Project Report:

"When does sexual violence occur in conflicts and when does is not"

A data-driven analysis of Conflict Related Sexual Violence in Armed Conflicts(CRSV)

Maria Teresa Lluesma Ballesteros, Linda Huber
Berlin and Madrid
September 2024

Introduction

The 2023 UN Report on Conflict-Related Sexual Violence highlights the increasing number of conflict zones where sexual violence has been reported globally this year, including Israel, Ethiopia, Sudan, Ukraine, the Democratic Republic of Congo, and Haiti.

Acts of conflict-related sexual violence (CRSV) not only violate the physical and mental well-being of victims but also contravene international humanitarian law and human rights standards. The repercussions go beyond the immediate trauma, affecting individuals, families, communities, and entire nations for years, even decades, to come. The scholarship on CRSV points to a couple important findings. One of them is particular important for this analysis: Common explanations for wartime reflect that emphasis: Rape is an effective strategy of war, particularly of ethnic cleansing; rape is one form of atrocity and occurs alongside other atrocities; war provides the opportunity for widespread rape. sexual violence in conflict is not inevitable. Research analysis of wars over the past 45 years by Yale University's Elisabeth Wood finds that not all wars result in sexual violence being used in conflict. This is important because it means there are ways to prevent the use of sexual violence in conflict. Dara Kay Cohen of Harvard demonstrated in her research that not only is there variation between wars, but even within the same conflict, some armed groups perpetuate sexual violence on a large scale and others do not. For example, rape was widespread in the civil wars of Sierra Leone and Timor-Leste but far less common during El Salvador's civil war or on the part of the Liberation Tigers of Tamil Eelam in Sir Lanka. If wartime-rape is neither ubiquitous nor inevitable, and if the extent of sexual violence varies greatly depending on the country, conflict and especially armed groups, the question for this analysis is: 'When does sexual violence occur in conflicts and what factors make sexual violence more likely?'

1. Dataset "Sexual Violence in Armed Conflict"

For our analysis, we utilized the Sexual Violence in Armed Conflict (SVAC) dataset, from the The Peace Research Institute Oslo and the John F. Kennedy School of Government at Harvard University. The Sexual Violence in Armed Conflict (SVAC) dataset measures reports of the conflict-related sexual violence committed by armed actors during the years 1989-2021. Sexual violence is defined by the International Criminal Court (ICC)[1] as rape, sexual slavery, forced prostitution, forced pregnancy and forced sterilization/abortion.

[1] International Criminal Court, Elements of Crimes, U.N. Doc. PCNICC/2000/1/Add.2 (2000). Article 8 (2)(e).

Following Elisabeth Wood (2009), the dataset also includes sexual mutilation and sexual torture. The SVAC dataset covers conflict-related sexual violence committed by the following types of armed conflict actors: Government/state military, Pro-government militias, and Rebel/insurgent forces. Peacekeeper and civilian perpetrators are not included as actors in the dataset. Additionally, only sexual violence by armed groups against individuals outside their own organization is included. The SVAC dataset covers all conflicts active in the years 1989-2021, as defined by the UCDP/PRIO Armed Conflict Database. Data is collected for all years of active conflict (defined by 25 battle deaths or more per year), for interim years when violence drops below the 25 battle-deaths threshold, but restarts before 5 years have passed, and for five years post-conflict. The dataset also includes post-conflict observations for conflicts that ended less than 5 years prior to 1989.[2] The three main sources used to code the data—annual reports issued by the State Department, Human Rights Watch and Amnesty International. The dataset includes three types of wars and conflicts, defined as Intrastate armed conflict, which occurs between the government of a state and one or more internal opposition groups without intervention from other states; Internationalized internal armed conflict, which occurs between the government of a state and one or more internal opposition groups with intervention from other states (secondary parties) on one or both sides; and Interstate conflicts, which occurs between the governments of two states.

In total, the dataset covers more than 184 conflicts in 86 countries, offering a comprehensive look at sexual violence over three decades.

1.1 Data Structure

The SVAC dataset is organized at the level of conflict-actor-year, meaning that each entry represents a particular actor (e.g., a government force, a rebel group) in a specific conflict year. This allows for a granular analysis of how different armed actors behave across time, especially with respect to their use of sexual violence.
We have a total of 19 columns and 11911 entries (rows). Memory usage: 1.7 MB.

[2] Additional information can be found on the project website: www.sexualviolencedata.org

In the following table, you can see the general variables of the dataset:

Variable Name	Type	Description
year	int64	The year of the observation.
conflictid	int64	ID assigned to the specific conflict.
actor	object	Name of the actor involved in the conflict.
actorid	float64	ID assigned to the actor involved.
actor_type	float64	Type of actor (e.g., state, rebel, etc.).
type	int64	Type of the conflict.
incomp	int64	Indicates the incompatibility (cause) of the conflict.
region	int64	Region code where the conflict took place.
location	object	Location (country) of the conflict.
gwnoloc	object	Gleditsch & Ward numerical country code.
gwnoloc2	float64	Secondary location code.
conflictyear	int64	Indicates if the year is an active conflict year.
interm	int64	Indicates if the year is an interim year.
postc	int64	Indicates if the year is a post-conflict year.
state_prev	float64	Prevalence score from State Department.
ai_prev	float64	Prevalence score from Amnesty International.
hrw_prev	float64	Prevalence score from Human Rights Watch.
child_prev	float64	Prevalence of sexual violence against children.
form	object	Form of sexual violence (e.g., rape, slavery).

Key features of the dataset include:

- Prevalence of Sexual Violence: An ordinal variable measuring the intensity of sexual violence, ranging from 0 (no violence) to 3 (massive, systematic violence). The dataset also accounts for missing data with a code of -99, ensuring transparency when information is not available.
- Actors Involved: The dataset includes information on whether the perpetrators were government forces, rebel groups, or pro-government militias, allowing for comparisons between different types of armed groups.
- Forms of Violence: The dataset captures different forms of sexual violence, including rape, sexual slavery, forced prostitution, and sexual mutilation, among others.

In the following table are the first 20 rows of the dataset:

year	conflictid	actor	actorid	actor_type	type	incomp	region	location	gwnoloc	gwnoloc2	conflictyea	interm	postc	state_prev	ai_prev	hrw_prev	child_prev	form
1989	205	Iran	114	1	3	1	2	Iran	630	0	0	1	0	0	0	0	0	-99
1989	205	KDPI	164	3	3	1	2	Iran	630	0	0	1	0	0	0	0	0	-99
1990	205	Iran	114	1	3	1	2	Iran	630	0	1	0	0	0	0	0	0	-99
1990	205	KDPI	164	3	3	1	2	Iran	630	0	1	0	0	0	0	0	0	-99
1991	205	Iran	114	1	3	1	2	Iran	630	0	0	1	0	0	0	0	0	-99
1991	205	KDPI	164	3	3	1	2	Iran	630	0	0	1	0	0	0	0	0	-99
1992	205	Iran	114	1	3	1	2	Iran	630	0	0	1	0	0	0	0	0	-99
1992	205	KDPI	164	3	3	1	2	Iran	630	0	0	1	0	0	0	0	0	-99
1993	205	Iran	114	1	3	1	2	Iran	630	0	1	0	0	0	0	0	0	-99
1993	205	KDPI	164	3	3	1	2	Iran	630	0	1	0	0	0	0	0	0	-99
1994	205	Iran	114	1	3	1	2	Iran	630	0	0	1	0	0	0	0	0	-99
1994	205	KDPI	164	3	3	1	2	Iran	630	0	0	1	0	0	0	0	0	-99
1995	205	Iran	114	1	3	1	2	Iran	630	0	0	1	0	0	0	0	0	-99
1995	205	KDPI	164	3	3	1	2	Iran	630	0	0	1	0	0	0	0	0	-99
1996	205	Iran	114	1	3	1	2	Iran	630	0	1	0	0	0	0	0	0	-99
1996	205	KDPI	164	3	3	1	2	Iran	630	0	1	0	0	0	0	0	0	-99
1997	205	Iran	114	1	3	1	2	Iran	630	0	0	0	1	0	0	0	0	-99
1997	205	KDPI	164	3	3	1	2	Iran	630	0	0	0	1	0	0	0	0	-99
1998	205	Iran	114	1	3	1	2	Iran	630	0	0	0	1	0	0	0	0	-99
1998	205	KDPI	164	3	3	1	2	Iran	630	0	0	0	1	0	0	0	0	-99
1999	205	Iran	114	1	3	1	2	Iran	630	0	0	0	1	0	0	0	0	-99
1999	205	KDPI	164	3	3	1	2	Iran	630	0	0	0	1	0	0	0	0	-99
2000	205	Iran	114	1	3	1	2	Iran	630	0	0	0	1	0	0	0	0	-99
2000	205	KDPI	164	3	3	1	2	Iran	630	0	0	0	1	0	0	0	0	-99
2001	205	Iran	114	1	3	1	2	Iran	630	0	0	0	1	0	0	0	0	-99
2001	205	KDPI	164	3	3	1	2	Iran	630	0	0	0	1	0	0	0	0	-99
2016	205	Iran	114	1	3	1	2	Iran	630	0	1	0	0	0	0	0	0	-99
2016	205	KDPI	164	3	3	1	2	Iran	630	0	1	0	0	0	0	0	0	-99
2017	205	Iran	114	1	3	1	2	Iran	630	0	0	1	0	0	0	0	0	-99
2017	205	KDPI	164	3	3	1	2	Iran	630	0	0	1	0	0	0	0	0	-99

1.2 Relevance of the Dataset for our Project

The SVAC dataset is particularly suitable for our project because it provides a detailed, actor-specific account of sexual violence in conflicts, over a long time period and across various regions. By using this data, we can explore whether specific factors—such as the actor type, conflict duration, or geopolitical region—can help predict the likelihood of sexual violence in future conflict years.

Our aim was to use this dataset to build an algorithm that could make predictions about the likelihood of sexual violence in conflicts, leveraging the detailed data on actors, their

behavior, and the context in which they operate. This data explores a particular pattern of wartime violence, the relative presence and absence of sexual violence on the part of many armed groups. But especially the second part – the absence of sexual violence – is interesting for us: Like Elisabeth Jean Wood argues “If some groups do not engage in sexual violence, then rape is not inevitable in war as is sometimes claimed and there are stronger grounds for holding responsible those groups that do engage in sexual violence”.
Working with data requires caution. At worst, data work/ journalism can oversimplify to the point of dehumanizing the subject of the data that their work is supposed to illuminate. We do not want to make people to genderless person icons but want to understand the underlying structure that leads to individual traumatizing and emotional stories.

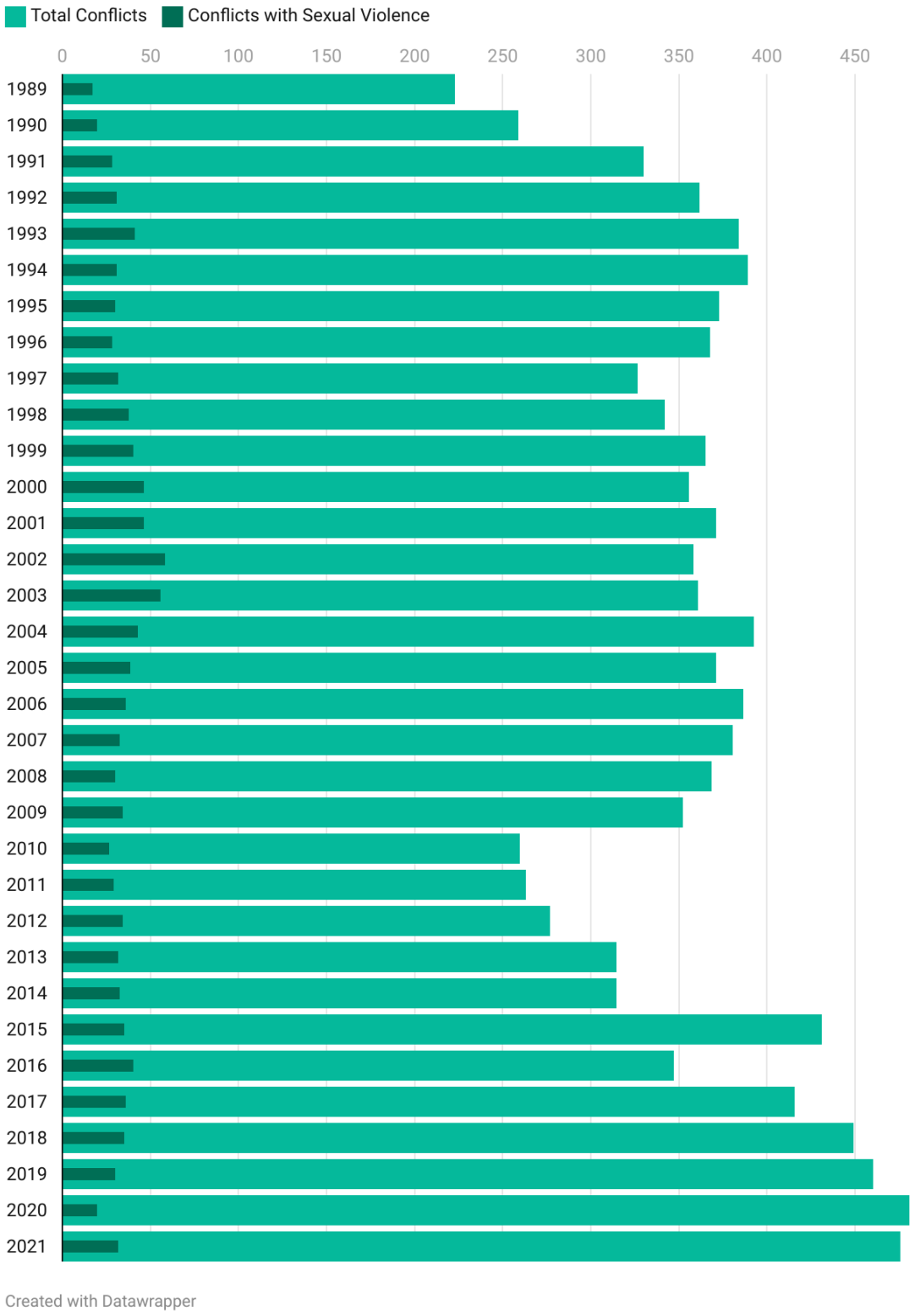
2. Data Exploration

In this section, we present key visualizations to explore and understand the patterns in our dataset before proceeding to the modeling phase. These visualizations offer insights into the temporal, geographical, and actor-specific trends of sexual violence in conflict settings.

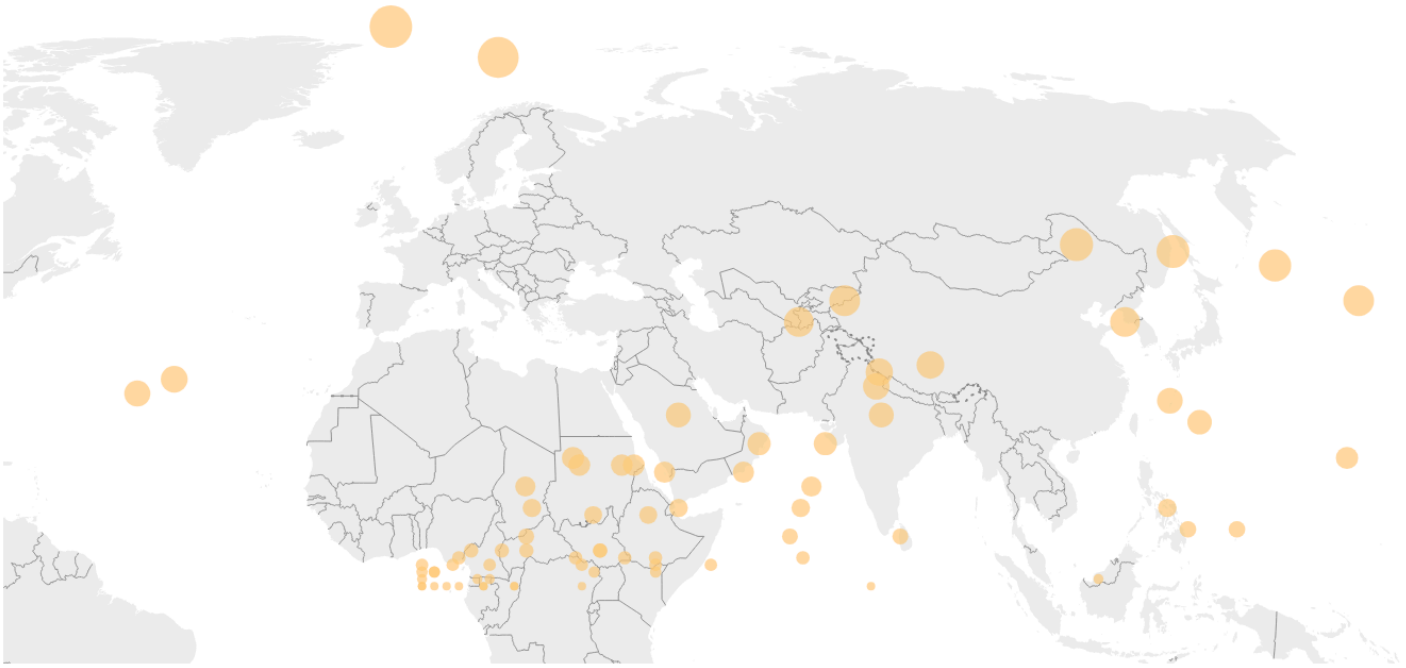
2.1 Exploratory Data Analysis (EDA)

This following visualization shows a stacked bar chart representing the total number of conflicts per year, with the proportion of conflicts that involved sexual violence highlighted.

The chart reveals that although conflicts fluctuate year to year, the proportion of conflicts involving sexual violence remains relatively stable, indicating that sexual violence is a persistent issue in conflict settings. Despite some years having more conflicts, this does not necessarily correlate with an increase or decrease in sexual violence incidents. This suggests that sexual violence may be driven by more specific factors than simply the prevalence of conflict.

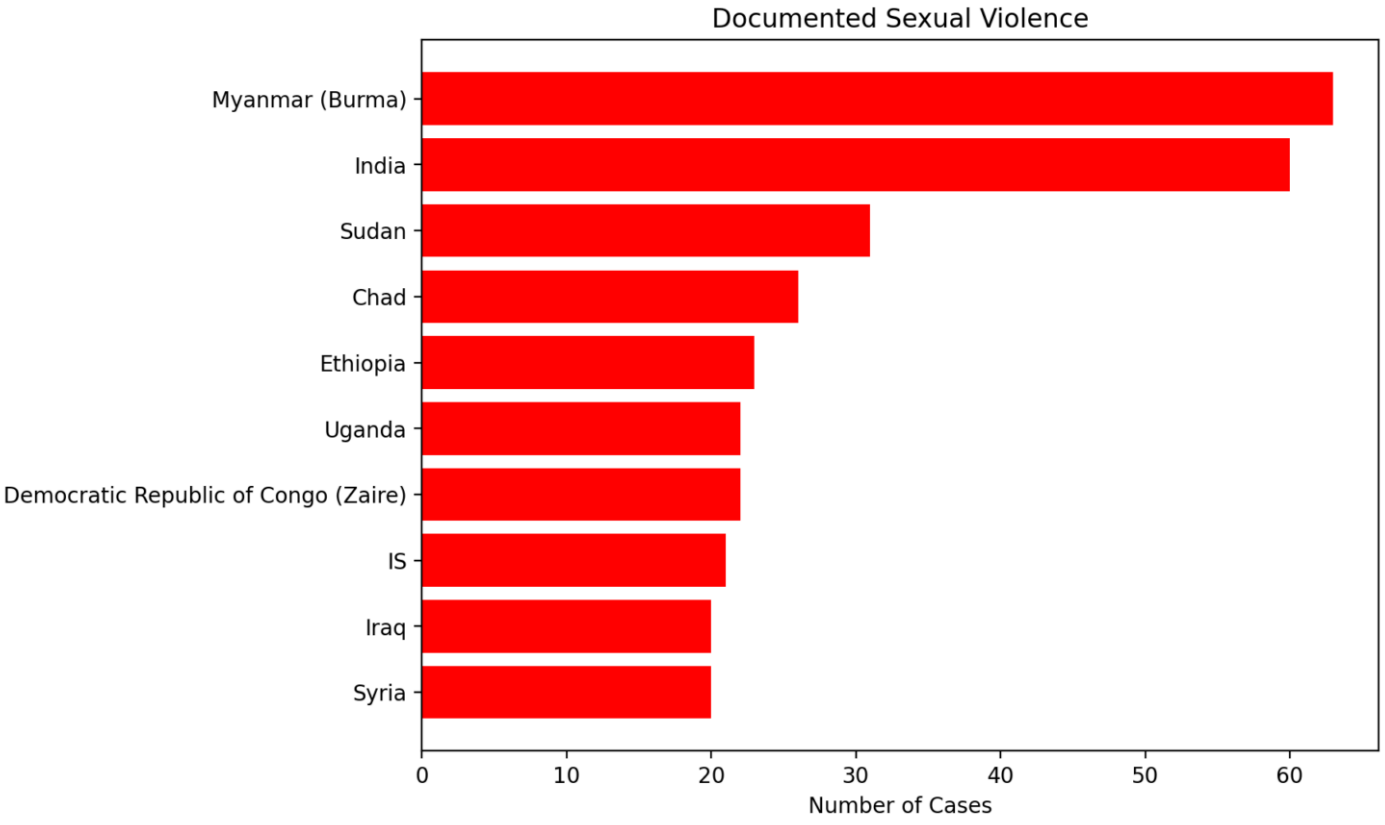


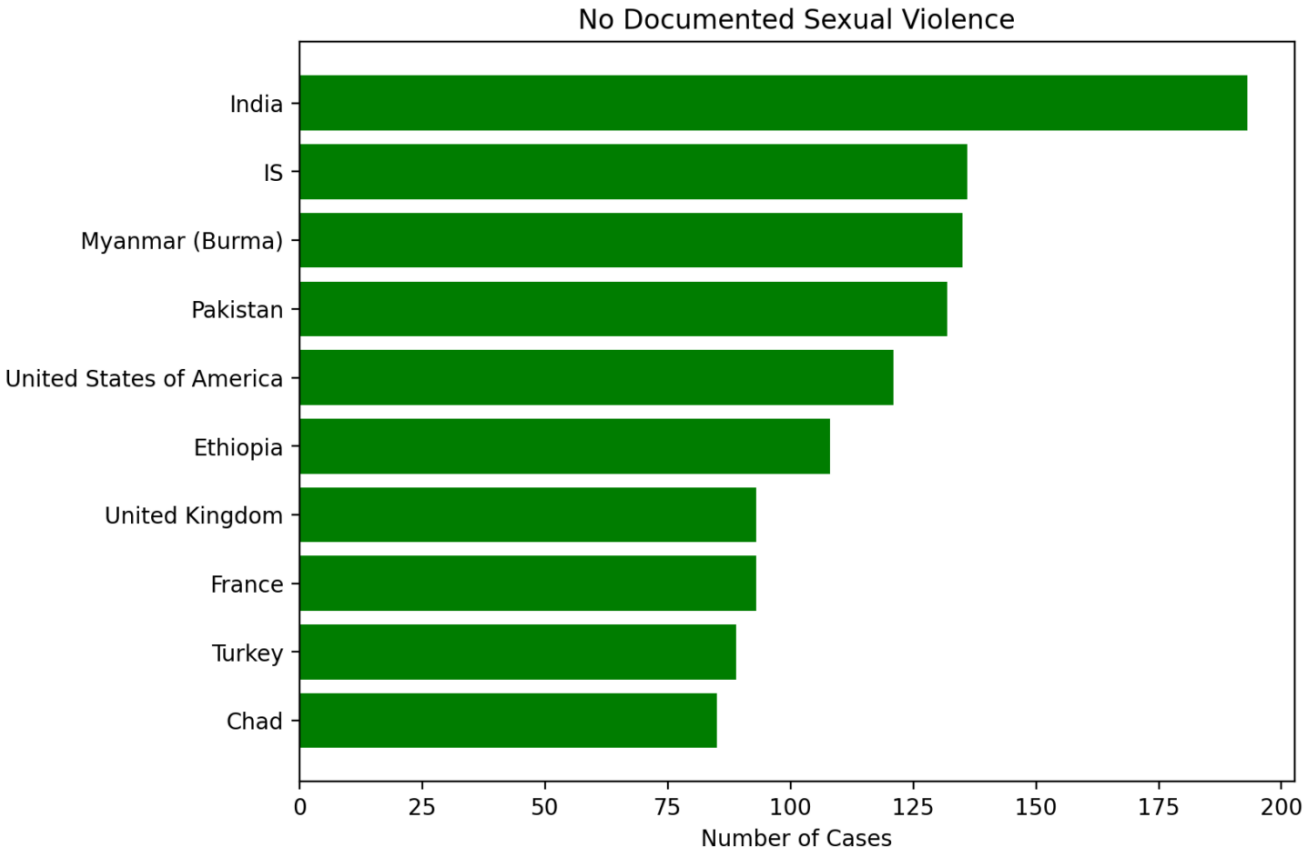
On the map visualization, we can see where sexual in armed conflicts took place between 1989 and 2021:



Source: CRSV dataset • Created with Datawrapper

The two following graphs show, respectively, the top ten actors with the most (on the left) and with the least (on the right) documented cases of sexualized violence.





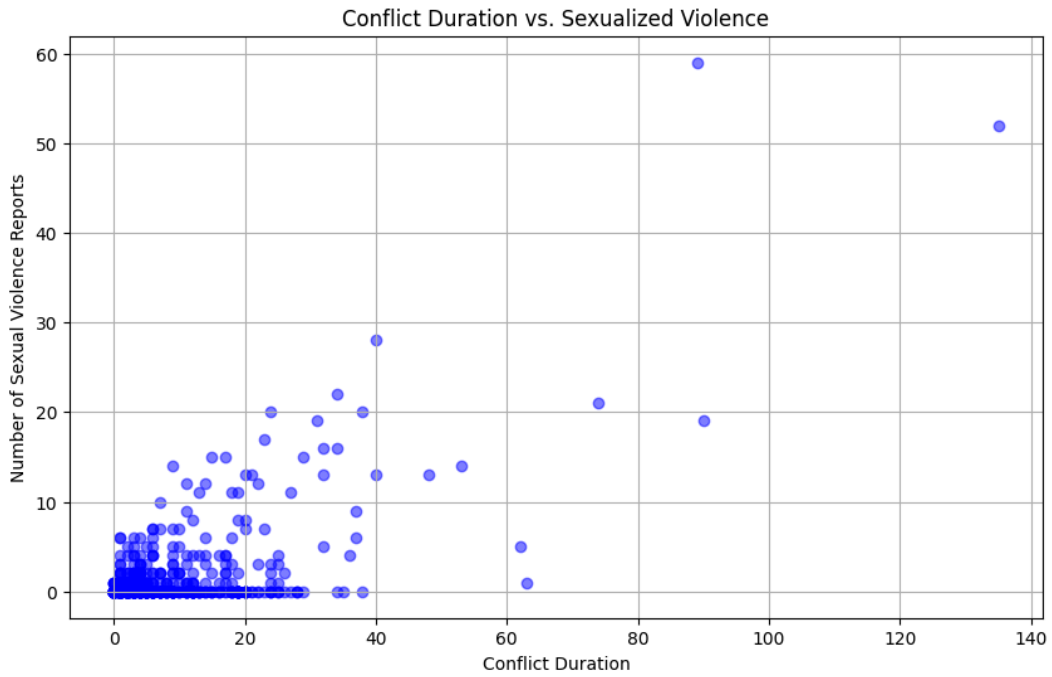
It is interesting to see that some actors (such as Myanmar or India) that appear among the top actors with documented sexual violence also appear among the top actors without documented sexual violence. This might be due to the fact that these countries probably accumulate a lot of conflicts and that not all conflicts are accompanied by sexual violence. In the lower graph, cases of violence were reported but with no mention of sexual violence. For this analysis we used a simple bar chart to show the frequency counts of each value, presented on the x-axis and the names of the actors on the y-axis, ordered from top to bottom.

2.2 Exploration: Correlation

In this section, we aim to uncover the relationships between different variables in the dataset, with a focus on the relationship between the occurrence of conflict and the likelihood of sexual violence.

2.2.1 Correlation Analysis

We began by examining the numerical correlations between variables such as conflict duration, number of actors involved, and sexual violence reports. This analysis provided an initial understanding of which variables might influence sexual violence in conflict settings.



2.2.2 Chi-square Test: Conflict and Sexual Violence

The Chi-square test was used to explore the relationship between two key categorical variables in our dataset: conflict occurrence and sexual violence occurrence. The results from this test provided important insights into the association between these variables, which influenced our approach to feature selection and model building.

The Chi-square test was performed to determine whether there was a statistically significant association between the presence of armed conflict and the occurrence of sexual violence.

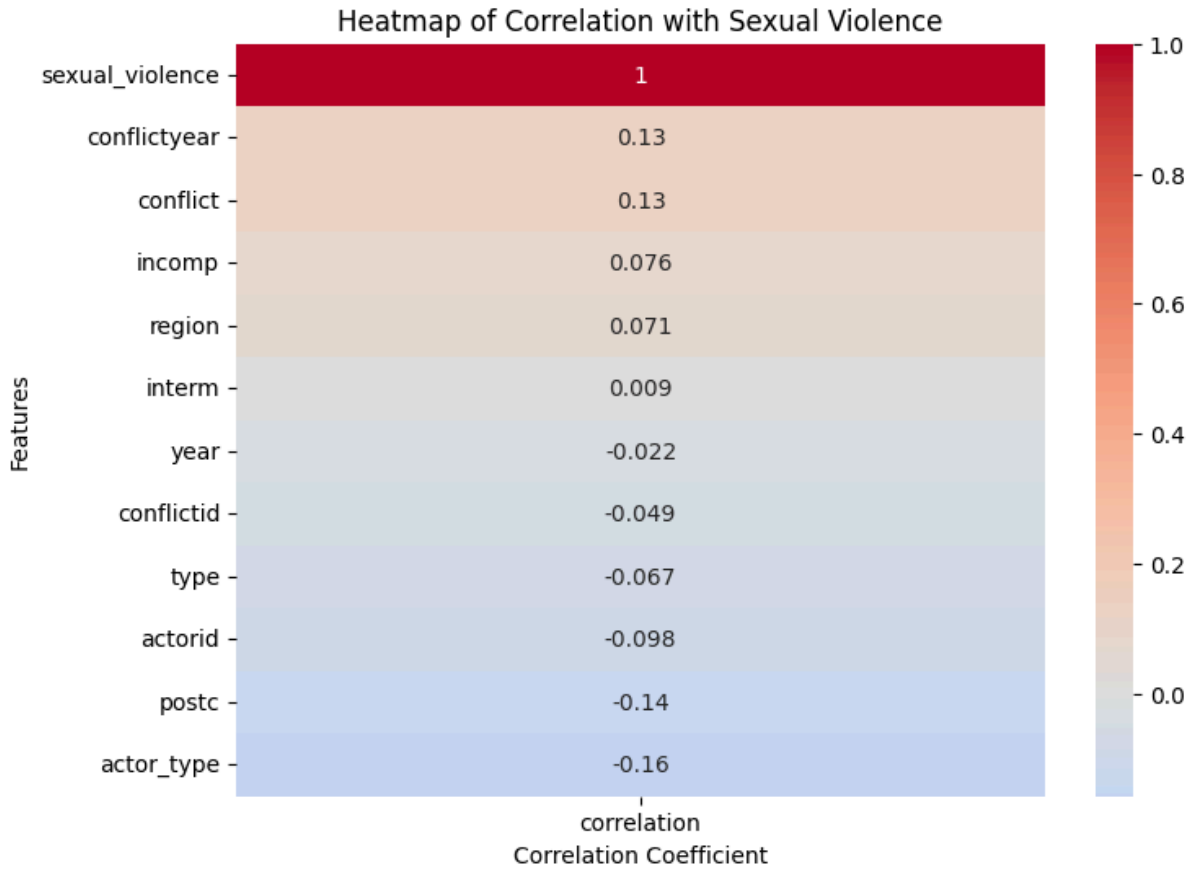
The hypothesis was that conflicts were more likely to be associated with sexual violence, and we needed to confirm this with statistical evidence before proceeding with modeling.

We created a contingency table using conflict presence and sexual violence occurrence, and then performed the Chi-square test to measure the strength of the association between these variables.

```
Chi-square statistic: 210.38710274976782, p-value: 1.1308664509522126e-47, Degrees of freedom: 1
Expected frequencies in contingency table:
 [[4610.07119469  487.92880531]
 [6160.92880531  652.07119469]]
```

A low p-value (typically < 0.05) from the Chi-square test indicated that the relationship between conflict and sexual violence was statistically significant. This suggested that the presence of a conflict had a meaningful association with the likelihood of sexual violence being reported. As a result, this variable (conflict occurrence) was deemed important and retained for further modeling.

We visualized the contingency table using a heatmap to provide a clearer representation of where the strongest associations lie:



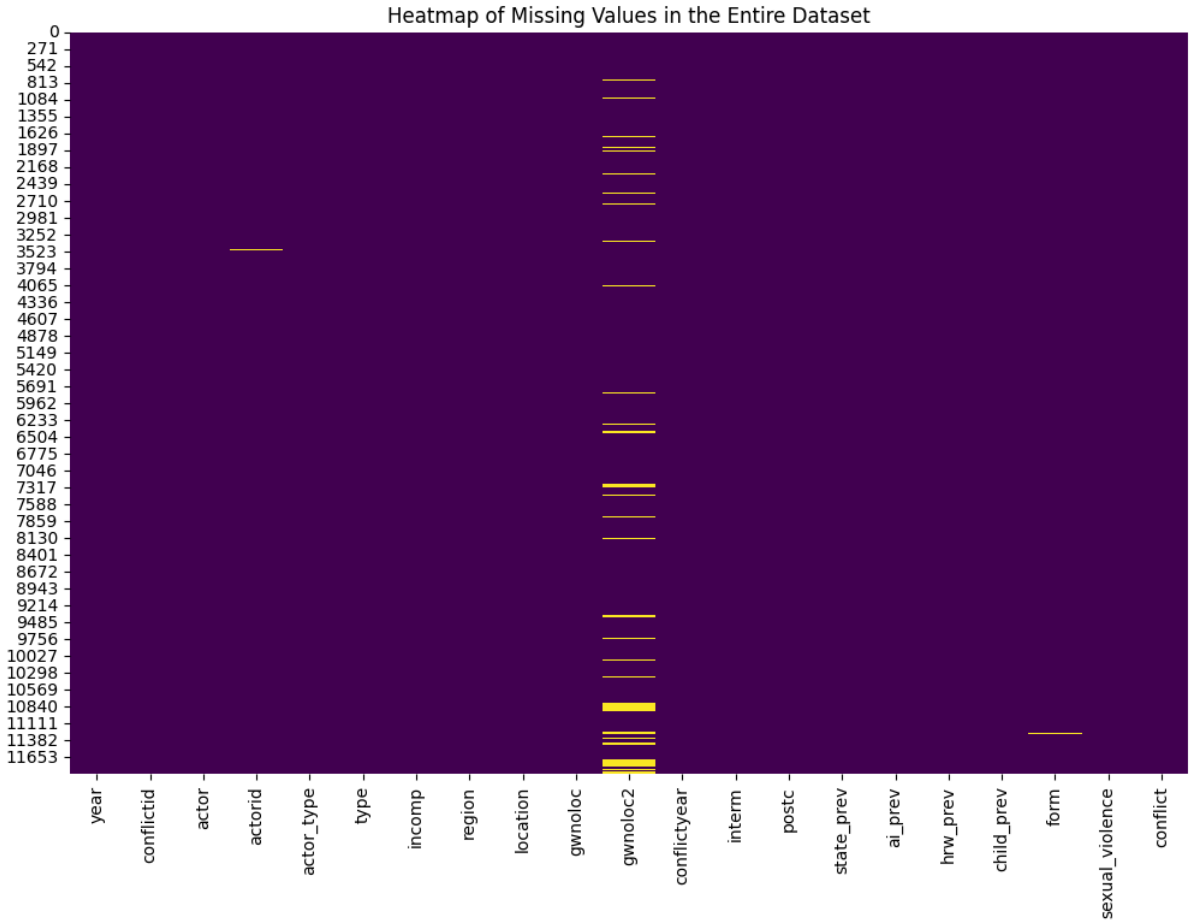
Additionally, the significant results of the Chi-square test justified the inclusion of actor type (e.g., government forces, rebel groups) in the model, as certain actor types were more prone to using sexual violence during conflicts.

3. Data and Pitfalls

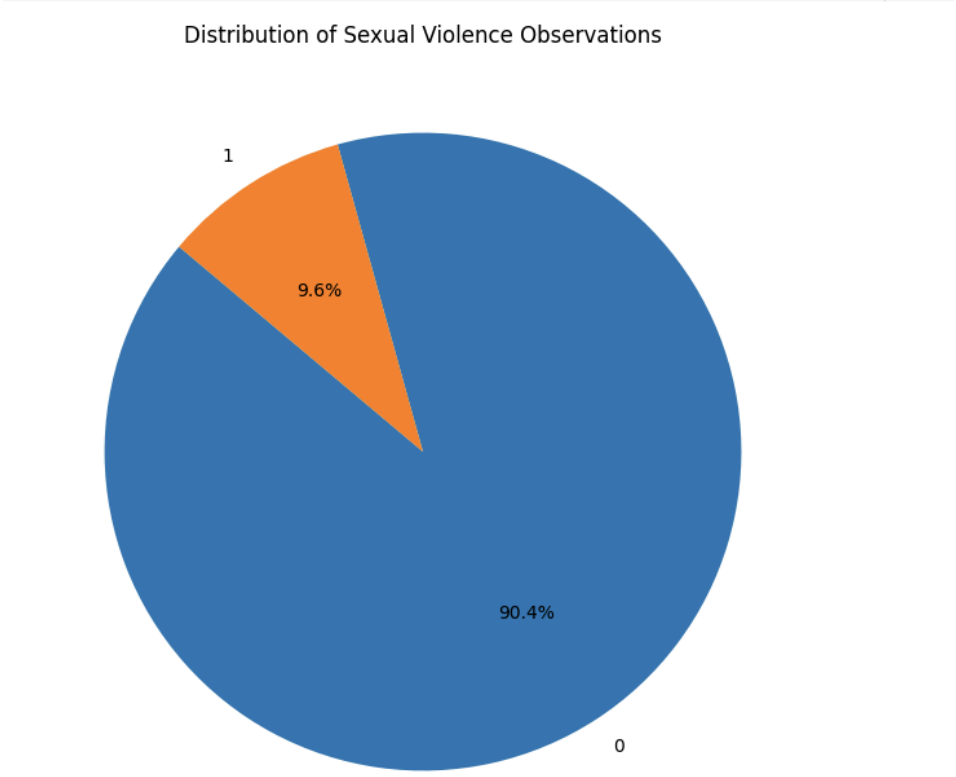
While the SVAC dataset provides invaluable insight, it presents certain challenges and limitations that needed to be addressed in order to build a robust predictive model. These challenges stem from several data quality issues and the nature of conflict-related sexual violence data. Below, we outline the key pitfalls and describe the modifications we applied to the dataset.

3.1. Missing Data

One pitfall was, as we can see, the ‘gwnoloc2’ column contains a large number of missing values, so its deletion will be considered in the data preprocessing part.



3.2 Class Imbalances



Another major pitfall was the imbalance in the target variable. The majority of conflicts in the dataset report no sexual violence (prevalence = 0), while only a minority involve isolated, numerous, or massive violence. This does not mean that there was no sexualized violence. It means that no data on this could be found. This class imbalance can lead to a model that heavily favors predicting the majority class (no violence), potentially leading to poor performance in identifying instances of sexual violence.

4. Data Preprocessing: Steps before Modeling

Before moving to the modeling phase, several key data preprocessing and exploratory steps were conducted to ensure the data was in optimal shape for training machine learning algorithms.

4.1 Handling missing data

The dataset had missing values for several key variables, particularly where data on sexual violence was not reported. This was a significant issue since machine learning models generally cannot handle missing values directly.

We replaced missing data, coded as -99 in the dataset, with NaN values for clarity. Depending on the variable, we used different strategies for handling missing data:

- Imputation: For variables with a relatively small amount of missing data, we used techniques such as mean or mode imputation. For categorical variables like actor types, the mode (most frequent value) was imputed.
- Dropping rows: In cases where certain features had a large amount of missing data and could not be imputed reliably, those features were excluded from the analysis to prevent bias in the model.

4.2 Feature Engineering

Feature engineering plays a key role in improving the predictive power of the model by transforming raw data into meaningful features. We carried out several feature engineering steps, including:

- Binary Transformation:

To simplify our modeling task, we binarized the target variable. This was done to reduce complexity and treat the problem as a binary classification. The regrouping process was as follows:

0: No reports of sexual violence (this corresponds to the original value of 0 in the dataset).

1: Any occurrence of sexual violence (this corresponds to the original values of 1, 2, or 3).

This means that we merged the categories for isolated, numerous, and massive violence (values 1, 2, 3) into one category representing the occurrence of sexual violence. This binary variable was essential for training the classification models (e.g., Random Forest and XGBoost). Since we're then trying to predict a specific class (the occurrence or non-occurrence of sexual violence), this is a typical binary classification problem.

- Aggregating Data by Year and Actor:

For actors with multiple conflict years, we created aggregated features such as total conflict duration and average levels of sexual violence. This allowed the model to capture historical behavior patterns of actors over time.

4.3 Addressing Class Imbalance

Without balancing the classes, the model would have been highly biased toward predicting the majority class, making it less likely to correctly predict instances of sexual violence. To mitigate this, we employed Random Oversampling of the minority class (conflict years with

sexual violence) to ensure that the model received sufficient training examples from both classes. By oversampling the minority class, we aimed to balance the dataset and provide the model with a more representative learning experience. This step was crucial for improving the model's ability to correctly predict the occurrence of sexual violence.

4.4 One-Hot Encoding for Categorical Variables

Several important features, such as actor type and region, were categorical variables that needed to be converted into numerical form for the machine learning algorithms to process them effectively. We applied One-Hot Encoding to transform categorical variables into binary columns. For example, the variable actor_type was transformed into multiple binary columns such as actor_type_1, actor_type_2, etc., representing each unique actor type.
Machine learning models like Random Forests and XGBoost can't process categorical variables directly, and this transformation is necessary to make those variables useful in the model training process.

4.5 Splitting the Dataset into Training and Testing Set

To evaluate the performance of the model fairly, we split the dataset into two parts:
Training Set: 70% of the data, used to train the model. Testing Set: 30% of the data, used to evaluate the model's performance. Splitting the data ensures that the model's evaluation, preventing overfitting and giving a more realistic assessment of how well the model will generalize to new data. By keeping the testing data separate, we could measure the model's ability to predict sexual violence in conflict years it hasn't seen before.

5. Modeling

The data was highly imbalanced, with far more observations for no sexual violence (class 0) than for occurrence of sexual violence (class 1). This imbalance can severely affect the performance of machine learning models, which tend to favor the majority class.

In addition to using traditional evaluation metrics (like precision, recall, and F1-score) to assess model performance, we leveraged the feature importance mechanisms provided by both Random Forest and XGBoost (eXtreme Gradient Boosting) to gain insights into which variables had the greatest impact on our predictions. These algorithms were chosen

because they are known to handle imbalanced classification tasks well, especially when combined with techniques like oversampling and class weighting.

5.1 Random Forest

We first trained a Random Forest model to predict the occurrence of sexual violence. It is a bagging algorithm that reduces overfitting by averaging multiple decision trees. It is robust to imbalanced datasets because it can handle both categorical and continuous variables and has mechanisms to deal with imbalance through techniques like class weighting and oversampling (which we implemented with RandomOverSampler) provides a built-in method to measure feature importance based on the Gini impurity or information gain. This is particularly useful for understanding which features contribute most to the model's decisions.

5.1.1 Feature Importance

The Feature Importance analysis revealed that actor type, conflict presence, and post-conflict status were among the most important predictors of sexual violence. These features significantly impacted the model's ability to correctly identify years and actors involved in sexual violence.

Certain actor types, particularly rebel groups, were identified as strong predictors of sexual violence. This aligns with conflict literature, which suggests that non-state actors may use sexual violence as a strategic tool during conflicts. The importance of post-conflict status highlights the vulnerability of populations in the immediate aftermath of conflict, where instances of sexual violence may persist or increase.

5.1.2 Model Evaluation

Before hyperparameter tuning, the Random Forest model achieved a high recall score for identifying sexual violence, though precision was lower due to the class imbalance in the dataset.

Classification Report:

	precision	recall	F1-score	support
0	0.98	0.85	0.91	3232
1	0.37	0.82	0.51	342
accuracy			0.85	3574

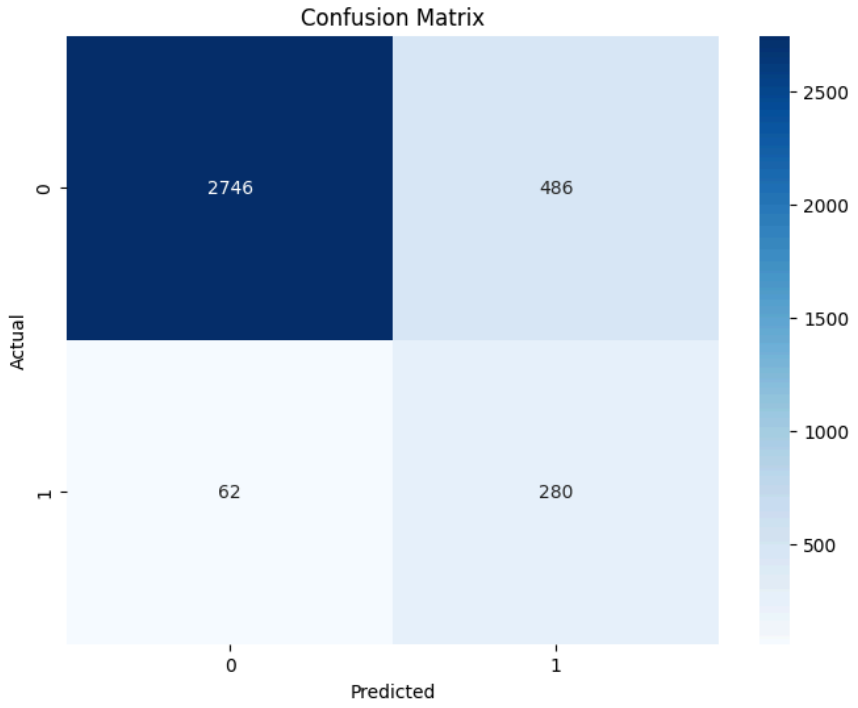
macro avg	0.67	0.83	0.71	3574
weighted avg	0.92	0.85	0.87	3574

Precision: For class 0 (no occurrence of sexual violence), the precision is very high (0.98), which means that the model is correct in most cases when it predicts 0. For class 1 (occurrence of sexual violence), the precision is lower (0.37), indicating that the model has many false positives for this class.

Recall: For class 1 is relatively high (0.82), indicating that the model recognises most actual occurrences of sexual violence. For class 0 is also high (0.85), which means that the model is also good at recognising non-occurrences.

F1 score: The F1 score for class 0 is very high (0.91), which means that the model is good at correctly predicting non-occurrences of sexual violence. The F1 score for class 1 is lower (0.50), indicating an imbalance in model performance.

The



The confusion matrix shows that the model predicted 2,748 cases of 0 correctly and 279 cases of 1 correctly. However, there are 484 false negatives (the model says 0, but it was actually 1) and 63 false positives (the model says 1, but it was actually 0).

To improve our model we try XGboost.

5.2 XGBoost

This boosting algorithm builds models sequentially and focuses on improving errors from previous models. It's particularly good for our imbalanced datasets because it allows you to weight classes or use custom loss functions to emphasize the minority class (sexual violence cases). XGBoost is often more efficient than Random Forest when dealing with complex, imbalanced data. In XGBoost, feature importance is determined by calculating the gain, which measures the contribution of each feature in improving the model's accuracy. It reflects how often a feature is selected to split the data and how much it reduces the model's error in each split.

5.2.1 Feature Importance

Similar to Random Forest, XGBoost identified actor type and conflict presence as the most influential features. Additionally, regional factors were found to play a role in predicting the likelihood of sexual violence, suggesting that geographical context affects conflict dynamics.

The prominence of actor type in XGBoost's feature importance analysis reinforces the model's focus on which groups are most likely to commit sexual violence. In particular, rebel groups (actor_type_3.0) and pro-government militias (actor_type_6.0) were among the highest contributors to the model's predictions.

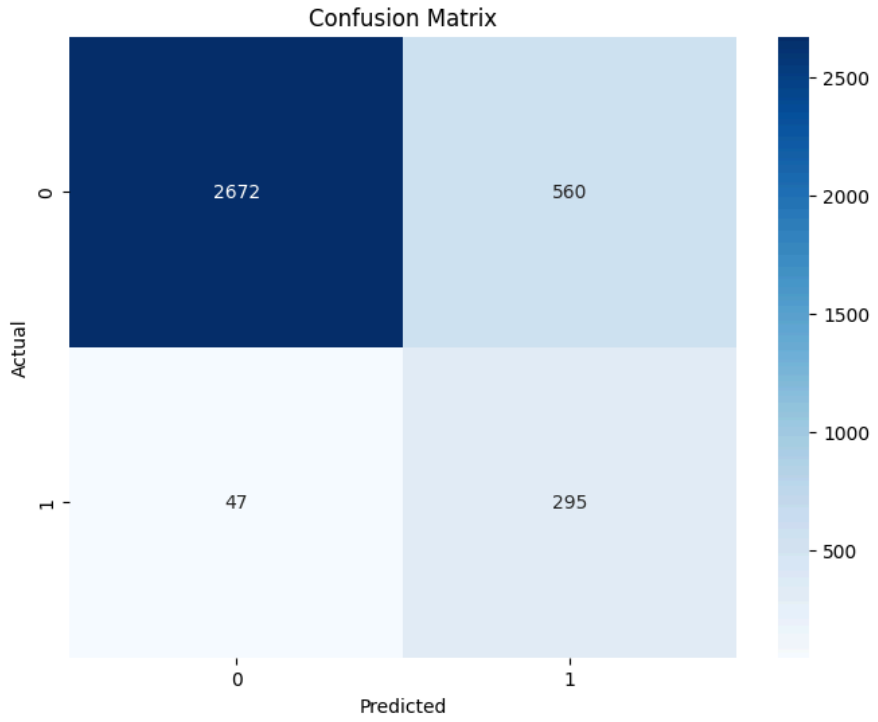
The inclusion of regional factors indicates that location-specific dynamics can influence the prevalence of sexual violence, emphasizing the need for geographically targeted interventions.

5.2.2 Model Evaluation

XGBoost performed well in both recall and precision before tuning, making it a strong candidate for hyperparameter optimization to further enhance its predictive power.

Classification Report:

	precision	recall	f1-score	support
0	0.98	0.83	0.90	3232
1	0.35	0.86	0.49	342
accuracy			0.83	3574
macro avg	0.66	0.84	0.70	3574
weighted avg	0.92	0.83	0.86	3574



We can see that while the model performed well in predicting no sexual violence (Class 0), it struggled with predicting sexual violence (Class 1). The precision for Class 1 was low, indicating many false positives. XGBoost performed slightly better than Random Forest, especially in precision and F1-Score for Class 1, but still had room for improvement.

5.3 Hyperparameter Optimization (Grid Search and Cross-Validation)

To improve the performance of both models, we employed Grid Search for hyperparameter optimization and used Stratified Cross-Validation to ensure that both classes (sexual violence and no sexual violence) were equally represented in each fold.

For Random Forest, we tuned parameters like:

- n_estimators (number of trees),
- max_depth (maximum depth of each tree),
- min_samples_split (minimum samples required to split a node),
- min_samples_leaf (minimum samples required to form a leaf).

For XGBoost, we tuned parameters such as:

- learning_rate (step size shrinkage used to prevent overfitting),
- max_depth (maximum depth of a tree),
- n_estimators (number of boosting rounds),

- subsample (fraction of training data to use for each boosting round).

Random Forest (after Grid Search):
Precision for Class 0: 0.99, Recall for Class 0: 0.88, Precision for Class 1: 0.42, Recall for Class 1: 0.85, F1-Score for Class 1: 0.57

After optimization, we saw a significant improvement in precision for Class 1 (sexual violence), meaning fewer false positives. The F1-Score also increased, reflecting a better balance between precision and recall.

XGBoost (after Grid Search):
Precision for Class 0: 0.98, Recall for Class 0: 0.89, Precision for Class 1: 0.45, Recall for Class 1: 0.87, F1-Score for Class 1: 0.61

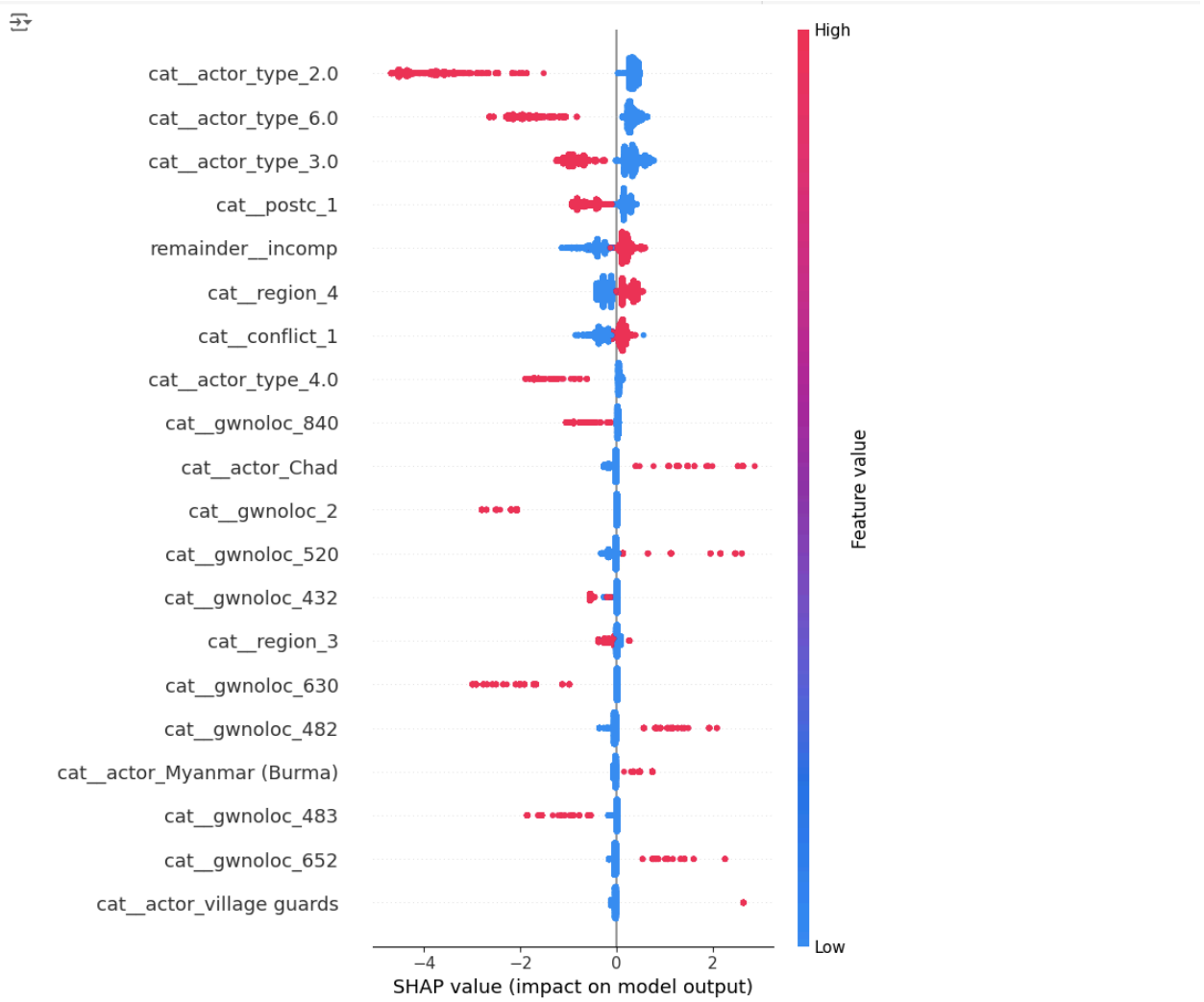
XGBoost improved further after hyperparameter tuning, showing better precision and a higher F1-Score for Class 1, indicating that it became more accurate at identifying real cases of sexual violence while avoiding many false positives

After hyperparameter tuning with Grid Search and Cross Validation, both models demonstrated improvements in precision, with XGBoost showing the most significant gains. This indicated that the tuning process helped the models better distinguish between instances of sexual violence and non-violence without sacrificing recall:

5.4 SHAP Analysis and Feature Importance

Both models provided insights into the most important features driving the predictions. The use of SHAP (SHapley Additive exPlanations) helped interpret the model's decision-making process by showing the contribution of each feature to individual predictions. SHAP allows us to determine which features push the model’s predictions towards sexual violence (class 1) and which push them towards no sexual violence (class 0).

The SHAP summary plot below shows the impact of each feature on the model's predictions, with dots representing individual predictions. Red dots indicate higher feature values, and blue dots indicate lower values. The x-axis shows the SHAP value, which represents the strength and direction of the feature’s influence on the model's prediction.



- **Actor Type**: Actor type played a pivotal role, with rebel groups (actor_type_3.0) and pro-government militias (actor_type_6.0) being significantly more likely to be associated with sexual violence. This is consistent with findings in conflict literature, which emphasize that certain non-state actors are more likely to use sexual violence as a tool of war.
- **Post-Conflict Status:** The importance of post-conflict status highlights the need for continued intervention and protection even after formal hostilities have ceased, as these periods are often prone to increased violence.

SHAP analysis confirmed that actor type and post-conflict status were key drivers in the model's predictions. SHAP values indicated that higher levels of conflict involvement and certain actor types consistently pushed predictions towards a higher likelihood of sexual violence. For example, rebel groups consistently showed higher SHAP values, correlating with an increased likelihood of sexual violence.

This interpretability is crucial for practitioners and policymakers, as it provides actionable insights into which actors and conflict phases require the most attention in terms of intervention and prevention efforts.

6. Conclusion

In this project, we aimed to predict the occurrence of sexual violence in armed conflicts using data from the SVAC dataset. We employed machine learning techniques, including Random Forest and XGBoost, and optimized them using GridSearchCV and Stratified Cross-Validation to address the class imbalance challenge. Here are the key conclusions drawn from the project.

6.1 Key Findings and Results

This project successfully applied machine learning models—Random Forest and XGBoost—to predict sexual violence in conflict zones. Several important insights were derived from the model's predictions and analysis:

- Actor Type as a Key Predictor: The type of actor involved in the conflict was the most influential predictor of sexual violence, with rebel groups and pro-government militias showing the highest likelihood of committing acts of sexual violence. This confirms findings from conflict studies that non-state actors often use sexual violence as a weapon of war.

- Post-Conflict Vulnerability: Our models highlighted the importance of post-conflict periods. Even after formal conflicts end, there remains a high risk of sexual violence, underlining the need for extended intervention and monitoring during these vulnerable times.

- SHAP Analysis and Model Interpretability: The use of SHAP (SHapley Additive exPlanations) allowed us to break down the contribution of each feature to the model's predictions. SHAP revealed how features such as actor type and conflict duration influenced individual outcomes, giving stakeholders clear, actionable insights into high-risk scenarios.

Overall, these findings validate the ability of machine learning models to assist in predicting and understanding conflict-related sexual violence.

6.2 Key Challenges and Improvements

While this project delivered significant insights, we encountered several challenges that required careful adjustments to improve model performance:

- Class Imbalance: A major obstacle was the imbalanced dataset, with far more instances of no sexual violence than cases of violence. Without adjustments, this would have skewed the model towards predicting non-violence. We addressed this challenge by using RandomOverSampler, which helped balance the classes and improved recall for instances of sexual violence.

- Hyperparameter Optimization: Initial model performance was limited by default hyperparameters. To improve precision and generalization, we applied GridSearchCV to optimize key hyperparameters such as the number of estimators and tree depth. This process significantly enhanced the model's ability to make accurate predictions.

- Cross-Validation and Model Robustness: To avoid overfitting and ensure robust results, we employed Stratified Cross-Validation, which maintained the balance of classes across training and testing splits. This was crucial in preventing bias toward the majority class and helped refine the final model.

These improvements made the models more reliable, especially in identifying high-risk conflict scenarios where sexual violence is more likely to occur.

6.3 Scientific Contributions and Future Directions

By using predictive models like Random Forest and XGBoost, we demonstrated the practical application of machine learning tools in identifying patterns of sexual violence in armed conflicts.

- Contributions to Conflict Prediction: This research shows that machine learning can be leveraged to predict conflict-related sexual violence by focusing on key features such as actor type and post-conflict status. These insights can help humanitarian organizations and policymakers better allocate resources to prevent and respond to such violence.

- Improving Decision-Making with Model Interpretability: The integration of SHAP for model interpretability is particularly valuable. SHAP provided a clear explanation of how different factors influenced the likelihood of sexual violence, which allows decision-makers to understand and trust the model's outputs. This interpretability is crucial for fields like humanitarian aid, where decisions must be transparent and data-driven.
- Future Research Directions: Despite the successes of this project, there are areas for future improvement. More advanced sampling techniques, such as SMOTE (Synthetic Minority Over-sampling Technique) or ADASYN, could further enhance model performance by addressing class imbalance more effectively. Additionally, exploring temporal features—such as conflict timelines or evolving actor behaviors over time—could add depth to the model's predictions. Future models could also benefit from integrating additional data sources, such as social media analysis or geospatial data, to create more comprehensive predictive tools.

By building on the findings from this study, future research can further refine these models and apply them to other forms of violence or conflict prediction, expanding the scope of data-driven humanitarian efforts.

6.4 Further Thoughts

In conclusion, this project has demonstrated the potential of machine learning to contribute meaningfully to the prediction and understanding of sexual violence in conflict zones. Through a combination of **feature importance analysis**, **SHAP-based interpretability**, and **model optimization**, we provided a comprehensive framework for identifying high-risk actors and conflict scenarios. The insights gained from this work can guide future research and practical applications in conflict monitoring and humanitarian decision-making, with the ultimate goal of reducing sexual violence in conflict-affected areas.

With further refinement and the incorporation of additional data sources, predictive models like those developed in this project could help to better understand how to prevent and respond to conflict-related sexual violence.

Literature:

Cohen, Dara Kay and Ragnhild Nordås. 2014. Sexual Violence in Armed Conflict Dataset. [Date Retrieved], from the Sexual Violence in Armed Conflict Dataset website: http://www.sexualviolencedata.org.

Wood, Elisabeth. 2009. “Armed Groups and Sexual Violence: When is Wartime Rape Rare?” Politics & Society 37(1): 131-161.

Authors of the report:

Maria Teresa Lluesma Ballesteros is a pharmacist with experience in research in the field of biotechnology. She has always been interested in social causes and feminism.

Linda Huber is a political scientist and journalist. During her time at the Free University in Berlin, she focussed on feminist approaches to peace and conflict research. As a journalist, she now covers criminal justice and power imbalances with the help of data-driven research.