Project Report: Predictive Modeling of Conflict Events in Africa using ACLED Data

Student: Ian Mbugua
Date: June 6, 2025
Course: Big Data

Abstract:

The project aim is to develop an end-to-end machine learning pipeline to probabilistically forecast conflict events at sub-national (Admin1). Test data The pipeline uses data Armed Conflict Location & Event Data Project (ACLED) between the years 2012 and 2023. It entails preprocessing of the data, training of XGBoost and baseline models, and adversive evaluation. The reason is that the initial XGBoost model, in which the spatial context was also taken into account, achieved a PR-AUC of 0.8363 and ROC-AUC of 0.8809 that is high and indicates the good predictive ability. The number of violent events and the diversity of actors-are notable predictive variables-long-term rolling averages of the number of violent events and the diversity of actors-as the literature on conflict momentum would have one anticipate. The machine learning misconception is a promising conflict early warning and the research is amongst the contributions to the emergent field of predictive conflict analytics (Hegre et al., 2019).

1. Introduction

Africa has been subjected to diversity and complicated trends of political violence. The fact that the threat of conflict needs to be recognized long before it actually happens is of utmost necessity so that the governments, humanitarian organizations and peacekeeping forces may step in time. As a part of this project, we have used a machine learning algorithm to augment the ACLED data to create a model of prediction of the monthly probability of conflict at Admin1 level. The project incorporates the existing frameworks, such as ViEWS (Hegre et al., 2019), which concern the historical pattern of warfare, engineering attributes, and the predictive model scoring.

2. Data and Methodology

2.1. Data Source and Scope

- **Data:** Armed Conflict Location & Event Data Project (ACLED).

In this project we used the ACLED dataset, which is a disaggregated data collection, analysis, and crisis mapping project that documents dates, actors, locations, fatalities, and type of all reported political violence and protest events occurring around the world

- **Geographic Scope:** African continent.
- **Time Period:** January 1, 2012, to December 31, 2023.
- **Unit of Analysis:** Country-Admin1-Month.

2.2. Pipeline Overview

The pipeline of the project is structured and organized by *acled_complete_pipeline.py* that comprised data acquisition and EDA, temporal feature engineering, spatial feature engineering, training of primary and baseline models, and finally prediction visualization and reporting.

2.3. Feature Engineering (acled_feature_engineering.py)

Since conflict dynamics are likely to be path-dependent, a large variety of time characteristics was constructed.

- **Aggregation & Base Metrics:** The data has been summarized to the country-admin1-month level and the following measures have been computed: the count of all and violent events (Battles, Riots, Explosions/Remote Violence, Violence Against Civilians), the count of fatalities, the count of different actors (based on actor1), the count of different actor types (based on inter1) and the diversity of events and sub-events.
- **Temporal Features:**
  - **Lagged Features:** One, two, three, six, and twelve-month values were added to reflect recent history.
  - **Rolling Window Features:** 3, 6, and 12-month rolling means and sums were computed to pick up the trends and filter out the short-term variations.
  - **Trend Features:** Current -lagged value differences (1-month, 3-month).
  - **days_since_last_violent_event:** Peace duration**.**

2.4. Feature Vector and Target Value Determinations

- **Feature Vector (X):** Held all the temporal engineered and lastly concatenated spatial features per country-admin1-month. Aggregate raw metrics of the same month leading up to it were excluded to prevent data leakage.
- **Target Value Changing_Towards (y):** Target variable conflict_occurs was constructed. (\text{int}) levels of conflict_compute at feature month t, whether violent_events_count in that region at t + 3 (prediction window of the last run) >= 1 or not.

- **Balance of Dataset**: The engineered dataset of the prediction window of 3 months (113,792 instances) was unbalanced with approximately 26.7% of positive (conflict_occurs = 1) instances and 73.3% negative instances. This asymmetry was considered when choosing metrics.

2.5. Spatial Feature Engineering (acled_spatial_model.py)

Spatial features were included to take into consideration the spatial dependencies, so that conflict in one region may have an effect on the surrounding regions (Ward & Gleditsch, 2008).

- The pipeline utilized a shapefile (Africa_Countries.shp) to identify contiguous Admin1 neighbours.
- **Neighbour-based Features:** For each focal region, features like the average violent events, average fatalities, and density of conflict in its neighbouring Admin1 regions during the previous month were calculated.

2.6. Modelling (acled_spatial_model.py, acled_baseline_model.py)

- **Primary Model:** we trained the XGBoost Classifier using both temporal and the shapefile-derived spatial features. We found that XGBoost is well-suited for handling complex interactions and non-linearities common in conflict data.
- **Baseline Models:** For comparison, two baseline models were trained using only the temporal features:
    1. Random Forest Classifier
    2. XGBoost Classifier (Baseline version without explicit spatial spillover features)
- **Validation:** A 80/20 temporal split was used, with older data for training and the most recent 20% for testing, ensuring a realistic out-of-sample evaluation.
- **Feature Scaling:** StandardScaler was applied to features before training baseline models; XGBoost can inherently handle unscaled features.

3. Results and Evaluation

3.1. Evaluation Metrics Used

Given the class imbalance and the importance of correctly identifying conflict, the primary evaluation metrics included:

- **Precision, Recall, F1-Score:** For the positive class ("Conflict Occurs").
- **ROC AUC:** Overall model discrimination ability.
- **PR AUC (Precision-Recall Area Under Curve):** Robust for imbalanced classes, focusing on positive class performance.

3.2. Model Performance

Model	PR AUC	ROC AUC	Conflict Precision	Conflict Recall	Conflict F1-Score	Accuracy
XGBoost(Spatial Features)	0.8363	0.8809	*~0.79*	*~0.62*	*~0.70*	*~0.81*
XGBoost_Baseline	0.8278	0.8722	0.80	0.65	0.71	0.82
RandomForest_Baseline	0.8265	0.8678	0.79	0.65	0.71	0.81

(Note: Precision/Recall/F1 of Spatial Model are assumed to be comparable to its XGBoost baseline twin brother in the earlier step-by-step log for conversation; the real numbers would be in its respective classification report logs).

Discussion of Performance:
The best PR AUC (0.8363) and ROC AUC (0.8809) were obtained with the XGBoost model using spatial features in the shapefile format. This is a small yet significant gain of around 1% in PR AUC and ROC AUC over the top performing baseline (XGBoost_Baseline). This implies that the explicit modeling of spatial dependencies has extra predictive capabilities. According to Muchlinski et al. (2016) when comparing Random Forest and Logistic Regression on class-imbalanced civil war data, obtaining good results on such data is hard, so this is a positive indication.

Precision of all models on the conflict class (approximately 0.79-0.80) was high, which means that in the cases where conflict is predicted, it is probably true. The recall (about 0.62-0.65) shows that the models predict about two-thirds of the actual future conflict instances which is the area of the future improvements.

3.3. Feature Importance

- **Baseline Models:** Analysis of feature importances from the RandomForest baseline and SHAP values from the XGBoost baseline consistently highlighted the critical role of **long-term (12-month) rolling average features**. The top predictors included violent_events_count_roll_mean12/sum12, distinct_actors_count_roll_mean12, and various event diversity rolling averages. This underscores the importance of sustained historical patterns of violence and actor complexity.
- **Spatial Model:** *(Detailed feature importances are available in output_pipeline_run/charts/spatial_feature_importances.csv).* It is anticipated that while temporal features remain crucial, the included spatial features contributed to the performance uplift.

3.4. Predicted High-Risk Regions
For the prediction window targeting December 2023 (using features up to September 2023), the model identified regions in Somalia, Burkina Faso, Madagascar, Mali, Cameroon, and Nigeria as having extremely high conflict probabilities. These largely align with contemporaneously known areas of significant conflict activity, lending face validity to the model's outputs.

4. Challenges and Future Work

- Data Granularity & Bias: ACLED data, while extensive, is subject to reporting variations.
- Spatial Feature Refinement: Further exploration of different spatial lag structures and contiguity definitions could be beneficial.
- Improving Recall: Techniques like advanced resampling (e.g., SMOTE), adjustments to class weights in models (like scale_pos_weight in XGBoost), or tuning the prediction threshold could be explored to improve the detection of true conflict events.
- Hyperparameter Optimization: Systematic tuning of XGBoost and other models.
- Integration of External Data: Incorporating socio-economic, political governance, or environmental stress data could enhance predictive power.

5. Conclusion

It is based on the project that end-to-end machine learning pipeline predicting conflict in Africa was developed and deployed successfully. Using extensive temporal and spatial features, the XGBoost model showed good predictive ability (PR AUC 0.8363, ROC AUC 0.8809) and surpassed non-spatial baselines. These results highlight the importance of long run historical conflict pattern and addition of value with the consideration of space. With any luck, this will be a contribution to the emerging discipline of computational conflict science and will provide a solid, expandable framework on which future research can be based and which might be useful in early warning systems.

6. References

- Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). ACM.
- Hegre, H., Allansson, M., Alsos, M. H., Hoelscher, K., & Vestby, J. (2019). ViEWS: A political violence early-warning system. *Journal of Peace Research, 56*(2), 155-174. https://doi.org/10.1177/0022343319823860
- Muchlinski, D., Siroky, D., He, J., & Kocher, M. (2016). Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis, 24*(1), 87-103. https://doi.org/10.1093/pan/mpv024
- Raleigh, C., Linke, A., Hegre, H., & Karlsen, J. (2010). Introducing ACLED: An armed conflict location and event dataset. *Journal of Peace Research, 47*(5), 651-660. https://doi.org/10.1177/0022343310378914
- Ward, M. D., & Gleditsch, K. S. (2008). *Spatial regression models* (Vol. 155). Sage Publications.
- ACLED. (2025, February 17). *ACLED (Armed Conflict Location and Event Data)*.

  https://acleddata.com/