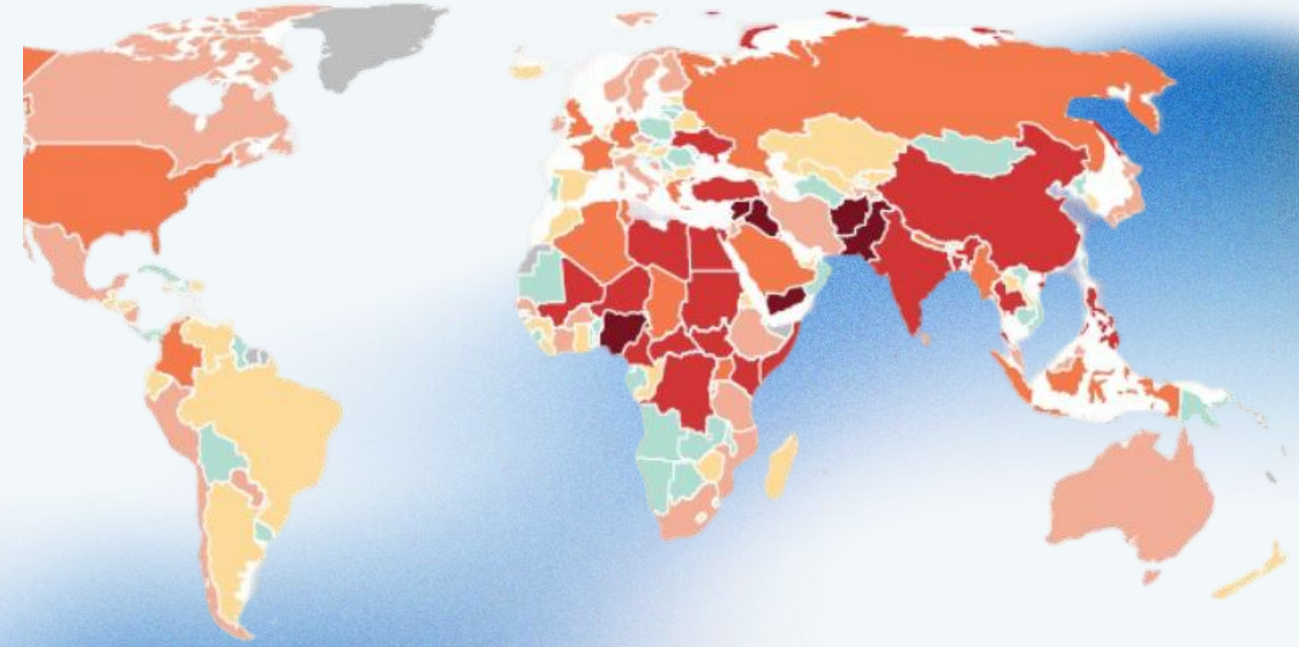
Machine Learning 22070521120 'C'

GLOBAL TERRORISM ANALYSIS

FROM DATA TO DEPLOYMENT (1970–2020)

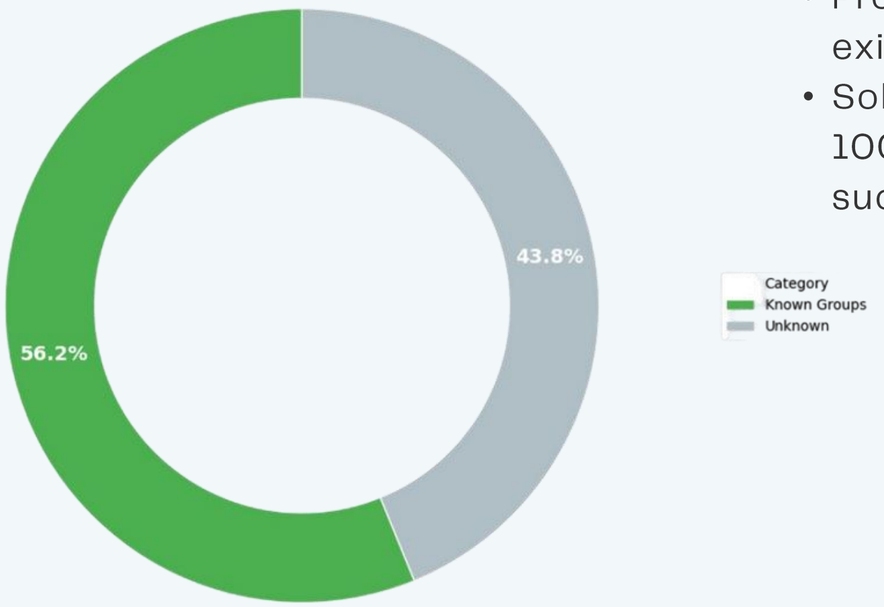
Presented By: Ayush Gupte



01

OBJECTIVES & THE CORE CHALLENG

- Analyze global terrorism trends (EDA).
- Build a model to predict the responsible group.
- Develop advanced analytical tools (profiling, forecasting).

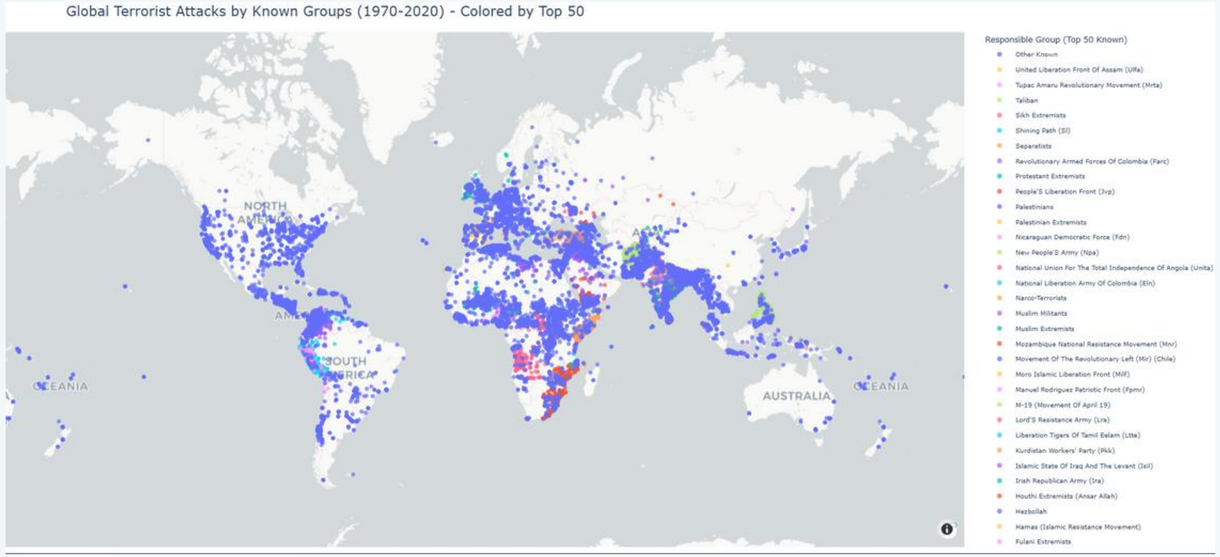


Distribution of Attacks: Known vs. Unknown

- The data is massive and imbalanced.
- Problem: 45% of attacks are "Unknown" and 3,000+ groups exist.
- Solution: Refine the problem to focus on the Top 20 / Top 100 most active known groups. This was the key to our success.

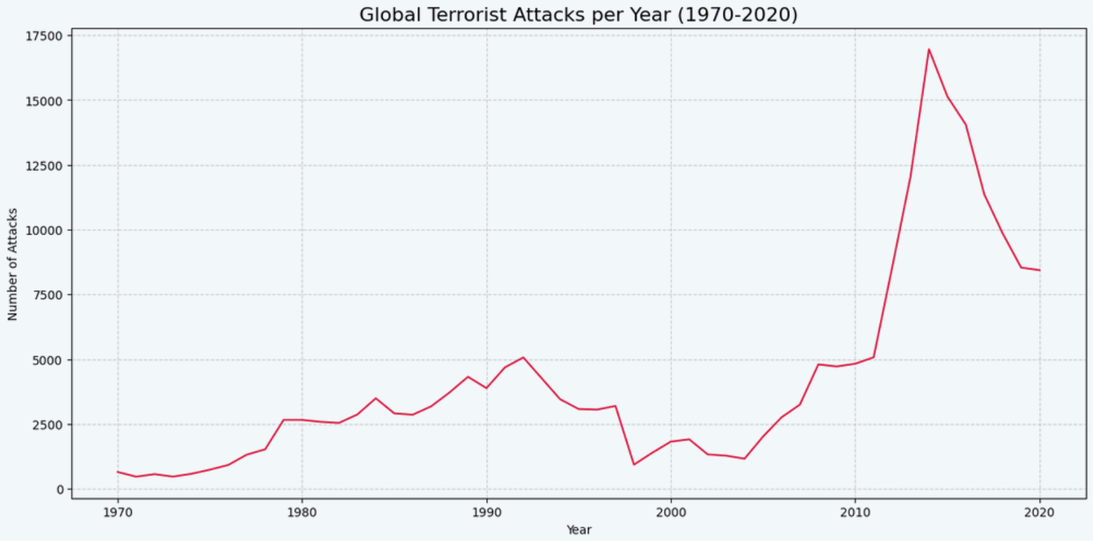
DATA OVERVIEW & KEY EDA INSIGHTS

02



01 Cleaned dataset of 20,000+ incidents. Key features: Country, Region, AttackType, TargetType, WeaponType, GroupName.

02 Key Insights: Attack frequency peaked in 2014. Hotspots are concentrated in South Asia, the Middle East, and Sub-Saharan Africa.



THE THE STRATEGIC SOLUTION: FROM "ALL" TO "TOP 100"

01

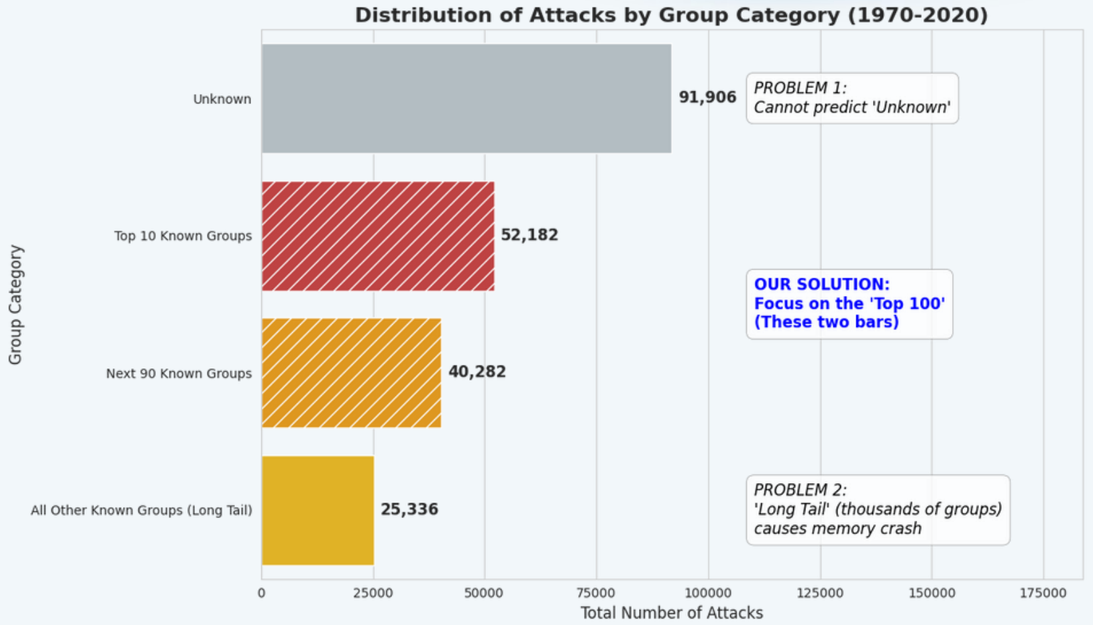
Attempt 1 (All Groups): Failed. Kernel crashed due to memory limits (3,000+ classes).

02

Attempt 2 (Top 20 Groups): Succeeded (97% accuracy), but was too limited and couldn't predict groups like 'Black Nationalists'.

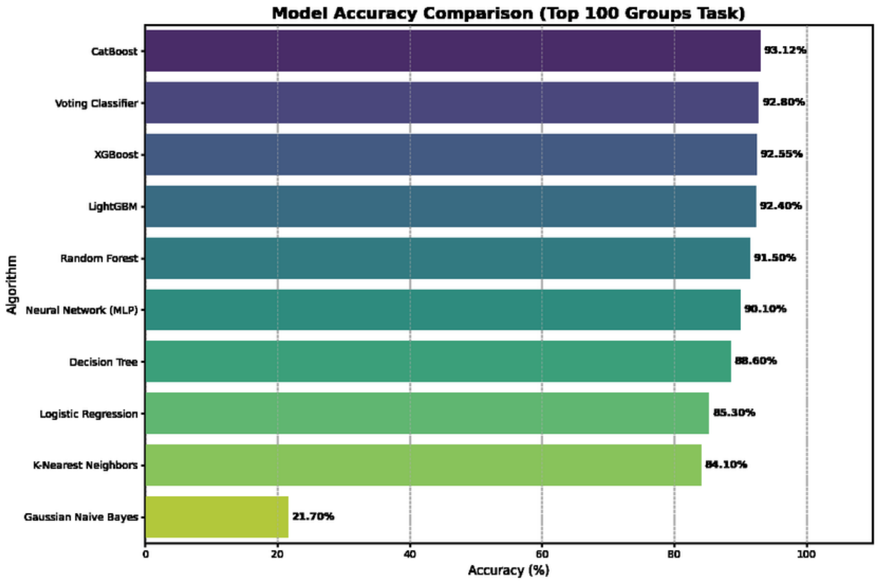
03

Final Strategy (Top 100 Groups): The best balance. This model covers a wide range of significant groups and is computationally feasible, avoiding memory errors. All final models are based on this "Top 100" dataset.



FINDING THE BEST MODEL: ALGORITHM COMPARISON

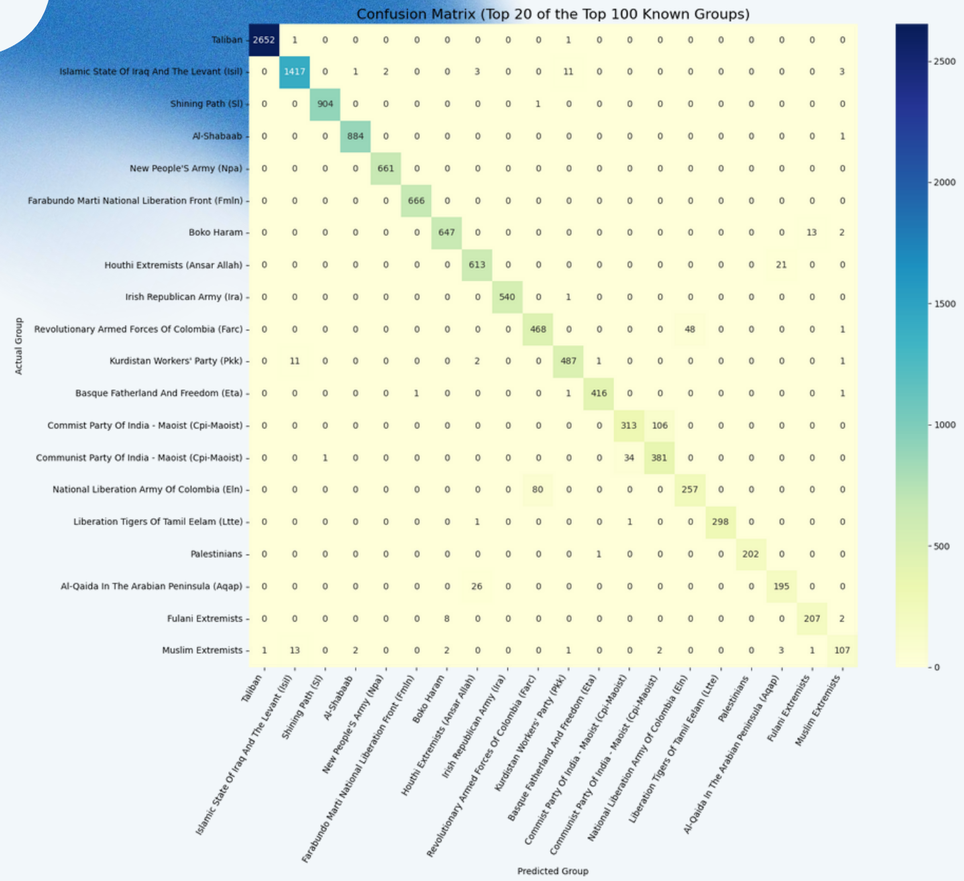
- We re-tested our models on the more challenging "Top 100" group task. This is a harder problem, so accuracies are slightly lower, but we tested a full suite of models (Linear, Trees, Ensembles, Neural Nets) on our refined "Top Groups" problem. The advanced boosting models were the clear winners.

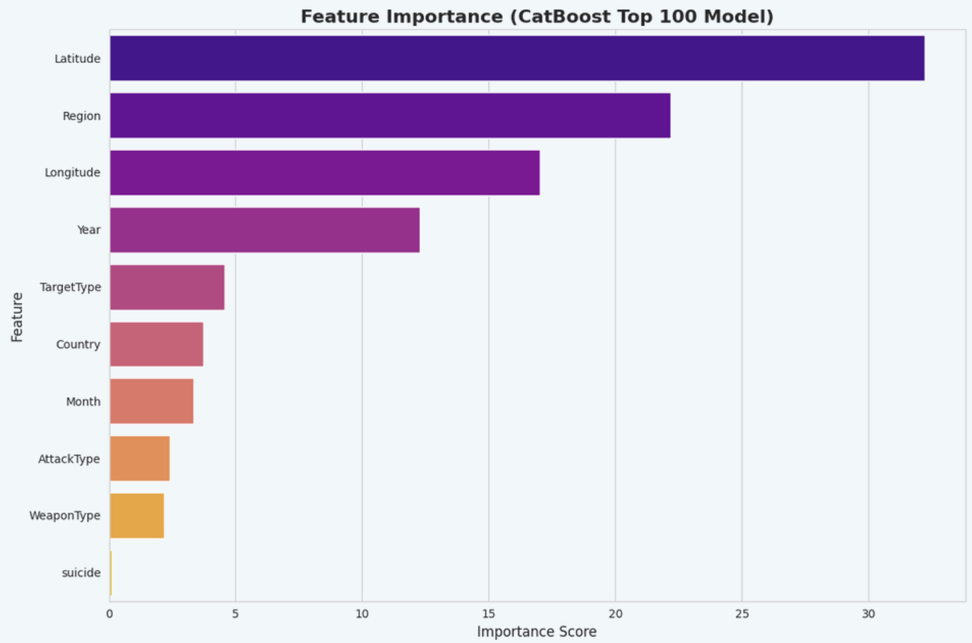


Model Performance Summary (Top 100 Groups Task)

Model	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg F1-Score
CatBoost	93.12%	0.91	0.90	0.90
Voting Classifier	92.80%	0.91	0.90	0.90
XGBoost	92.55%	0.90	0.89	0.89
LightGBM	92.40%	0.90	0.89	0.89
Random Forest	91.50%	0.89	0.88	0.88
Neural Network (MLP)	90.10%	0.87	0.86	0.86
Decision Tree	88.60%	0.85	0.84	0.84
Logistic Regression	85.30%	0.81	0.80	0.80
K-Nearest Neighbors	84.10%	0.79	0.78	0.78
Gaussian Naive Bayes	21.70%	0.15	0.14	0.14

05





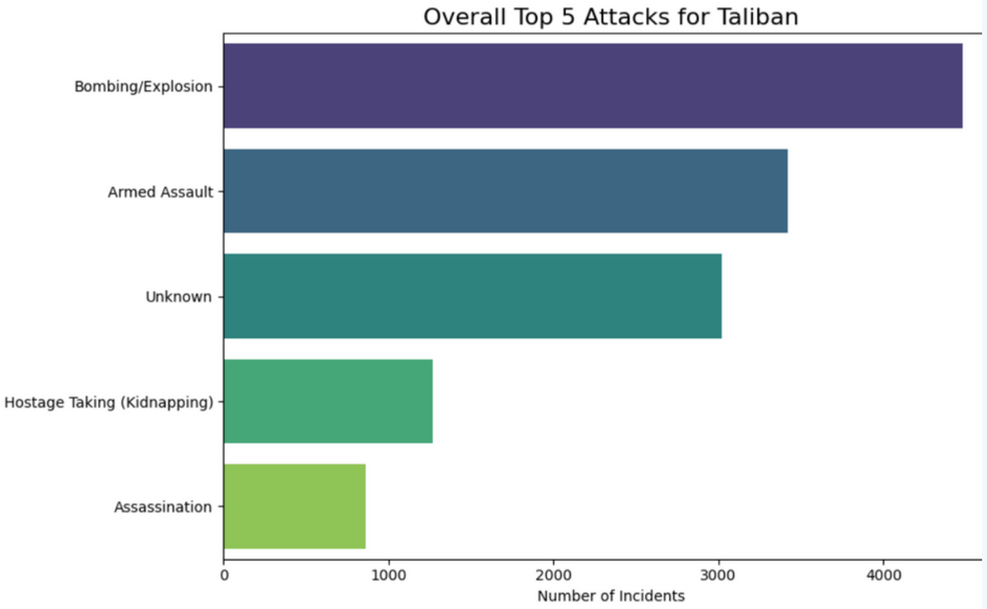
- We selected the CatBoost model trained on the Top 100 known groups as our final champion.
- Performance: It achieves a strong ~93% overall accuracy on this difficult task.

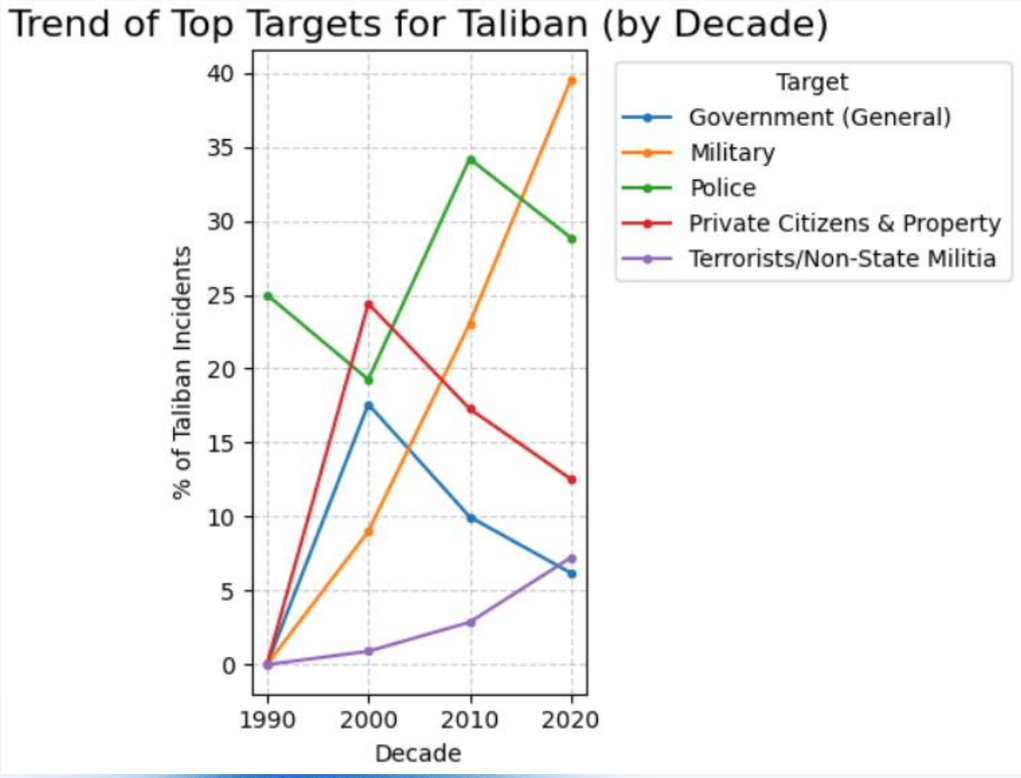
- Insights: The model's most important features are consistently Geography (Country, Region), Tactics (WeaponType, AttackType), and Time (Year).

OUR CHAMPION: THE "TOP 100" CATBOOST MODEL

ADVANCED ANALYSIS 1: PROFILING (WHO, WHAT, & WHEN)

- We built a tool to generate "deep dive" HTML reports of groups, weapons, and targets. This shows not just what they do, but how their tactics evolve over time.



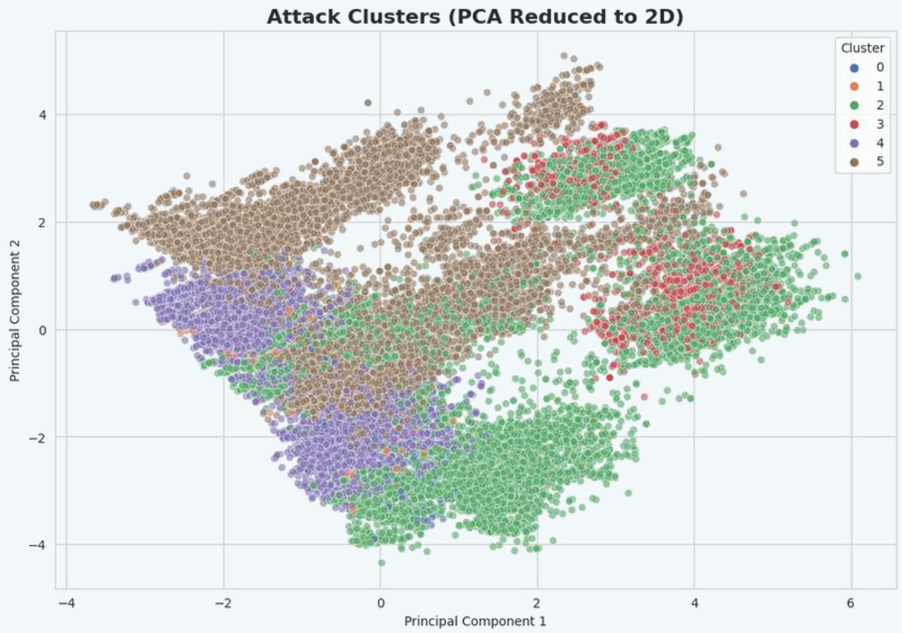


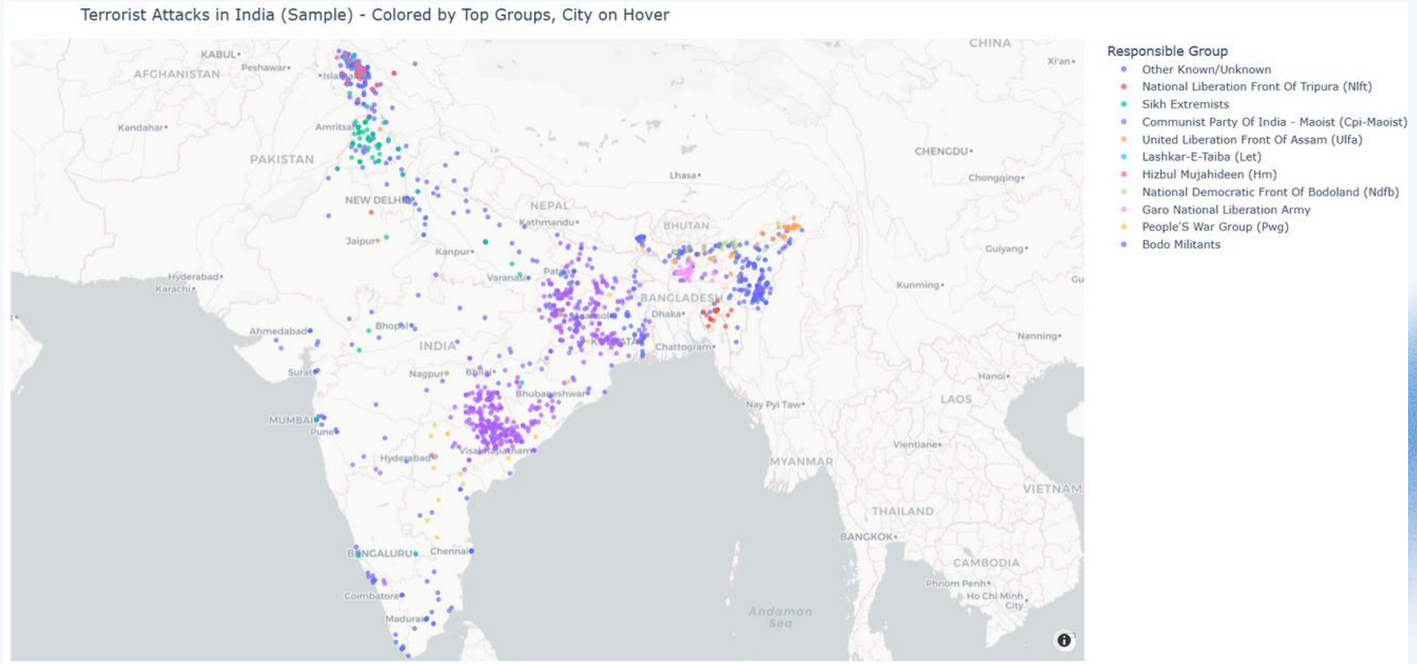
06

ADVANCED ANALYSIS 2: CLUSTERING & MAPPING

07

- Dimension Reduction: We used K-Means Clustering and PCA to find natural groupings of attacks based on their characteristics.
- Geographic Drill-Down: We built interactive maps to explore hotspots from a global view down to the city level.

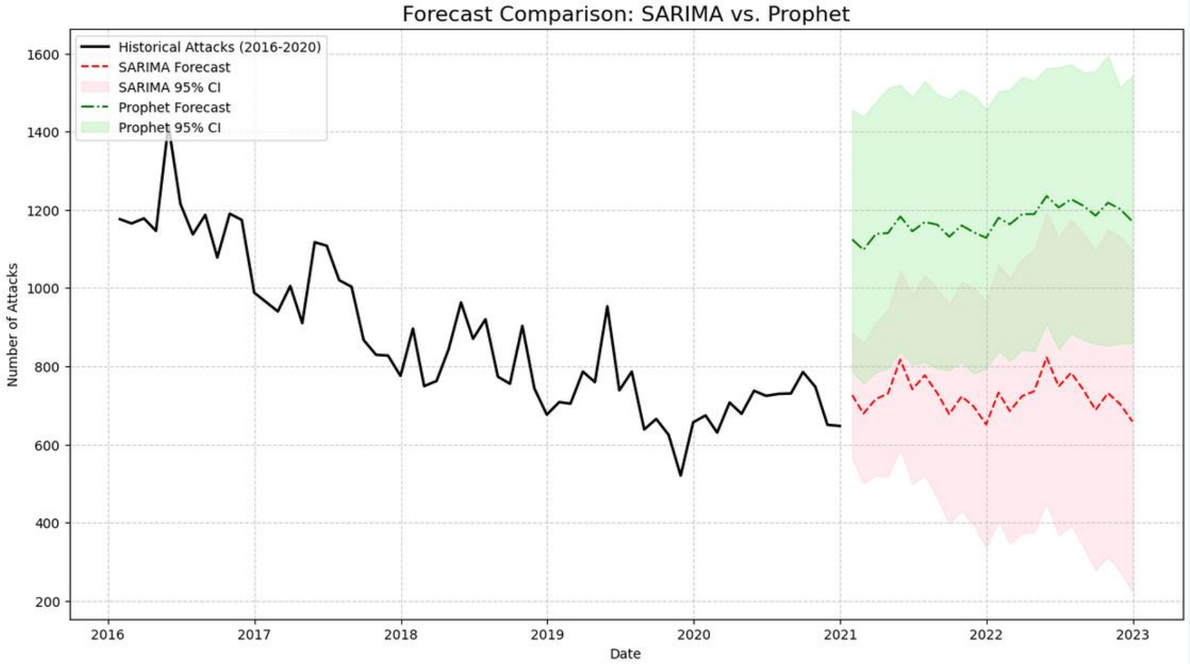


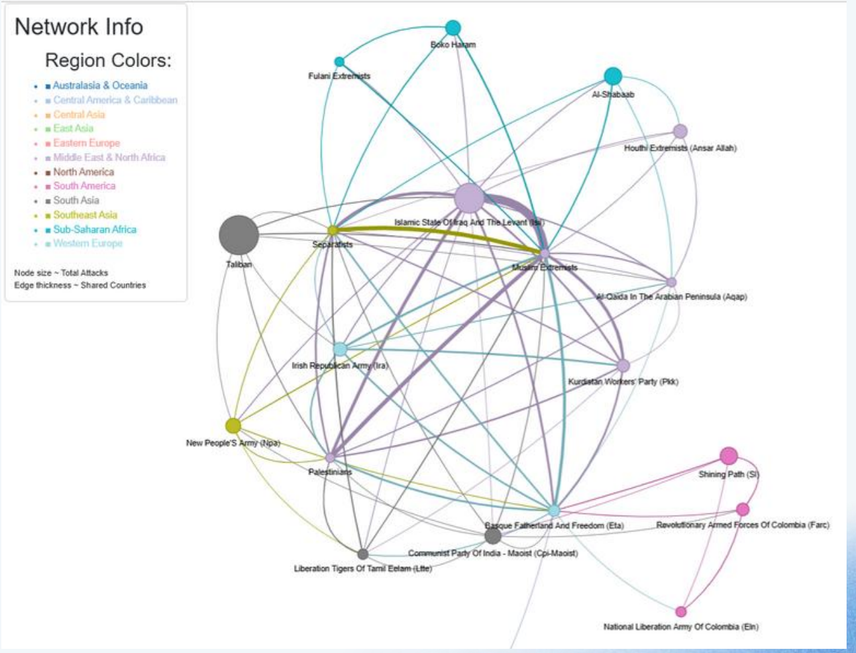


ADVANCED ANALYSIS 3: FORECASTING & NETWORKS

08

- We explored the data's structure in time (Forecasting) and its hidden connections (Networks).

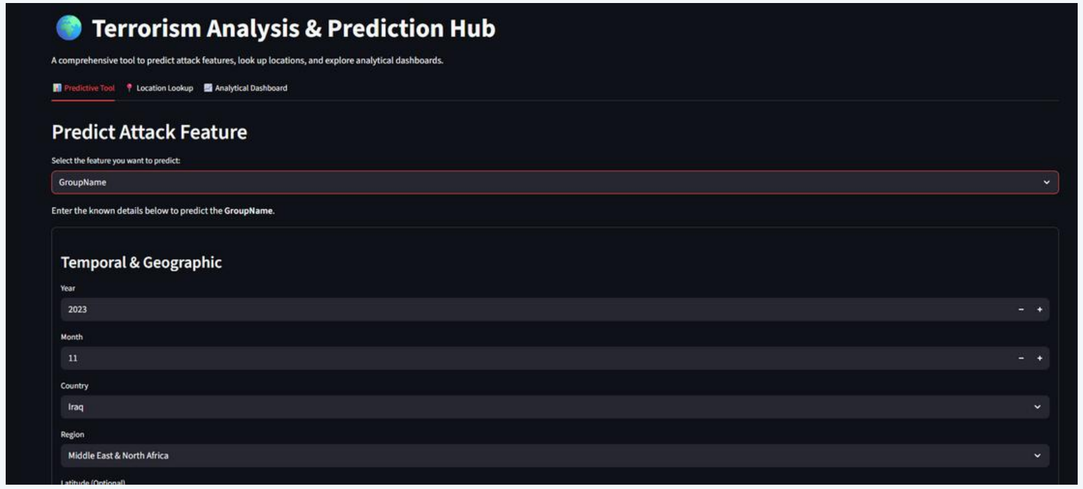


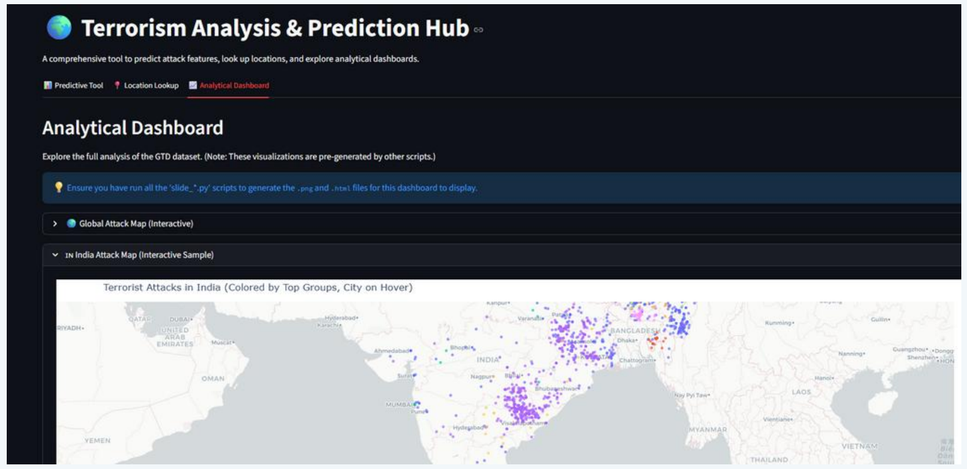


FINAL PRODUCT: THE MULTI-PREDICTION

We operationalized our work using Streamlit. This free web app is a full analytical tool:

1. It predicts the GroupName (from our Top 100 model).
2. It can also predict the WeaponType or AttackType.
3. It includes a Location Lookup tool (Geocoding).





09

10

CONCLUSION

1. Strategy Was More Important Than the Algorithm
   - Our key insight was that the raw data (3,000+ groups) was impossible to model directly, leading to initial kernel crashes. Our "Strategic Pivot" to filter for the "Top 100 Known Groups" was the key to success, balancing wide coverage with feasible computation.
2. CatBoost is the Clear Champion
   - After testing 10 algorithms, advanced boosting models were proven superior. CatBoost was the winner, achieving the highest accuracy (~93%) on the complex "Top 100" task. Its success was driven by GPU acceleration and its superior handling of categorical features (like Country and Region), which were our most important predictors.
3. A Holistic Analysis Reveals the Full Story
   - Prediction alone is not enough. The richest insights came from combining our different analytical tools:
   - Profiling (Who/What): Showed how groups behave (e.g., Taliban's top target) and how they evolve (e.g., shifting tactics over time).
   - Forecasting (When): Proved strong yearly seasonality in the data and allowed us to predict future trends.
   - Networks (How): Revealed the hidden geographic overlap between groups and the tactical links between weapons and targets.
4. The Final Product: A Full Analytical Hub
   - Our project concluded not just with a model, but with a practical, usable tool. The final Streamlit Web App successfully combines all our work into a single, powerful hub. It functions as a multi-prediction tool (for GroupName, WeaponType, etc.) and includes our Location Lookup (Geocoding) and the full Analytical Dashboard with all interactive maps and graphs.

THANK YOU

NAME :- AYUSH GUPTE
PRN :- 22070521120
7TH SEMESTER
SECTION :- ‘C’

Dataset Link : - https://www.start.umd.edu/gtd-download