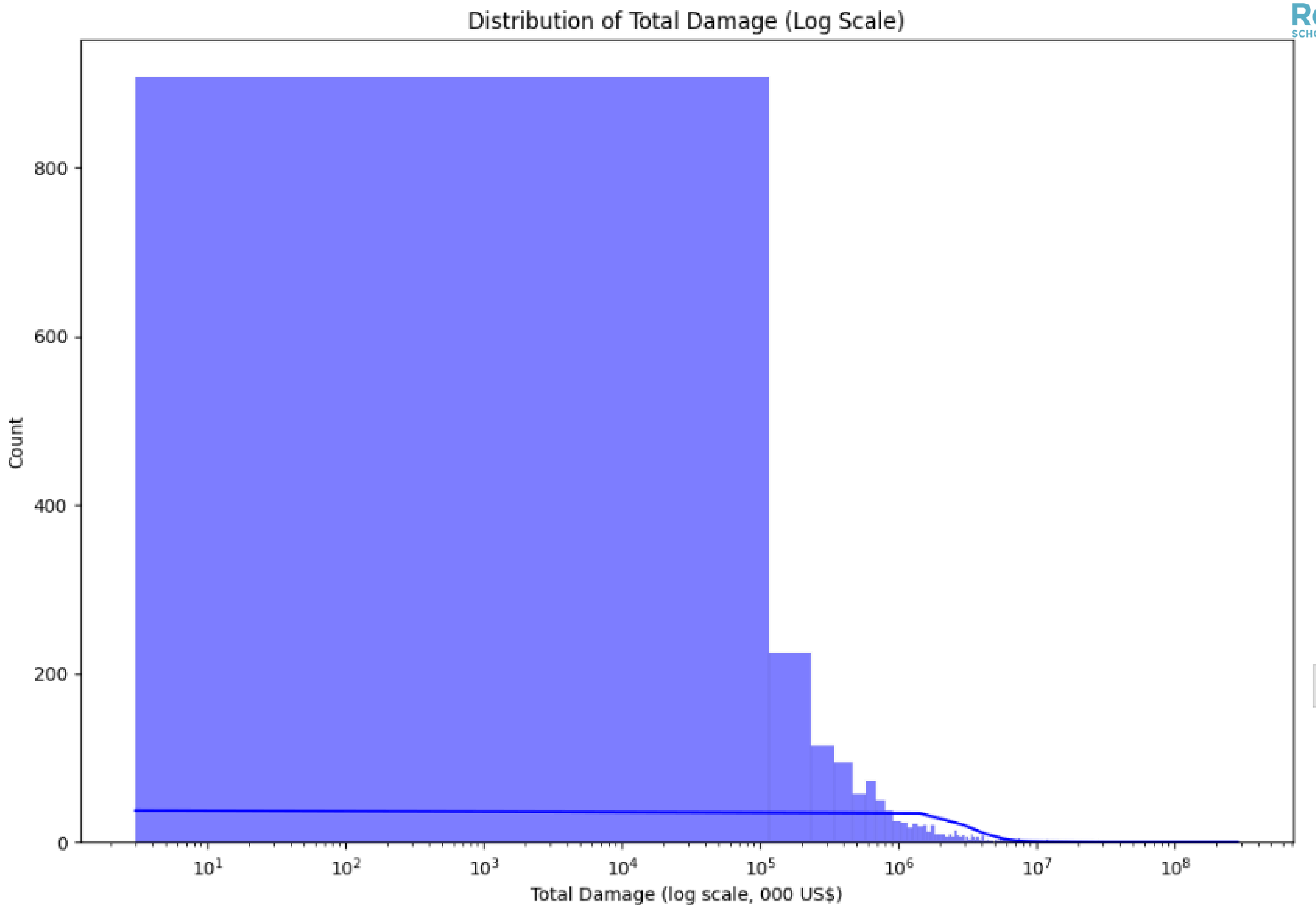
1. Load, clean, and merge datasets
2. Exploratory Data Analysis (EDA)
3. Feature Engineering (incl. categorical encoding)
4. Train and evaluate regression models
5. Detect economic outliers using HDBSCAN clustering
6. Reflect on limitations and ethical considerations
7. Conclusion and future work

Features & Functionality

- Predict total economic damage from disasters using regression models (Linear Regression, Random Forest (Default),Random Forest with Log-transformed Target, XGBoost (Default), AGBoost Regression with Tuned Hyperparameters, CatBoost, Neural Network)
- Identify unusual disaster events using unsupervised anomaly detection (HDBSCAN)
- Analyze feature importance to understand key factors influencing damage
- Visualize actual vs. predicted damage and anomaly clusters (with PCA)
- Integrate and clean multi-source real-world data (EM-DAT, EC-JRC INFORM Risk Index, World Bank)
- Apply machine learning pipelines for both supervised and unsupervised tasks

Target Variable Distribution: Total Damage

- Damage is right-skewed.
- Most events are small-scale, but a few disasters cause extremely high losses.
- Only events with ≥ $10,000 USD in reported damage were included to focus on significant economic impact.



Regression Models Compared

Models Used:

Linear Regression, Random Forest (Default & Log Target), XGBoost (Default & Tuned), CatBoost, Neural Network

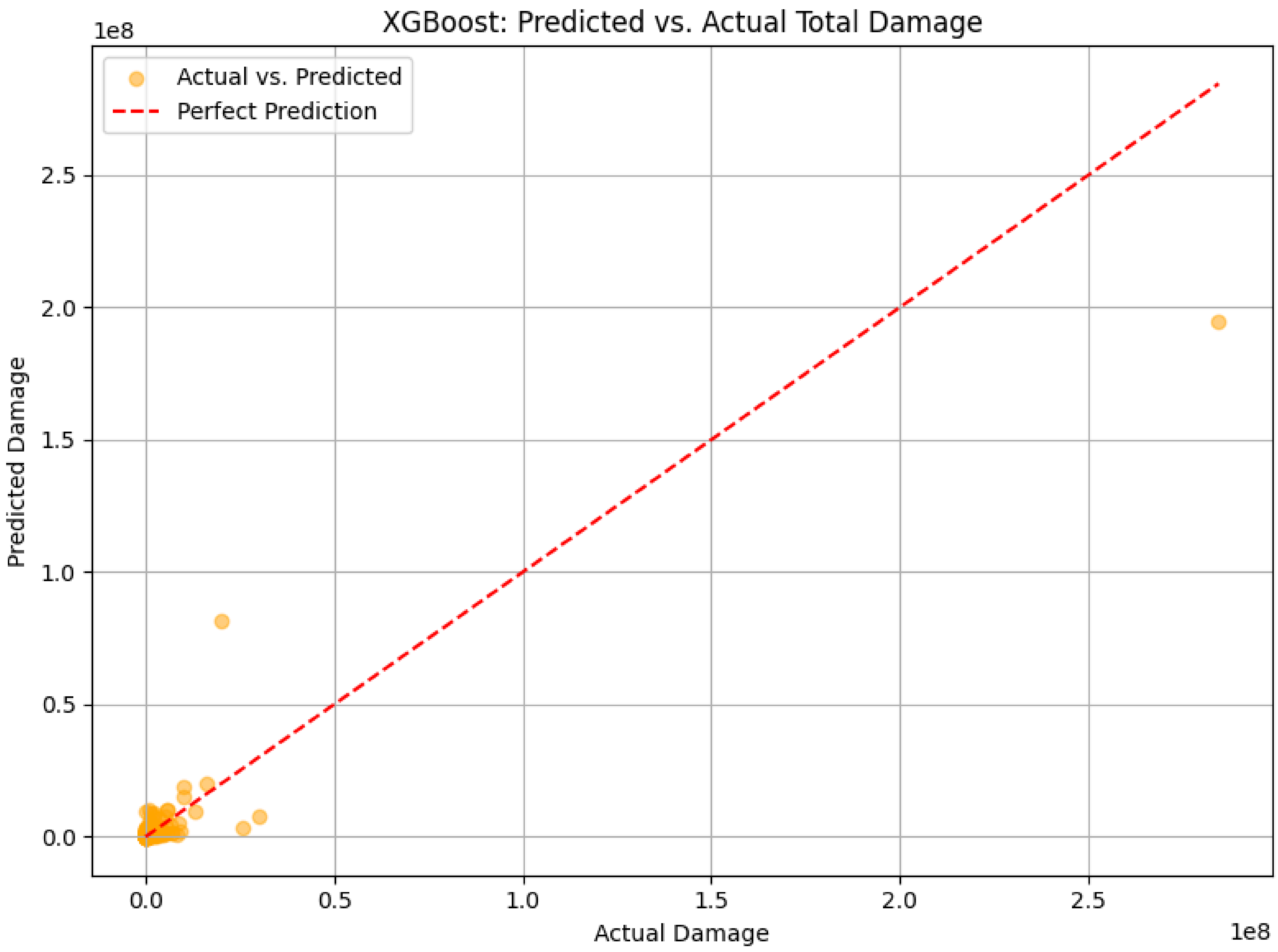
Model Evaluation Comparison

Model	MSE	RMSE	R²	Performance	Speed	Complexity
Linear Regression	210,359,085,633,349.09	14,503,761.09	0.03	Low	Fast	Low
Random Forest (Default)	114,348,645,549,993.84	10,693,392.61	0.47	Medium	Medium	Medium
Random Forest (Log Target)	187,336,454,312,772.50	13,687,090.79	0.14	Good	Medium	Medium
XGBoost (Default)	46,339,333,284,240.18	6,807,300.00	0.79	Good	Medium	High
XGBoost (Tuned)	36,445,385,285,835.54	6,037,001.35	0.83	Best	Slow-Medium	High
CatBoost	81,913,858,951,514.50	9,050,627.54	0.62	Good	Slow-Medium	High
Neural Network	209,178,092,060,085.38	14,462,990.43	0.04	Low	Slow	High

🏆 XGBoost (Tuned) achieved the best performance (R^2 = 0.83, RMSE ≈ 6M USD)

🌲 Tree-based models (XGBoost, CatBoost) outperformed linear and neural models

The Best Regression Model: XGBoost (Tuned)



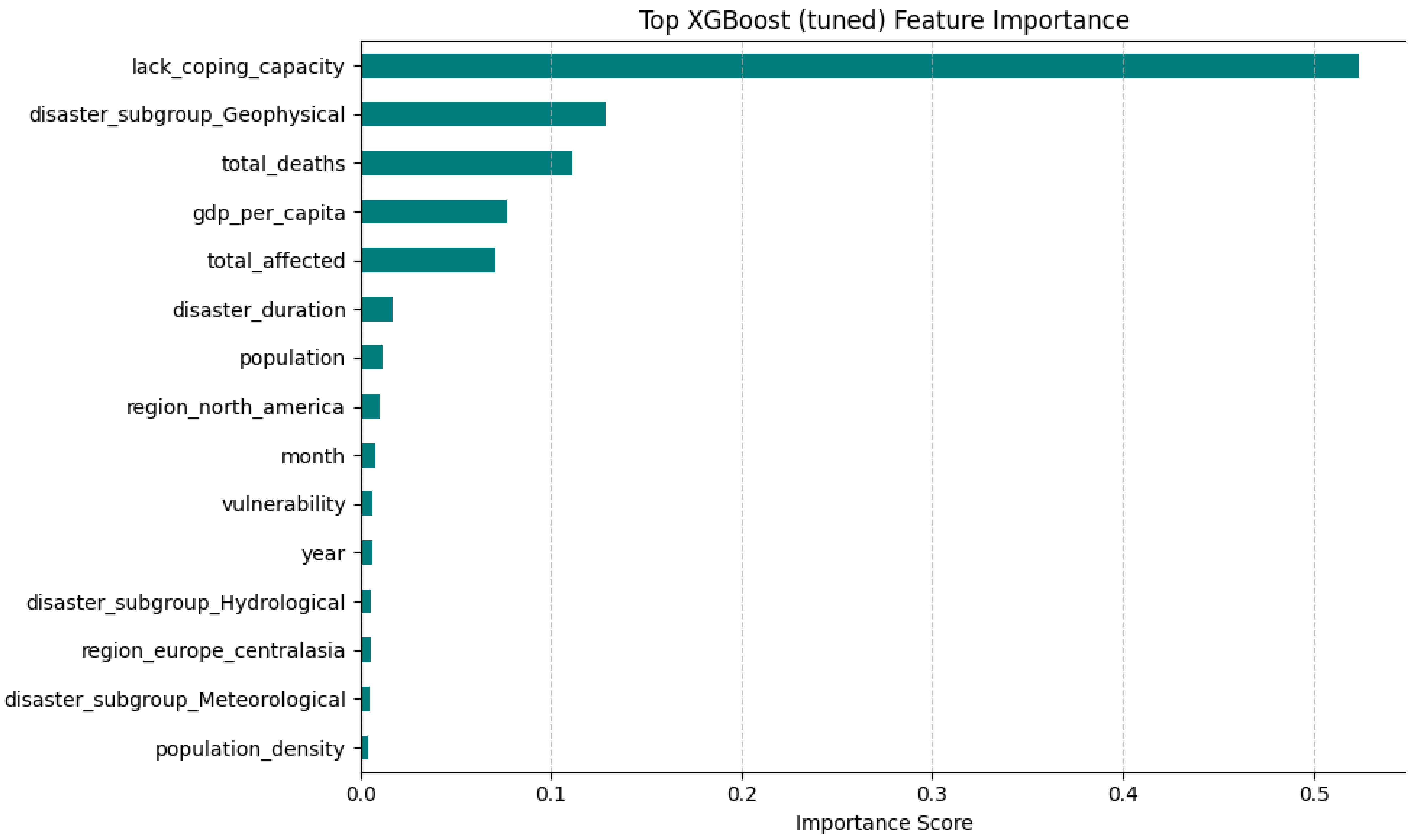
Result:

- MSE: 36445385285835.54
- RMSE: 6037001.35
- R^2: 0.83 (83%) --> Best among all models

The model captures 83% of the variance in disaster damage, with the lowest error among all models.

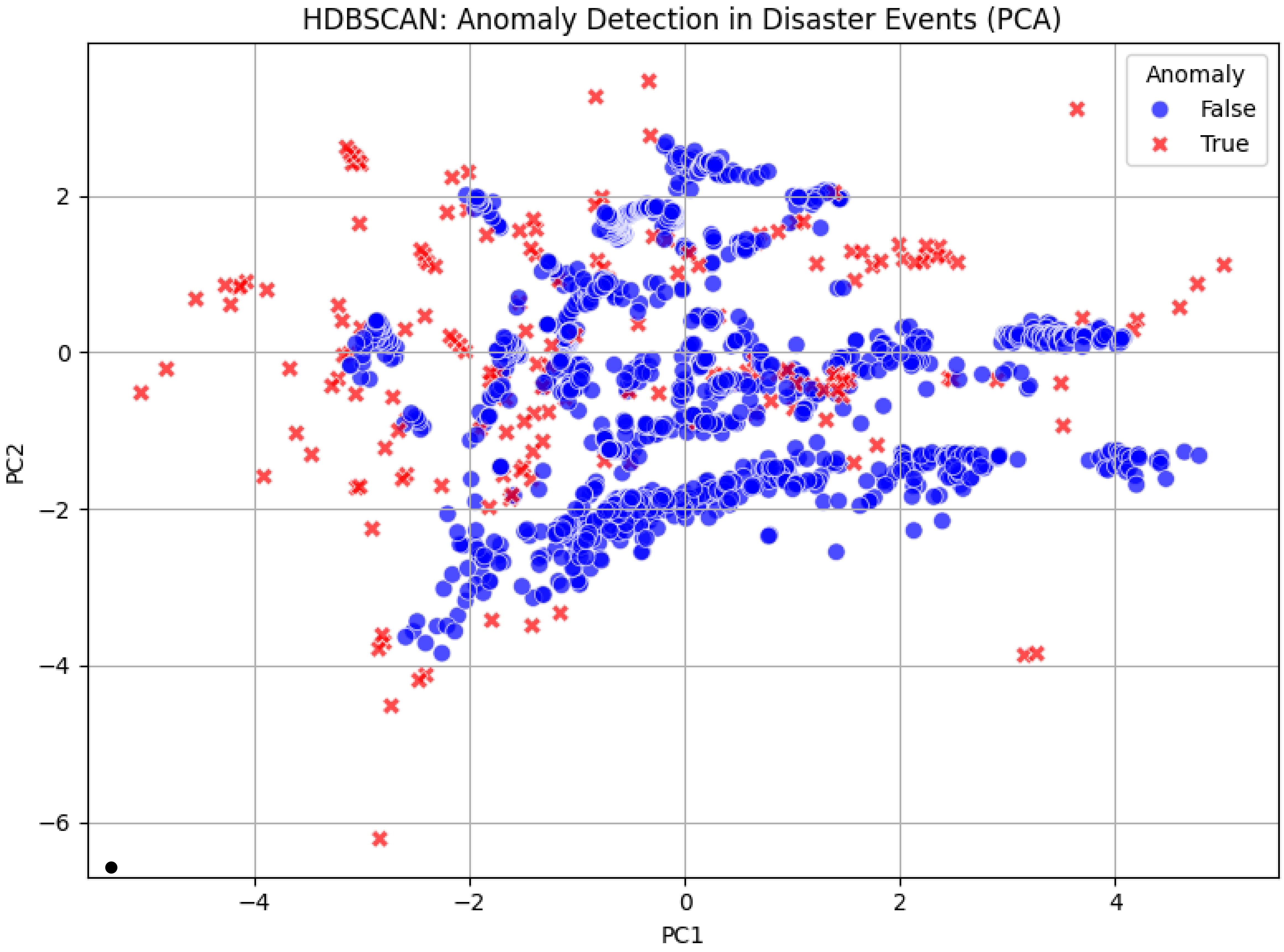
Most events have low damages, with a few high-impact outliers.

Feature Importance



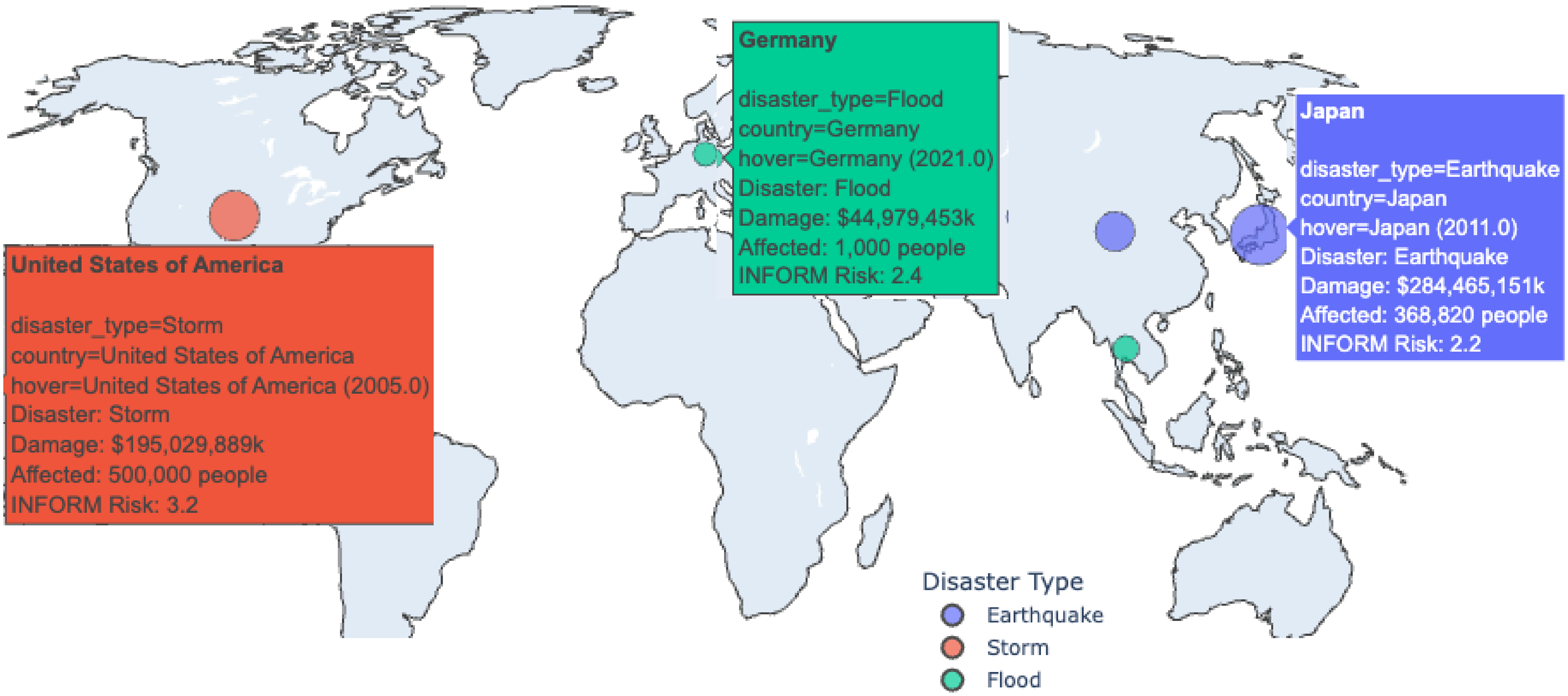
- Coping capacity is the most influential factor in predicting disaster damage.
- Geophysical disasters and fatalities also play a major role.
- Socioeconomic factors matter more than region or population size.

Anomaly Detection using HDBSCAN (PCA Projection)



- Red points represent disaster events flagged as anomalies.
- PCA reduces dimensionality to 2D for visualization only.
- Anomalies reveal events with unusual damage patterns or unexpected severity.

Top Disaster Anomalies (Max Damage per Country)



Real-World Use Cases & Impact of this Project

- Identifies countries with high-impact disasters to help prioritize global attention and funding
- Supports policy decisions on disaster preparedness and climate adaptation
- Informs NGOs and humanitarian agencies where to focus relief efforts and pre-position supplies
- Assists governments in planning and allocating emergency response budgets
- Aids insurers and risk analysts in evaluating disaster exposure and financial risk
- Enables urban planners to design more resilient infrastructure using data-driven insights
- Can be extended with machine learning to detect future anomalies and support early warning systems

Challenges & Solutions

⚠️ Challenges

- Data Bias: Incomplete reporting may underrepresent vulnerable regions.
- Model Bias: Socioeconomic factors may skew predictions unfairly.
- Interpretability: Complex models are hard to explain.
- Temporal Limitations: Past data may not reflect future risks (e.g., climate change).
- Ethical Risks: Misuse could affect funding or aid decisions.

🛠️ Solutions

- Ensure transparency in data filtering and assumptions.
- Add complementary datasets or weighting for fairness.
- Treat models as support tools, not final decision-makers.
- Update models regularly to reflect changing risk.

Conclusions & Future Improvements

Conclusions

- ML models like XGBoost predicted disaster damage with strong accuracy ($R^2 \approx 0.83$).
- Lack of coping capacity was the most important damage predictor.
- Geophysical disasters and total deaths were also strong indicators.
- HDBSCAN detected anomalies, revealing extreme or underreported events.

Future Improvements

- Integrate weather and spatial data for deeper insights.
- Explore time-series models to capture temporal patterns.
- Expand to more countries and smaller-scale disasters to reduce bias and improve generalizability.

Potential Uses

🌍 Potential Use

- Help build early warning and planning systems
- Support real-time monitoring of disaster anomalies
- Create dashboards for governments and NGOs
- Guide funding and aid decisions
- Support climate resilience planning
- Improve disaster insurance models
- Test “what-if” disaster scenarios

Thank you!

Munich, 1st June 2025