Mechanical & Industrial Engineering
UNIVERSITY OF TORONTO

MIE1624: Alternate Course Project Report

Enhancing Transnational Intelligence: Predicting Conflict Fatalities in Africa

Insights from the 2023/24 ViEWS Prediction Competition

Group 8: Bader Al-Hilawani, Omar Al-Hilawani, Sebastian Villada, Max Beggs, Yasser Sleiman

December 2nd, 2024

1. Background and Objective

The prediction of conflict-related fatalities has become a critical area of focus worldwide, as it plays a key role in enhancing early warning systems to mitigate political violence effects. In 2017, the United Nations called for 'early-warning early-action' procedures to support armed conflict impact mitigation, as these conflicts uproot millions of lives worldwide annually [1]. Armed conflicts in Africa are especially devastating, driven by deep-rooted historical, economic, and political inequities [2]. The persistent outbreaks of violence in the region stress the pressing need for effective predictive tools to inform policymakers and stakeholders in resource allocation and the development of preventive strategies.

The ViEWS (Violence Early-Warning System) 2023/24 Prediction Competition focused on forecasting the number of fatalities in organized political violence in Africa using data from the Uppsala Conflict Data Program (UCDP). Last year's competition will form the basis for this project. The annual ViEWS competition rewards those who excel in both point predictions and uncertainty estimation.

As such this project aims to:

- Analyze political, socio-economic, and environmental contributory factors to conflicts in Africa by examining structured data sources.
- Evaluate the performance of state-of-the-art regression and classification machine learning models for predicting conflict fatalities.

2. Literature Review

Forecasting conflict fatalities presents unique challenges, particularly due to the zero-inflated and right-skewed nature of fatality distributions. Previous iterations of the ViEWS competition have demonstrated the strengths and limitations of various methodologies in this domain. Ensemble averaging methods, which combine predictions from multiple models, typically perform well in stable contexts where patterns in historical data persisted. However, these methods often struggled with sudden conflict outbreaks or escalations in the data, highlighting the need for robust evaluation metrics and models that can handle extreme variability and uncertainty. Additionally, these results emphasize the importance of interpretability, as models must not only predict accurately but also offer insights into the factors driving conflict fatalities.

To address these limitations, our team explored a range of machine learning models that balance predictive accuracy with transparency. XGBoost was one such method, a distributed gradient-boosted decision tree classifier known for its efficiency and predictive performance. XGBoost excels in handling structured data with missing values, making it suitable for datasets like UCDP, which have imbalanced target values. XGBoost's ability to model interactions between variables and reduce overfitting through regularization techniques makes it a robust choice for capturing nuanced relationships in conflict data. Moreover, XGBoost's importance ranking features provide a level of interpretability, allowing analysts to identify the most influential predictors and signals of conflict fatalities. Our aim is to create a comprehensive

framework that balances accuracy, interpretability, and flexibility. As such, this framework could address the core challenges of conflict prediction, enabling robust and actionable insights for policy and humanitarian interventions.

3. Methodology

3.1 Data Cleaning

The Uppsala Conflict Data Program (UCDP) is a long-standing initiative with nearly 40 years of data on organized violence, including conflict metrics, fatalities, and various demographic, economic, political, and environmental indicators [3]. UCDP provides open-source conflict fatality data covering 191 countries. This project focuses on monthly armed conflict fatalities in Africa from January 1990 to April 2024. As the first step of dataset preprocessing, we confirmed there were no missing values, and key identifier columns (such as country names) which were numerically encoded were changed to text and made more readable.

Next, the features in the data were examined for correlation, and it was determined that several features were highly correlated with one another. Forty-eight correlated features were identified and removed. Additionally, one feature was detected with over 98% identical values and was thus discarded as it would not provide useful information.

3.2 Exploratory Analysis

The ViEWS dataset for this project includes over 120 features, including statistics from the World Development Index (population, gender disparity), and Varieties of Democracy (level of equality, corruption). The dataset's target variables are three metrics describing the monthly state-based, non-state-based, and one-sided conflict fatalities for each country (respectively: at least one side of the conflict is a state government, no side is a state government, violence targeting unarmed civilians). The features were visualized and statistically compared with the target variables to gain some insight into the dataset.

Initial analysis focused on the spatial and temporal distribution of conflict fatalities in Africa. By visualizing the density plots of total fatalities since 1990 [Figure 1] and their geographical distribution [Figure 2], it becomes evident that fatalities are unevenly distributed across regions, with a higher concentration in East Africa, while some countries have experienced little to no fatalities during this period. In fact, the top five countries account for close to 80 percent of all the continent's fatalities.

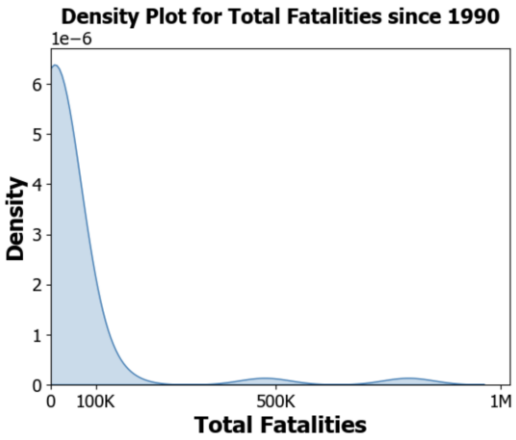

Figure 1: Distribution of total country fatalities since 1990.

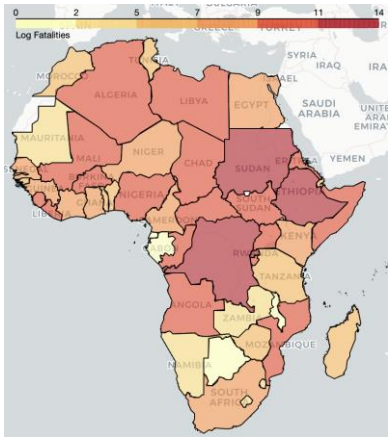

Figure 2: Logarithmic scale of total country fatalities since 1990.

Additionally, the temporal distribution of fatalities indicates similar patterns with very high concentrations of fatalities occurring at specific points in time. The temporal distribution of fatalities in Africa [Figure 3] shows extreme peaks relating to devastating civil conflicts, such as the 1994 Rwanda Genocide, and the 2020 to 2023 Ethiopia Civil Conflict. All these findings align with the expectation of zero-inflated and right-skewed fatality data.

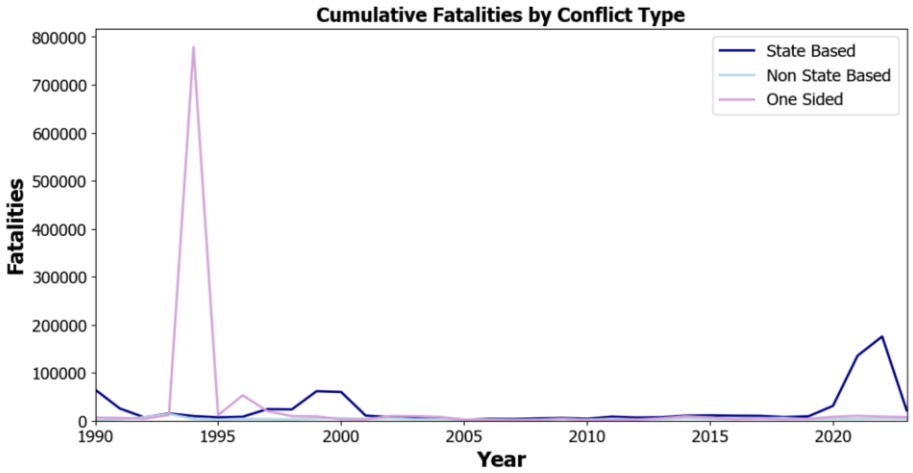

Figure 3: Cumulative continent fatalities by year and conflict type

The most relevant features in the dataset were chosen for the model using LightGBM, a gradient-boosting framework that can be used as an automated feature selection tool, calculating importance rankings by evaluating how much each feature contributes to reducing prediction error. By training a LightGBM model on the dataset, the framework provided a feature importance score for each variable, and this was used to limit the training data to the 50 most important features, improving computation times and reducing the potential negative effect from irrelevant features.

3.3 Feature Engineering

Weather has been shown to play a significant role in predicting conflicts, as seen in a paper on gang violence in Chicago which showed that warmer months had more violent crime [4]. With this in mind, a feature was created to incorporate weather into our analysis. Countries in the Northern Hemisphere, Southern Hemisphere, and those near the equator experience different seasonal patterns however, so we found an external dataset that provides the average temperature for every month for every country. While this dataset does not give a comprehensive view of weather conditions, with factors such as rainfall, wind, and humidity, it does still offer some insight that can be utilized by our models. The frequency distribution of the recorded temperatures is shown in Appendix A.

3.4 ARIMA

Insights on the dataset were also obtained through our exploration into forecasting models, specifically when ARIMA (AutoRegressive Integrated Moving Average) models were examined. These time-series forecasting models attempt to detect patterns in the variations of the target variable and are usually well-suited for capturing trends and seasonal patterns in time-series data. ARIMA models rely solely on the target variable and do not incorporate additional features, which limits its ability to capture the sometimes-sudden escalations in conflict seen in the data. To demonstrate this, we built an ARIMA model using Nigeria's data, and as Figure 4 shows, there were substantial residuals in the predictions at times, highlighting the model's limitations. While ARIMA helped us realize the need for ML techniques to better handle such data, it also provided valuable insights into the patterns of the data.

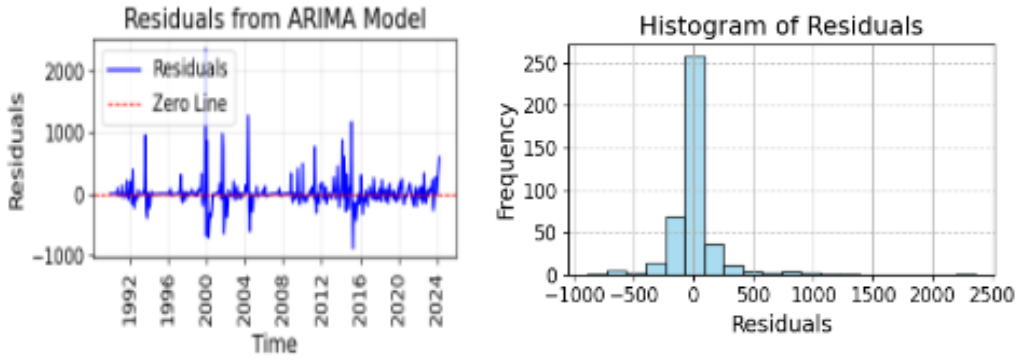

Figure 4: The residuals from the ARIMA model and their histogram.

By building the ARIMA model, we identified some trends. We observed a slight upward trend in fatalities in the 2000s, potentially due to population growth. We also calculated the average fatalities for each month and found some months had higher averages. However, after constructing confidence intervals, we noted significant overlap and concluded that seasonality (time of year) was not a major factor in predicting fatalities. Finally, using the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF), we determined that lag terms (the value of the target

variable from a previous time period) up to five months were significant, but beyond that, their influence diminished, as seen in Figure 5.

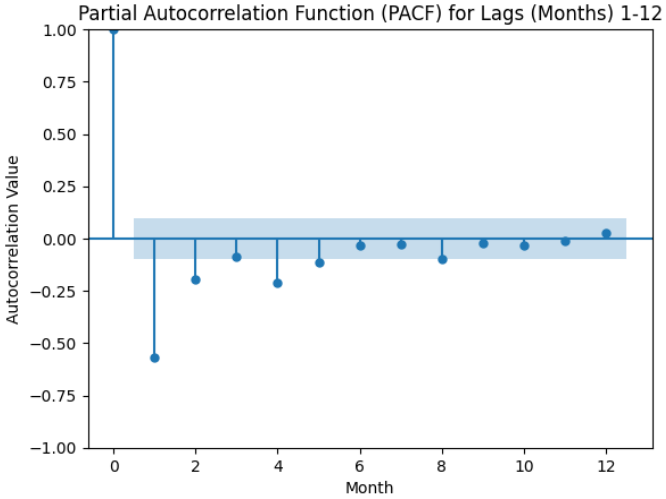

Figure 5: PACF for the ARIMA model. Notice that the ticks exceed the shaded area for the first 5 months, meaning those lag terms are significant predictors.

3.5 Final Model Implementation

The team chose to train tree-based ensemble models for forecasting conflict fatalities. These models benefit from automatic feature selection, and iteratively improve their predictions based on residual error terms. As previously discussed, ensemble tree models have demonstrated their effectiveness in forecasting tasks and, having been introduced during the course, emerged as a natural choice for this project. Consequently, the team trained an XGBoost (Extreme Gradient Boosting) model.

Preliminary model training focused on producing forecasts for 2022 fatalities using training data from 1990 to 2021. The data was divided into time-based folds using TimeSeriesSplit from Sci-Kit Learn, allowing to train on earlier data and validate using later data, as seen in Table 1.

Table 1: Example visualization of Train (Green), Validation (Yellow), Test (Red) split

2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
2011	2012	2013	2014	2015	2016	2017	2018	2019	2020

Model hyperparameters were chosen via grid search with cross-validation, these include a learning rate of 0.01, a maximum depth of 10, and early stopping was used to prevent overfitting. Root Mean Squared Error (RMSE) was calculated to assess the model's performance, and the average RMSE across all time folds was computed to provide an overall evaluation. The final model was re-trained on the entire training set, and performance was evaluated using Mean Absolute Error (MAE) and RMSE. These metrics offered insight into the model's ability to generalize and predict future data.

4. Results

To interpret the results, the actual and forecasted values of the three target variables were plotted for Africa as a whole (predictions were made at a per country level but these results were aggregated for visualization). At initial glance, while the predicted forecasts do not fully match the expected values, the results are still hopeful. For example, in the state-based conflict graph, while the model incorrectly predicted a spike in conflicts in the middle of 2022, which in real-world applications would diminish decision-makers' trust in the model, it also successfully predicted the second and third significant spikes toward the end of the year. Even if the model did not fully capture the magnitude of the spikes, the model's ability to anticipate these events remains valuable and indicates that through additional training it may eventually capture the full shape of the forecast more accurately.

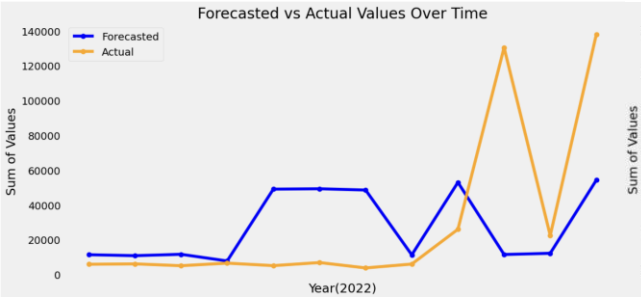

Figure 6: State Based Conflict Fatalities across Africa

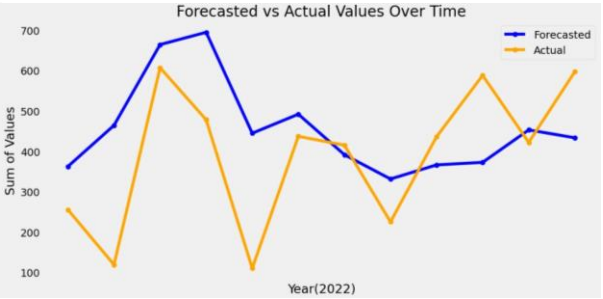

Figure 7: Non-State Based Conflict Fatalities across Africa

Figure 8: One-Sided Conflict Fatalities across Africa

5. Conclusion and Future Directions

This project demonstrated the potential of machine learning models, particularly XGBoost, in forecasting conflict-related fatalities across Africa. Moreover, it has contributed to our understanding of the dynamics of conflict fatalities by identifying correlated features and visualizing their spatial and temporal distribution.

While attempting to address the challenges posed by zero-inflated and right-skewed fatality distributions, our analysis identified significant temporal and spatial patterns, with a few regions and time periods accounting for most fatalities. While ARIMA provided insights into lag dependent trends, its limitations underscored the need for advanced models that implemented more data features in its predictions. XGBoost, combined with innovative feature engineering via lightGBM, successfully captured most of the critical aspects of the conflict trends, although occasional mispredictions highlighted areas for improvement.

Future efforts should focus on expanding datasets by searching for additional contributory factors, and in turn refining our models. This will allow us to support proactive peacekeeping efforts across Africa and propose actionable strategies to guide resource distribution to areas of high predictive uncertainty. It is also crucial to remember that the contest creators require the model to justify its predictions, providing a clear rationale for why it is making those predictions and thus helping build confidence in its decision-making process. As such, delving into the use of explainable AI to improve the interpretability of complex models would be beneficial.

References

[1] United Nations. (2024, May 21). We Must Go Above, Beyond Compliance, Fully Protect Civilians against 'Harms They Are Suffering on Our Watch', Senior Humanitarian Official Tells Security Council. *United Nations.* https://press.un.org/en/2024/sc15702.doc.htm#:~:text=Syria%20and%20Ukraine.-,In%20total%2C%20the%20United%20Nations%20alone%20recorded%20more%20than%2033%2C000,transport%20and%20patients%20were%20recorded.

[2] Aremu, J. O. (2010). Conflicts in Africa: Meaning, Causes, Impact and Solution. *African Research Review*, *4*(4), Article 4. https://doi.org/10.4314/afrrev.v4i4.69251

[3] Description of Dataset Features https://drive.google.com/file/d/1IMq4FzsREUwFba-kBLyl1KGHc02Gjvnv/view?usp=sharing

[4] Washburn, K. (2024, September 12). *Tips for contextualizing summer violence*. Association of Health Care Journalists. https://healthjournalism.org/blog/2024/08/tips-for-contextualizing-summer-violence/#:~:text=The%20University%20of%20Chicago%20Crime,incidents%2C%20the%20Crime%20Lab%20found.

Appendix A: Temperature Data

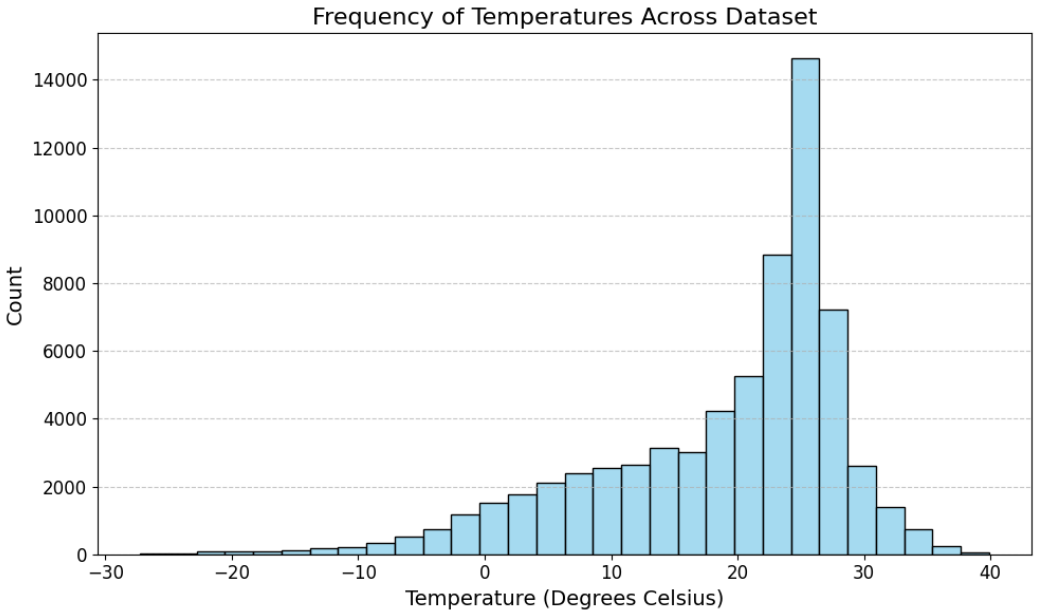

Figure A1: Average temperature frequency distribution.
Most temperatures are between 20 to 30 degrees.