UNIVERSITY OF GHANA

COLLEGE OF BASIC AND APPLIED SCIENCES

DSCD 611 PROGRAMMING FOR DATA SCIENTIST

PREDICTIVE ANALYTICS FOR IDENTIFYING FUTURE TERRORISM HOTSPOTS:

A MACHINE LEARNING APPROACH USING THE GLOBAL TERRORISM

DATABASE AND PYTHON

BY

GROUP A5

JOEL DADI-KLUTSE LEADER	JUSTICE MOSES	RICHMOND MARTEY
22424513	22425107	22424349

Predictive Analytics for Identifying Future Terrorism Hotspots: A Machine Learning Approach Using the Global Terrorism Database and Python

Joel Dadi-Klutse * Richmond Martey * Justice Moses

February 1, 2026

1 Introduction and Topic Overview

This project investigated the application of predictive analytics and machine learning techniques to forecast future terrorism hotspots by analyzing historical attack patterns from the Global Terrorism Database (GTD). The central research question examined whether spatial and temporal patterns in over 180,000 recorded terrorist incidents spanning 1970 to 2017 could be leveraged to identify geographic locations at elevated risk of future terrorist activity. The study specifically focused on developing a robust predictive model capable of identifying "hotspots", defined as geographic grid cells that would experience three or more attacks within a subsequent 90-day window. By framing terrorism prediction as a supervised classification problem, the project sought to bridge the gap between historical pattern analysis and forward-looking risk assessment, ultimately aiming to provide actionable intelligence that could inform security planning and resource allocation decisions.

2 Background and Prior Knowledge

Terrorism exhibits well-documented patterns of spatial and temporal clustering, where one incident significantly increases the likelihood of subsequent nearby attacks (Braithwaite2012, Townsley et al., 2008). This spatial autocorrelation has been explored using regression models, spatial analysis, and classification algorithms (Mohler et al., 2013). However, critical challenges limit practical application. The GTD dataset, maintained by START at the University of Maryland, contains substantial data quality issues including missing coordinates, uncertain dates, and inconsistent casualty reporting (LaFree Dugan, 2007). Most prior studies inadequately address class imbalance and lack rigorous temporal validation strategies that prevent information leakage—using future information to predict past events—leading to inflated metrics that fail to generalize (Bergmeir Benítez, 2012; Tashman, 2000).

The choice of this dataset was strongly influenced by alarming increases in terrorist activities across sub-Saharan Africa's Sahel region (Byrne, 2020; Nsaibia Weiss, 2020). Countries including Mali, Burkina Faso, Niger, and Nigeria have experienced escalating violence from extremist groups, raising concerns about spillover to historically stable coastal West African nations (International Crisis Group, 2020).

3 Relevance and Importance

Accurately predicting terrorism hotspots carries profound implications for security policy, humanitarian response, and strategic resource allocation (Bowers et al., 2004; Perry Hasisi, 2015). Predictive models enable proactive responses, potentially saving lives and reducing societal disruption (Sherman et al., 1989). Beyond security, this demonstrates how analytics can be applied to spatial-temporal risk forecasting including epidemic prediction and crime prevention (Chainey et al., 2008). Of particular interest was understanding terrorist activities near Ghana's borders and identifying possible hotzones within Ghana. As a stable nation in an increasingly volatile region, Ghana faces growing security concerns as violence in Burkina Faso—sharing a 602-kilometer northern border—has intensified dramatically since 2015 (Nsaibia Weiss, 2020). Multiple attacks have occurred within 20-30 kilometers of Ghana's border, with documented cross-border militant movements (Aning Abdallah, 2013). This project sought evidence-based assessment of emerging risks in Ghana's northern regions to inform targeted interventions including enhanced border surveillance and community engagement programs (Lacher, 2012; Aning Pokoo, 2014).

4 Data Description and Preparation

The analysis utilized the Global Terrorism Database containing 181,691 recorded terrorist incidents across 135 variables spanning nearly five decades (National Consortium for the Study of Terrorism and Responses to Terrorism, 2018). The dataset encompassed temporal markers, geographic coordinates, attack characteristics, perpetrator information, and outcome measures including fatalities and injuries. The raw dataset presented numerous data quality challenges requiring extensive preprocessing. After systematic cleaning, the final working dataset comprised 177,133 incidents across 33 carefully validated variables. The data cleaning pipeline addressed five critical dimensions following best practices (García et al., 2015): temporal validity involved handling uncertain dates; geospatial cleaning removed 4,558 records with missing coordinates while flagging 8,534 low-precision records; casualty data required imputation and creation of severity categories; categorical variables needed text standardization; and binary indicators required validation. The cleaned dataset covered 204 countries with 911,142 total casualties. Memory optimization reduced the dataset from 600 MB to 126 MB while preserving analytical integrity (McKinney, 2010). Click here to access the GitHub repository.

5 Methodology and Technical Approach

The project employed a comprehensive machine learning pipeline implemented in Python using scikit-learn, imbalanced-learn, pandas, and visualization libraries (Pedregosa et al., 2011; Lemaître et al., 2017). The analytical workflow began with extensive exploratory data analysis revealing critical temporal trends, geographic concentrations, and attack patterns. Feature engineering represented a crucial phase where domain knowledge guided creation of predictive variables across five categories (Domingos, 2012): temporal features captured yearly, quarterly, and seasonal patterns; geographic features discretized the world into degree-by-degree latitude-longitude grid cells; attack history features quantified recent activity through rolling windows calculating attacks and casualties over the preceding 30, 90, 180, and 365 days within each grid cell, inspired by research on near-repeat victimization (Johnson et al., 2007); attack type features created binary indicators for specific tactics; and perpetrator features aggregated group-level statistics (LaFree et al., 2010).

The critical methodological innovation involved implementing strict temporal validation that prevented information leakage (Bergmeir Benítez, 2012). Rather than random train-test splits, the project partitioned data chronologically: training on 1970-2012 data (60.5%), validating on 2013-2014 data (14.2%), and testing on held-out 2015-2017 data (20.9%), following best practices for time series forecasting. The percentages do not sum to 100% because observations near temporal boundaries were deliberately excluded so that hotspot labels, defined using 90-day future observation window, could be computed without information leakage between datasets. The construction of the target variable incorporated the availability logic of the label, ensuring that observations could only be classified as hotspots if sufficient future data existed to validate that designation, mimicking real-world forecast conditions. To address severe class imbalance (71% hotspots in training data), the project implemented SMOTE for tree-based models while using class weighting for logistic regression. Three model architectures were evaluated: logistic regression with L2 regularization and balanced class weights (Hosmer et al., 2013); random forest with 200 trees and maximum depth of 14 (Breiman, 2001); and gradient boosting with 200 estimators and learning rate of 0.1. All hyperparameter selections were justified through validation set performance.

6 Results and Model Performance

The validation-driven model selection process identified logistic regression as the final production model based on superior generalization performance and interpretability, despite tree-based models showing marginally higher training metrics . On the held-out 2015-2017 test period, the final model achieved precision of 0.9120, recall of 0.9543, F1-score of 0.9327, and ROC-AUC of 0.9445, demonstrating robust predictive capability while maintaining temporal integrity. When the model predicted a location would become a hotspot, it was correct 91.2% of the time, while successfully identifying 95.4% of actual hotspots. Feature importance analysis revealed that historical attack frequency dominated predictive power, with attacks in the preceding 365, 180, 90, and 30 days collectively accounting for approximately 67% of model importance (Breiman, 2001). This validated the central hypothesis that terrorism exhibits strong persistence effects where past violence predicts future violence in specific geographic cells, consistent with spatial-temporal clustering theories (Bowers Johnson, 2004; Townsley et al., 2008). Notably, medium-term windows (90-365 days) were more predictive than very short-term indicators (30 days), suggesting hotspots develop over months rather than days (Braithwaite Johnson, 2012). While random forest and gradient boosting achieved slightly higher validation scores, logistic regression demonstrated superior test generalization, underscoring the value of simpler decision boundaries when forecasting complex social phenomena subject to structural breaks (Wolpert, 1996).

7 Societal Impact and Policy Implications

This predictive framework has practical applications across security planning and humanitarian response. Early warning systems could integrate risk scores into intelligence workflows, enabling proactive resource positioning before violence escalates. Humanitarian organizations could use hotspot probabilities to adjust safety protocols and plan evacuations (Fast, 2014). The model's interpretability facilitates communication with policymakers and demonstrates feasibility of rigorous predictive analytics while respecting ethical constraints (Barocas Selbst, 2016). However, significant caveats exist: the model predicts where violence may occur but cannot explain why or prescribe interventions. It may fail to anticipate novel tactics or unprecedented disruptions (Taleb, 2007). Responsible deployment requires continuous monitoring, periodic retraining, and integration with human expertise rather than automated decision-making.

8 Team roles and tasks

The project was completed through collaborative efforts of three team members. Joel Dadi-Klutse served as Team Lead, identifying the Global Terrorism Database, exploratory data analysis, graph visualizations, and preparing the LaTeX report and PowerPoint presentation. Justice Moses focused on conducting data loading, data cleaning, feature engineering and the analytical pipeline. Richmond Martey concentrated on machine learning analysis namely temporal train/val/test splits, model training, and evaluation metrics and establishing the GitHub repository for version control. This division of labor enabled efficient completion of data preparation, model development, evaluation, and documentation while maintaining methodological rigor and reproducibility standards.

9 Reflections and Lessons Learned

Several critical insights emerged from this project. Temporal validation proved paramount—initial models using random splits achieved unrealistic performance (less than 98% accuracy) that collapsed with proper chronological holdouts, illustrating how information leakage misleads (Bergmeir Benítez, 2012). Python's comprehensive libraries enabled efficient implementation of complex analytical pipelines (Pedregosa et al., 2011). Feature engineering with domain expertise and simpler models generalized better on non-stationary data. Ethically, predicting terrorism risk requires transparent communication and human oversight to avoid unintended consequences like community stigmatization.

Appendix

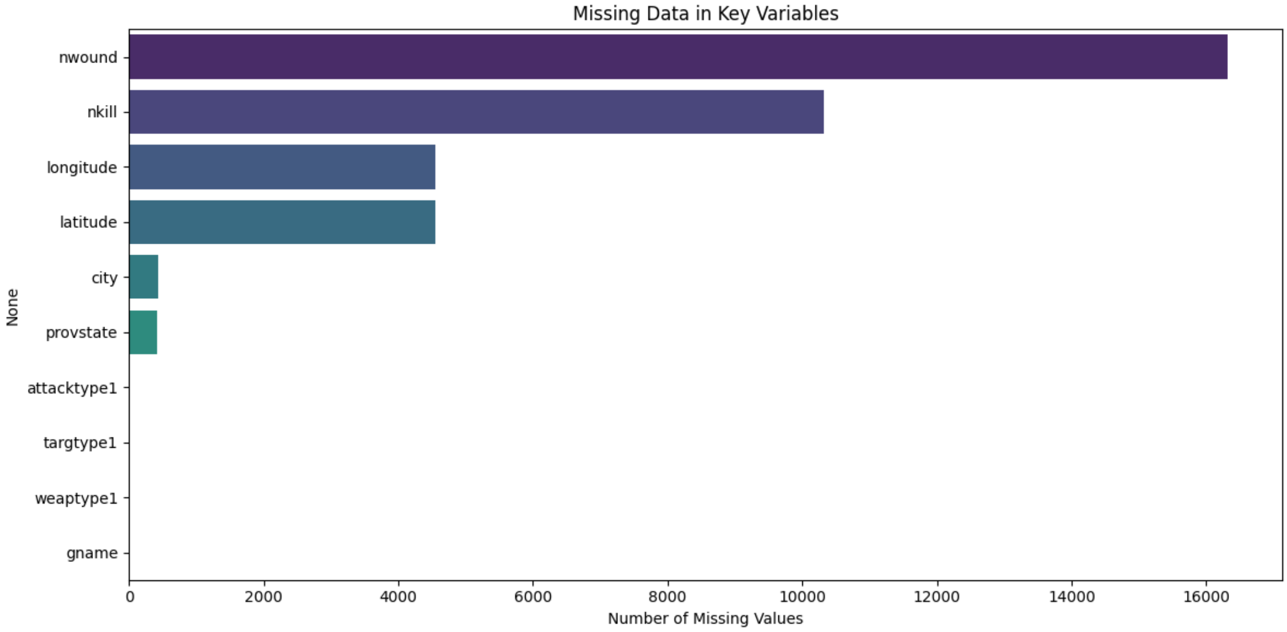

Figure 1: Missing Data in Key Variables

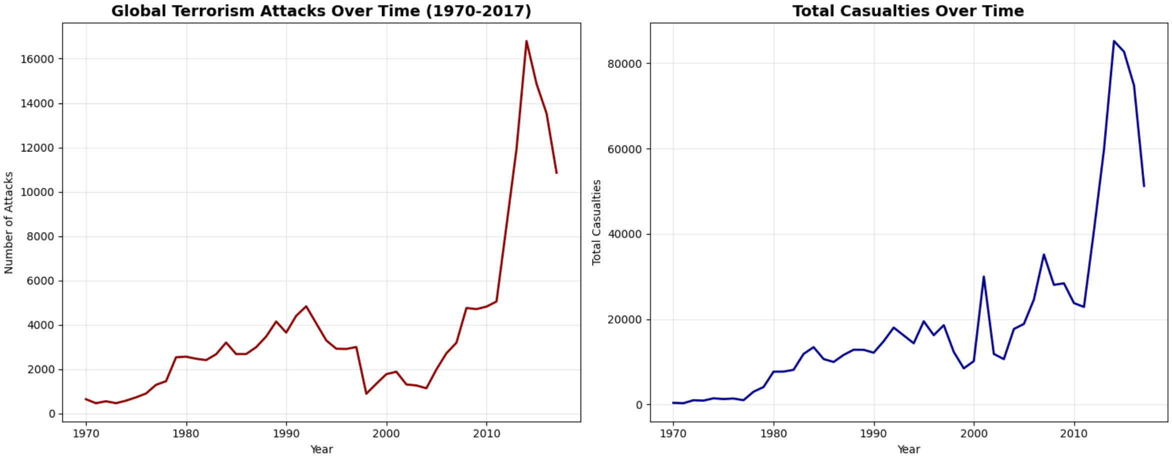

Figure 2: Terrorism Trend and Casualties over time

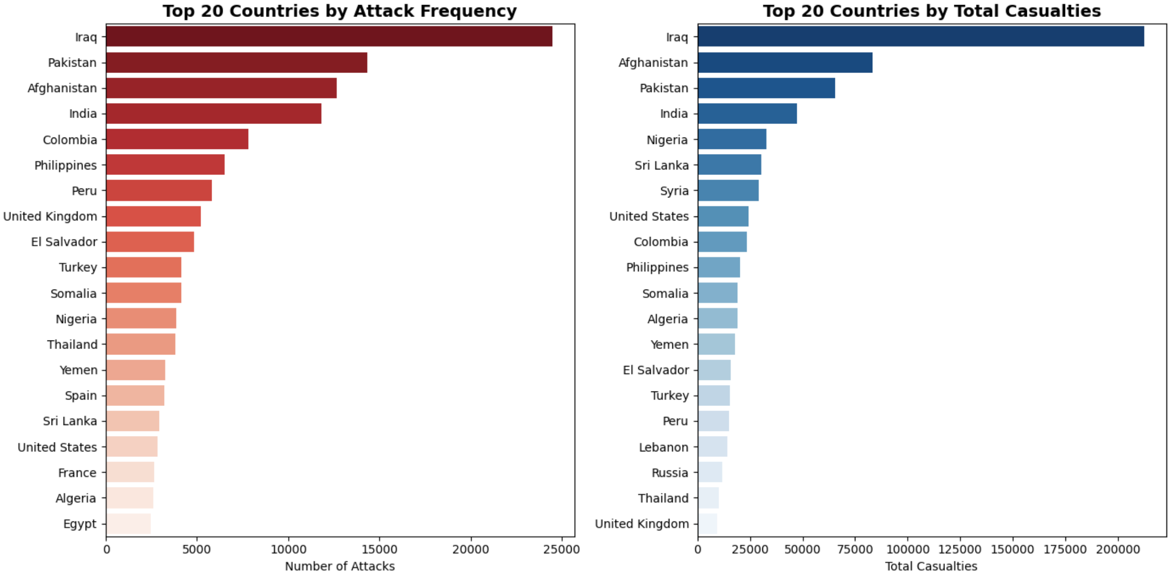

Figure 3: Country attacks

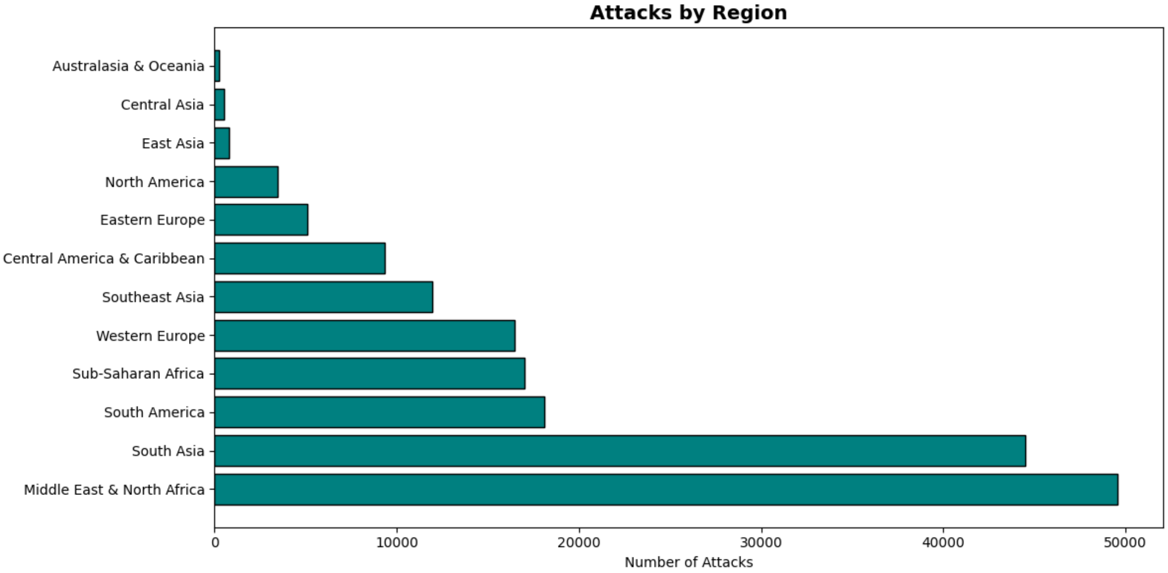

Figure 4: Attack by region

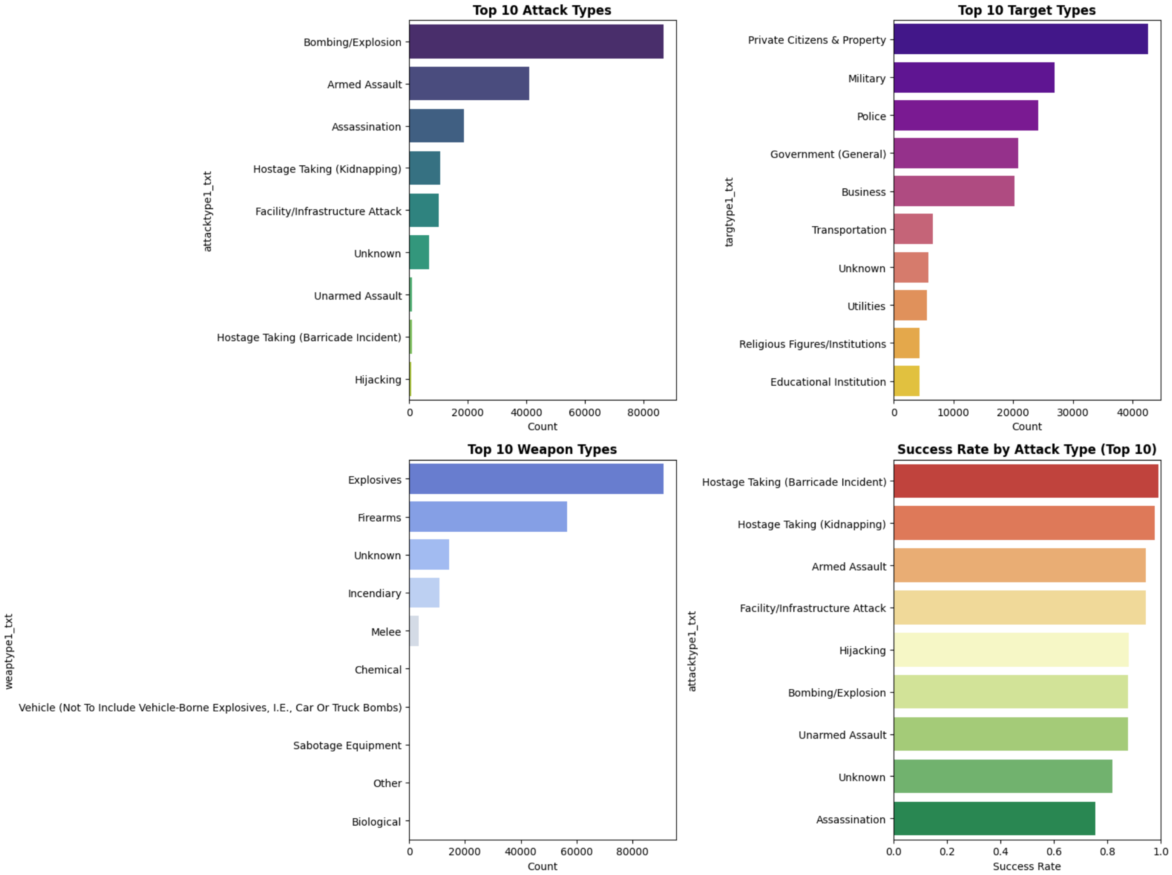

Figure 5: Types of attacks

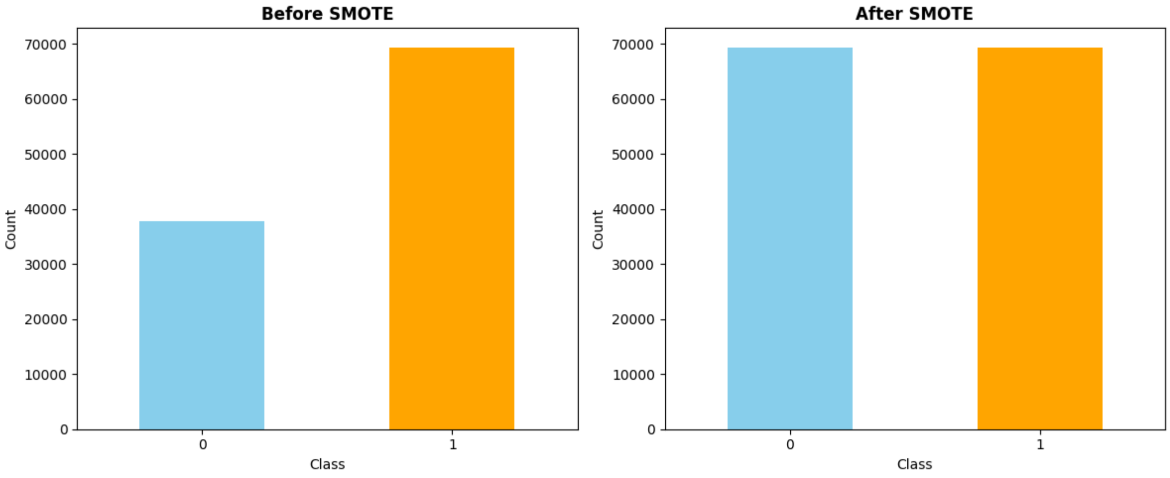

Figure 6: Data balancing Distribution

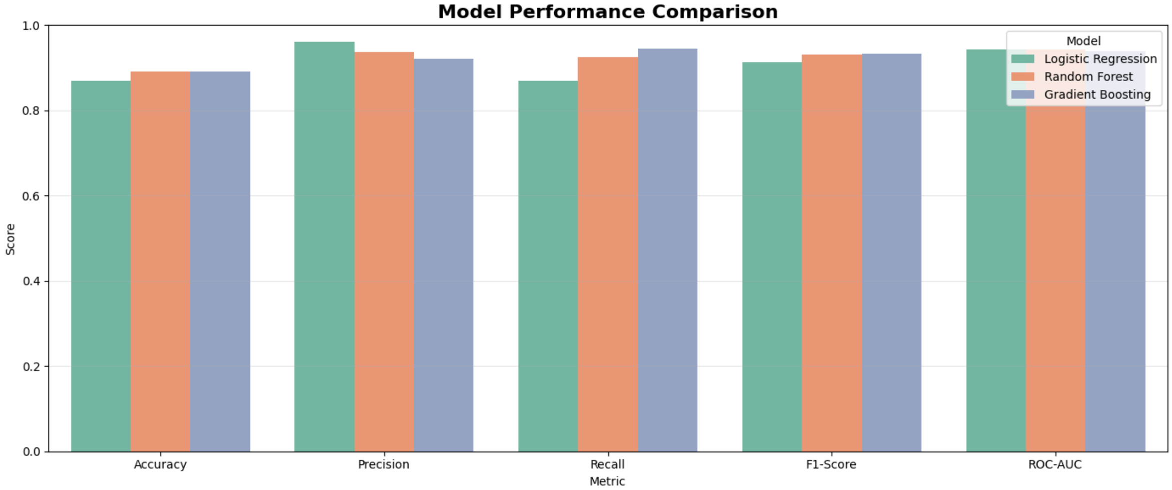

Figure 7: Model Perfromance

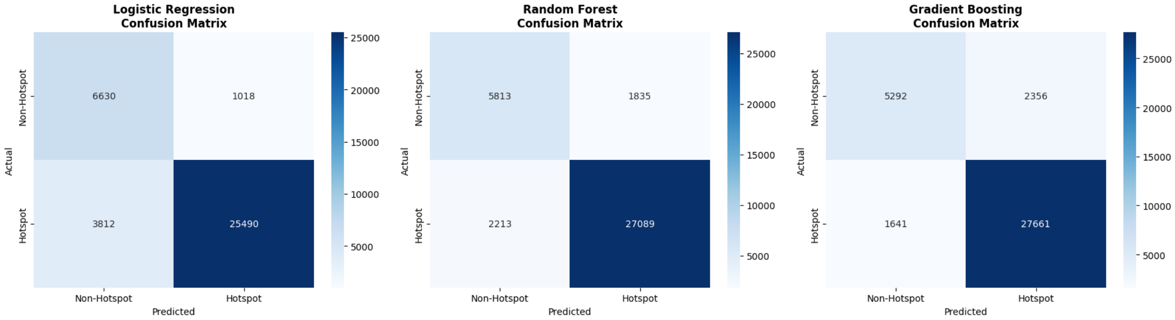

Figure 8: Confusion Matrix

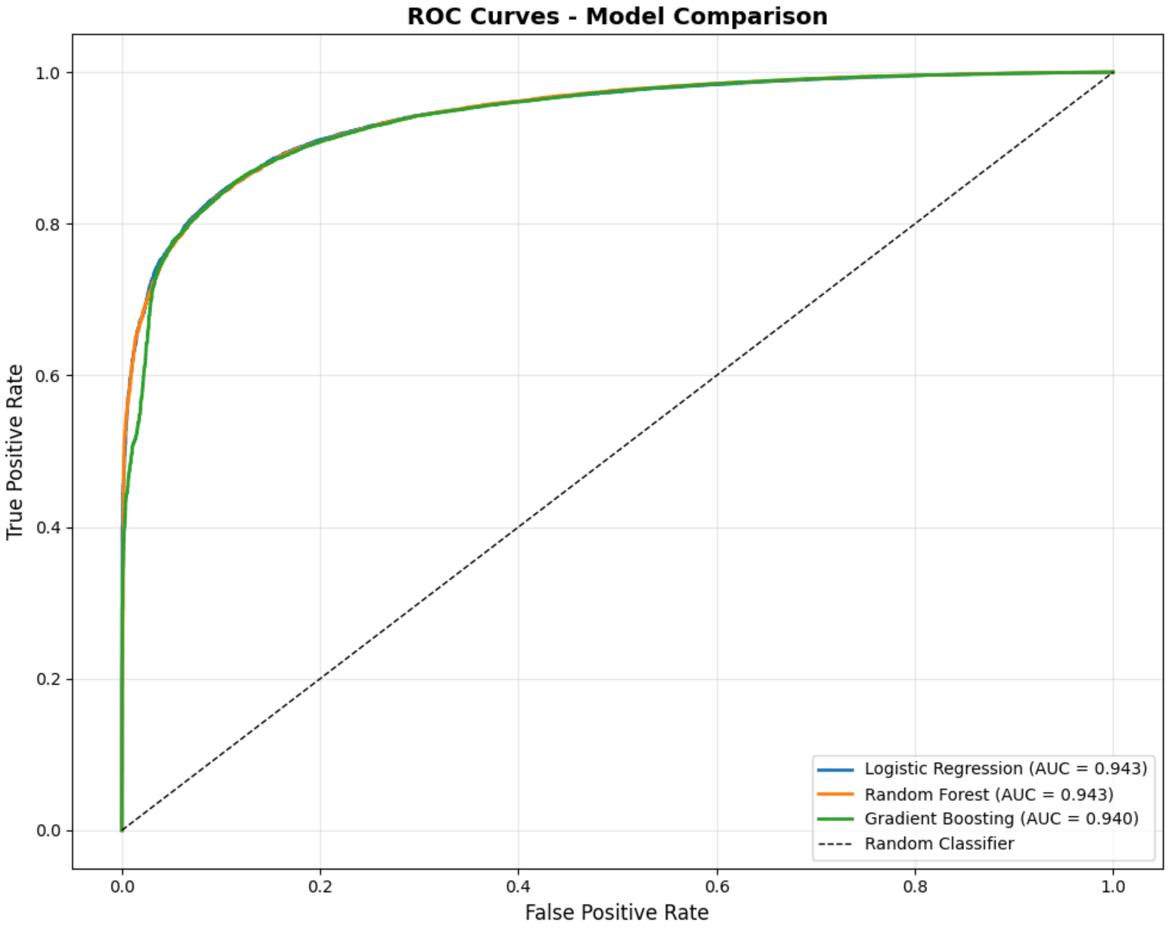

Figure 9: ROC curve

References

Aning, K., Pokoo, J. (2014). Understanding the nature and threats of drug trafficking to national and regional security in West Africa. Stability: International Journal of Security and Development, 3(1), 1-13.

Barocas, S., Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104(3), 671-732.

Bowers, K. J., Johnson, S. D., Pease, K. (2004). Prospective hot-spotting: The future of crime mapping? British Journal of Criminology, 44(5), 641-658.

Braithwaite, A., Johnson, S. D. (2012). Space–time modeling of insurgency and counterinsurgency in Iraq. Journal of Quantitative Criminology, 28(1), 31-48.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

Byrne, E. (2020). Declining security in the Sahel and the rise of Jama'at Nasr al-Islam wal Muslimin. Small Wars Insurgencies, 31(6), 1142-1171.

Domingos, P. (2012). A few useful things to know about machine learning. Communications of the ACM, 55(10), 78-87.

Fast, L. (2014). Aid in danger: The perils and promise of humanitarianism. University of Pennsylvania Press.

García, S., Luengo, J., Herrera, F. (2015). Data preprocessing in data mining. Springer.

Green, B., Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1-24.

Hosmer, D. W., Lemeshow, S., Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). John Wiley Sons.

International Crisis Group. (2020). The risk of jihadist contagion in West Africa (Africa Briefing No. 149). Brussels: International Crisis Group.

Lacher, W. (2012). Organized crime and conflict in the Sahel-Sahara region. Carnegie Endowment for International Peace.

Lemaître, G., Nogueira, F., Aridas, C. K. (2017). Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research, 18(17), 1-5.

McKinney, W. (2010). Data structures for statistical computing in Python. Proceedings of the 9th Python in Science Conference, 445, 51-56.

National Consortium for the Study of Terrorism and Responses to Terrorism (START). (2018). Global Terrorism Database [Data file]. University of Maryland. https://www.start.umd.edu/gtd

Nsaibia, H., Weiss, C. (2020). The end of the Sahelian anomaly: How the global conflict between the Islamic State and al-Qa'ida finally came to West Africa. CTC Sentinel, 13(7), 1-14.

Pearl, J., Mackenzie, D. (2018). The book of why: The new science of cause and effect. Basic Books.

Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2(1), 37-63.

Sherman, L. W., Gartin, P. R., Buerger, M. E. (1989). Hot spots of predatory crime: Routine activities and the criminology of place. Criminology, 27(1), 27-56.

Taleb, N. N. (2007). The black swan: The impact of the highly improbable. Random House.

Townsley, M., Johnson, S. D., Ratcliffe, J. H. (2008). Space time dynamics of insurgent activity in Iraq. Security Journal, 21(3), 139-146.

Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. Neural Computation, 8(7), 1341-1390.