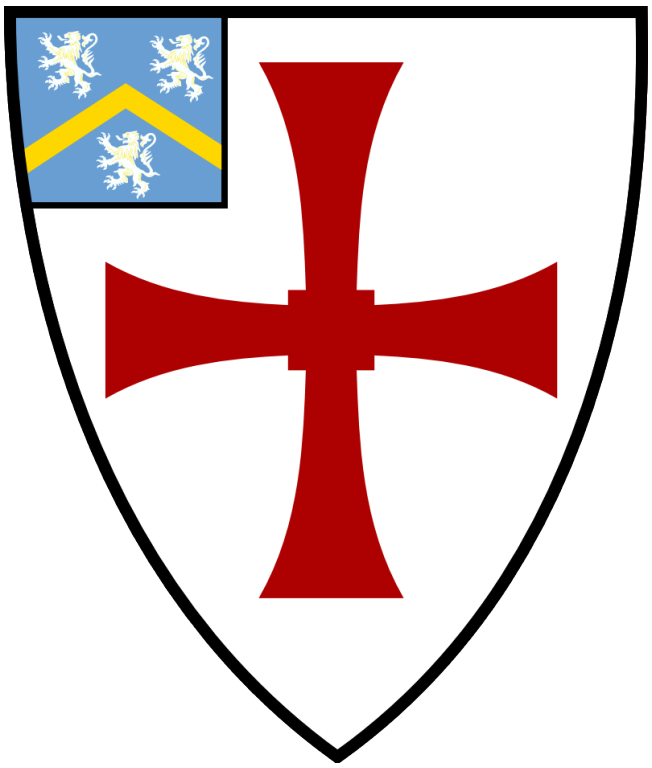
From Battles to Bytes: Assessing Machine Learning's Prowess in Predicting Global Conflicts

Ziyi Cui

A thesis presented for the degree of

Master of Data Science



Department of Natural Sciences

Durham University

August 2023

From Battles to Bytes: Assessing Machine Learning's Prowess in Predicting Global Conflicts

Ziyi Cui

Abstract

In the contemporary era, global conflicts pose significant threats to international peace and security. With the evolution of technology, predicting such wars using data-driven approaches has gained prominence. This study embarks on the journey of forecasting wars utilizing machine learning models, emphasizing the challenges presented by imbalanced datasets. Drawing inspiration from the 2022 Russian invasion of Ukraine, the research seeks to address two primary questions: identifying the most precise machine learning model for war prediction in the face of data imbalances and discerning the most influential features in the prediction process. Our findings indicate that the Random Forest model outperforms others in terms of balancing precision and recall, while features like GDP ratio and Democracy differential play pivotal roles. However, variables like Military expenditure showcased less importance, potentially due to imputation discrepancies. This research underscores the potential of machine learning in geopolitical predictions while highlighting the need for meticulous data handling and the significance of balancing precision in predictions to avoid false alarms.

Keywords: War Prediction, Machine Learning, Class Imbalance, Rare Events

Supplementary code and datasets used in this research are provided in the **GitHub** repository: https://github.com/zoeycui94/MDS2023DissertationProject

Supervisors: Dr Juraj Medzihorsky

Contents

Chapter 1

Introduction

1.1 Motivation

Throughout history, wars and conflicts have consistently left indelible marks, shaping the course of civilisations and nations. The events of 2022, particularly the invasion of Ukraine by Russia, provided a stark reminder of the profound consequences of warfare. Beyond the immediate casualties and displacements, the socio-economic and political aftermath of such conflicts can be extensive. Economically, wars can lead to significant damage to infrastructure, disruption of trade, and imposing financial burdens. Social consequences often include population displacement, loss of life, psychological trauma, and the emergence of refugee crises (Stein & Russett, 1980). Politically, wars have the potential to reshape international relations, shift power dynamics, and change territorial boundaries.

The 2022 conflict in Ukraine sent ripples through international relations, straining diplomatic relationships, instigating sanctions, and influencing global economies and energy markets (Khudaykulova, Yuanqiong & Khudaykulov, 2022). Given the wide-ranging implications of wars, predicting their onset becomes crucial. Accurate predictions can guide diplomatic actions, inform policy-making, and aid in strategic planning, potentially averting or at least mitigating the effects of such crises.

In a world rapidly advancing in technology, utilising machine learning to predict conflicts

emerges as a promising avenue. This research is motivated not only by the events of 2022 but by a more encompassing vision. By comprehending the complex precursors of wars and tapping into the capabilities of predictive analytics, this study aims to enhance knowledge that could potentially prepare societies better and, with hope, guide efforts toward conflict prevention.

Thus, the objective of this research is to predict the onset of conflicts with a minimised risk of false alarms, a goal deeply influenced by the personal experience of observing the Ukraine conflict and its worldwide ramifications.

1.2 Research Questions

The primary objective of this study is to utilise machine learning techniques to predict the onset of wars. Due to the complexities and the data imbalances intrinsic to this endeavour, the research is anchored by the following question:

"*In datasets where one class (e.g., peace) significantly outnumbers the other (e.g., war), which machine learning model provides the most accurate predictions for the onset of wars?*"

Predicting rare events, such as wars, in machine learning poses unique challenges. Skewed data can significantly affect the accuracy and reliability of predictions. Thus, it becomes imperative to identify models that are robust and effective in such scenarios.

"*Within the context of the machine learning model, which features hold the most significance when predicting the onset of wars?*"

Determining the importance of distinct features not only refines the predictive power of the model but also sheds light on potential precursors or indicators of wars. This research question seeks to identify patterns or factors that have historically played a crucial role in the buildup to conflicts.

1.3 Essay Structure

The essay is organised into five main chapters. The **Introduction** sets the stage, providing context, motivation, key research questions, and an overview of the essay's structure. This is followed by the **Literature Review,** which delves into the existing body of work on the causes of war, the role of machine learning in social science research, and the inherent challenges in utilising machine learning for war forecasting. In the **Methodology** chapter, the processes of data collection and preparation are detailed, along with the selection, building, and refinement of the machine learning models employed. The **Results and Discussion** chapter offers a deep dive into the data, visualising and interpreting the findings to provide insights related to the research questions. The essay concludes with the **Conclusion** chapter, summarising the primary takeaways, addressing research limitations, and suggesting avenues for future research in this domain.

Chapter 2

Literature Review

This chapter offers a comprehensive review of existing literature related to interstate war, elucidating its causes and the conditions precipitating its occurrence. Subsequently, the integration of machine learning in social science research is examined. The focus then shifts to specific empirical studies that harness machine learning for conflict prediction, especially when navigating the challenges presented by rare event data. The chapter concludes by addressing the limitations and challenges intrinsic to using machine learning for war forecasting.

2.1 Comprehending Interstate War and its Underlying Causes

The propensity for interstate wars, marked by armed conflict between two or more nations, hinges on a myriad of factors, deeply entrenched and multifaceted in nature. In the complex tapestry of causes, the threads of economy, democracy, military strength, and geopolitics often intertwine and surface as pivotal catalysts for such conflicts (Blainey, 1988).

Economic elements have consistently emerged as significant precursors to many interstate wars throughout history. Economic disputes may span a wide spectrum from competition over scarce resources to disagreements over trade policies. A notable example is the Falklands War of 1982, where Argentina, plagued by economic turmoil under a military dictatorship, aimed

to assert control over the Falklands, a territory teeming with potential offshore oil deposits. This war served as a strategic diversion from Argentina's domestic economic issues (Freedman & Gamba-Stonehouse, 1991).

Simultaneously, the chasm in political systems and ideological frameworks often underpins interstate conflicts. Democratic nations, driven by their core principles, have frequently found themselves embroiled in conflicts with the aim to either protect existing democracies or promote democratic ideals. The ideological tussle manifested vividly during the Korean War from 1950 to 1953, pitting democratic South Korea, backed by the US, against communist North Korea, supported by the Soviet Union and China, reflecting the larger ideological skirmish of the Cold War era (Stueck, 1997).

The calculation of military capabilities and the perception of threats are inherent factors that significantly influence a nation's decision to wage war. Nations possessing superior military prowess may yield to the temptation of employing force to realize their objectives. The Six-Day War of 1967 bears testament to this, as Israel initiated the conflict against Egypt, Jordan, and Syria, driven predominantly by perceived military threats against itself. This episode underscores how military considerations can trigger interstate wars (Bar-On, 2016).

Geopolitics, which evaluates the influence of geographical factors on international relations, also assumes a crucial role in the inception of interstate wars. Elements such as territorial disputes, strategic locations, and spheres of influence are typical geopolitical factors that can instigate conflict. The ongoing South China Sea dispute exemplifies this situation, involving China and several Southeast Asian nations vying for control over a region of immense strategic importance due to its significant oil and natural gas reserves and its role as a major trade route. China's aggressive stance on its territorial claims, contested by other nations, has amplified tensions in the region, increasing the potential for an armed conflict (Buszynski, 2012).

The triggers for interstate wars are often complex and interrelated, with economic, democratic, military, and geopolitical factors frequently at the forefront. Recognizing and understanding these causes is instrumental in informing policies and initiatives aimed at conflict prevention and

resolution. As we navigate the modern global landscape, the importance of fostering economic interdependence, promoting democratic norms, encouraging military restraint, and facilitating geopolitical dialogue cannot be overstated, as they remain critical strategies in mitigating the risk of interstate wars (Levy & Thompson, 2011).

2.2 Harnessing Machine Learning in Social Science Research

Machine learning (ML) is a rapidly advancing field that has been increasingly adopted by various scientific disciplines, including the social sciences. It is a subfield of artificial intelligence, which refers to systems capable of improving their performance by learning from data without being explicitly programmed (Russell & Norvig, 2016). The incorporation of these techniques in social science research signifies a transformative shift from traditional statistical methods to predictive and interpretative models (Mullainathan & Spiess, 2017).

Traditional social science research often employs statistical techniques that involve hypothesis testing and p-values. However, the advent of Big Data has resulted in an increase in the volume and variety of data, necessitating more sophisticated tools for data analysis. In response to this, many social scientists are turning to ML to analyse complex, high-dimensional datasets (Ij, 2018).

Machine learning algorithms differ from traditional statistical methods in several ways. They prioritize predictive accuracy over explanatory power and are more flexible, and capable of handling non-linear relationships and interactions among variables (Blei & Smyth, 2017). In order for flexible models to learn patterns and achieve the specified goal, ML has established study design techniques that assist in selecting usable attributes that describe experiences. As with research centred on traditional statistical methods, ML's concentration is on increasing its performance on a clearly defined task and related performance metric rather than worrying about bias relative to an unobserved parameter or model. Traditional methods assume that

the model is, in some way, accurately stated a priori and conditional to what we learn from the data on this premise. Conversely, ML techniques depend on what we discover about a model representation's capacity to raise an identifiable performance metric (Colaresi & Mahmood, 2017).

In recent years, several studies in the social sciences have begun to utilize ML algorithms. For example, Muchlinski, Siroky, He and Kocher (2016) used a random forest algorithm to predict instances of civil war with higher accuracy than traditional logistic regression models. Similarly, machine learning has been applied to study political behaviour. Beauchamp (2017) used topic modelling, a form of unsupervised ML, to analyse a large corpus of Twitter data, identifying the primary political issues concerning US Twitter users during the 2012 Presidential election. Further, ML can also facilitate the automation of research tasks. In a study on human rights abuses, Fariss (2019) used supervised machine learning to analyse and categorise a large volume of textual data, significantly reducing the time needed for manual coding.

2.2.1 Utilising Machine Learning for Conflict Prediction

As ML has gained significant traction in the prediction of geopolitical and social conflicts, a particularly promising application of ML in this context lies within the domain of war forecasting.

Muchlinski et al. (2016) used a random forest algorithm to predict civil war onset, utilising data from Uppsala Conflict Data Program (UCDP) and Peace Research Institute Oslo (PRIO). They compared the random forest approach to traditional logistic regression. While both methodologies achieved similar precision, the random forest had a better recall, indicating it was more successful in identifying true positives, i.e., actual conflict occurrences. In a study focusing on international conflict, Chadefaux (2014) used support vector machines (SVM) to analyse a corpus of over 100 million news articles from the Global Data on Events, Location, and Tone (GDELT) project. The SVM model was trained to detect precursor signals of conflict within the data, outperforming traditional time-series analysis.

Yan, Nuttall and Ling (2006) used a combination of ML techniques, including k-nearest neighbours, decision trees, and SVM, for predicting crime hotspots in Los Angeles. They used historical crime data from the LA Police Department, alongside demographic and geospatial data. Their models achieved an accuracy of up to 72%, significantly higher than traditional hotspot prediction methods.

In another innovative application, Zhukov and Stewart (2013) utilised text mining techniques to analyse data from millions of news articles, aiming to predict political protests. Their method was based on detecting changes in the volume and tone of news related to potential conflict areas. However, when forecasting these types of rare events such as wars, military coups, pandemics, and governmental vetoes, class imbalance is an inevitable issue that cannot be overlooked as it would largely affect the performance of the prediction models.

2.2.2 Addressing Class-Imbalance Issues and Potential Solutions

ML algorithms typically assume a balanced class distribution or equal misclassification costs, but real-world data often violate these assumptions (Japkowicz & Stephen, 2002). In the presence of class imbalance, most algorithms tend to be biased towards the majority class, resulting in a poor predictive performance for the minority class (Krawczyk, 2016). Class imbalance is indeed a significant challenge in conflict prediction, as wars and other major conflicts are relatively rare events. This leads to a dataset where the non-conflict class is substantially larger than the conflict class, potentially skewing predictions (Bhattacharyya, Jha, Tharakunnel & Westland, 2019).

One common approach to address class imbalance is resampling, which involves modifying the dataset to create a more balanced class distribution. Techniques include oversampling the minority class, undersampling the majority class, or a combination of both (Chawla, Bowyer, Hall & Kegelmeyer, 2002).

Developed by Chawla et al., Synthetic Minority Over-sampling Technique (SMOTE) is a popular oversampling technique specifically designed to tackle the problem of class imbalance in

machine learning. The principal idea behind SMOTE is to create synthetic examples of the minority class, rather than replicating existing instances, which is the case with random oversampling. This approach helps alleviate the overfitting problem associated with simple oversampling.

Muchlinski et al. applied SMOTE to their civil war onset data and found a marked improvement in the model's ability to predict the minority class. They specifically noted that the usual logistic regression model struggled to make meaningful predictions due to this imbalance. By creating synthetic instances of conflict cases using SMOTE, the researchers sought to provide the random forest model with more instances from which to learn. This augmentation of minority class examples helped in overcoming the model's initial bias towards the majority class, that is, peace events. As a result, they observed a marked improvement in the model's predictive power for conflict onsets. The study found that the model's recall improved significantly, which meant that it was better at identifying true conflict cases in the sea of peace events. This study stands as a testament to the efficacy of SMOTE in mitigating the challenges of class imbalance in machine learning, particularly in complex, real-world scenarios.

Another approach is to use cost-sensitive learning, which assigns a higher misclassification cost to the minority class. By doing so, it creates an incentive for the machine learning algorithm to correctly classify the minority instances, thus helping to mitigate the problem of class imbalance. This essentially nudges the learning algorithm to focus more on the minority class and less on the majority class, which results in better minority class predictions.

A study by He and Garcia (2009) demonstrated that cost-sensitive learning, when combined with SVM, significantly improved the recall for minority class prediction. They argued that standard SVM, while an effective and widely used technique, does not inherently consider the cost of misclassifications. By implementing cost-sensitive learning, they incorporated an adjusted cost function that penalized the SVM more heavily for misclassifying minority class instances. The result of this method was a significant improvement in the model's ability to predict the minority class instances, as demonstrated by an increase in recall, a metric

that quantifies the model's ability to correctly identify actual positives. This work serves as a compelling case for the power of cost-sensitive learning in addressing the issue of class imbalance in ML.

Random Forest, an ensemble learning method that operates by building multiple decision trees and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees, is another commonly used algorithm. However, it can be biased towards the majority class in the presence of class imbalance. To address this issue, Chen, Liaw and Breiman (2004) proposed a modification to the Random Forest algorithm, which involves adjusting the class distribution in bootstrap samples. The algorithm creates bootstrap samples (random samples with replacement) for each tree.

However, in their modification, instead of allowing all classes to be equally likely to be sampled, they adjusted the probabilities to make the minority class instances more likely to be included in the samples. This modification yielded an improvement in the classification performance of the minority class, as shown in Chen et al.'s research, which illustrates the potential effectiveness of adjusting existing algorithms at an intrinsic level to better handle class imbalance. This success underlines the importance of continuing exploration and adaptation of machine learning methodologies in line with the nature and distribution of the data they handle.

While these techniques have proven effective, each comes with its drawbacks. For example, oversampling might lead to overfitting, while undersampling might discard potentially useful data (Fernández, García, Herrera & Chawla, 2018). The choice of approach often depends on the nature of the data and the problem context.

Future research could focus on developing hybrid methods that combine different approaches for handling class imbalance. Moreover, developing advanced metrics that go beyond accuracy, such as G-mean, F-measure, and Area Under the Receiver Operating Characteristic (AUROC), would provide a more comprehensive evaluation of model performance in the presence of class-imbalance (Sokolova & Lapalme, 2009).

2.3 The Challenges and Limitations in the Use of Machine Learning for War Prediction

As we can see, the use of ML in predicting war and conflict has opened up new avenues in the realm of social science research. Yet, despite promising advancements, there are significant challenges and limitations that researchers must confront.

As mentioned above, one of the most fundamental challenges in using ML for war prediction is the problem of class imbalance. Wars, being relatively rare events when compared to periods of peace, are the minority class in most datasets, leading to an imbalance that traditional ML algorithms struggle to handle effectively. They tend to be biased towards the majority class, thereby failing to accurately predict the onset of wars (Japkowicz & Stephen, 2002; Muchlinski et al., 2016).

Several techniques have been proposed to address class imbalance, such as Synthetic Minority Over-sampling Technique (SMOTE), cost-sensitive learning, and algorithmic-level methods. While these methods have shown promise, they are not without limitations (Chawla et al., 2002; Chen et al., 2004; Liu & Zhou, 2006). Oversampling techniques like SMOTE might lead to overfitting while undersampling could result in the loss of potentially important information (Fernández et al., 2018).

Besides, predicting war also requires access to high-quality, accurate, and detailed data. The process of collecting, cleaning, and curating this data for use in machine learning models can be challenging and time-consuming. Additionally, the availability of data might be limited due to political, legal, and ethical constraints. Moreover, historical conflict data can be inconsistent and incomplete, with varying levels of reliability and accuracy. A machine learning model is only as good as the data it's trained on, and poor-quality data can lead to misleading or inaccurate predictions.

Another inherent challenge is the risk of overgeneralisation. While machine learning algorithms

can identify patterns within training data and apply these patterns to new data, they might fail to account for unique or context-specific factors. This overgeneralisation can be particularly problematic when predicting complex social phenomena like wars, which are influenced by a host of geopolitical, cultural, economic, and historical factors that might not be fully captured in the data.

Utilising ML for war prediction also raises critical ethical questions. For instance, if a model predicts an increased likelihood of conflict, how should policymakers respond? Predictive models might also be misused or misinterpreted, potentially leading to unwarranted fear or even provoking the conflicts they are designed to predict.

Therefore, while the application of machine learning in predicting war presents promising possibilities, these are not without significant challenges and limitations. Issues of class imbalance, data quality, overgeneralisation, and ethical considerations need to be carefully addressed. Only through such critical engagement with these challenges can we responsibly harness the potential of machine learning for social science research and policymaking.

Chapter 3

Methodology

In this chapter, we will detail the processes by which the data was collected, prepared, and cleaned. Subsequently, we will provide a rationale for the selection of the three base models, namely 1) Logistic Regression, 2) Decision Tree, 3) Random Forest, and three optimised models with base models constructed from this data 4) Bagging on Decision Tree, 5) Synthetic Minority Over-sampling Technique on random forest, and 6) Cost-sensitive learning on Random Forest. Finally, we will outline the methodology used to evaluate the performance of these models.

3.1 Data Collection

In the endeavour to predict war, it's paramount to discern the underlying causes of conflicts to select relevant indicators for dataset construction. As discussed in the Literature Review section, the catalysts for interstate conflicts are multifaceted, with intricate interplays of various factors. Notably, economic, democratic, military, and geopolitical elements are often at the forefront of these causes, as highlighted by Blainey (1988). Drawing from this understanding, my dataset incorporates the following pivotal indicators shown in Table 3.1: The study's timeline, from 1970 to 2018, was primarily chosen due to consistent data availability across the consulted sources. This 48-year span encompasses a variety of geopolitical events and shifts, making it apt for accurate war prediction given the dynamic nature of interstate relations.

Table 3.1: Variables List

Variable Names	**Explanation**	**Data Type**
Year	Representing the year of each observation from 1970-2018	Discrete
Country A	Representing the names of two countries in each dyad	Categorical
Country B	Representing the names of two countries in each dyad	Categorical
War or not	A binary response variable where 1 indicates the occurrence of war between Country A and Country B, and 0 indicates no war	Binary
gdp_ratio	Representing the GDP ratio between the two countries	Continuous
gdp_per_capita_ratio	Representing the GDP per Capita ratio between the two countries	Continuous
Military_expenditure_ratio	Representing the Military Expenditure in Percentage of GDP ratio between the two countries	Continuous
Democracy_differential	Representing the Difference of Democracy Level between the two countries	Discrete
Share Border or not	A binary indicator where countries pair that share borders marked as 1, otherwise marked as 0	Binary
Political Instability	0 indicates both countries are politically stable; 1 stands for one of the countries is unstable and 2 means both countries are unstable	Categorical

The binary "War or not" indicator serves as the dependent variable, where "0" represents no war and "1" indicates its occurrence. Data is derived from the *UCDP/PRIO Armed Conflict Dataset version 23.1*, a record of armed conflicts from 1946-2022 involving at least one state actor (Davies, Pettersson & Öberg, 2023). Within this dataset, a conflict is termed a "war" if there are 1,000 or more casualties. Accordingly, conflicts meeting this threshold and involving two governmental entities were extracted to serve as the dependent variable in my analysis.

The independent variables include "gdp_ratio" and "gdp_per_capita_ratio," representing GDP and GDP Per Capita ratios between countries. This data comes from the *Country-Year: V-Dem Full+Others version 13* dataset by V-Dem (Coppedge et al., 2023; Pemstein et al., 2023), covering 1789-2022.

"Military_expenditure_ratio" provides a ratio of military spending between countries relative

to GDP, sourced from The SIPRI Military Expenditure Database. Notably, global military expenditure aggregates pre-1988 are unavailable due to missing Soviet Union statistics. Handling any missing values in this dataset will be addressed in the Data Preparation section.

Both "Democracy_differential" and "Political Instability" are based on the *Polity 5* dataset, which rates countries' democracy levels (Marshall & Gurr, 2020). The dataset employs a "Polity Score" spanning -10 (hereditary monarchy) to +10 (consolidated democracy). For this study, the differential is calculated by subtracting one country's score from another. "Political Instability" considers scores of -66, -77, and -88.

Lastly, the "Share Border or not" indicator, denoting if two nations share a border, is taken from *GeoDataSource's "Country Borders"* dataset.

The broad process of data collection is visualised in Figure 3.1.

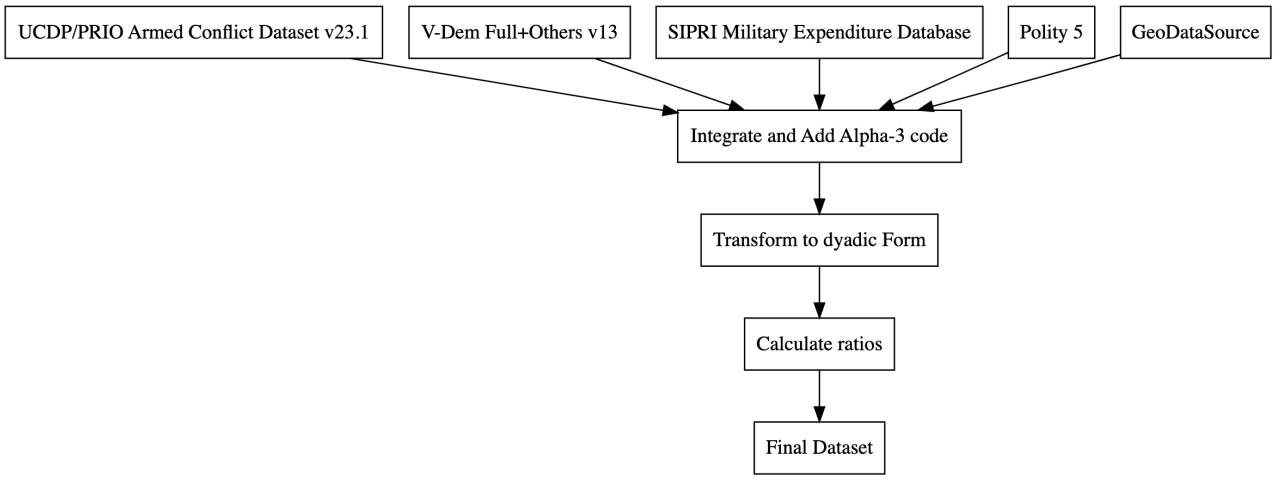


Figure 3.1: Visualisation of Dataset Collection and Management

3.2 Data Preparation

To forecast interstate wars, the dataset was structured in dyadic form. This approach, wherein every country is paired with every other country, facilitates the comparative analysis of the

indicators and has been previously endorsed for its utility in conflict studies (M. Ward, Siverson & Cao, 2007). Organising data dyadically also allows for systematic observation of potential missing values, which are known to impact the performance of predictive models (Little & Rubin, 2019).

During the initial exploration phase, it became evident that certain smaller nations, like the Bahamas, Comoros, and Grenada, presented numerous missing values in most indicators. Such missing data issues, if not addressed, can compromise the validity and reliability of statistical analyses (Gelman & Hill, 2006). Also, these nations hadn't experienced interstate wars during the study timeframe. The combination of their absence from the main event of interest (war) and the prevalence of missing values suggests that their inclusion might introduce noise rather than clarity, a concern noted in past conflict prediction studies (Schrodt, 2014). To enhance the robustness of subsequent models, rows corresponding to these nations, and their dyadic counterparts, were removed. Furthermore, any country exhibiting over 5 years of consecutive missing data for any indicator had the respective years excluded, consistent with recommendations on handling longitudinal data (Singer & Willett, 2003).

Addressing sporadic missing values required a more nuanced approach. When dealing with time-series data, imputation methods that respect temporal structures are vital (Honaker & King, 2010). If a country presented missing values for an indicator within a 5-year span but had valid data both preceding and following this gap, imputation methods were employed. This approach, using average values from proximate data points for imputation, offers a reasonable estimate based on the country's historical and subsequent trends and has been validated in prior research (Enders, 2010). The same method was also applied to other indicators like military expenditure, except for Soviet Union's military expenditure. In the SIPRI Military Expenditure Database, data for the Soviet Union is all missing. But as an influential and important country at the time, there would be some literature (Busch, 1997; Harrison, 2003; Ricón, 2016; Steinberg, 1990) that did research on its military expenditure during the period, so the estimation numbers were filled into the gaps.

An essential step in our data preparation was the normalisation of several key variables, namely 'gdp_ratio', 'gdp_per_capita_ratio', 'Military_expenditure_ratio', and 'Democracy_differential'. Normalisation, a technique of scaling all variables to the same range, often [0,1], ensures that no variable disproportionately influences the model due to its scale (Singh & Singh, 2020). This step is especially vital for models sensitive to feature magnitude, leading to more robust results across our diverse set of modelling techniques.

3.3 Model Building

This section elucidates the methodologies used to construct predictive models. Our primary aim is to forecast the likelihood of conflict emergence between specific country pairs, relying on various socio-economic and geopolitical indicators.

3.3.1 Logistic Regression

Logistic regression, a generalised linear model, was selected as the primary analytical tool. This model is especially pertinent for scenarios with a binary dependent variable, which in this study is Conflict or not. The methodological prowess of logistic regression lies in its capability to estimate the logarithm of odds contingent upon a set of predictors. Over the years, logistic regression has seen extensive application across disciplines, including conflict studies and international relations (Ward et al., 2003). Its methodological robustness and empirical validity, when assumptions are met, make it a reliable choice for such complex studies.

One of the primary strengths of logistic regression lies in its inherent interpretability (Hosmer Jr, Lemeshow & Sturdivant, 2013). Given the study's focus on understanding factors influencing conflicts between countries, the coefficients yielded by the logistic regression model provide an intuitive understanding of the effect size and direction of each predictor on the likelihood of a conflict occurring. This assists in deriving policy insights and actionable strategies based on individual predictor variables.

Unlike some classification techniques, logistic regression computes the probability of a particular instance belonging to a category (Peng, Lee & Ingersoll, 2002). In the geopolitical context, having an understanding of the risk or likelihood of conflict, rather than a sheer binary prediction, can be crucial for nuanced policymaking and conflict prevention.

Besides, logistic regression is specifically designed for binary classification problems, making it apt for predicting outcomes like 'Conflict or not' (Cramer, 2002). Its assumption of the linearity of log odds often aligns with empirical phenomena where influencing factors incrementally increase or decrease the odds of an event occurring. Given the dataset's potential size, encompassing multiple countries over multiple years, the efficiency of logistic regression becomes advantageous. Its computational simplicity relative to more complex models ensures quicker iterations and validations (Le Cessie & Van Houwelingen, 1992).

The procedure of building a logistic regression model would first be a partitioning of the dataset. The data corpus was systematically bifurcated into training (80%) and testing (20%) subsets. Stratified sampling, achieved by setting `stratify=y`, ensured a representative distribution of conflict instances across both subsets, mirroring the distribution in the original dataset. Then the model would be initialised with parameters tailored for balanced class weights utilising `class_weight= 'balanced'`. This adjustment compensates for any inherent class imbalance, ensuring that the model isn't unduly influenced by the majority class. The model was then meticulously calibrated using the training dataset. It aims to discern intricate relationships between predictors and the likelihood of conflict manifestation. Post-calibration, the model underwent rigorous evaluation using the testing dataset. Metrics such as accuracy, precision, recall, F1 score, and AUC were computed to comprehensively assess the model's proficiency. A confusion matrix provided a vivid visual representation of prediction veracity. Coefficients derived from the logistic regression were analysed to fathom the relative importance of predictors. Finally, to ensure robustness against overfitting and to ascertain generalisability, a 5-fold cross-validation was implemented. This procedure provides an encompassing view of the model's performance over varied data partitions.

3.3.2 Decision Tree

Machine learning often employs a category of algorithms known as decision trees. These trees are versatile, catering to both regression and classification challenges, and are especially apt for extreme event studies, as noted by Frohwein and Lambert (2000). Unlike many traditional statistical methods, decision trees don't rely on any probability density function, meaning they don't assume any specific underlying distribution. As a result, they fall under the realm of nonparametric statistics. In techniques rooted in trees, the predictor space is divided into several straightforward regions. For every such region, the predicted response value is determined independently.

For a more hierarchical and intuitive understanding of the factors affecting geopolitical conflicts, a Decision Tree model is employed. Decision Trees offer a visual representation of decision rules, making them one of the most interpretable machine-learning algorithms (Quinlan, 1986). Given the diverse audience that might be interested in the outcomes of geopolitical studies — ranging from policymakers to academics — an intuitive, graphical representation can be particularly impactful. While logistic regression assumes a linear relationship in the log odds of the dependent variable, decision trees make no such assumption, making them adept at capturing non-linear relationships and interaction effects between predictors without explicit feature engineering (Breiman, Friedman, Olshen & Stone, 1986). In addition, it inherently prioritises certain features at higher nodes, which can provide insights into which variables play a more dominant role in predicting conflicts. This hierarchical structure can offer a nuanced perspective into the sequence or hierarchy of geopolitical factors leading to conflicts (Kaur & Wasan, 2006). The depth and complexity of decision trees can be controlled through pruning strategies, enabling the model to be adapted to different levels of granularity as per the study's requirements (Esposito, Malerba & Semeraro, 1997).

Similar to the procedure of building logistic regression, the dataset was partitioned into training and test sets as well, retaining 80% of the observations for training and the remaining 20% for testing. The `stratify=y` parameter ensured that both sets maintained the same proportion

of positive and negative conflict cases as the original dataset. To prevent overfitting and ensure the decision tree's performance was optimised, we conducted a grid search over multiple hyperparameters. The hyperparameters under consideration included:

`max_depth`: Maximum depth of the tree. The considered values were unrestricted depth, 5, 10, and 20.`min_samples_split`: The minimum number of samples required to split an internal node. Values tested included 2, 5, and 10. `min_samples_leaf`: The minimum number of samples required to be at a leaf node. We evaluated the tree for leaf sizes of 1, 2, and 4 samples.

Then the decision tree classifier was initialised with a random state for reproducibility and `class_weight= 'balanced'` to address a potential class imbalance in the conflict cases. After training, the model's performance was evaluated on the test set. Metrics used for evaluation included accuracy, precision, recall, F1 score, and the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC). Additionally, a confusion matrix was plotted to visualise the true positive, true negative, false positive, and false negative rates. Following that, the trained Decision Tree's feature importance was visualised to provide insights into which geopolitical variables had the most significant influence in predicting conflicts.

3.3.3 Random Forest

Random Forest, an ensemble learning method, has proven itself to be a potent classifier that addresses many limitations present in individual decision trees, like overfitting and high variance. The main idea behind the Random Forest is to combine the outputs of multiple trees, thereby capitalizing on their collective power to achieve better generalization and predictive performance (Breiman, 2001).

An intrinsic challenge in predicting rare events, such as conflict in geopolitical studies, is the underrepresentation of these events in the data. In scenarios where events like war and government budget shifts are sparse, a simple majority vote might misclassify significant observations. For instance, when predicting major budget shifts, the base rate might indicate a 10 percent probability. Still, if the model suggests a 40 percent probability, this representation is four

times the random chance, suggesting an anomaly worth noting. However, a strict majority vote might disregard this probability and classify the observation as a regular event (Chen et al., 2004).

The essence of this challenge is mirrored in predicting conflicts, as they, too, are extreme and rare events. Thus, relying solely on majority voting could potentially lead to overlooking important precursors or signs of an emerging conflict. Random Forest provides flexibility in this regard. By tweaking the ensemble rule — say by classifying an observation as a conflict if a certain percentage of trees predict so, instead of a strict majority — the model can be tuned to detect these rare events more effectively. While this might compromise the overall classification accuracy slightly, in the context of conflict prediction, it is often more crucial to detect potential threats than to achieve perfect classification.

Given the importance and consequences associated with geopolitical conflicts, it becomes pivotal to employ a model that can be tailored to serve as an efficient "detector" for such events. Thus, Random Forest, with its inherent capability of handling imbalanced data and its flexibility in classification rules, was deemed an appropriate choice for this study.

As for the procedures of building the model, data partitioning and model evaluation are the same as the last two models built. However, the random forest classifier was initialised with the Gini criterion, a set random state for reproducibility, and `class_weight= 'balanced'` argument to adjust for potential class imbalances in the dataset. An essential aspect of maximizing the random forest's predictive power is hyperparameter tuning. A comprehensive grid search was conducted over several hyperparameters: 1). `n_estimators`: The number of trees in the forest. The evaluated values were 50, 100, and 200. 2). `max_depth`: The maximum depth of the trees. Values considered were unrestricted depth, 5, 10, and 20. 3). `min_samples_split`: Minimum number of samples required to split an internal node. The evaluated values were 2, 5, and 10. 4). `min_samples_leaf`: Minimum number of samples required at a leaf node. Tree models were assessed for leaf sizes of 1, 2, and 4.

3.3.4 Bagging with Decision Tree

This section returns to the roots of ensemble modelling, emphasising the basic idea behind bagging – reducing the variance of a decision tree model by averaging out numerous trees. Consistency being paramount in comparative research, the dataset was once again partitioned into a training and test set using an 80-20 split, ensuring stratification to retain class proportions. A Decision Tree serves as our base model. Its inherently high variance, susceptible to data changes, makes it a prime candidate for bagging. By employing `class_weight= 'balanced '`, we proactively tackle potential class imbalances, and then BaggingClassifier is initialised with the Decision Tree as its base estimator. This Bagging Classifier is then trained on our dataset, creating an ensemble of Decision Trees, using this aggregated model, predictions are made on the test set. Finally, the model's performance is assessed employing familiar metrics: accuracy, precision, recall, F1 score, and AUC-ROC. To visualise the classification accuracy and misclassifications, a confusion matrix is plotted and to understand which predictors hold the most weight in this ensemble model, feature importance from individual trees is aggregated. By averaging this importance, we derive a consolidated understanding of pivotal variables.

Bagging with Decision Trees is a classic ensemble approach. Each individual tree might capture different nuances of the data, and by aggregating them, the ensemble aims to combine their strengths and mitigate their individual weaknesses. Given its foundational nature, this approach offers a cleaner, more interpretable model compared to more complex ensembles.

This model serves as a baseline to understand how basic ensembles perform against the more intricate ones, like Bagging with Random Forests. It's crucial to gauge if the additional complexity introduced in the previous section genuinely amplifies model performance or if simpler aggregations, such as this, are equally, if not more, effective.

3.3.5 Synthetic Minority Over-sampling Technique (SMOTE) on random forest

Transitioning from basic and advanced ensemble methods, this section introduces a pre-processing technique aimed at addressing class imbalance—a common challenge in predicting rare events like conflicts. Given the highly imbalanced nature of our dataset, where the instances of war (positive class) are significantly outnumbered by non-war instances (negative class), there's a risk that models might be biased towards predicting the majority class. This can lead to high accuracy but poor recall, which isn't ideal since we're particularly interested in correctly predicting the onset of wars.

SMOTE addresses this issue by generating synthetic samples in the feature space. It achieves this by:

1) Choosing a minority instance, a. 2) Randomly picking one of its k-nearest neighbors, b. 3) Constructing a new point, c, along the line segment connecting a and b.

By doing so, SMOTE effectively increases the number of minority class instances, balancing the class distribution and allowing the model to learn more representative decision boundaries.

After splitting the dataset into training and test sets with an 80-20 ratio. SMOTE is then initialised and employed on the training set. By picking instances that are close to the feature space, SMOTE introduces synthetic points to create a balanced dataset. With the newly balanced training data, the Random Forest classifier is then trained with the hyperparameters of the RandomForest model optimised using RandomizedSearchCV. After training, the importance of each feature in predicting conflicts is visualised in a bar chart. Then the same model evaluation as mentioned before would be performed.

3.3.6 Cost-sensitive learning on random forest (CSL)

One of the challenges in predictive modelling, especially with imbalanced datasets, is the risk of the model becoming insensitive to the minority class. While techniques like SMOTE adjust the dataset itself, cost-sensitive learning adjusts the algorithm's behaviour to make misclassification of the minority class more "expensive". By introducing penalties or weights, this method ensures that the algorithm learns the importance of correctly classifying minority class instances.

As with earlier models, the dataset is split into training and test sets with an 80-20 ratio. Stratified sampling ensures a proportional representation of the classes in both sets. Then random forest model is weighted. the RandomForestClassifier is initialised with different weights for the two classes. In this case, misclassifying a "Conflict" instance (label 1) is ten times as costly as misclassifying a "No Conflict" instance (label 0). A RandomizedSearchCV was employed to tune the hyperparameters of the RandomForest model, emphasizing the recall metric. Among the hyperparameters optimised, the `class_weight` parameter was varied to explore different cost-sensitive settings, such as balanced and custom weights like `{0: 1, 1: 100}` and `{0: 1, 1: 500}`. By doing so, the Random Forest algorithm now has an incentive to be more cautious about missing conflict events. After the model's understanding of variable significance is visualised, it is evaluated on the untouched test set using various metrics, including accuracy, recall, F1 score, the AUC-ROC score as well as a confusion matrix.

Cost-sensitive learning offers a refined approach to enhancing the model's sensitivity towards the minority class. By penalising misclassifications, the model tends to predict more "Conflict" instances, but this could also lead to an increase in false positives. While accuracy is an integral metric, the primary focus remains on recall. Given the importance of predicting conflicts, high sensitivity is crucial. However, it's important to remember that with higher sensitivity, there may be an accompanying decrease in precision. By increasing the weight for "Conflict" instances, there's a conscious acceptance of the risk of predicting more false positives. In contexts like conflict prediction, this trade-off might be acceptable given the potential consequences of missing an actual conflict.

if we always predict True no matter what - our Recall would be 100% because there would be cases where model predicted False for a instance that was True. But Precision would be low as there would be multiple cases where the instance was False but model classified it as True
If we always predict False no matter what - Recall would be 0% with unknown precision (cause TP is 0)

FP - Type 1 error. The model predicted True but correct label is False
FN - Type 2 error. The model predicted False but correct label is True

Table 3.2: Evaluation Metrics Definitions

Metric	**Explanation**
Accuracy	Measures the proportion of correctly predicted classifications in the total predictions. = TP + TN / all 4
Recall	Focuses on the positive class and measures the proportion of actual positives the model correctly identified. = TP / TP + FN.
Precision = TP/TP+FP	Measures the proportion of true positives in the positive predictions.
F1 Score	Harmonic mean of precision and recall, providing a balance between the two metrics.
AUC-ROC	A comprehensive metric that evaluates the ability of a model to differentiate between the classes.
Confusion Matrix	In this matrix, True Positive represents the count of accurately classified positive instances, False Negative corresponds to the number of positive instances mistakenly labelled as negative, False Positive indicates the count of negative instances incorrectly identified as positive, and True Negative signifies the number of correctly classified negative instances.

3.4 Model Evaluation

After building various models, evaluating their performance is crucial to determine which one best meets the requirements of the project. This section systematically presents a comparative analysis of the models constructed.

3.4.1 Evaluation Metrics

Each model was evaluated using several metrics and the detailed explanations for some important indicators will be presented in Table 3.2. Within these evaluation metrics, *Accuracy* is the most commonly used method to evaluate models, however, in this study, the dataset is highly class imbalanced, as accuracy puts more weight on the majority class, it should not be the best evaluation indicator to measure the models.

In the realm of conflict prediction, the primary objective is often to ensure that potential conflicts are not overlooked. Given the severe implications and consequences associated with armed conflicts, the goal is to capture every possible event, even at the risk of predicting false positives. Consequently, the most crucial evaluation metric for this study is *recall*, or the true

positive rate. A high recall ensures that the maximum number of actual conflicts are identified, underscoring the significance of not missing out on any potential conflict. While other metrics such as accuracy and precision are still relevant and offer insights into the model's overall performance, they take a backseat to recall in this specific context. The rationale behind this prioritisation is straightforward: the potential cost of a false negative, or missing out on predicting a real conflict, is far more consequential than a false positive, where a peaceful scenario might be inaccurately flagged as conflict-prone.

Another commonly used metric in the medical field is the receiver operating characteristic (ROC) analysis and its associated area under curve (AUC). The ROC represents a balance between a false positive rate (FPR) and a true positive rate (TPR), also known as sensitivity. FPR points are plotted on the x-axis, while TPR points are plotted on the y-axis. Therefore, an effective classifier would generate points located in the upper left corner of the diagram. On the other hand, random guessing would be represented along the main diagonal line (Fawcett, 2006). AUC does not exhibit bias towards any specific class and thus remains impartial even for rare classes.

3.4.2 Comparative Analysis

In the multifaceted domain of conflict prediction, the choice of modelling technique holds profound implications for the efficacy of predictions. This section undertakes a systematic dissection of six models, categorised into two distinct groups: baseline models and optimised models.

3.4.2.1 Baseline Models

- **Logistic Regression:** As a probabilistic linear classifier, logistic regression serves as a fundamental technique for modelling the log odds of a dichotomous outcome. Its simplicity offers transparent interpretability, though it may lack the flexibility needed to capture intricate patterns in the data.

- **Decision Trees:** Offering non-linear decision boundaries, decision trees can capture more complex relationships in the dataset. However, their susceptibility to overfitting necessitates caution in their deployment.
- **Random Forest:** By aggregating multiple decision trees, the Random Forest model aims to achieve enhanced generalization and reduced variance, providing a more robust baseline against which optimizations can be benchmarked.

3.4.2.2 Optimised Models

- **Bagging with Decision Trees:** This approach amalgamates the power of multiple decision trees through bootstrap aggregating. The primary objective is to achieve a reduction in variance and an enhancement in model stability.
- **Random Forest with SMOTE:** Addressing the challenge of class imbalance, this model employs Synthetic Minority Over-sampling Technique (SMOTE) in tandem with Random Forest. The aspiration is to bolster the recall metric, even if it invites a slight increase in false positives.
- **Random Forest with Cost-sensitive Learning:** Here, misclassification costs are asymmetrically allocated to accentuate the penalty for incorrectly classifying "Conflict" instances. This approach seeks a careful balance, emphasizing the importance of correct conflict predictions over the inadvertent misclassification of peaceful instances.

Chapter 4

Results and Discussion

The heart of this research lies in the extensive modelling and evaluation processes we embarked upon. This section aims to revisit the methodologies we've employed, casting light on their results and deriving pertinent insights from them. The exploration began with a diverse set of modelling techniques – from logistic regression, a foundational statistical method, to decision trees which offer intuitive decision-making insights. We extended our exploration to ensemble methods like Random Forest, which harnesses the power of multiple decision trees, and Bagging. Additionally, we probed into techniques to handle class imbalance, an oft-neglected yet critical aspect in conflict prediction. This included methods like Synthetic Minority Over-sampling Technique (SMOTE) and cost-sensitive learning. Each method was meticulously evaluated against a set of performance metrics to ensure a holistic understanding.

4.1 Exploratory Data Analysis (EDA)

The dataset under investigation encompasses an extensive collection of 447,078 individual data entries. During the rigorous data cleansing process delineated in the Methodology section, modifications were made to the categorical response variable. This variable initially possessed two distinct categorical labels: "War" and "No War". For the sake of computational efficiency and to facilitate straightforward interpretation, these labels were systematically transformed

Table 4.1: Features and Definitions

Features	**Description**
GDP Ratio	This metric encapsulates the relative economic magnitudes of two countries. A pronounced disparity might be indicative of latent economic frictions or imbalances.
GDP per Capita Ratio	This parameter offers a nuanced understanding of the economic prosperity of the populace in respective countries.
Military Expenditure Ratio	A salient indicator, it potentially signifies power imbalances or a veiled arms escalation between juxtaposed nations.
Democracy Differential	Disparities in this metric can suggest divergent political paradigms and governance modalities.
Share Border or not	Geographical contiguity, by virtue of shared borders, could be a harbinger of conflicts, especially in the context of territorial altercations.
Political Instability	This factor, by mirroring the internal political dynamism, provides insights into a nation's propensity for external confrontations.

Table 4.2: Missing Value Counts and Proportions

Variable	**Missing Count**	**Missing Proportion (%)**
Year	0	0.0
Country A	0	0.0
Country A Code	0	0.0
Country B	0	0.0
Country B Code	0	0.0
Conflict or not	0	0.0
gdp_A	29051	4.97
gdp_per_capita_A	29051	4.97
gdp_B	8868	1.52
gdp_per_capita_B	8868	1.52
Military Expenditure in GDP_A	112788	19.30
Military Expenditure in GDP_B	107225	18.34
Democracy_level_A	35535	6.08
Democracy_level_B	34710	5.94

into standardised numerical indicators, with "War" represented by the integer '1' and "No War" by '0'. A brief introduction of each independent Variables is shown in Table 4.1

Before further EDA, it was found that there were some noticeable missing values in each variable as shown in Table 4.2. After dealing with the missing values as mentioned in the Methodology chapter, the remaining countries can be seen in Figure 4.1. Countries covered in this research are coloured with purple meaning "War" and blue meaning "No War", while countries that are

Table 4.3: Statistical Descriptions

	count	mean	std	min	25%	50%	75%	max
Year	447078	1997	13	1970	1988	1999	2009	2018
War or not	447078	0.00015	0.012422	0	0	0	0	1
gdp_ratio	447078	15	102	0.00005	0.1223	0.7478	4.5031	7534.6139
gdp_per_capita_ratio	447078	3	6	0.0047	0.2762	0.8980	2.8726	154.8462
Military_expenditure_ratio	447078	48	3157.92	0	0.4793	0.9567	1.9390	961947.8036
Democracy_differential	447078	7	6	0	2	6	13	20
Share Border or not	447078	0.0231	0.1501	0	0	0	0	1
Political Instability	447078	0.0751	0.2729	0	0	0	0	2

removed would be in grey. Also, the statistical descriptions of each feature are presented in Table 4.3.

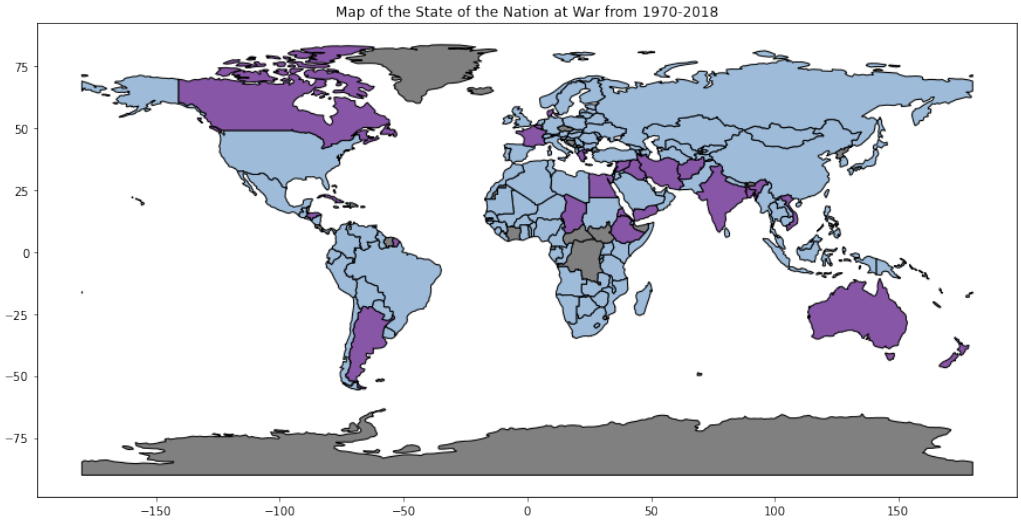


Figure 4.1: Map of the State of the Nation at War from 1970-2018

A numerical summary of the data provides a stark observation regarding the distribution of conflict instances. As vividly depicted in Figure 4.2, there exists a pronounced imbalance between the two classes, the vast majority of dyadic country pairs in the dataset did not go to war, while only a tiny fraction of the pairs went to war. This disparity is not just a numerical distinction but has methodological implications, potentially influencing the predictive capabilities of the models trained on such imbalanced data.

Moving on to the predictor variables, a preliminary assessment of their distribution, as visualised in Figure 4.3, underscores that none of them adheres strictly to a normal distribution. This deviation from normality is crucial, as the assumptions of many statistical procedures

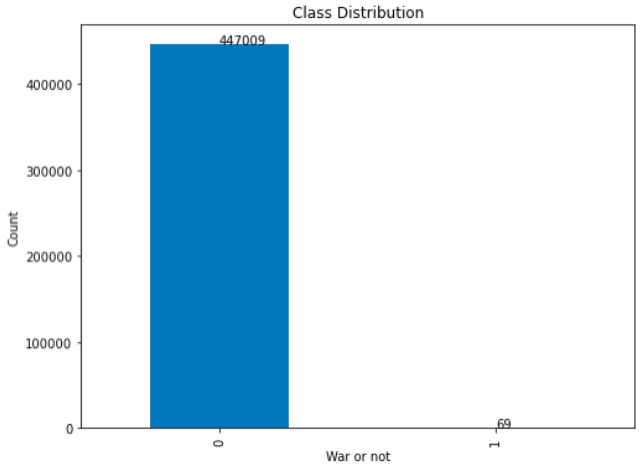


Figure 4.2: Response Variable Distribution

and algorithms are predicated on the data being normally distributed (Johnson & Wichern, 2007). The distribution of GDP ratios and GDP per Capita ratios between the two countries is highly right-skewed. Most country pairs have low GDP and GDP per Capita ratios, with few exceptions. The variable "Military expenditure ratios" is also right-skewed. Most country pairs seem to have comparable military expenditures as a percentage of their GDPs. The distribution of democracy differential between the two countries is more spread out, indicating variations in democracy scores between countries. For the last two variables, most country pairs do not share borders, which is evident from the tall bar at 0 and exhibit no political instability, with fewer showing instability in one country and even fewer showing instability in both.

To discern potential relationships or associations among the variables in our dataset, we constructed a heatmap visualising the correlation coefficients for each variable pairing. As exhibited in Figure 4.4, the results of this exercise are quite revealing. Most variables have low correlations with the target variable 'War or not'. 'Share Border or not' variable has the highest correlation with it, although it's still quite low. Some predictors, such as 'gdp_ratio' and 'gdp_per_capita_ratio', are weakly positively correlated. Predominantly, the correlation values gravitate around zero, suggesting a lack of strong linear relationships between most pairs of variables. This insight can have implications for the modelling process, especially for models that assume or thrive on multicollinearity (Field, 2013).

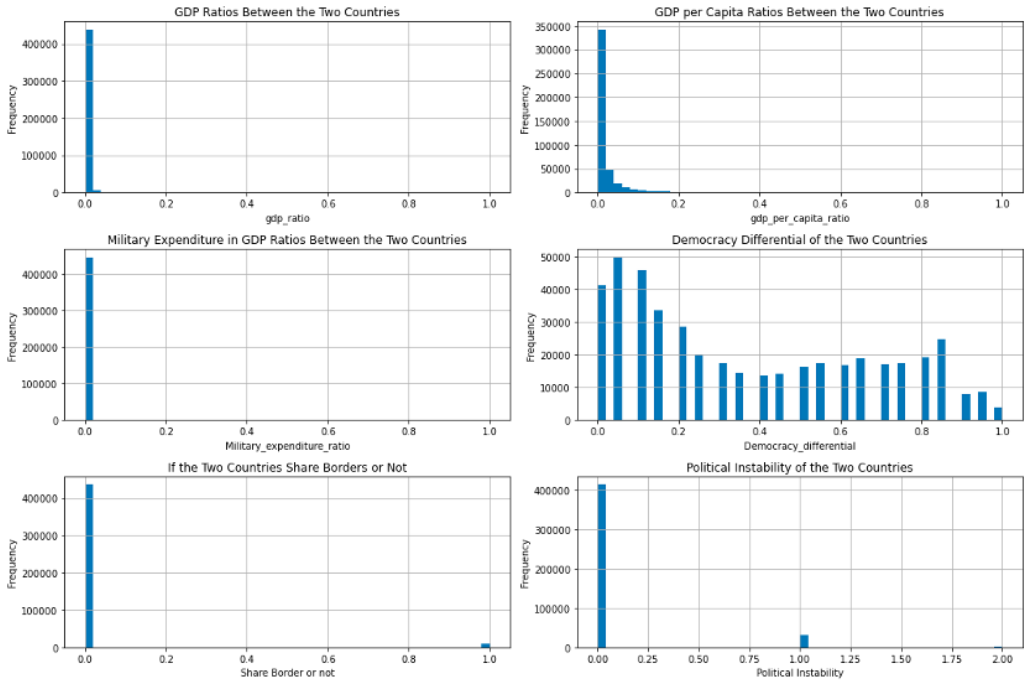


Figure 4.3: Predictors Distribution

As shown in Figure 4.5, the "gdp_ratio" and "gdp_per_captia_ratio" for country pairs with wars had a slightly lower median, but both groups had wide distributions with many high outliers, same for "Military_expenditure_ratio", while the "Democracy_differential" was relatively similar for both groups, although country pairs without wars showed a slightly lower median. To better understand the characteristics and differences between country pairs with war happened, a histogram (Figure 4.6) is plotted.

Upon examining the distributions of selected economic and political indicators for country pairs (dyads) that have engaged in wars, several pertinent observations emerge. The majority of the dyads exhibit a skewed distribution towards lower GDP ratios. However, there's a noticeable peak around the ratio of 1, suggesting that a subset of these warring dyads comprises countries with comparable GDPs. The plot for GDP per capita displays a relatively uniform distribution across the range. Notably, peaks are observed at the extremities, near ratios of 0 and 1. This indicates the presence of dyads where one country has a significantly higher GDP per capita compared to the other, as well as dyads where the two countries have similar GDP per capita values. As for the military expenditure ratio, a pronounced skewness towards the lower end of

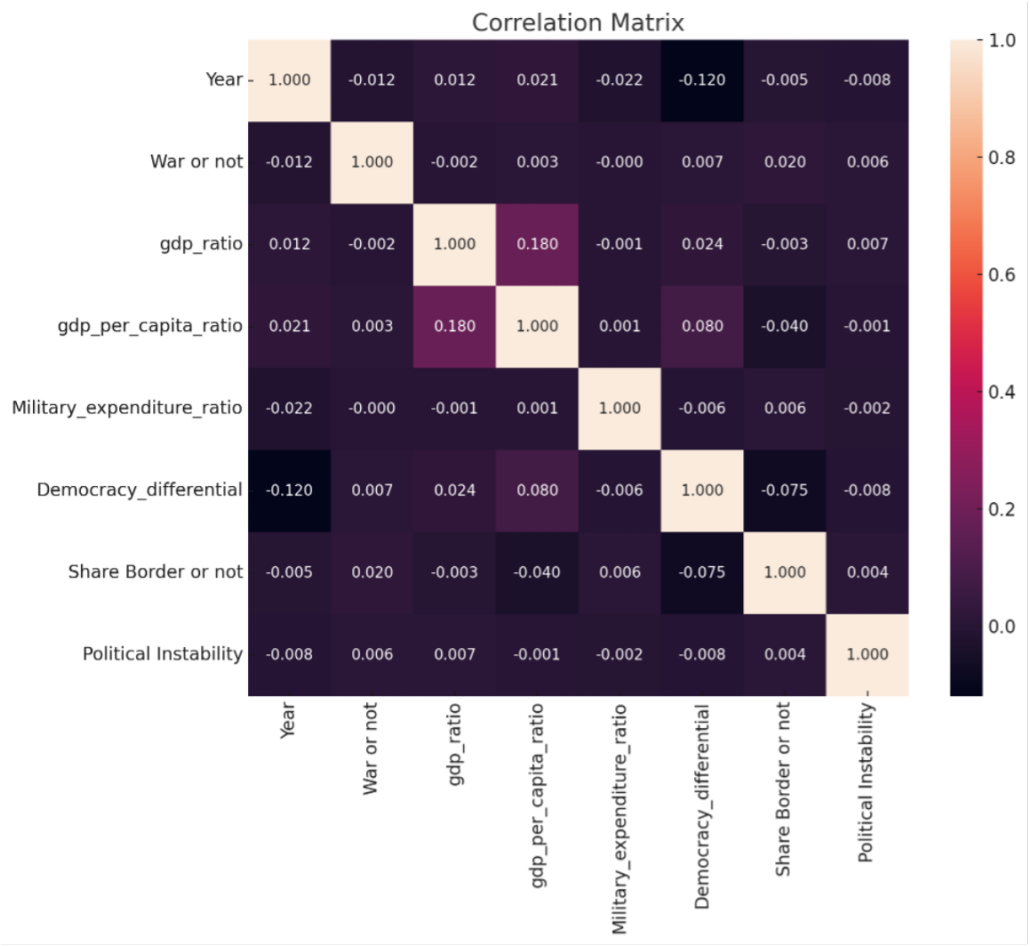


Figure 4.4: Coefficients Between Each Variables

the spectrum characterises the distribution. A limited number of dyads approach a ratio of 1, implying that only a few warring dyads have countries with analogous military expenditures. The distribution for democracy differential is relatively balanced across its range. Nonetheless, a higher concentration of dyads is observed in the mid to lower values of the spectrum.

Examining Figure 4.7, we observe the interplay between countries sharing borders and instances of war. The horizontal axis delineates whether countries share a border or not, while the vertical axis indicates the occurrence of war. The slender bar at the top underscores a notable insight: among country pairs that have engaged in war, there is a higher prevalence of neighbouring countries than non-neighbouring ones.

Transitioning to Figure 4.8, the emphasis shifts to the dynamics of political stability. Here, the horizontal axis represents the political instability status, indicating whether both countries

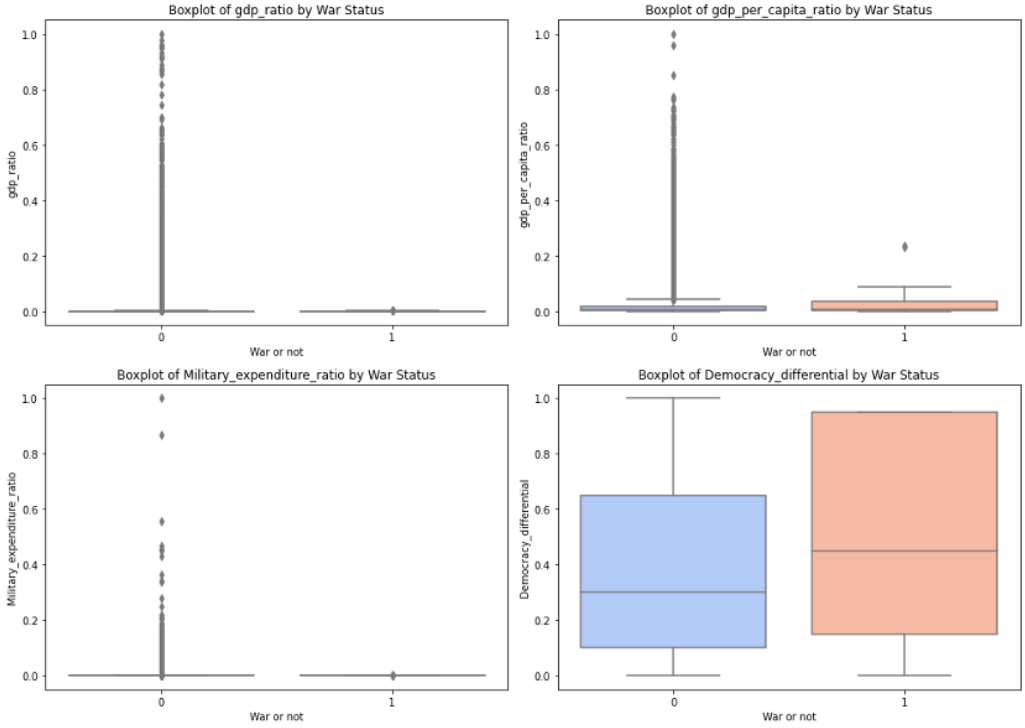


Figure 4.5: Boxplot of Numerical variables

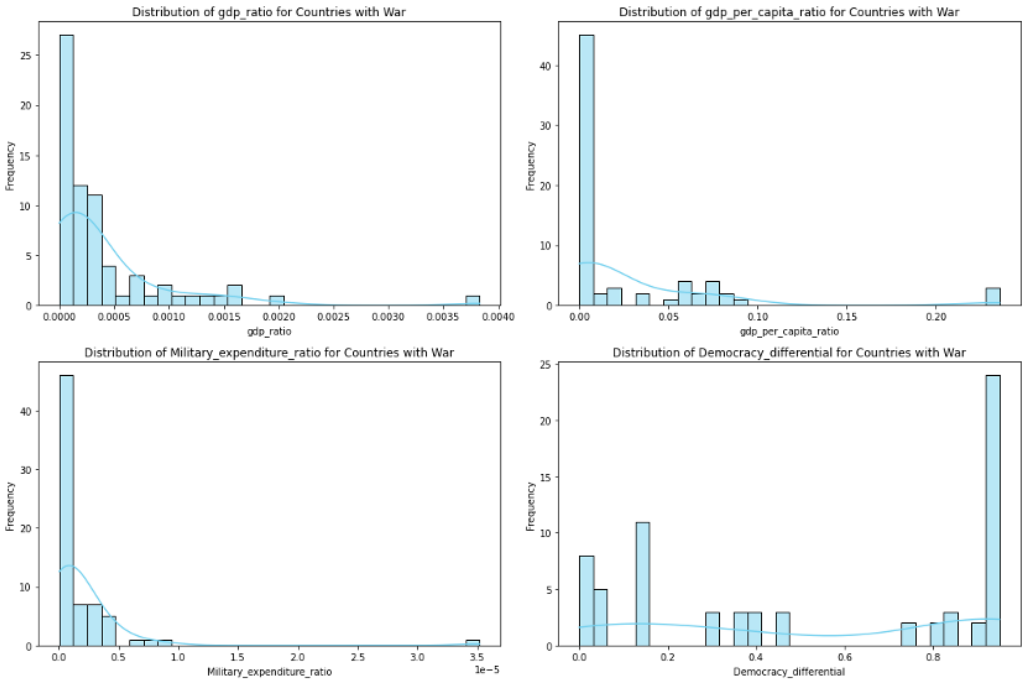


Figure 4.6: Histogram for Numerical Variables Based on the Happened of War

are either politically stable or unstable. The vertical axis, similar to Figure 6, represents the occurrence of war. A discernible observation from this plot is that among country pairs at war, there tends to be greater political instability in one of the countries, as opposed to both being

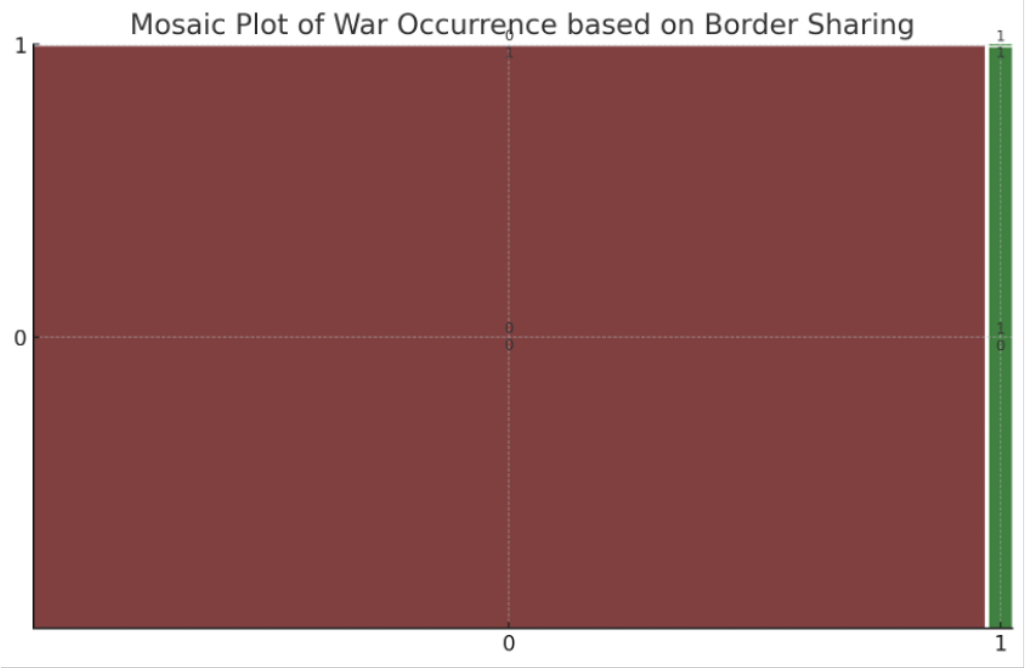


Figure 4.7: Mosaic Plot for Binary Variable “Share Border or not”

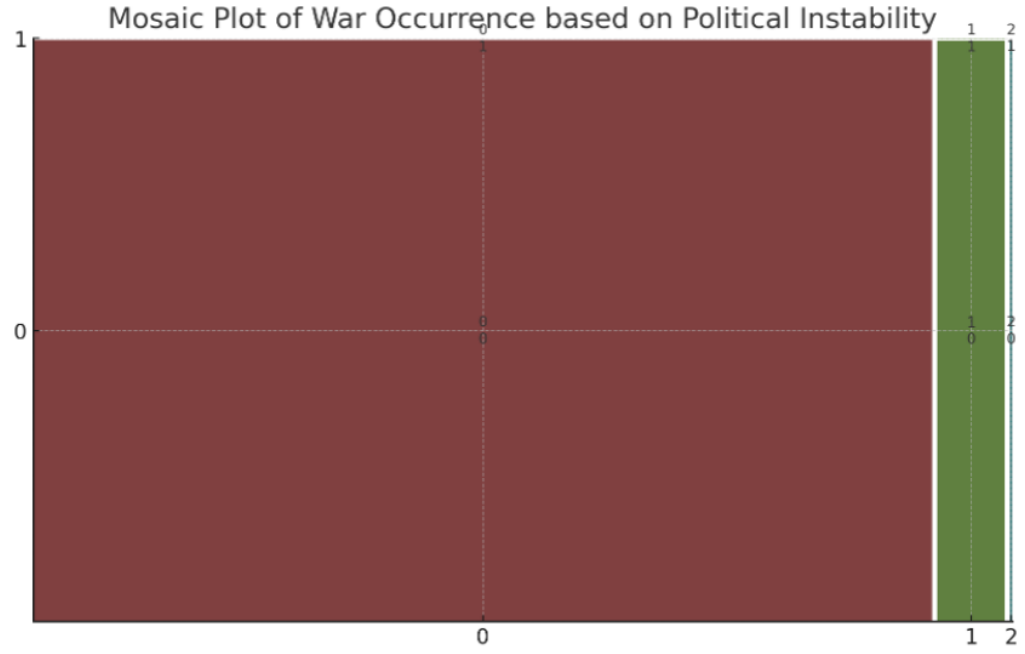


Figure 4.8: Mosaic Plot for Categorical Variable “Political Instability”

simultaneously stable or unstable.

Lastly, Figure 4.9 shows that within the countries that go into war, the total number of occurrences in War data. We can see that Iraq has the highest occurrences of other countries.

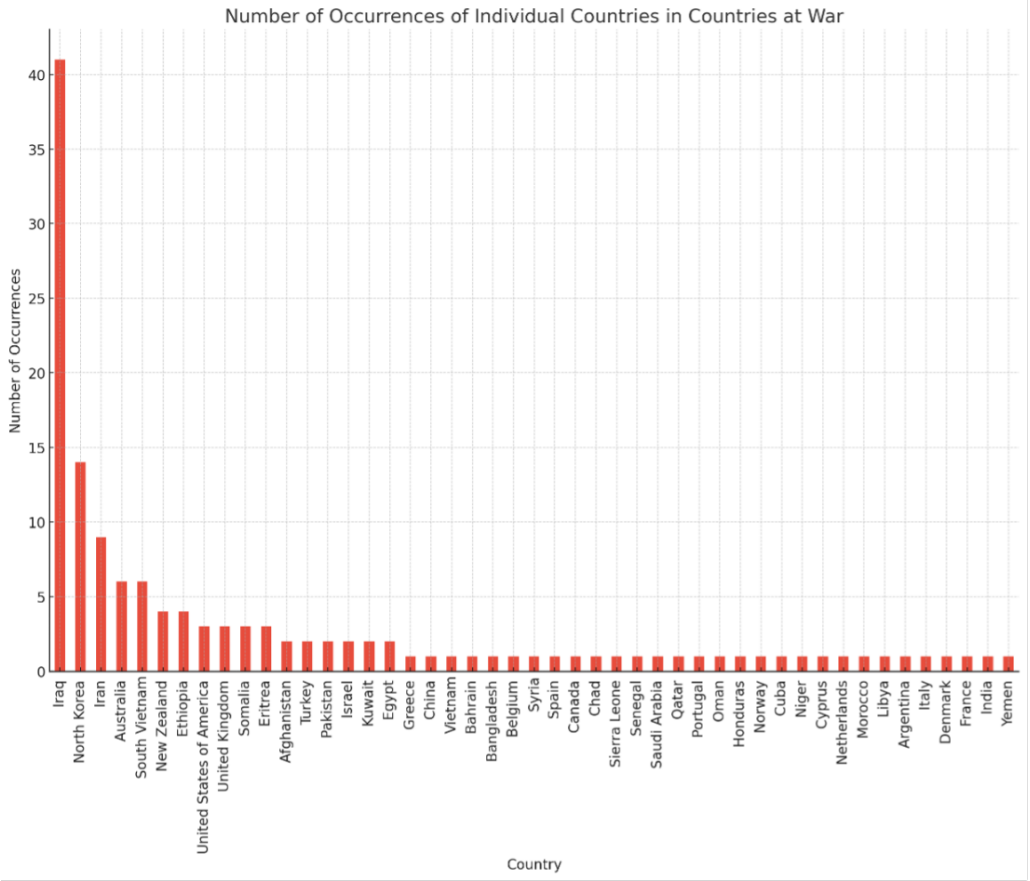


Figure 4.9: Number of Occurrences of Individual Countries in Countries at War

4.2 Modelling and Analysis

4.2.1 Logistic Regression

The first model we implemented was Logistic Regression(LR). This algorithm is particularly well-suited for binary classification problems like ours. Understanding the relationship between our independent variables and the dependent binary outcome, provided us with a baseline performance and also allowed for the interpretation of feature coefficients, offering insights into the importance of each predictor. The evaluation metrics of LR are shown in Table 4.4. An accuracy of 75.82% means that in about three out of every four instances, the model correctly predicted the outcome. While this seems like a reasonable metric at first glance, given the severe class imbalance in the dataset, accuracy might not be the most informative metric. It's

Table 4.4: Results from the Logistic Regression Model

Metric	Value
Accuracy	75.82%
Precision	99.98%
Recall	75.82%
F1 Score	86.23%
AUC	82.61%

possible to achieve high accuracy by merely predicting the majority class. Therefore, while the accuracy is decent, it's essential to consider other metrics for a comprehensive evaluation.

A precision of nearly 100% indicates that almost every time the model predicted a war, it was correct. This means there were very few false positives. However, given the class imbalance, this high precision came at the cost of recall, which is evident from the number of false negatives.

Next, the recall value is the same as the accuracy in this context. This means that out of all the actual wars, the model could correctly predict approximately three out of every four. Given the potential consequences of not predicting a war, maximising recall might be a priority, even at the expense of precision.

Moving on, an F1 Score of 86.23% suggests that the model found a decent balance between precision and recall, leaning more towards precision. In datasets with an uneven class distribution, the F1 score can be more informative than accuracy.

Finally, an AUC of 82.61% is considered good and indicates that the model has a good measure of separability. It means there's an 82.61% chance that the model will be able to distinguish between a random positive and a random negative observation. This suggests that the model, overall, is capable of identifying patterns that differentiate wars from non-wars.

However, the confusion matrix (Figure 4.10) reveals some challenges. While the model was capable of correctly predicting 11 out of 14 wars, it also produced 21,618 false positives. This high false positive rate might be concerning in certain contexts, as predicting a war when there is not one can have significant implications.

The Feature Coefficients plot (Figure 4.11) provides a graphical representation of these relation-

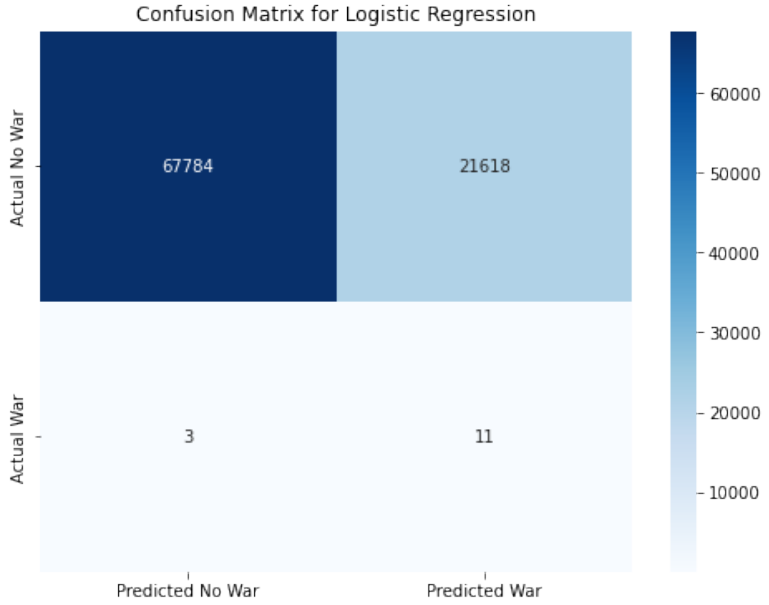


Figure 4.10: Confusion Matrix for Logistic Regression

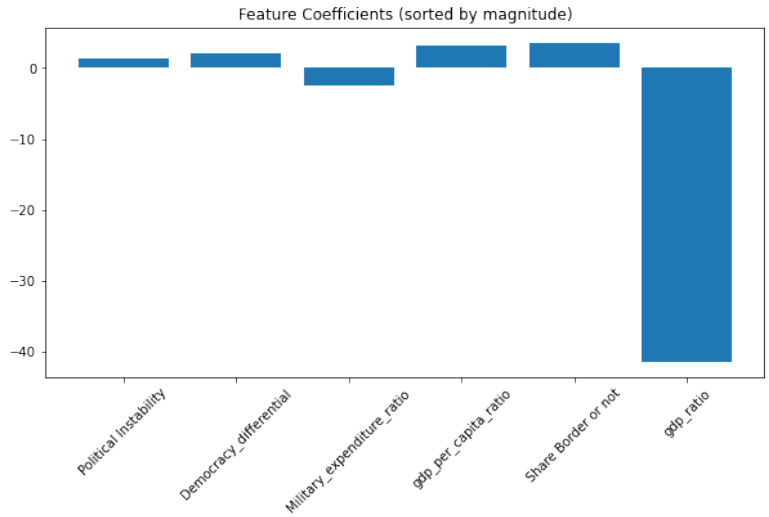


Figure 4.11: Feature Importance for Logistic Regression

ships. The magnitude and direction (positive or negative) of each feature's coefficient can be clearly seen, giving a straightforward interpretation of each feature's influence on the outcome. Specifically, the 'gdp_ratio' had the most substantial negative coefficient, suggesting that as this variable increases, the likelihood of a war occurring decreases. Moreover, 'Share Border or not' and 'gdp_per_capita_ratio' had positive coefficients, indicating that an increase in these variables might increase the probability of war.

4.2.2 Decision Tree Classifier

The Decision Tree Classifier (DT) was chosen due to its interpretability and ability to handle both numerical and categorical variables. To address the class imbalance, the `class_weight='balanced'` parameter was utilised. This ensures that the algorithm is more sensitive to the minority class.

Table 4.5: Results from the Decision Tree Classifier

Metric	Value
Accuracy	99.99%
Precision	99.98%
Recall	99.99%
F1 Score	99.98%
AUC	64.28%

Presented in Table 4.5, DT achieved an accuracy of 99.99%, indicating that the model correctly predicted the outcome nearly all the time. However, given the class imbalance, it's crucial to investigate other metrics. 99.98% of precision suggests that almost every time it predicted a war, it was accurate. This minimises the number of false alarms (false positives). Next, the recall value suggests that the model identified almost every actual war instance in the test set. This is crucial for the problem at hand as failing to predict an actual war can have crucial implications. The AUC score, representing the model's ability to differentiate between the positive and negative classes, is considerably lower compared to other metrics, which can be attributed to the class imbalance issue.

In addition, the confusion matrix (Figure 4.12), a staple in our performance evaluation, offers a granular view of the model's predictions. It shows that the model correctly identified 4 out of 10 instances of wars, and correctly identified a significant number of occasions as non-wars. While the number of false negatives is relatively low, in the context of predicting wars, each missed instance can have important implications. This once again underscores the importance of recall in our analysis, a sentiment echoed by our discussion on the LR model.

Presented in Figure 4.13, gdp_per_capita_ratio had the highest importance, suggesting that

Figure 4.12: Confusion Matrix for Decision Tree Classifier

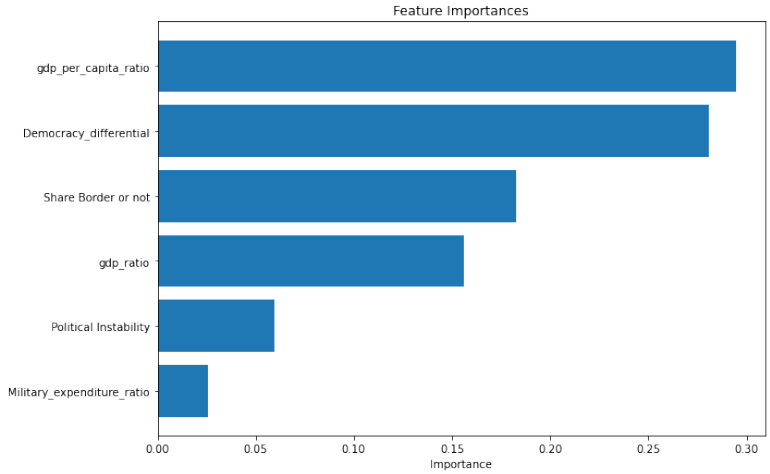


Figure 4.13: Feature Importance for Decision Tree Classifier

the GDP per capita ratio between countries plays a significant role in predicting the occurrence of wars. Then the difference in the democracy level between countries was the second most influential factor for DT. Besides, whether countries share a border also substantially influenced the model's decisions. Unlike LG, other features like gdp_ratio, Political Instability, and Military_expenditure_ratio had a lesser but non-trivial influence on the model's predictions.

Figure 4.14 shows the tree structure of DT. The tree structure visualisation provides a graphical representation of the decision-making process. The DT's root node split on the "Democracy_differential" feature, which had a threshold value of $<= 0.925$. This means that the very

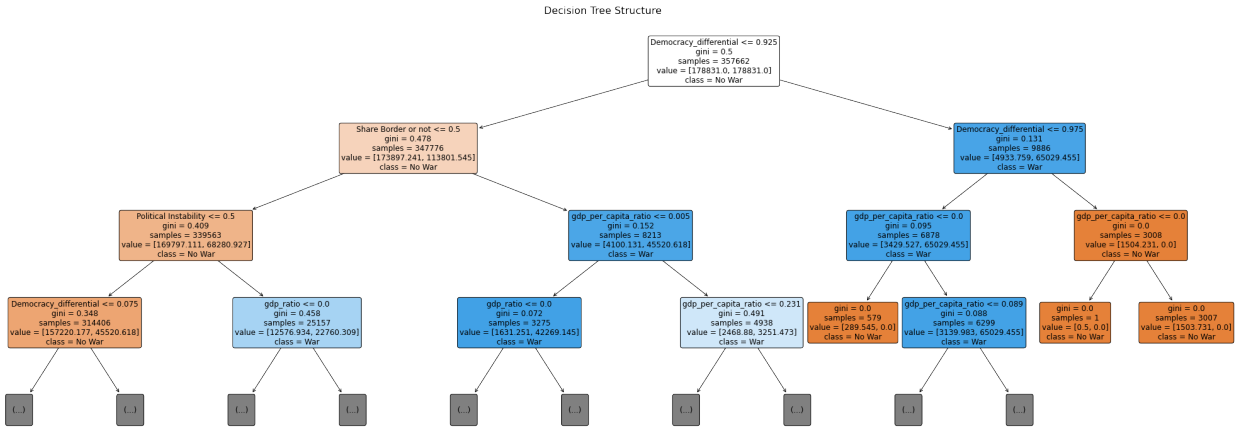


Figure 4.14: Tree Structure for Decision Tree Classifier

first decision the tree made was based on this feature. As descended the tree, other splits were made on features such as "Political Instability" and "gdp_per_capita_ratio".

Interestingly, while "gdp_per_capita_ratio" emerged as the most important feature in terms of global feature importance, it was not the feature chosen for the root node. This leads to an essential understanding of feature importance versus decision tree splits. The root node's selection is based on the feature that offers the maximum reduction in impurity for the initial split. In this dataset's context, "Democracy_differential" provided the best initial split. On the other hand, "gdp_per_capita_ratio" might have frequently contributed to splits at various positions in the tree, each time significantly reducing impurity. Cumulatively, this made it the most important feature.

4.2.3 Random Forest Classifier

The Random Forest classifier (RF) is an ensemble method that aggregates the predictions of multiple decision trees to produce a more stable and accurate model. This method benefits from the strengths of decision trees while mitigating their tendency to overfit, as individual trees in the ensemble are trained on random subsets of the data.

After employing the Gini impurity criterion and a grid search approach to optimise hyper-parameters, the performance metrics are shown in Table 4.6. These metrics highlight RF's

if we always predict True no matter what - our Recall would be 1 because there would be cases where model preditealse

Table 4.6: Results from the Random Forest Classifier

Metric	Value
Accuracy	97.95%
Precision	99.98%
Recall	97.95%
F1 Score	98.95%
AUC	94.72%

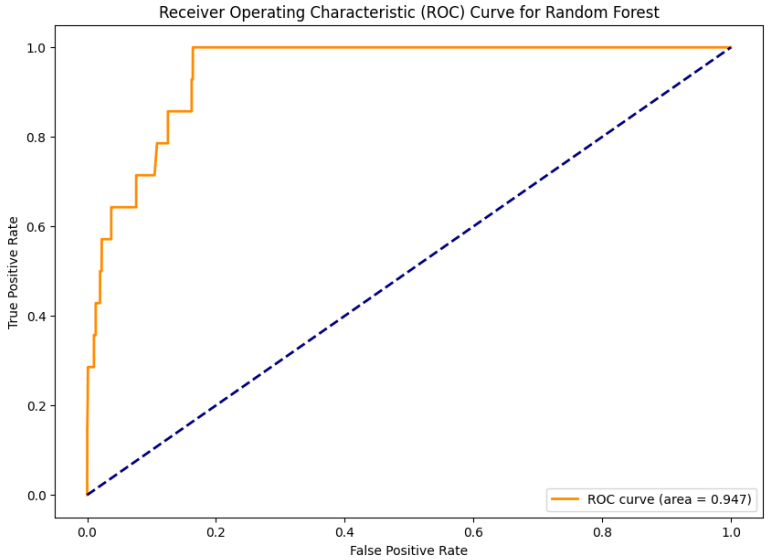


Figure 4.15: Receiver Operating Characteristic (ROC) Curve for Random Forest

effectiveness in differentiating between the two classes, with a commendable AUC of 94.72% indicating its capability to distinguish between positive and negative classes. The ROC curve (Figure 4.15) further reinforced the model's efficacy. With an area under the curve (AUC) of 0.947, it showcases RF's robustness in accurately predicting the likelihood of a war over various decision thresholds.

From the confusion matrix (Figure 4.18), we observed that the model correctly predicted "No War" 87,580 times and "War" 7 times. However, there were 1,822 false positives, where the model wrongly predicted "War" when there was none, and 7 false negatives, where actual wars were missed.

Moving on to the feature importance (Figure 4.19), it is notable that "Democracy_differential", "gdp_per_capita_ratio", and "Share Border or not" emerged as the most crucial predictors in this model as well, aligning with our findings from DT. This consistent importance across

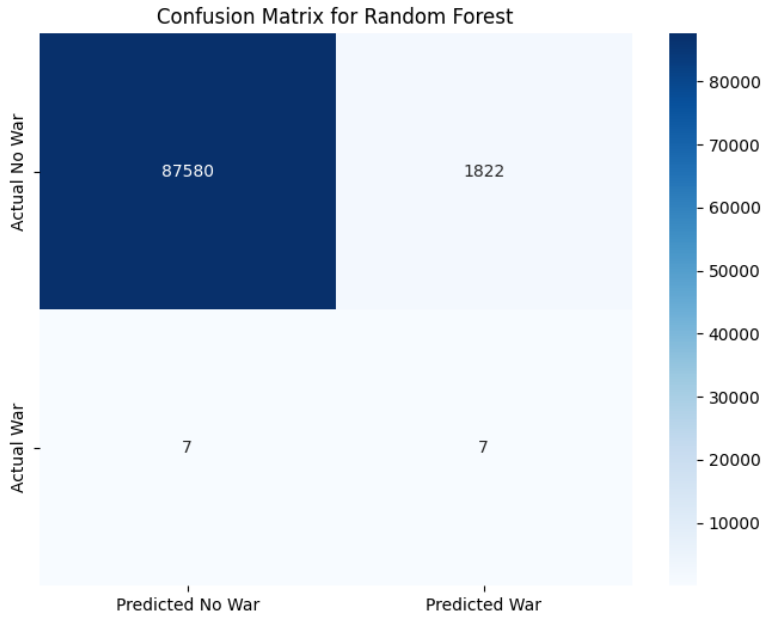


Figure 4.16: Confusion Matrix for Random Forest Classifier

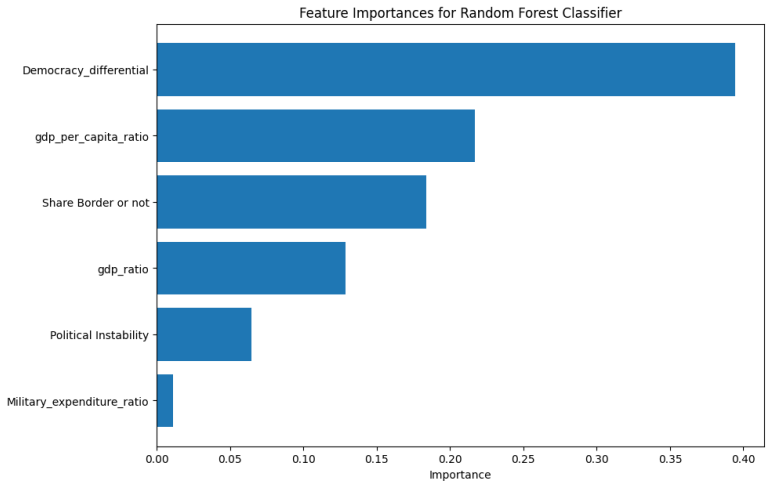


Figure 4.17: Feature Importance for Random Forest Classifier

models underscores its significance in predicting wars.

4.2.4 Bagging on Decision Tree

Bagging, or Bootstrap Aggregating, is an ensemble method that aims to reduce the variance of an individual model by constructing multiple versions of it using random subsamples of the training data. For this study, we applied bagging on a decision tree, leveraging its benefits to enhance the performance of the inherently high-variance decision tree model. By creating

multiple Decision Trees on random subsets of data and averaging their predictions, this ensemble method aims to improve the robustness and accuracy of single-tree predictions.

The base model for bagging was chosen as a decision tree with balanced class weights to handle class imbalance. This tree was then instantiated 100 times using different random subsets of the training data, and their collective output was aggregated to produce the final prediction.

Table 4.7: Results from the Bagging on Decision Tree

Metric	Value
Accuracy	99.99%
Precision	99.98%
Recall	99.99%
F1 Score	99.98%
AUC	71.36%

The model's high accuracy, precision, recall, and F1 score suggest an excellent performance in classifying the instances. However, the AUC, while still good, indicates a modest capability of the model in differentiating between the positive and negative classes.

Figure 4.18: Confusion Matrix for Bagging on Decision Tree

The confusion matrix (Figure 4.18) provides further insights: out of the test instances, the model made only one false positive prediction and missed 12 actual wars. This is quite commendable, but the false negatives (missed wars) highlight areas for potential improvement.

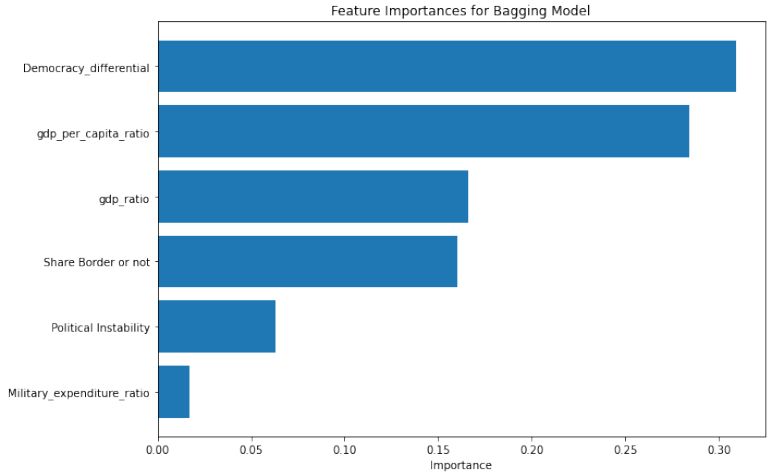


Figure 4.19: Feature Importance for Bagging on Decision Tree

The aggregated feature importance from all decision trees in the ensemble reveals the hierarchy shown in Figure 4.19. The "Democracy_differential" and "gdp_per_capita_ratio" continue to emerge as vital predictors, indicating their consistent importance across different models.

4.2.5 Synthetic Minority Over-sampling Technique (SMOTE) on Random Forest

Given the class imbalance in the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) was utilised in conjunction with the Random Forest model. This method generates synthetic samples in the feature space, aiming to balance out the distribution of our target classes and potentially enhance the model's sensitivity to the minority class.

Table 4.8: Results from SMOTE on Random Forest

Metric	Value
Accuracy	25.83%
Precision	0.02%
Recall	92.86%
F1 Score	0.039%
AUC	71.85%

As shown in Table 4.8, very different from other models mentioned earlier, the SMOTE model's accuracy is notably low, but its recall is exceptionally high. This shows the model's capability

to identify most of the wars correctly, albeit at the cost of a high number of false positives. This trade-off is evident from the precision score and further emphasised by the F1 score.

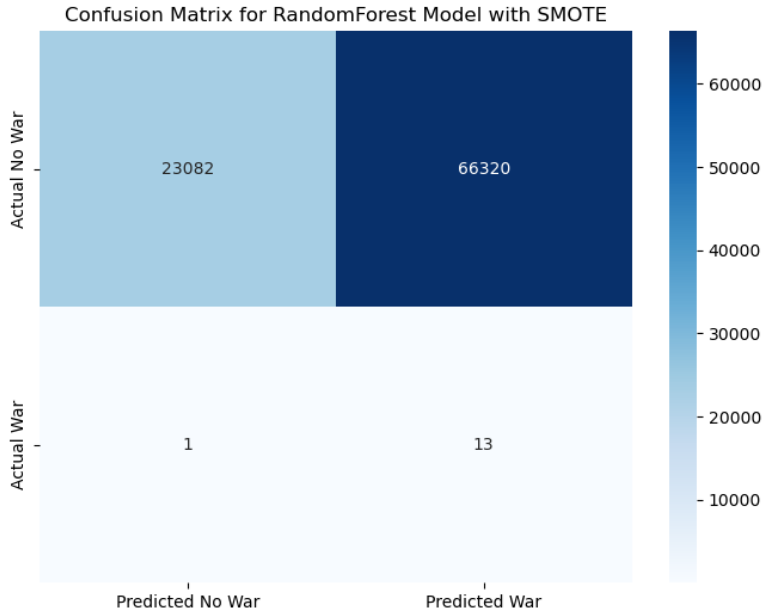


Figure 4.20: Confusion Matrix for SMOTE on Random Forest

The confusion matrix (Figure 4.20) indicates a high number of false positives (66,320), corroborating the low precision. However, it managed to correctly identify 13 out of 14 wars, resulting in its high recall.

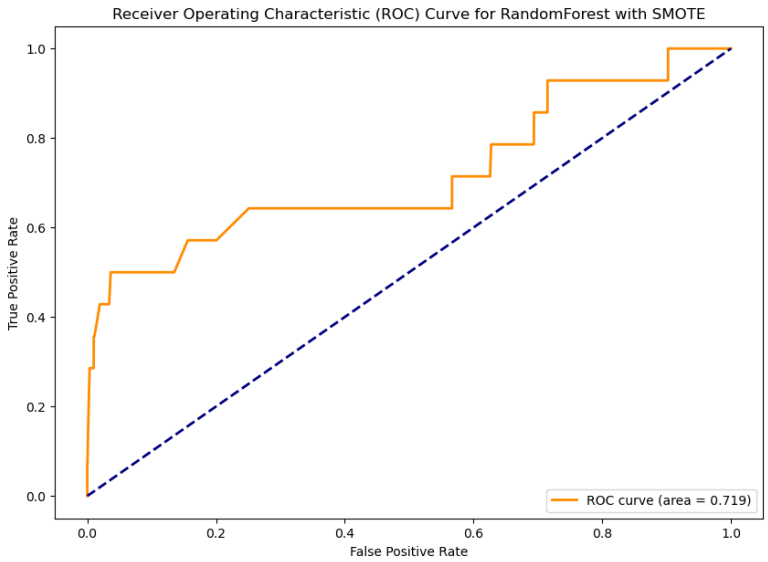


Figure 4.21: Receiver Operating Characteristic (ROC) Curve for SMOTE on Random Forest

The ROC curve (Figure 4.21), with an area of 0.719, showcases the model's proficiency in

predicting the likelihood of war across various decision thresholds, an impressive feat given the dataset's imbalanced nature.

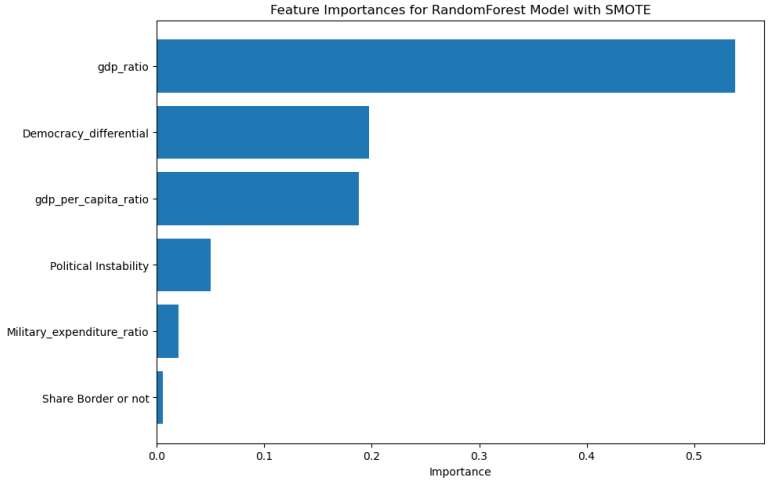


Figure 4.22: Feature Importance for SMOTE on Random Forest

4.2.6 Cost-sensitive learning on Random Forest

Another approach to tackle the class imbalance issue was to implement cost-sensitive learning (CSL). CSL is an essential approach when the misclassification costs of different classes are unequal. In the context of our dataset, the cost of misclassifying a war event (a positive class) is significantly higher than misclassifying a non-war event. This is because failing to predict a war could have dire consequences, whereas falsely predicting a war might only result in increased vigilance. Thus, the aim is to reduce the false negatives by making the model more sensitive to the positive class.

Table 4.9: Results from CSL on Random Forest

Metric	Value
Accuracy	99.94%
Precision	1.32%
Recall	5.0%
F1 Score	2.57%
AUC	86.75%

The accuracy of the model is notably high, registering at 99.41%. However, the recall score of 50% indicates that the model was able to correctly predict only half of the actual war events.

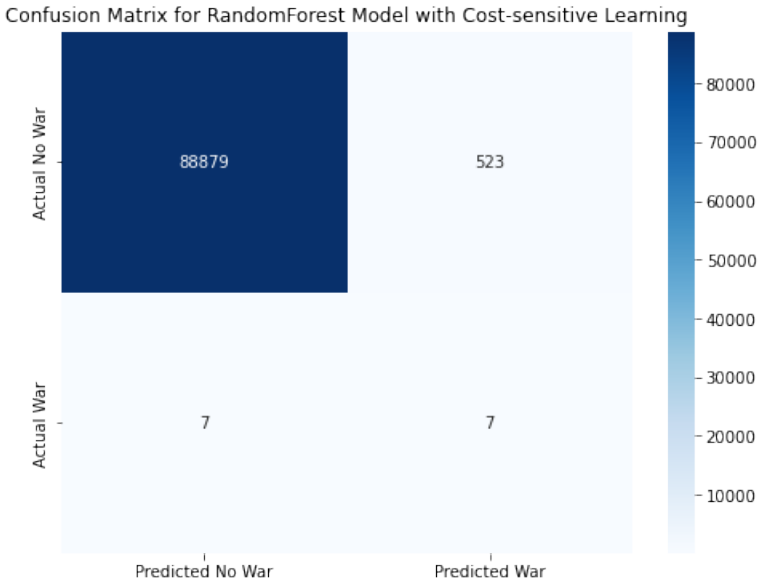


Figure 4.23: Confusion Matrix for CSL on Random Forest

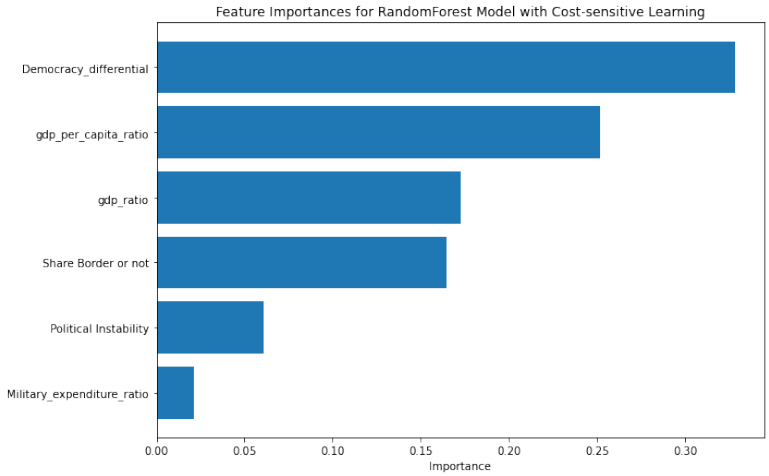


Figure 4.24: Feature Importance for SMOTE on Random Forest

The confusion matrix (Figure 4.23) further elucidates the model's predictions, showing that while it has made significant strides in predicting true wars, it has also over-predicted, leading to a number of false positives. An AUC of 0.867, showcases the model's ability to differentiate between the classes effectively.

Observing Figure 4.24, the 'Democracy_differential' feature remains the most important, followed closely by 'gdp_per_capita_ratio'. This aligns with the notion that political and economic factors play a significant role in the onset of wars.

4.3 Comparative Analysis

4.3.1 Model Performance Metrics Comparison

Upon a detailed analysis of each individual model, we can draw comparisons to gain a holistic understanding of their performance.

As illustrated in Figure 4.25, DT, RF, Bagging with Decision Tree, and CSL models all boast nearly perfect accuracy scores, hovering close to 1. This indicates that these models make correct predictions for both positive and negative classes most of the time. LR presents a slightly diminished accuracy of approximately 0.76, hinting at potential misclassifications. SMOTE with Random Forest lags with the lowest accuracy, around 0.26, indicative of frequent erroneous predictions, possibly due to the oversampling of the minority class.

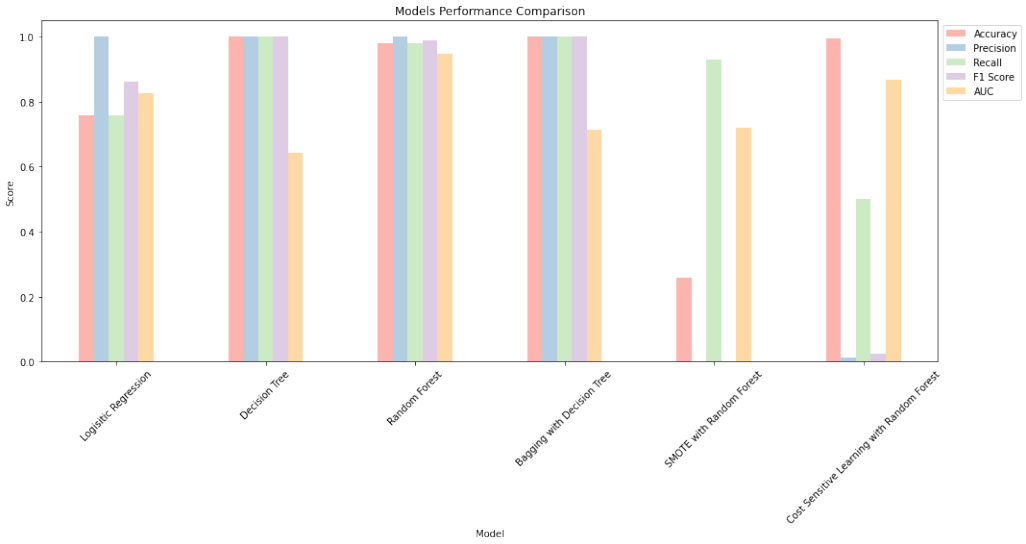


Figure 4.25: Models Performance Comparison

DT, RF, and Bagging with Decision Tree excel with almost perfect precision scores, implying their near-faultless predictions, especially in identifying non-war instances. LR follows closely. In stark contrast, SMOTE with Random Forest struggles with a nearly zero precision score, suggesting a high rate of false positives in its war predictions. Similarly, CSL has a precision of 1.3%.

DT, RF, and Bagging with Decision Tree again have very high precision scores, close to 1, which means that when these models predict a war, they are almost always right, at least in predicting countries that would not have war. Same, for LR nearly perfect. However, SMOTE with Random Forest has extremely low precision, almost 0, suggesting that while it might predict many wars, the majority of them are false positives. Similarly, CSL has a precision of 0 as it never predicts a single war correctly.

A pivotal metric, recall measures the model's adeptness in accurately identifying all relevant instances. The importance of recall is paramount for predicting actual wars. SMOTE with Random Forest outperforms others in recall, although at the expense of precision. LR offers a balanced recall, harmonising with its accuracy.

Representing the harmonic mean of precision and recall, F1 score balances the two metrics. RF claims the top spot with the highest F1 score, epitomising an exemplary equilibrium between precision and recall. In contrast, both SMOTE with Random Forest and CLS models have lower F1 scores, signalling potential imbalances in their precision and recall dynamics.

Finally, AUC represents the model's ability to distinguish between positive and negative classes. RF model has the highest AUC, suggesting excellent discriminative power and LR and CSL also display commendable AUC values.

4.3.2 Feature Importance

The success of a machine learning model not only depends on its accuracy but also on its interpretability. Understanding the significance of features (or variables) that the model uses to make predictions can provide insights into the underlying mechanisms of the decision-making process. In our study, which aims to predict wars, we compared the importance of various features across different models.

"Democracy_differential" consistently ranks high in importance across all models, underscoring the role of democratic differentials in potentially influencing conflict. A larger differential in

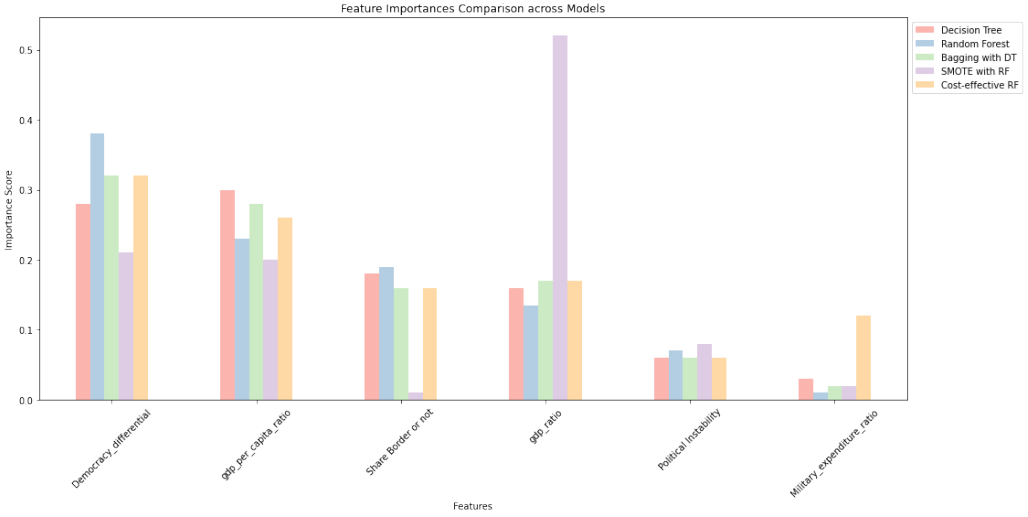


Figure 4.26: Feature Importances Comparison across Models

democracy scores between two countries could lead to misunderstandings, differing ideologies, and ultimately, conflict.

Economic indicators, such as "gdp_per_capita_ratio", hold substantial weight in almost all models. This suggests that disparities in economic prosperity could be a significant factor in driving conflicts. Countries with vast economic differences might view each other with suspicion or competition, potentially leading to tensions.

The significance of the feature "Share Border or not" varies across models. It holds high importance in DT and RF but is much less significant in the SMOTE with Random Forest model. This suggests that while sharing a border can be a potential cause for conflict, other factors may overshadow its significance depending on the model's configuration.

Another economic feature, "gdp_ratio", is deemed most crucial in the SMOTE with Random Forest model. This feature's consistent ranking across other models signifies the role of economic power and disparities in predicting wars.

While political instability can lead to internal conflicts, our models suggest it's not a primary driver for wars between nations. Its importance is relatively lower across all models. While not the most significant feature in any model, political instability consistently appears as a factor, indicating that internal political turmoil can be a precursor or contributor to external conflicts.

Interestingly, the amount a country spends on its military does not hold as much weight as one might expect, given that military strength and expenditure are often directly linked to a country's propensity for conflict. However, it's still a factor, especially in the Cost-effective Learning with Random Forest model.

In conclusion, while certain features like 'Democracy_differential' and 'gdp_per_capita_ratio' are consistently deemed important across models, others like 'Share Border or not' and 'Military_expenditure_ratio' vary in their significance. This emphasises the need for a multifaceted approach to conflict prediction, considering a range of economic, political, and geographical factors. It also underscores the importance of model selection and tuning in determining feature importance.

4.3.3 Trade-offs

To predict actual war without sacrificing too much precision, there would certainly be a trade-off between Precision and Recall. As depicted in Figure 4.27, LR achieves a harmonious balance between precision and recall, making it effective in terms of predicting actual wars correctly without flagging too many false positives. Both DT and Bagging with Decision Tree showcase nearly impeccable precision and recall scores. Their coinciding points on the plot underline their similar performance metrics. This resemblance in their outcomes is anticipated since bagging aims to diminish the variance of a decision tree by averaging the results from multiple decision trees.

Meanwhile, RF slightly trails behind the DT in terms of recall but sustains commendable precision. This hints at RF's capability to predict actual wars accurately while keeping false positives in check. SMOTE with Random Forest exhibits a recall rate approaching 1, implying its efficiency in identifying most of the actual war instances. However, its precision is remarkably low, signifying a substantial number of false positives. Such a trade-off arises due to the employment of the SMOTE technique, which focuses on addressing imbalanced datasets by generating synthetic samples for the minority class. While CSL maintains a reasonably high

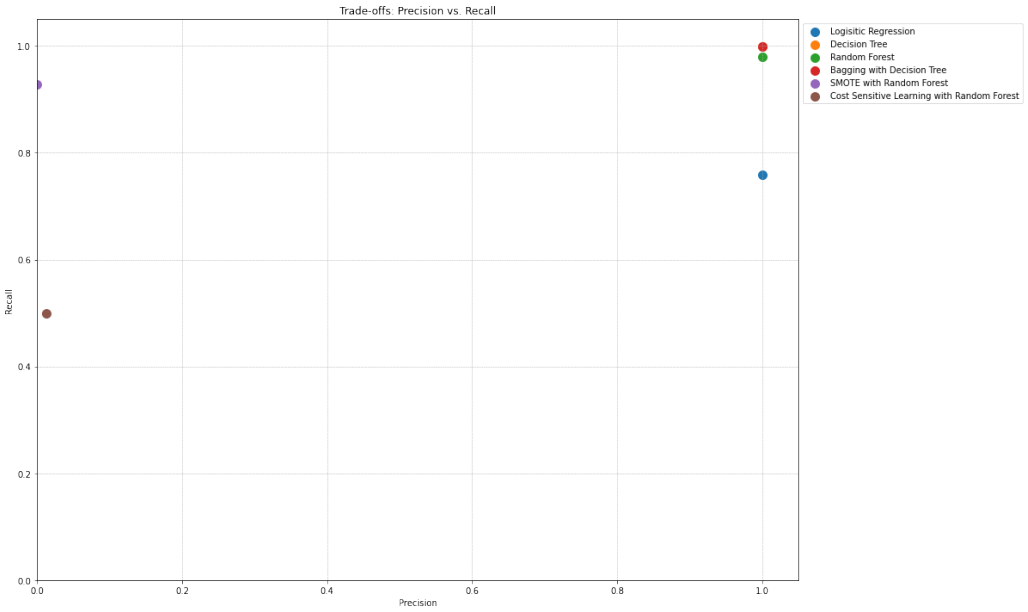


Figure 4.27: Precision vs. Recall for Different Models

precision of 1.3%, its recall stands at 50%. This means that while the model can predict actual wars, it might still miss half of them.

4.3.4 Discussion on Special Techniques

4.3.4.1 Bagging

Decision Trees, while interpretable, can sometimes overfit to the training data, capturing noise and reducing their generalisation capability. While Bagging involves creating multiple subsets of the original data (with replacement) and training a decision tree on each subset. The final prediction is an aggregation (majority vote) of predictions from all trees. This reduces variance, prevents overfitting, and improves the model's overall robustness and accuracy.

Table 4.10: Confusion Matrix Comparison for Decision Tree and Bagging on Decision Tree

Metric	Decision Tree Model	Bagging on Decision Tree Model
True Negatives (TN)	89399	89401
False Positives (FP)	3	1
False Negatives (FN)	10	12
True Positives (TP)	4	2

When comparing the two models in this study, the performance of decision tree Bagging is very close to the basic decision tree model, but its AUC is improved, which indicates that Bagging technology does enhance the generalization ability of the model. However, if comparing their confusion matrix (Table 4.10), we can see that the single decision tree model correctly identified more positive samples (war events) than the bagging model (4 vs. 2). The bagging model made fewer mistakes in falsely identifying negative samples as positive (1 vs. 3). Nevertheless, as mentioned before, the bagging model has a higher AUC, suggesting its capability to rank positive instances higher than negative instances is slightly better, even though it made more mistakes in classification.

As the main purpose of this research is to predict as many wars as possible without compromising too much precision, the single Decision Tree model seems to be a better choice here.

4.3.4.2 SMOTE

SMOTE addresses the class imbalance issue by generating synthetic samples in the feature space. It does this by selecting two or more similar instances (using a distance measure) and perturbing an instance one attribute at a time by a random amount within the difference to the neighbouring instances. This approach increases the number of war cases, balancing the dataset and enabling the model to learn better.

Comparing the evaluation metrics to its baseline model RF, although SMOTE does not seem to perform better, and its precision and F1 Score are nearly 0, the confusion matrix (Figure 4.20) suggests that it predicted almost all the war correctly. However, in the meantime, it also predicted too many False Positives which might give policymakers some wrong ideas.

4.3.4.3 CSL

Standard models assume equal costs for false positives and false negatives, which is not always the case. In predicting wars, a false negative (not predicting a war when it happens) could be far more costly than a false positive. By assigning different costs to different types of errors,

cost-sensitive learning allows the model to focus on reducing the more costly error type. This is done by introducing weights for the classes or changing the algorithm to be aware of the costs associated with misclassification.

Table 4.11: Confusion Matrix Comparison for Random Forest and CSL on Random Forest

Metric	Random Forest Model	CSL on Random Forest
True Negatives (TN)	87580	88879
False Positives (FP)	1822	523
False Negatives (FN)	7	7
True Positives (TP)	7	7

CSL has shown improved results than its base model RF (Table 4.11). While it has the same number of True positives, i.e. at predicting actual wars, it tends to less over-predict the number of false positives than the base model. This could be a result of the model being trained to be more sensitive to the minority class (wars) in the imbalanced dataset. Adjusting the cost-sensitive parameters or further tuning might help in achieving a better balance between precision and recall.

Given the study's primary objective—to predict actual wars accurately without generating too many false alarms— RF emerges as the most suitable choice. It demonstrated a well-rounded performance, striking a balance between high precision and recall, and showcased excellent class-separation capability. While DT and Bagging with Decision Tree models also delivered commendable performances, they had slightly lower AUC scores. Depending on specific contexts and the associated costs of false predictions, one might prioritise certain metrics over others. However, for the aims of this study, RF aligns best with the objective, ensuring precise war predictions without a propensity for over-predicting.

Chapter 5

Conclusion

This chapter will first revisit the research questions to provide a succinct summary of the findings and answer the two questions. Then broader implications of the findings and recommendations based on it will be presented. Lastly, the potential limitations of this research and areas that future studies can explore will be discussed.

5.1 Discussion on Research Questions

5.1.1 First Research Question

Our research embarked on the quest to determine the most suitable machine learning model to predict wars, especially given the challenges posed by datasets with severe class imbalances. Based on the comparative analysis of various models, including Logistic Regression, Decision Trees, Random Forests, Bagging with Decision Trees, SMOTE with Random Forest, and Cost-sensitive Learning with Random Forest, the Random Forest model emerged as the most promising candidate. This model not only demonstrated a high degree of accuracy but also maintained an impressive balance between precision and recall. The ability of the Random Forest model to achieve such performance, especially in the face of class imbalances, underscores its robustness and adaptability.

5.1.2 Second Research Question

Predicting wars is a multifaceted problem, and identifying the most influential predictors is crucial for understanding the underlying dynamics and for improving prediction accuracy. The analysis of feature importance across different machine learning models provided significant insights into the variables that play pivotal roles in war prediction.

This study found that the democracy differential between the two countries showed high importance across models. That is to say, the bigger the difference between the two countries' political systems, the more likely they would have war. It's widely recognised in political science literature that differences in political systems, especially those related to democratic values and institutions, can influence conflict likelihood. Democratic peace theory suggests that democracies are less likely to go to war with each other (Russett & Oneal, 2001). Thus, the difference in democratic values between two countries can indeed be a significant predictor of potential conflicts.

Economic factors have long been linked to war propensity. A country's economic strength or weakness can be both a reason for initiating conflict or avoiding it. Countries with significant economic disparities might have underlying tensions, potentially leading to conflicts (Gartzke, 2007). The GDP per capita can also be indicative of the economic development stage and overall well-being of the country's population, which can influence its foreign policies and war decisions.

Geographical proximity has been a traditional factor in inter-state conflicts. Countries sharing borders have historically had territorial disputes and conflicts, making this a significant feature (Vasquez & Henehan, 2001). However, this feature is not as important as the previous two mentioned above in this research.

Internal political turmoil and instability can have external repercussions. Although countries undergoing political transitions, civil unrest, or other forms of instability might be involved in external conflicts as a way to divert internal tensions or due to weakened diplomatic negotiations

(Hegre, Ellingsen, Gates & Gleditsch, 2001), it turned out to be the second least important factor in this study.

Finally, it seems to be a very important factor as military capabilities and expenditure have been vital factors in war and peace decisions, and a country's military strength, often reflected in its military expenditure, can be a deterrent or a provocation, influencing the likelihood of conflicts (Colaresi & Thompson, 2002). This study surprisingly indicates that this is the least important factor to consider. This might have something to do with the data processing on these variables, which will be discussed later in the limitation section.

5.2 Implications and Recommendations

The significance of economic indicators, particularly GDP ratios and GDP per capita ratios, in predicting the likelihood of wars has been underscored by this study. Countries with pronounced economic disparities or those experiencing rapid economic changes might be more susceptible to conflicts. As supported by Gleditsch (2002), policymakers and international organizations should closely monitor such economic indicators as potential early signs of conflicts.

The relevance of features like democracy differential and political instability between two countries corroborates the assertions made by Hegre et al. (2001) about the importance of socio-political factors in war prediction. This suggests that military and economic indicators, while crucial, need to be complemented with data that captures the socio-political dynamics of countries.

For recommendations, leveraging the predictive prowess of the models, especially the Random Forest model, there's potential to develop early warning systems. Such systems can be instrumental in alerting policymakers and international organizations about potential conflict zones and facilitating proactive interventions (Brandt, Freeman & Schrodt, 2011). Besides, the importance of accurate data collection, especially in the context of "Military_expenditure_ratio", cannot be understated. As highlighted by M. D. Ward, Greenhill and Bakke (2010), refined

data collection methodologies, consistent metrics, and judicious handling of missing values can significantly bolster model accuracy.

Moreover, the dynamic nature of geopolitics necessitates continuous model updates. Models trained on historical data might not always remain relevant for predicting future conflicts, emphasising the need for periodic retraining. Last but not least, the interdisciplinary nature of conflict prediction has been emphasised by scholars like Beck, King and Zeng (2000). Collaborations between data scientists and domain experts can foster more comprehensive and insightful models.

5.3 Limitations and Future Research Directions

There are some limitations in this research in the field such as data imputation and feature engineering. The imputation of missing values, particularly for 'Military_expenditure_ratio', might have introduced some inaccuracies into the model predictions. Imputing with average values can sometimes lead to a loss of variance, which in turn can affect the model's performance. Also, while the study incorporated a range of socio-political and economic indicators, there might be other unaccounted variables, both quantitative and qualitative, that could play a crucial role in predicting wars. In addition, the study primarily focused on a static dataset, but geopolitical dynamics are in constant flux, and the factors that may lead to conflicts can change over time.

Based on the limitations, future research direction can experiment with different imputation techniques, like model-based imputation or using algorithms like KNN, which might result in more accurate datasets and improved model performance. Besides, a mixed-methods approach, integrating both quantitative and qualitative data, can offer a more comprehensive understanding. Interviews, expert opinions, and ground reports can provide context that pure numbers might miss. What is more, engaging in collaborative, interdisciplinary research can bring together experts from the fields of international relations, history, and data science, fostering a

holistic approach to conflict prediction.

In addition, incorporating time-series data can be used to capture the temporal dynamics of geopolitical conflicts. This would allow for more nuanced predictions that take into account evolving geopolitical scenarios.

With the rapid advancements in AI, deep learning models, especially recurrent neural networks (RNN) or long short-term memory networks (LSTM), can be explored for predicting conflicts given their prowess in handling sequences and time-series data.

References

Bar-On, M. (2016). *Never-ending conflict: Israeli military history.* Yale University Press.

Beauchamp, N. (2017). Predicting and interpolating state-level polls using twitter textual data. *American Journal of Political Science*, *61*(2), 490-503.

Beck, N., King, G. & Zeng, L. (2000). Improving quantitative studies of international conflict: A conjecture. *American Political science review*, *94*(1), 21-35.

Bhattacharyya, S., Jha, S., Tharakunnel, K. & Westland, J. C. (2019). Data mining for credit card fraud: A comparative study. *Decision Support Systems*, *50*(3), 602-613.

Blainey, G. (1988). *Causes of war.* Simon and Schuster.

Blei, D. M. & Smyth, P. (2017). Science and data science. *Proceedings of the National Academy of Sciences*, *114*(33), 8689-8692.

Brandt, P. T., Freeman, J. R. & Schrodt, P. A. (2011). Real time, time series forecasting of inter-and intra-state political conflict. *Conflict Management and Peace Science*, *28*(1), 41-64.

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

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1986). *Classification and regression trees.* CRC press.

Busch, A. E. (1997). Ronald reagan and the defeat of the soviet empire. *Presidential Studies Quarterly*, *27*(3), 451-466.

Buszynski, L. (2012). *The south china sea maritime dispute: Political, legal and regional*

perspectives. Routledge.

Chadefaux, T. (2014). Early warning signals for war in the news. *Journal of Peace Research*, *51*(1), 5-18.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321-357.

Chen, C., Liaw, A. & Breiman, L. (2004). *Using random forest to learn imbalanced data* (Vol. 110; Tech. Rep.). University of California, Berkeley.

Colaresi, M. & Mahmood, Z. (2017). Do the robot: Lessons from machine learning to improve conflict forecasting. *Journal of Peace Research*, *54*(2), 193-214.

Colaresi, M. & Thompson, W. R. (2002). Strategic rivalries, proximity, and crisis escalation. *Journal of Peace Research*, *39*(3), 303-319.

Coppedge, M. et al. (2023). *V-dem [country-year/country-date] dataset v13.* Varieties of Democracy (V-Dem) Project. Retrieved from `https://doi.org/10.23696/vdemds23`

Cramer, J. S. (2002). *The origins of logistic regression* (Discussion Paper). Tinbergen Institute.

Davies, S., Pettersson, T. & Öberg, M. (2023). Organized violence 1989-2022 and the return of conflicts between states? *Journal of Peace Research*, *60*(4).

Enders, C. (2010). *Applied missing data analysis*. Guilford Press.

Esposito, F., Malerba, D. & Semeraro, G. (1997). A comparative analysis of methods for pruning decision trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(5), 476-491.

Fariss, C. J. (2019). *Human rights treaty compliance and the changing standard of accountability*. Cambridge University Press.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*.

Fernández, A., García, S., Herrera, F. & Chawla, N. V. (2018). Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, *61*, 863-905.

Field, A. (2013). *Discovering statistics using ibm spss statistics*. sage.

Freedman, L. & Gamba-Stonehouse, V. (1991). *Signals of war: The falklands conflict of 1982*. Princeton University Press.

Frohwein, H. I. & Lambert, J. H. (2000). Risk of extreme events in multiobjective decision trees part 1. severe events. *Risk Analysis*, *20*(1), 113-124.

Gartzke, E. (2007). The capitalist peace. *American Journal of Political Science*, *51*(1), 166-191.

Gelman, A. & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gleditsch, K. S. (2002). Expanded trade and gdp data. *Journal of Conflict Resolution*, *46*(5), 712-724.

Harrison, M. (2003). How much did the soviets really spend on defence? new evidence from the close of the brezhnev era.

He, H. & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, *21*(9), 1263-1284.

Hegre, H., Ellingsen, T., Gates, S. & Gleditsch, N. P. (2001). Toward a democratic civil peace? democracy, political change, and civil war, 1816–1992. *American Political Science Review*, *95*(1), 33-48.

Honaker, J. & King, G. (2010). What to do about missing values in time-series cross-section data. *American Journal of Political Science*, *54*(2), 561-581.

Hosmer Jr, D. W., Lemeshow, S. & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley Sons.

Ij, H. (2018). Statistics versus machine learning. *Nat Methods*, *15*(4), 233.

Japkowicz, N. & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent data analysis*, *6*(5), 429-449.

Johnson, R. A. & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th ed.). Upper Saddle River, NJ: Prentice-Hall.

Kaur, H. & Wasan, S. K. (2006). Empirical study on applications of data mining techniques in healthcare. *Journal of Computer Science*, *2*(2), 194-200.

Khudaykulova, M., Yuanqiong, H. & Khudaykulov, A. (2022). Economic consequences and implications of the ukraine-russia war. *International Journal of Management Science and Business Administration*, *8*(4), 44-52.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221-232.

Le Cessie, S. & Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied Statistics*, 191-201.

Levy, J. S. & Thompson, W. R. (2011). *Causes of war.* Wiley-Blackwell.

Little, R. & Rubin, D. (2019). *Statistical analysis with missing data.* John Wiley Sons.

Liu, X. Y. & Zhou, Z. H. (2006). The influence of class imbalance on cost-sensitive learning: An empirical study. In *sixth international conference on data mining (icdm'06)* (p. 970-974). IEEE.

Marshall, M. G. & Gurr, T. R. (2020). *Polity5: Political regime characteristics and transitions, 1800-2018* (Vol. 2). Center for Systemic Peace.

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.

Mullainathan, S. & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87-106.

Pemstein, D., Marquardt, K. L., Tzelgov, E., Wang, Y.-t., Medzihorsky, J., Krusell, J., ... von Römer, J. (2023). *The v-dem measurement model: Latent variable analysis for cross-national and cross-temporal expert-coded data.* V-Dem Working Paper No. 21. 8th edition. University of Gothenburg: Varieties of Democracy Institute.

Peng, C. Y. J., Lee, K. L. & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The journal of educational research*, *96*(1), 3-14.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*(1), 81-106.

Ricón, J. L. (2016). *The soviet union series.* Retrieved from `https://nintil.com/the-soviet-union-series/`

Russell, S. & Norvig, P. (2016). *Artificial intelligence: A modern approach*. Pearson.

Russett, B. & Oneal, J. R. (2001). *Triangulating peace: Democracy, interdependence, and international organizations*. Norton Company.

Singer, J. & Willett, J. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press.

Singh, D. & Singh, B. (2020). Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, *97*, 105524.

Sokolova, M. & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing Management*, *45*(4), 427-437.

Stein, A. A. & Russett, B. M. (1980). Evaluating war: Outcomes and consequences. In *Handbook of political conflict: theory and research* (p. 399-422).

Steinberg, D. (1990). Trends in soviet military expenditure. *Europe-Asia Studies*, *42*(4), 675-699.

Stueck, W. (1997). *The korean war: An international history*. Princeton University Press.

Vasquez, J. A. & Henehan, M. T. (2001). Territorial disputes and the probability of war, 1816–1992. *Journal of Peace Research*, *38*(2), 123-138.

Ward, M., Siverson, R. & Cao, X. (2007). Disputes, democracies, and dependencies: A reexamination of the kantian peace. *American Journal of Political Science*, *51*(3), 583-601.

Ward, M. D., Greenhill, B. D. & Bakke, K. M. (2010). The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research*, *40*(4), 363-375.

Yan, R., Nuttall, J. & Ling, C. X. (2006). Application of machine learning to short-term equity return prediction. *Available at SSRN 888778*.

Zhukov, Y. M. & Stewart, B. M. (2013). Choosing your neighbors: Networks of diffusion in international relations. *International Studies Quarterly*, *57*(2), 271-287.