Conflicts, Cause of War in 19th and 20th Century

Capstone Project by Mehmet J Oguz

INTRODUCTION

Why do wars occur and recur? Question is as old as war itself, and many answers already existed. If we want to understand or explain how peace can be achieved, we have to understand war and its causation. It is essential to have a perspective on the various sources of conflict. The vital causes of war are found in the nature and behavior of man. There are two prerequisites for a war between actors. At least one sides that involved has to expect that the anticipated gains from a war in terms of resources, power, glory, territory. So forth exceed the expected costs of conflict, including expected damages to property and life. Second, there has to be a failure in bargaining. So that for some reason, there is an inability to reach a mutually advantageous and enforceable agreement. If there is a lack of ability to enforce or commit to an agreement, then war may last a long time. It will continue until one side has emerged victorious, or the situation has changed so that the costs of continued conflict have become overwhelmingly high for all parties. Such a lack of enforceable agreements is often one of the main ingredients leading to protracted wars.

This analysis aims to analyze the conflict and cause of war in the 19th and 20th century. I used the Correlates of War Project datasets to predict future disputes and to prevent war before it happens. Other sources are Oslo Peace Research Institute armed conflict data, Stockholm International Peace Research Institute's war data, and Our World in Data websites. The Conflict Project goal is to create a better model that screen the red conflict zones with daily updates and reports the weekly connected events analysis. Possible interested parties are international relation based think-tank's, international relation departments in universities, defense industries, human rights organizations, political groups.

DATA

The data from the Correlates of War Projects website has data from 1816 to 2018, but not all datasets have input for all years. So it is expected that there are null values throughout the datasets. Missing data must be kept in mind throughout the entire process. The datasets in other sources are also missing values.

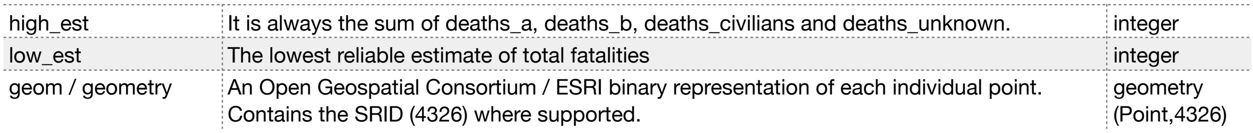
After the identification of data sources below list gives the details for each data source.

1. Correlates of War has the best datasets of religion, interstate war, dyadic interstate wars, territorial data, alliance, military data. The site offers csv files of each dataset. Csv files can then be read into a pandas data frame very easily.
2. Oslo Peace Research Institute's armed conflict data is in xls and pdf. Conflict data is between 1946 and 2008.
3. Stockholm International Peace Research Institute's war data is in xls format and allow you to download countries separately.
4. Our World in Data provides data in csv format that can be read into the pandas data frame very easily.

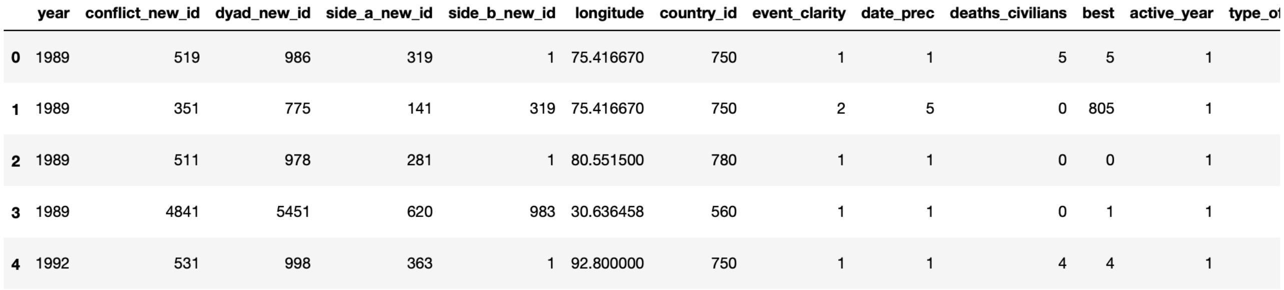
Variables included in the dataset: The variables included in the UCDP GED version 18.1:

Variable name	Content	Type
id	A unique numeric ID identifying each event.	integer
year	The year of the event	integer
active_year	1: if the event belongs to an active conflict/dyad/actor- Integer year 0: otherwise	integer
type_of_violence	1: state-based conflict 2: non-state conflict 3: one-sided violence	integer
conflict_new_id	A unique conflict identification code for each individual conflict in the dataset.	integer
conflict_name	Name of the UCDP conflict to which the event belongs. For non-state conflicts and one-sided violence this is the same as the dyad name.	string(9999)
dyad_new_id	A unique conflict identification code for each individual dyad in the dataset.	integer
dyad_name	Name of the conflict dyad creating the event. A dyad is the pair of two actors engaged in violence (in the case of one-sided violence, the perpetrator of violence and civilians).	string(9999)
side_a_new_id	A unique ID of side A.	integer
gwnoa	The Gleditsch and Ward number for Side A if the side is a state. Empty if Side A is not a state.	string(9999)
side_a	The name of Side A in the dyad. In state-based conflicts always a government. In one-sided violence always the perpetrating party.	string(9999)
side_b_new_id	A unique ID of side B.	integer
gwnob	The Gleditsch and Ward number for Side B if the side is a state. Empty if Side B is not a state.	string(9999)
side_b	The name of Side B in the dyad. In state-based always the rebel movement or rivalling government. In one- sided violence always “civilians”.	integer
number_of_sources	Number of total sources containing information for an event that were consulted.	
source_article	References to the names, dates and titles of the source material from which information on the event is gathered. This variable is highly streamlined for information collected since 2013, and is less so for older data. For such older data, abbreviations are sometimes used for source agencies. The most frequent are: R: Reuters News, BBC: BBC Monitoring AP: Associated Press Newswires AFP: Agence France Presse, X: Xinhua DOW: Dow Jones Wires	text
source_office	The name of the organizations publishing the source materials.	text
source_date	The dates the source materials were published on.	text
source_headline	The titles of the source materials.	text
source_original	The name or type of person or organization from which the information about the event originates in the original report.	string(9999)

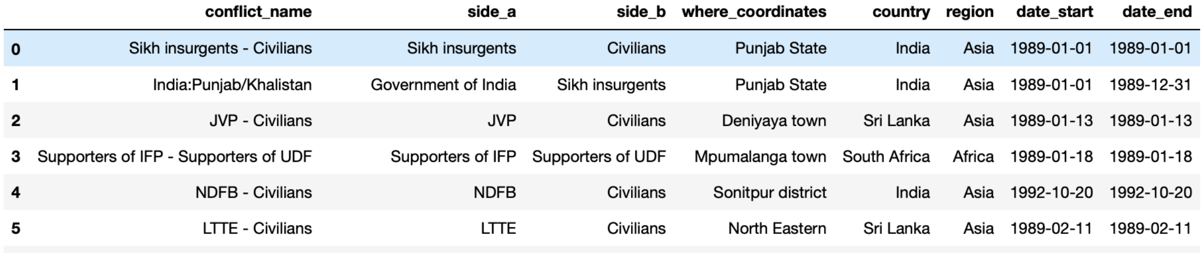
where_prec	The precision with which the coordinates and location assigned to the event reflects the location of the actual event. 1: exact location of the event known and coded. 2: event occurred within at maximum a ca. 25 km radius around a known point. The coded point is the known point. 3: only the second order administrative division where an event happened is known. That administrative division is coded with a point representing it (typically the centroid). 4: only the first order administrative division where an event happened is known. That administrative division is coded with a point representing it (typically the centroid). 5: the only spatial reference for the event is neither a known point nor a known formal administrative division, but rather a linear feature (e.g. a long river, a border, a longer road or the line connecting two locations further afield than 25 km) or a fuzzy polygon without defined borders (informal regions, large radiuses etc.). A representation point is chosen for the feature and employed. 6: only the country where the event took place in is known. 7: event in international waters or airspace.	integer
where_coordinates	Name of the location to which the event is assigned. Fully standardized and normalized.	string(9999)
adm_1	Name of the first order (largest) administrative division string(9999) where the event took place	string(9999)
adm_2	Name of the second order administrative division where the event took place	string(9999)
latitude	Latitude (in decimal degrees)	numeric(9,6)
longitude	Longitude (in decimal degrees)	numeric(9,6)
geom_wkt	An Open Geospatial Consortium textual representation of the location of each individual point. Formatted as OGC WKT (well known text) without SRID.	string(9999)
priogrid_gid	An Open Geospatial Consortium textual representation of the location of each individual point. Formatted as OGC WKT (well known text) without SRID.	integer
country	The PRIO-grid cell id (gid) in which the event took place. Compatibility with PRIO-grid (T ollefsen, 2012) is guaranteed for both PRIO-grid 1 and 2.	string(9999)
country_id	Gleditsch and Ward number of the country in which the event takes place.	integer
region	Region where the event took place. One of following: {Africa, Americas, Asia, Europe, Middle East}	string(9999)
event_clarity	1 (high) for events where the reporting allows the coder to identify the event in full. 2 (lower) for events where an aggregation of information was already made by the source material that is impossible to undo in the coding process.	integer
date_prec	How precise the information is about the date of an event. 1: exact date of event is known; 2: the date of the event is known only within a 2-6 day range. 3: only the week of the event is known 4: the date of the event is known only within an 8-30 day range or only the month when the event has taken place is known 5: the date of the event is known only within a range longer than one month but not more than one calendar year.	integer
date_start	The earliest possible date when the event has taken place.	Date YYYY-MM- DD
date_end	The last possible date when the event has taken place.	Date YYYY-MM- DD
deaths_a	The best estimate of deaths sustained by side a. Always 0 for one-sided violence events.	integer
deaths_b	The best estimate of deaths sustained by side b. integer Always 0 for one-sided violence events.	integer
deaths_civilians	The best estimate of dead civilians in the event. For non-state or state-based events, this is the number of collateral damage resulting in fighting between side a and side b. For one-sided violence, it is the number of civilians killed by side a.	integer
deaths_unknown	The best estimate of deaths of persons of unknown integer status.	integer
best_est	The best (most likely) estimate of total fatalities resulting from an event.	integer



high_est	It is always the sum of deaths_a, deaths_b, deaths_civilians and deaths_unknown.	integer
low_est	The lowest reliable estimate of total fatalities	integer
geom / geometry	An Open Geospatial Consortium / ESRI binary representation of each individual point. Contains the SRID (4326) where supported.	geometry (Point,4326)



	year	conflict_new_id	dyad_new_id	side_a_new_id	side_b_new_id	longitude	country_id	event_clarity	date_prec	deaths_civilians	best	active_year	type_o
0	1989	519	986	319	1	75.416670	750	1	1	5	5	1	
1	1989	351	775	141	319	75.416670	750	2	5	0	805	1	
2	1989	511	978	281	1	80.551500	780	1	1	0	0	1	
3	1989	4841	5451	620	983	30.636458	560	1	1	0	1	1	
4	1992	531	998	363	1	92.800000	750	1	1	4	4	1	



	conflict_name	side_a	side_b	where_coordinates	country	region	date_start	date_end
0	Sikh insurgents - Civilians	Sikh insurgents	Civilians	Punjab State	India	Asia	1989-01-01	1989-01-01
1	India:Punjab/Khalistan	Government of India	Sikh insurgents	Punjab State	India	Asia	1989-01-01	1989-12-31
2	JVP - Civilians	JVP	Civilians	Deniyaya town	Sri Lanka	Asia	1989-01-13	1989-01-13
3	Supporters of IFP - Supporters of UDF	Supporters of IFP	Supporters of UDF	Mpumalanga town	South Africa	Africa	1989-01-18	1989-01-18
4	NDFB - Civilians	NDFB	Civilians	Sonitpur district	India	Asia	1992-10-20	1992-10-20
5	LTTE - Civilians	LTTE	Civilians	North Eastern	Sri Lanka	Asia	1989-02-11	1989-02-11

Preparation

All datasets diagnosed for the inconsistent column names, missing data, duplicate rows, untidy, and unexpected data values. All datasets were explored with pandas methods such as .head(), .info(), and .describe(), and DataFrame attributes like .columns and .shape . I concatenated the needed datasets, and I created a single dataframe and saved for later analysis. I spent most of the time cleaning and ensuring data quality. After concatenated the datasets missing data and duplicate rows appeared. I spent considerable time looking at what data was missing and how best to fill it. One way is using additional datasets.

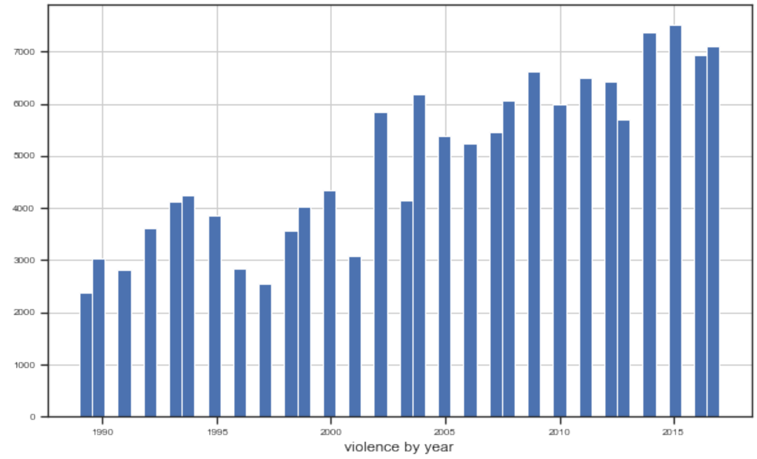
I also use the left join to merge the two dataframes because not all countries involved in a conflict. Any country missing in one dataset mean they might not be engaged, and thus have null values in the new dataset. After combining columns, a complete dataframe is now forming. The last step is to fill in the null data throughout the dataset. There are various ways to do this. I fill those values are set to -1, though this may change later with using other resources.

Among the many columns in the datasets I have combined, I only added one that describes the reason for the war based on conflicts, involved parts, and type of the war. Additional columns may also be added relating to war, state, and territory in the future.

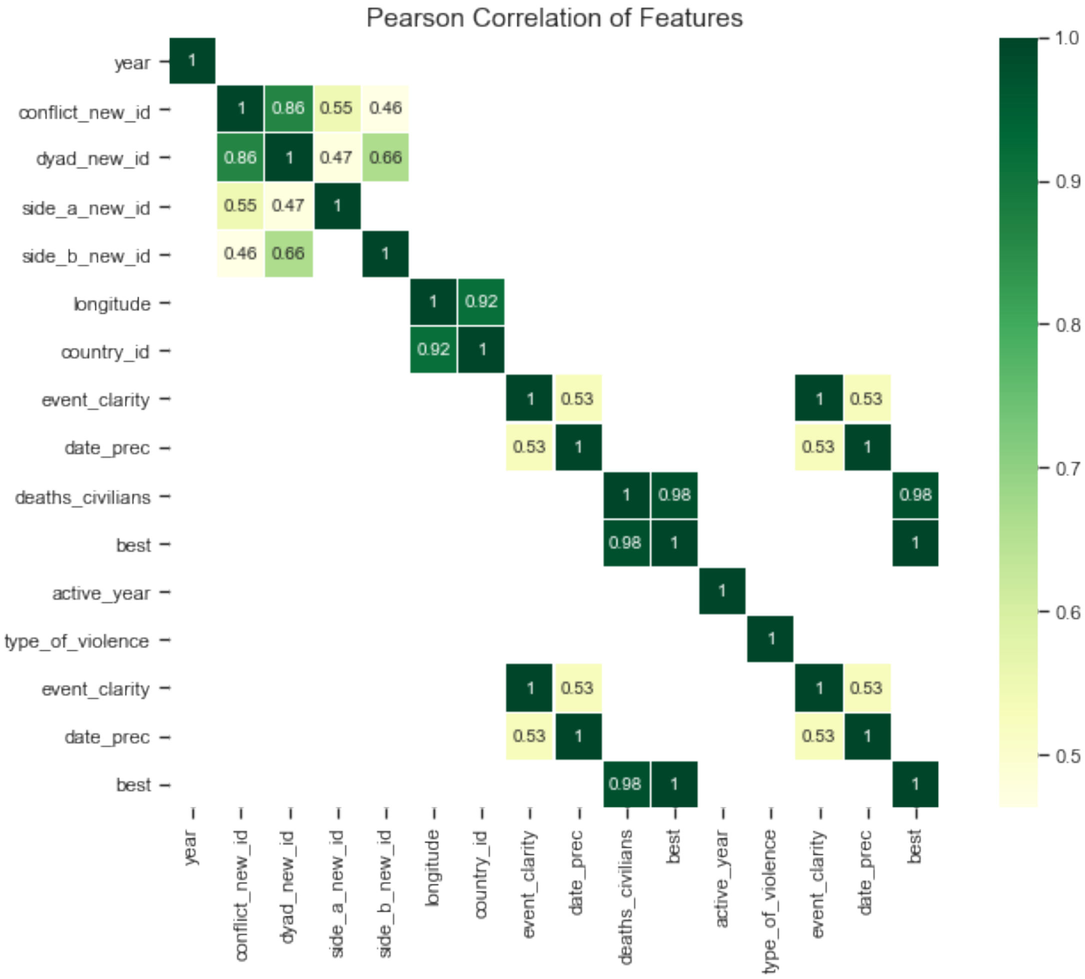
Some changes may have to be made later on the analysis process, but for now, this is the first draft of the dataframe. This cleaned dataframe is saved for further analyze.

Preliminary Exploration and Findings

As seen in below graphic over the years, there is an increase in violence. There might be many reasons that caused a rise in violence. One and underlying reason might be the better reporting of the event and better communication channels nowadays.



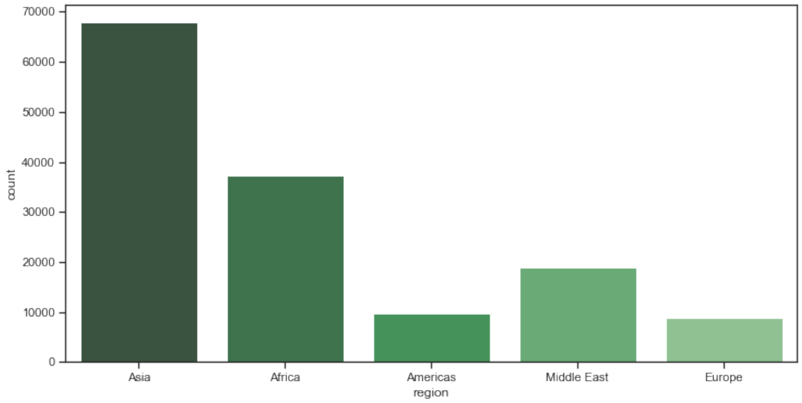
Trying to plot all the numerical features in a seaborn pairplot will take us too much time and will be hard to interpret. We can try to see if some variables are linked between each other and then explain their relationship with common sense. Our Pearson Correlation graph shows that there is a strong relationship between the events clarity and reporting time. I take in consideration of correlation that is greater or equal to 0.5 or less than or equal to -0.4.

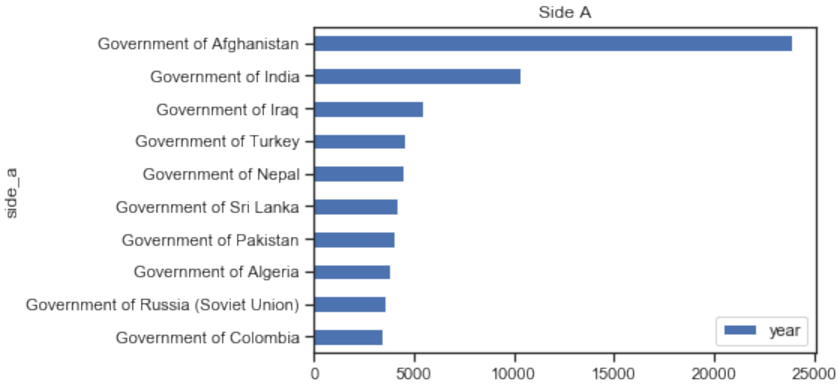


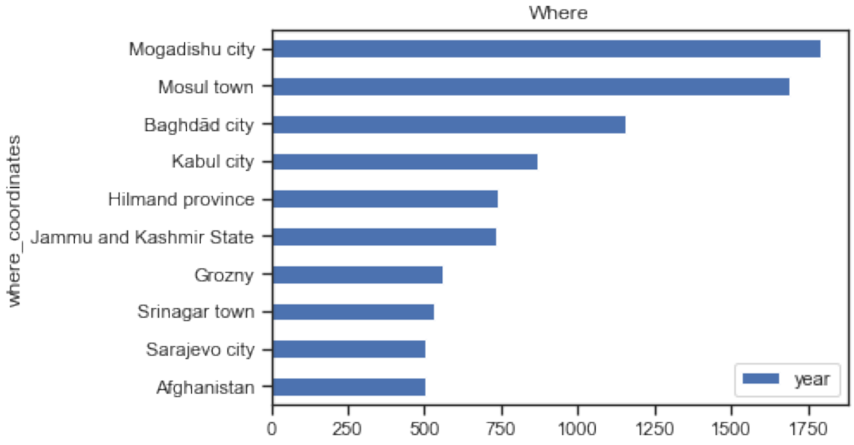
When I did exploratory data analysis with the help of Pearson Correlation, a lot of features seems to be correlated. I found that there is a strong correlation between the country and it's longitude. The state and longitude are particularly significant in terms of explaining the answer to my project question. When I check the correlation between year and other features, there are 15 correlated values with the year, but none of the correlation is strong.

```
year                1.000000
dyad_new_id         0.238214
conflict_new_id     0.192456
side_b_new_id       0.188353
country_id          0.139166
longitude           0.134292
side_a_new_id       0.071549
active_year         0.036479
event_clarity       0.010542
event_clarity       0.010542
deaths_civilians   -0.007568
best               -0.012103
date_prec          -0.065405
date_prec          -0.065405
type_of_violence   -0.099923
```

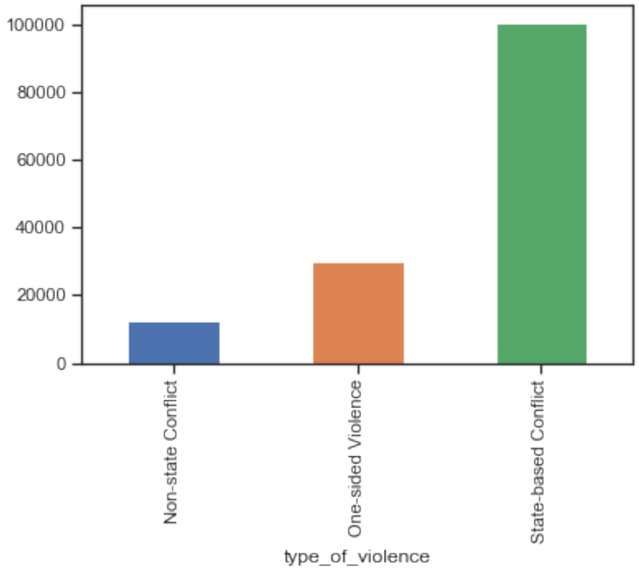
Now I need to add more questions to my research. How the geographic position of a country affected people's relation and behavior? As seen in below graphic, 8 out of the first ten countries that involved in a conflict are in Asia. What is the main reason behind Asian conflicts? Is it economic reasons, religion or something else?

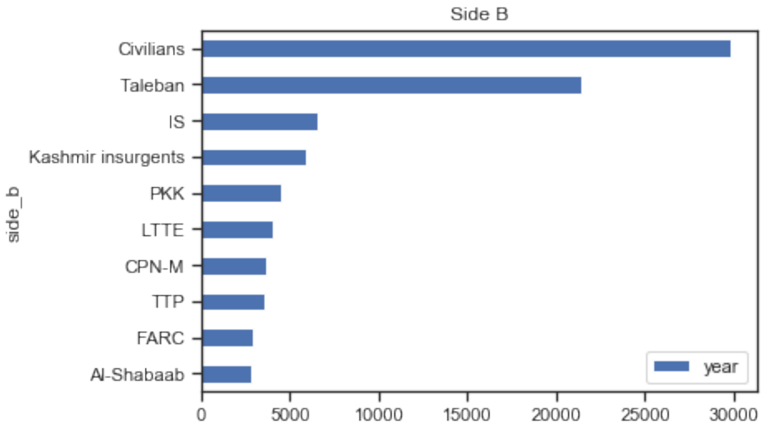






When we look at the type of violence that occurred, we see that state-based conflict is the one mostly happened. State-based conflict is any non-governmental group of people having announced a name for their group and using armed force against a government. Civilians are the largest group that used the armed force against the government. Since it is the sum of all civilians that involved in a conflict, we need to do a separate analyze to examine the reasons, sides, and active years.





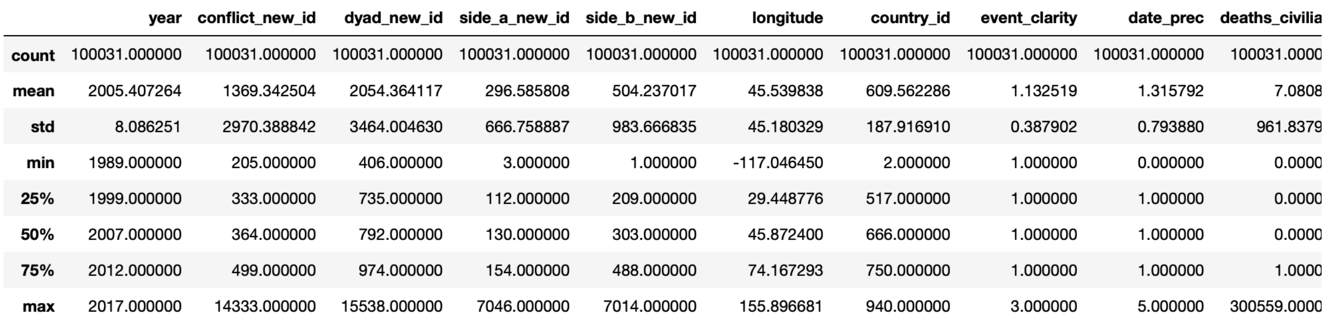
MODEL APPROACH

Feature engineering and selection often provide the highest return on time invested in a machine learning problem. Feature engineering is creating additional features from the raw data. In feature selection, I remove elements to help to generalize my model better to new data and create a more interpretable model. The machine learning model will learn from the conflicting data, includes all the relevant information for my task. If I don’t feed a model to the correct data, then I am setting it up to fail, and I should not expect the model to learn.

The final step I took before getting started with modeling is establishing a naive baseline. This is essentially a guess against which we can compare our results. If the machine learning models do not beat this guess, then we might have to conclude that machine learning is not acceptable for the task, or we might need to try a different approach. The metric I will use is mean absolute error (MAE), which measures the average absolute error on the predictions. The mean absolute error is easy to calculate and is interpretable.

Before calculating the baseline, I split my data into a training and a testing set. I used 70% of the data for training and 30% for testing.

The baseline guess is a score of 45.8 and baseline performance on the test set: is MAE = 32.3763. This shows our average estimate on the test set is off by about 32 points. The scores are between 1 and 100, so this means the average error from an original method is about 32%. The naive method of guessing the median training value provides us a low baseline for our models to beat!



	year	conflict_new_id	dyad_new_id	side_a_new_id	side_b_new_id	longitude	country_id	event_clarity	date_prec	deaths_civilia
count	100031.000000	100031.000000	100031.000000	100031.000000	100031.000000	100031.000000	100031.000000	100031.000000	100031.000000	100031.0000
mean	2005.407264	1369.342504	2054.364117	296.585808	504.237017	45.539838	609.562286	1.132519	1.315792	7.0808
std	8.086251	2970.388842	3464.004630	666.758887	983.666835	45.180329	187.916910	0.387902	0.793880	961.8379
min	1989.000000	205.000000	406.000000	3.000000	1.000000	-117.046450	2.000000	1.000000	0.000000	0.0000
25%	1999.000000	333.000000	735.000000	112.000000	209.000000	29.448776	517.000000	1.000000	1.000000	0.0000
50%	2007.000000	364.000000	792.000000	130.000000	303.000000	45.872400	666.000000	1.000000	1.000000	0.0000
75%	2012.000000	499.000000	974.000000	154.000000	488.000000	74.167293	750.000000	1.000000	1.000000	1.0000
max	2017.000000	14333.000000	15538.000000	7046.000000	7014.000000	155.896681	940.000000	3.000000	5.000000	300559.0000

I worked on a supervised regression task to develop a model that can predict the conflict. My focus is on both the accuracy of the predictions and interpretability of the model. I evaluated five different models covering the complexity spectrum by mean absolute error.

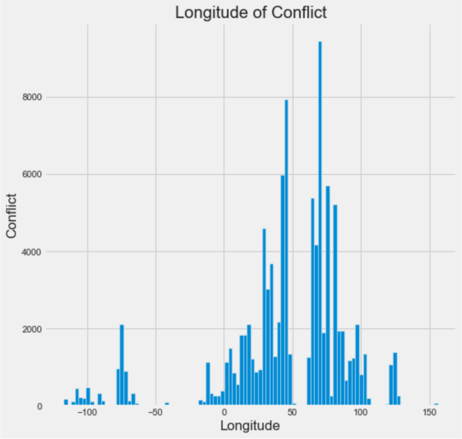
1. Linear Regression

2. K-Nearest Neighbors Regression

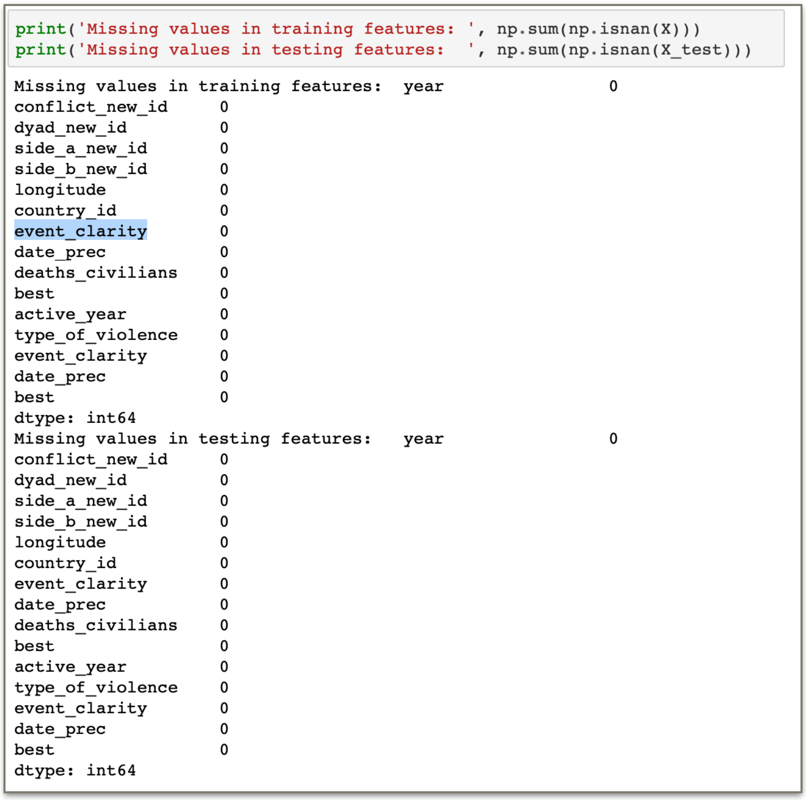
3. Random Forest Regression

4. Gradient Boosted Regression

5. Support Vector Machine Regression



Machine learning models cannot deal with any absent values, so I checked all data one more time. As seen below, I don't have any missing data.



```
print('Missing values in training features: ', np.sum(np.isnan(X)))
print('Missing values in testing features:  ', np.sum(np.isnan(X_test)))
```

```
Missing values in training features:  year                0
conflict_new_id     0
dyad_new_id         0
side_a_new_id       0
side_b_new_id       0
longitude           0
country_id          0
event_clarity       0
date_prec           0
deaths_civilians    0
best                0
active_year         0
type_of_violence    0
event_clarity       0
date_prec           0
best                0
dtype: int64
Missing values in testing features:   year                0
conflict_new_id     0
dyad_new_id         0
side_a_new_id       0
side_b_new_id       0
longitude           0
country_id          0
event_clarity       0
date_prec           0
deaths_civilians    0
best                0
active_year         0
type_of_violence    0
event_clarity       0
date_prec           0
best                0
dtype: int64
```

After all the work I spent cleaning and formatting the data, training, and predicting with the models is relatively simple. I used the Scikit-Learn library in Python. I made one model in Scikit-Learn and implemented others. Model creation, training, and testing are each one line! To build the other models, I used the same syntax, with the only change the name of the algorithm. The results presented below:

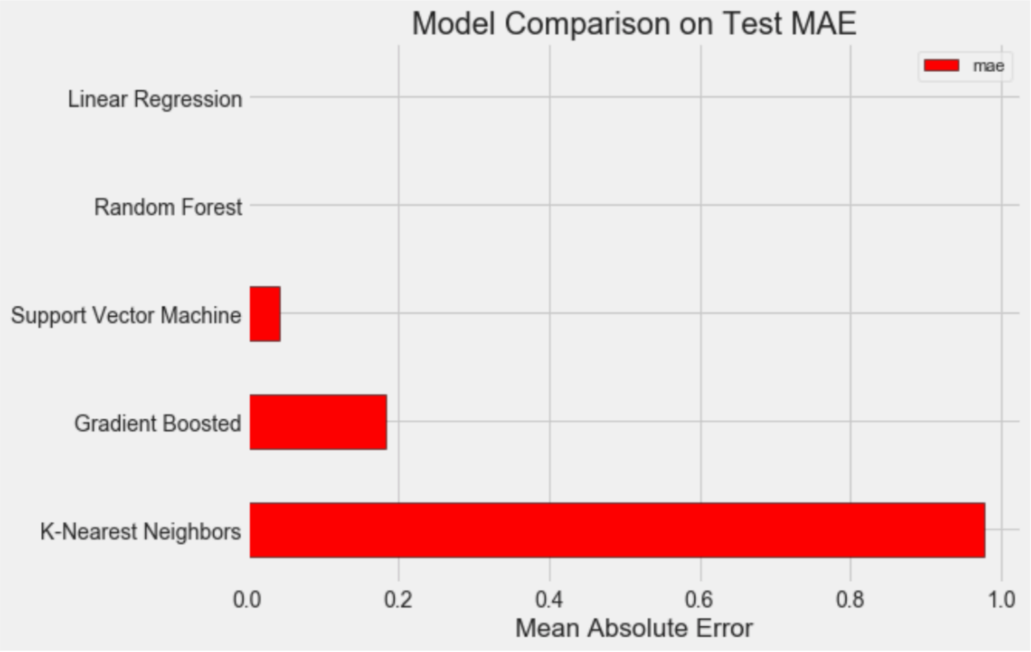
1. Linear Regression Performance on the test set: MAE = 0.0000

2. Support Vector Machine Regression Performance on the test set: MAE = 0.0424

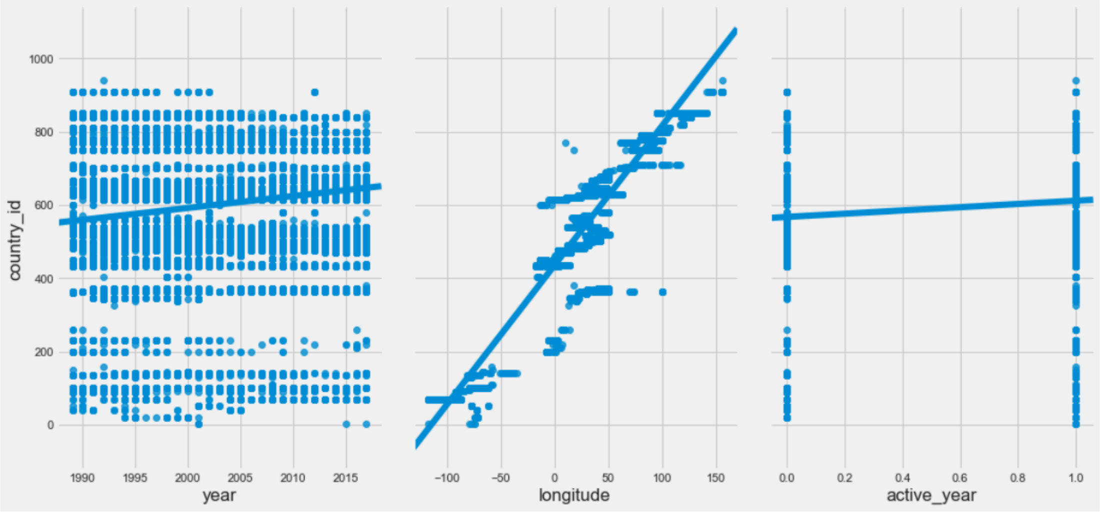
3. Random Forest Regression Performance on the test set: MAE = 0.0018

4. Gradient Boosted Regression Performance on the test set: MAE = 0.1832

5. K-Nearest Neighbors Regression Performance on the test set: MAE = 0.9770



Depending on the run the Linear Regression and Random Forest performs the best followed by the Support Vector Machine. I will use Linear Regression as a model optimization, and as seen in the graph, there is a strong relationship between longitude and country _id.



The data split into 70% of training, 30% testing set, and the slope and intercept of the data contained in the model's fit parameters. The interception of the model is 435.83437, and the slope is 3.80887. Means that for every one unit of change in country_id, the change in the longitude is about 3.80%.

Now let's look at other parameters to compare the accuracy of our model.

Mean Absolute Error: 53.00359

Mean Squared Error: 5611.81805

Root Mean Squared Error: 74.91206

The root mean squared error is 74.91, which is higher than the mean value of the longitude of all states, which is 45.58. The result means that our algorithm was not very accurate but can still make reasonably good predictions.

CONCLUSION

Many factors may have contributed to this inaccuracy like we need more data, or we made the wrong assumption that this data has a linear relationship. Another reason might be the poor features that used may not have had a high enough correlation to the values we were trying to predict.

The essential causes of war are found in the nature and behavior of man. Conflicts result from misdirected aggressive impulses, from stupidity, and selfishness. If these are the leading causes of war, then the elimination of war must come through uplifting and enlightening men or securing their psychic-social adaptation. If human nature can't be changed (good or bad), then we can't diminish the occurrence of war by trying to change it. Human nature can be changed. If the the longitude is effecting human's behavior then the behavior of people and the reasons behind such acts need to analyze with the collection of more data. Political and social institutions should be seen as a factor which is changing circumstance to decrease chances of war.

CITATION

1- COW: Sarkees, Meredith Reid, and Frank Wayman (2010). Resort to War: 1816 - 2007. Washington, DC: CQ Press.

2- Territorial change data: Tir, Jaroslav, Philip Schafer, Paul Diehl, and Gary Goertz. 1998. "Territorial Changes, 1816-1996: Procedures and Data "Conflict Management and Peace Science 16:89-97.

Alliance: Gibler, Douglas M., 2009. International military partnerships, 1648-2008. CQ Press.

3- WRD: Zeev Maoz and Errol A. Henderson. 2013. "The World Religion Dataset, 1945-2010: Logic, Estimates, and Trends." International Interactions, 39: 265-291.

4- National material capabilities: Singer, J. David, Stuart Bremer, and John Stuckey. (1972). "Capability Distribution, Uncertainty, and Major Power War, 1820-1965." in Bruce Russett (ed) Peace, War, and Numbers, Beverly Hills: Sage, 19-48.

5- Military Interstate dispute: Palmer, Glenn, Vito D'Orazio, Michael Kenwick, and Matthew Lane. 2015. "The Mid4 Dataset, 2002–2010: Procedures, Coding Rules, and Description." Conflict Management and Peace Science 32: 222-42.

6- 'What causes wars? ", Jan Tudovic https://iapss.org/2014/11/26/what-causes-wars/

7- "The Reasons for Wars – an Updated Survey "Matthew O. Jackson and Massimo Morelli Revised: December 2009, Handbook on the Political Economy of War, Elgar Publishing