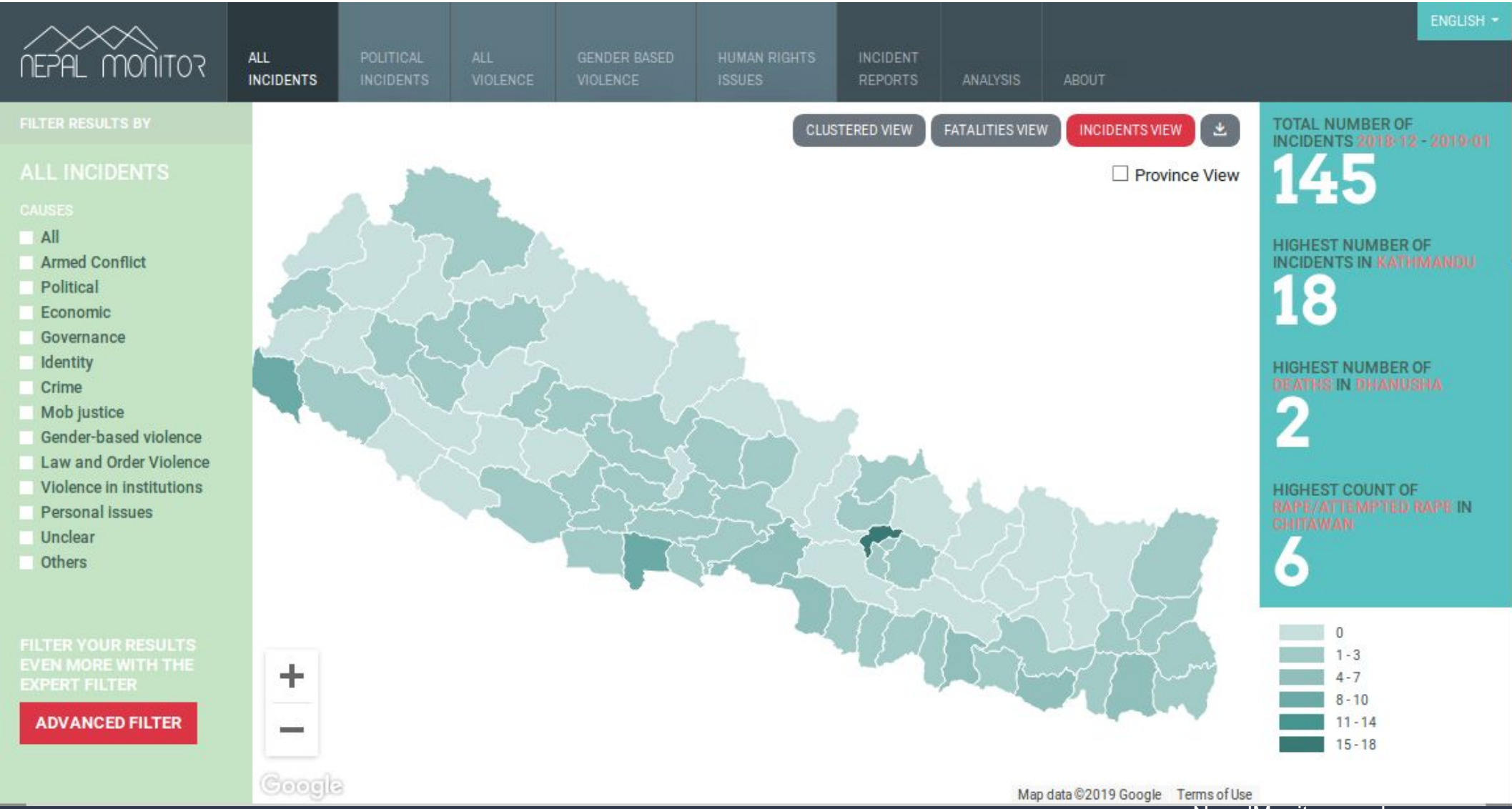
Nepal Monitor Predictive Analysis

SpringBoard Capstone 1

Neil Horning

Contents

Client - Nepal Monitor



NepalMonitor.org homepage

The Nepal Monitor Project (NMP) comprehensively maps human rights and violence data across Nepal to improving understanding, better respond to it, and promote peace.

To be an effective early warning system, Nepal Monitor should answer two main questions:

1. What violence is likely to happen where, and when?

2. What variables predict that?

Exploratory Data Analysis

Event Date	#	Title	Nepali Title	Event Date	Publication Date	Location	Description	Nepali Description	Tags	Latitude	...	Youth raped
2018-12-31	21793	Complaint registered against rape	चकलेट दिन्छु भन्दै ५ वर्षीया बालिका बलात्कार।	2018-12-31	2018-12-31	Province 2, Parsa, Pakahamainpur	A complaint against 19 years old Dev Kumar Pat...	पर्साको पकहामैनपुर गाउँपालिका ३ मा भोरमा ५ वर...	GBV\nChildren's Rights	27.03046158726636	...	1.0
2018-12-31	21803	Man held on charge of attempted rape	बलात्कार प्रयास घटनाका आरोपित गिरफ्तार।	2018-12-31	2019-01-01	Province 1, Morang, Biratnagar, Ward 10	Police have arrested 25 years old Gyanendra Sa...	१८ वर्षीया युवतीलाई बलात्कार प्रयास गरेको आरोप...	GBV	26.454562034856874	...	1.0
2018-12-31	21804	Agitating civil servants take to streets	समायोजनविरुद्ध सडकमै उत्रिए निजामती कर्मचारी।	2018-12-31	2019-01-01	Province 3, Kathmandu	Agitating civil servants today took to the str...	कर्मचारी समायोजन अध्यादेशको विरोधमा निजामती कर...	Governance	27.7058766168406	...	NaN
2018-12-31	21810	Complaint registered against rape	बलात्कार घटनाका आरोपितविरुद्ध उजुरी दर्ता।	2018-12-31	2019-01-02	Province 2, Parsa	A complaint against 19 years old Dev Kumar Pat...	पाँच वर्षीया बालिकालाई बलात्कार गरेको आरोपमा प...	GBV\nChildren's Rights	27.173588000000002	...	1.0

Event Date	Total killed	Female killed	Youth killed	Total injured	Female injured	Total raped	Female raped	Youth raped	Total abducted	Female abducted	...	Actor 2 - Affiliation_security forces - apf	Actor 2 - Affiliation_security forces - army
2017-01-01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
2017-01-08	2.0	0.0	0.0	2.0	1.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
2017-01-15	2.0	0.0	1.0	4.0	3.0	2.0	2.0	2.0	1.0	0.0	...	0.0	0.0
2017-01-22	1.0	0.0	0.0	33.0	1.0	1.0	1.0	1.0	0.0	0.0	...	0.0	0.0
2017-01-29	3.0	1.0	2.0	18.0	1.0	1.0	1.0	1.0	0.0	0.0	...	0.0	0.0

5 rows × 321 columns

Data:

- Nepal Monitor Export
  - Individual incidents
  - Dated, Geocoded, categorized
  - Subject to federal restructuring and website revamp
  - Boundary Hierarchy 2017: Ward < VDC < District < Zone
  - ‘’ 2018: Ward < Palika< District < Province
- 2011 Ward level population census
- Ward Boundary Shapefile
  - Exported to calculate area and population density

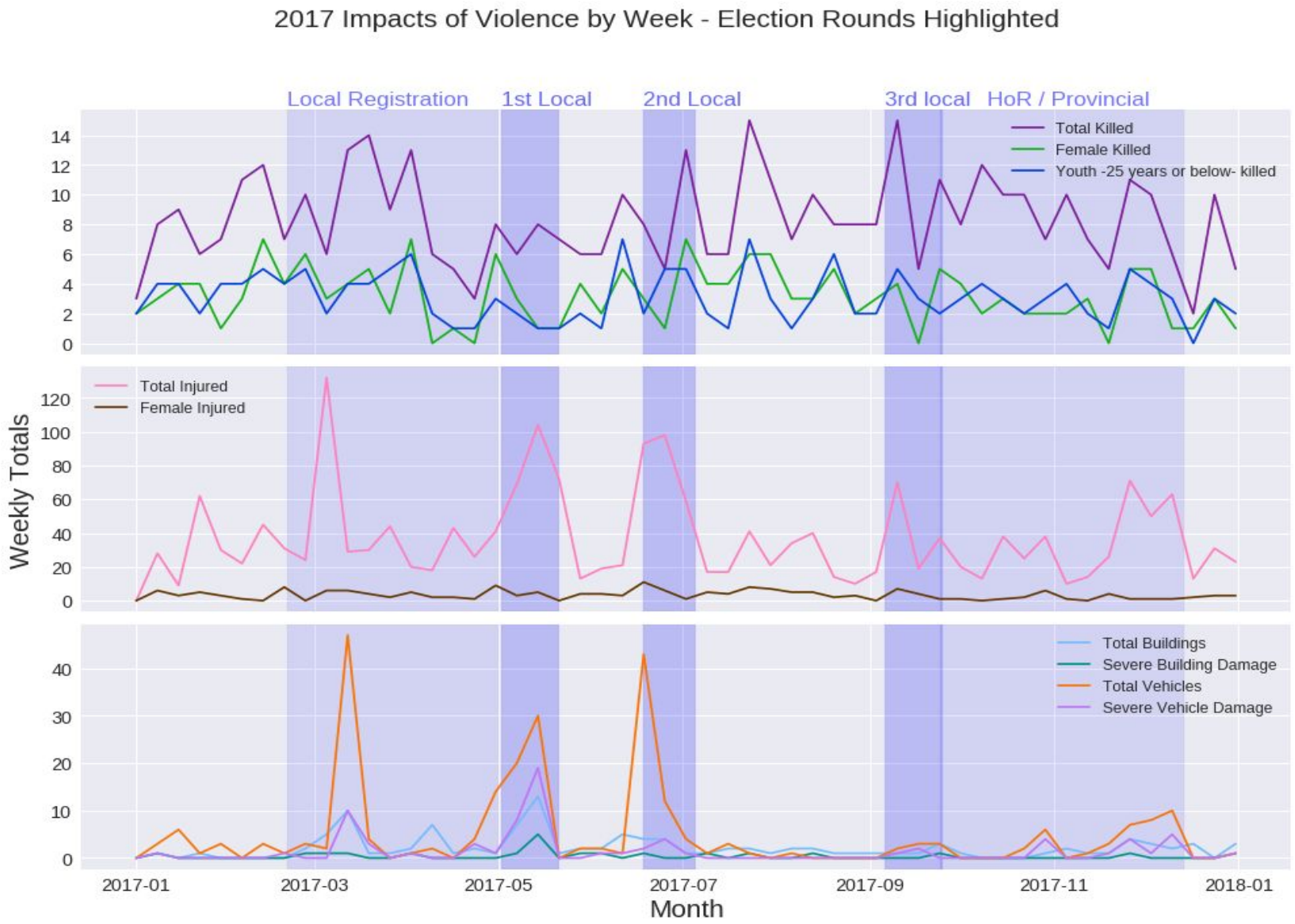
Data Cleaning:

- Normalized timestamps so that only dates remained
- Removed unnecessary columns
- Converted yes and no values to 1’s and zeros

Data Wrangling:

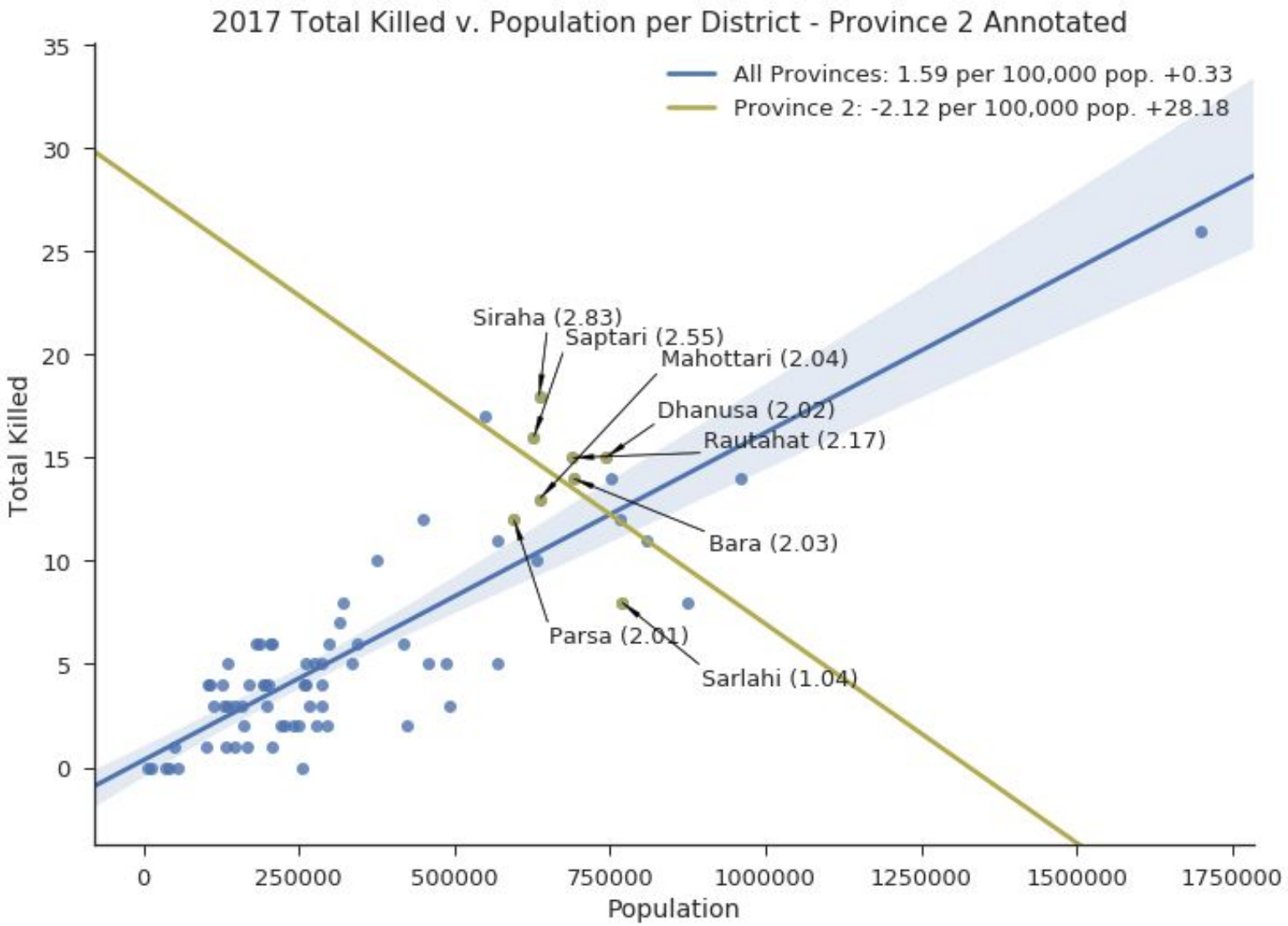
- Separated out impacts of 2017 incident data for analysis
- Created dataframe of impacts grouped by date
- Created dataframe of impacts grouped by district joined with province number and district population data.
- Created a dataframe of incidents grouped by VDC and joined with VDC level census data
- Created “dummy columns” for variables tracked by the NMP project, making the dataset 375 columns wide.
- Aggregated by week and processed into timeshifted dataframe for time series analysis

Is violence correlated with specific time periods?



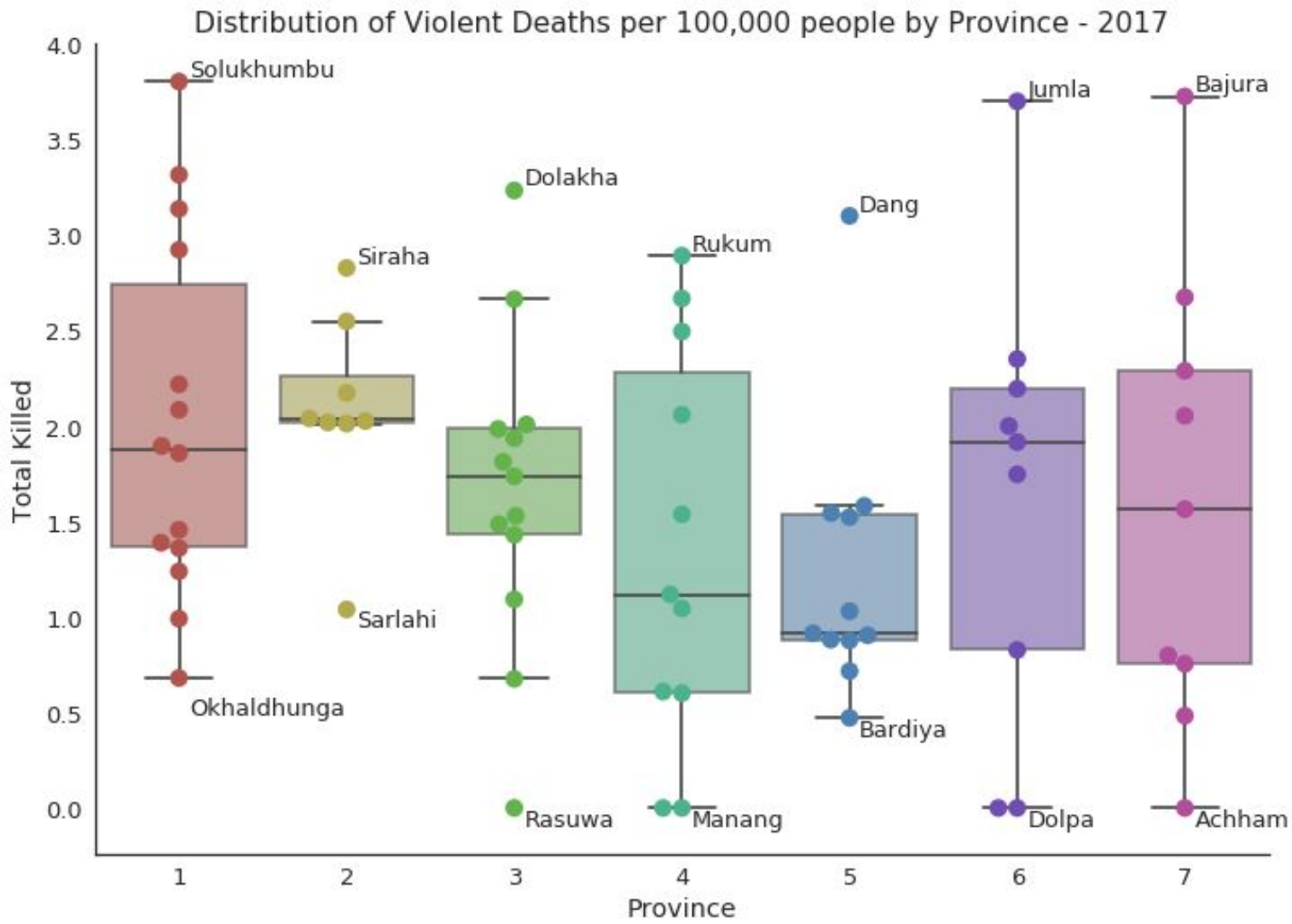
Injuries and physical damage appear correlated with electoral rounds, while other impacts have no significant correlation with these periods.

How does the distribution of violence differ between provinces?



Province 2 has the highest Total Killed among all the provinces and the highest population. It also has the tightest distribution of Total Killed per 100,000 people. And a negative regression line of violence per-capita.

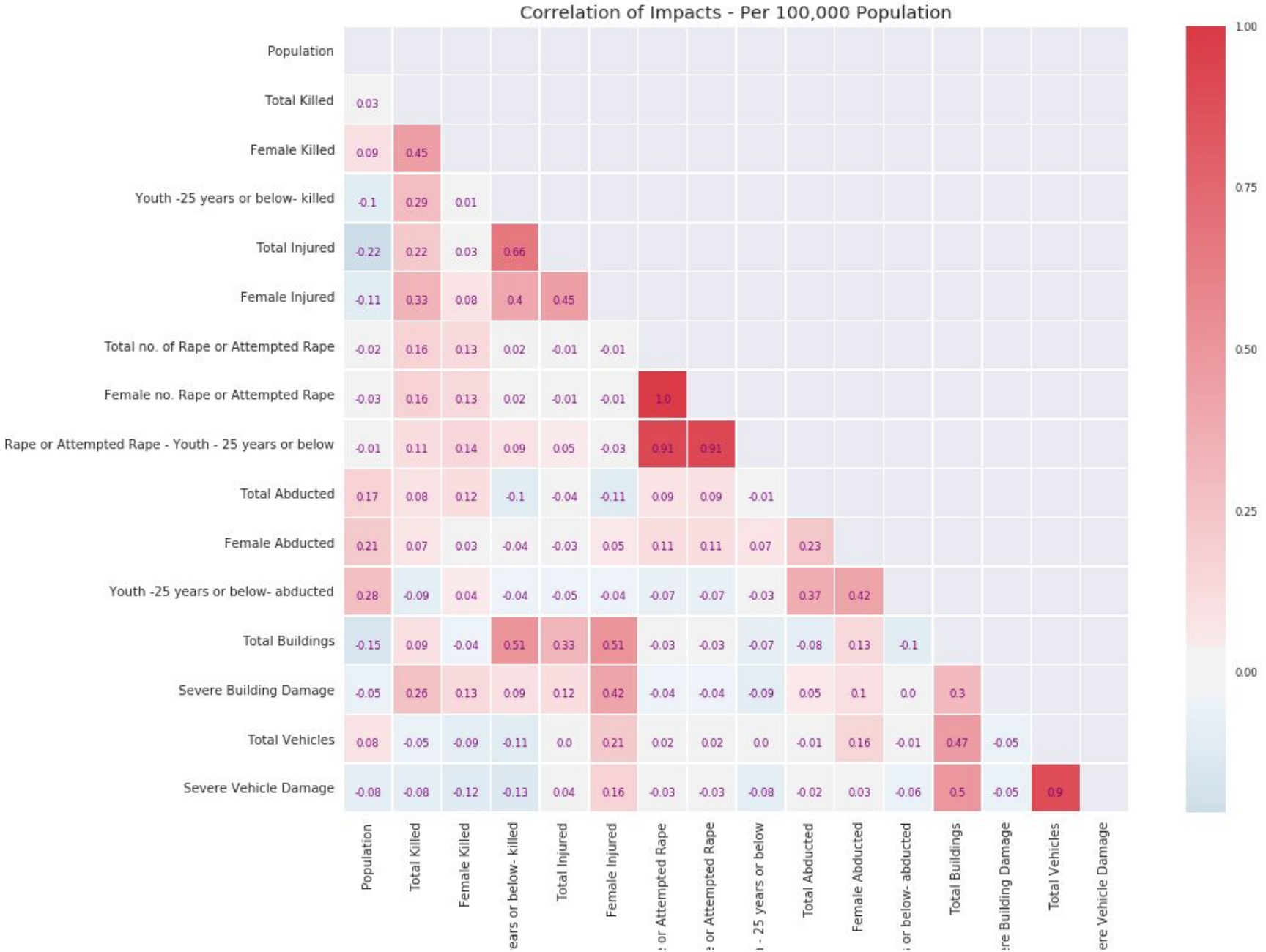
How does the distribution of violence differ between provinces?



The reason for the abnormal regression line of Province 2 is the outlier status of Sarlahi district. It has both the highest population and the lowest number of deaths.

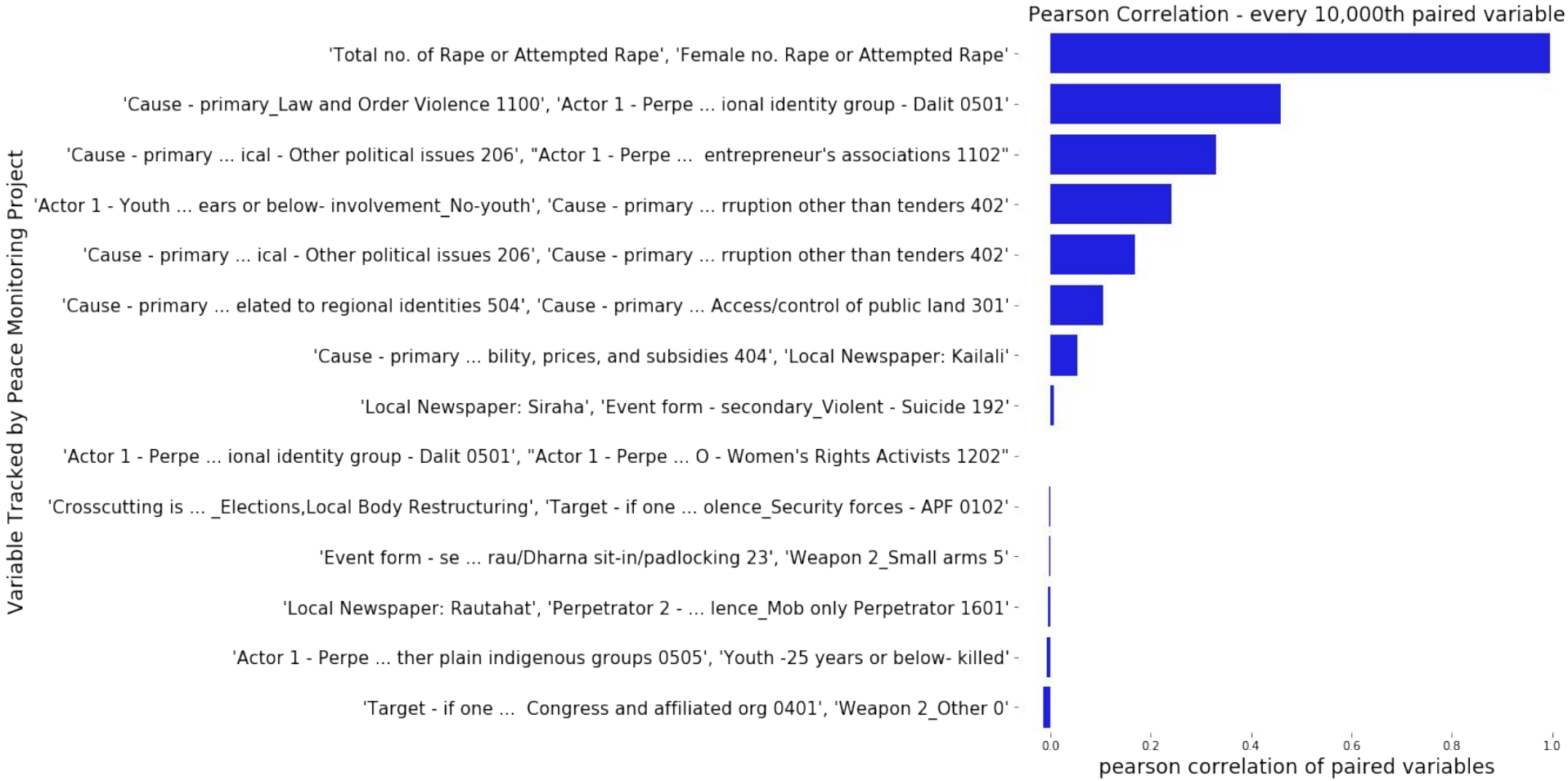
What variables predict violence?

There are significant relationships between district population and certain impacts of violence, the highest being with Total Killed



Plotting pairs of impacts in a heatmap reveals a relationship between injuries, youth killed, and and property damage (Buildings, Vehicles, etc) in 2017 - likely an effect of where demonstrations are taking place.

What variables predict violence? cont.



However, roughly 50% of all 132,076 possible variable pairs have some positive correlation. This highlights the need for machine learning in sorting out meaningful relationships.

Time Series Analysis

To predict what will happen one week in the future, the model is fed what has happened several weeks prior (X) to a given set of weeks (y) and trains itself to find any relationship. It is then asked to guess what will happen in the next week (y_pred).

This is done in sequence for every week using multiple variables.

X:

Event Date	Total killed (t-004)	Total killed (t-003)	Total killed (t-002)	Total killed (t-001)
2018-12-09	8.0	2.0	6.0	3.0
2018-12-16	2.0	6.0	3.0	4.0
2018-12-23	6.0	3.0	4.0	3.0

Train

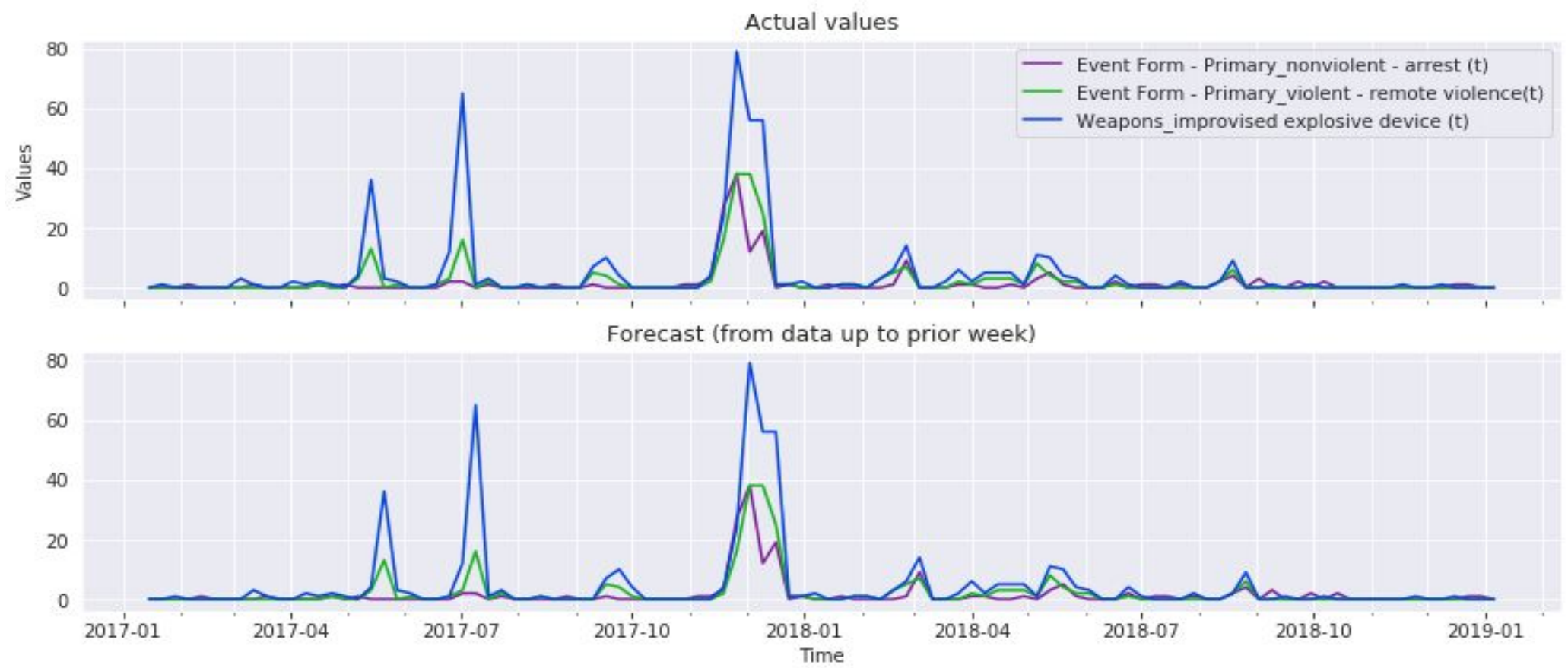
y:

Event Date	Total killed(t)
2018-12-09	4.0
2018-12-16	3.0
2018-12-23	4.0

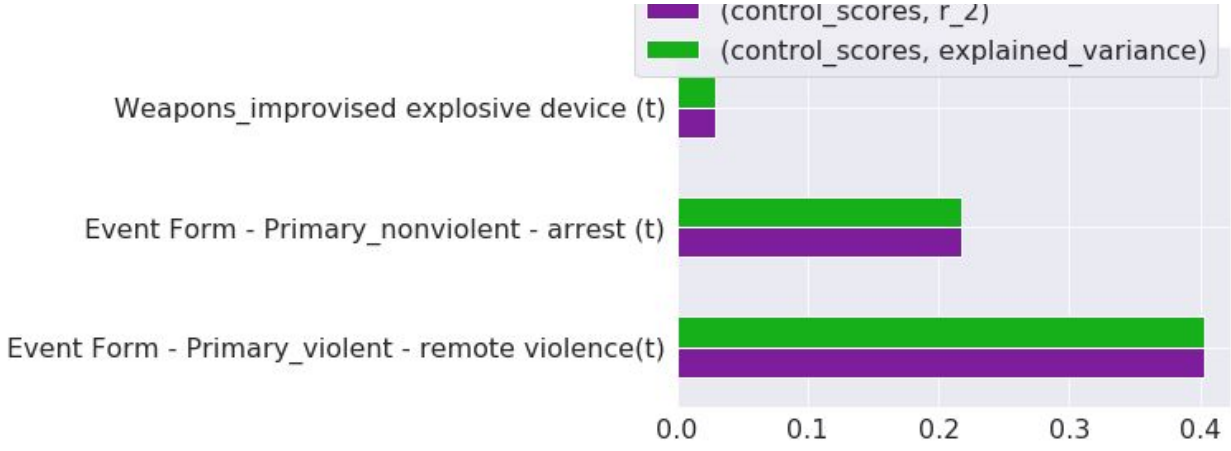
Predict

y_pred:

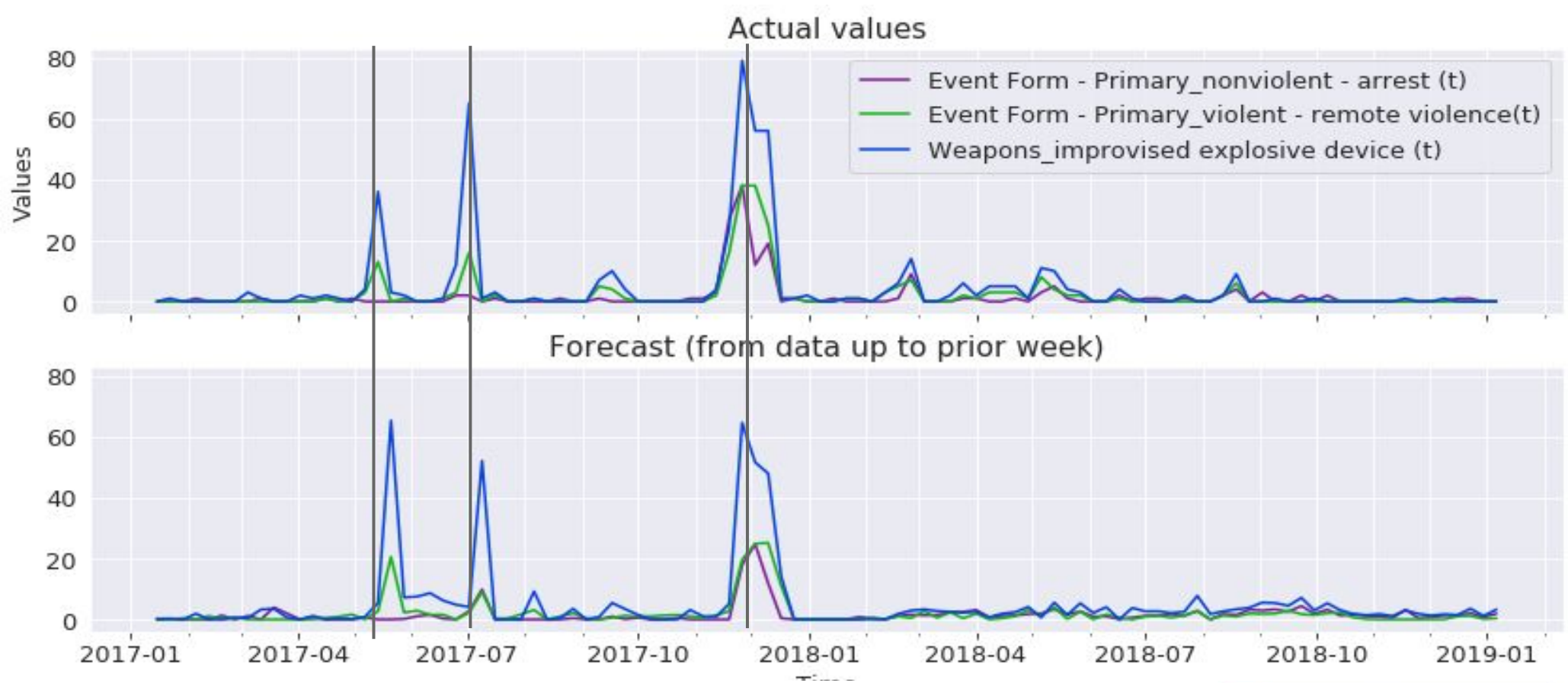
Event Date	Total killed(t)
2019-01-01	4.2



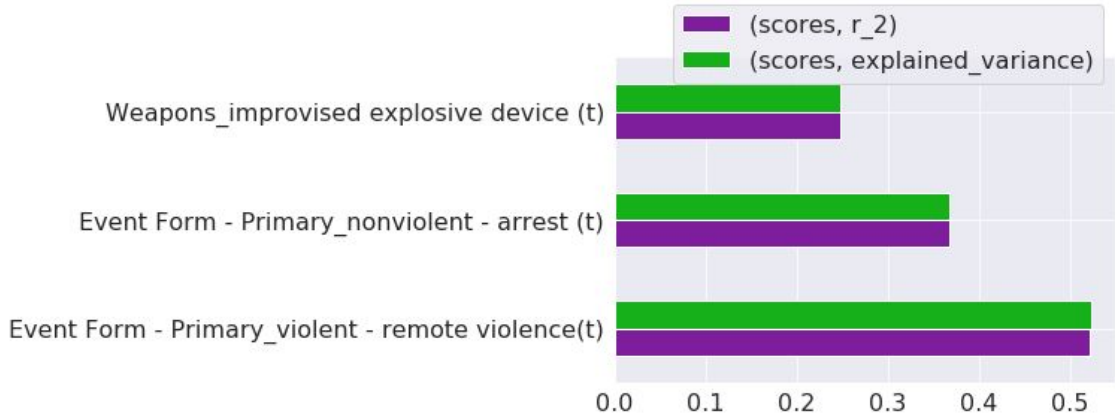
Baseline



Baseline scores are calculated by assuming that each variable will have the same value as the previous week, such as in the example variables above. Note the gridlines show the forecast lagging behind the ground truth by a week. r^2 and explained variance measure the degree actual values are predicted by the forecast, with 1 being the highest possible score.

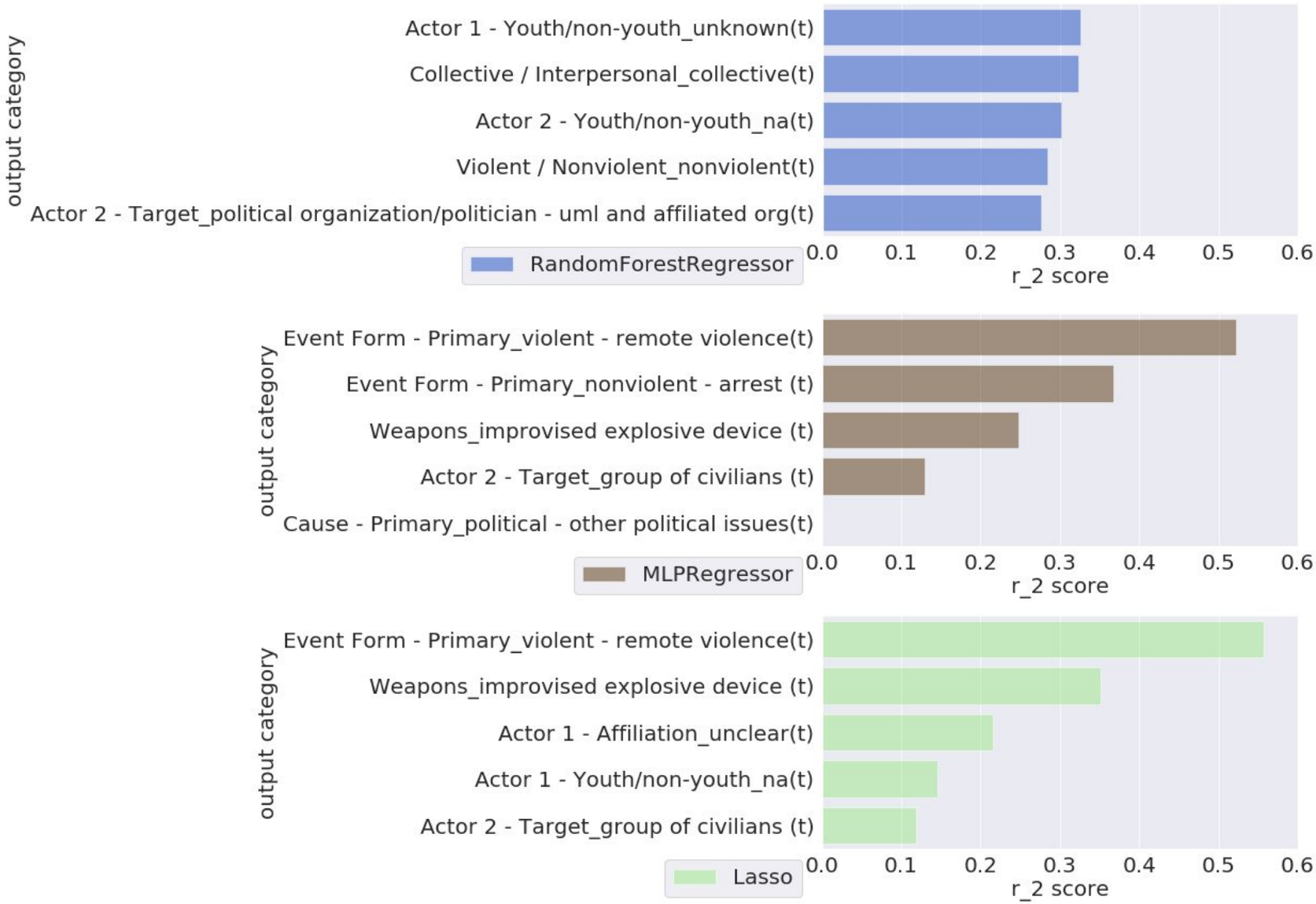


Improvement



In the run above, MLPRegressor (a neural net), significantly improves the accuracy in forecasts over the baseline in multiple tracked variables. Note how the forecast lags a week behind the first two peaks, but the model seems to have “learned” enough to predict the 3rd one.

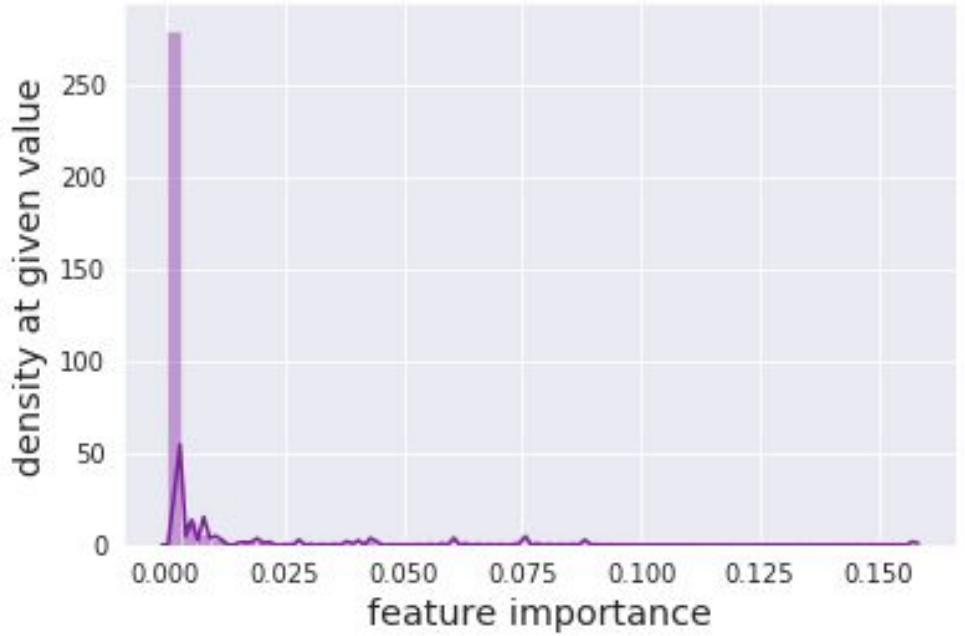
Best Models

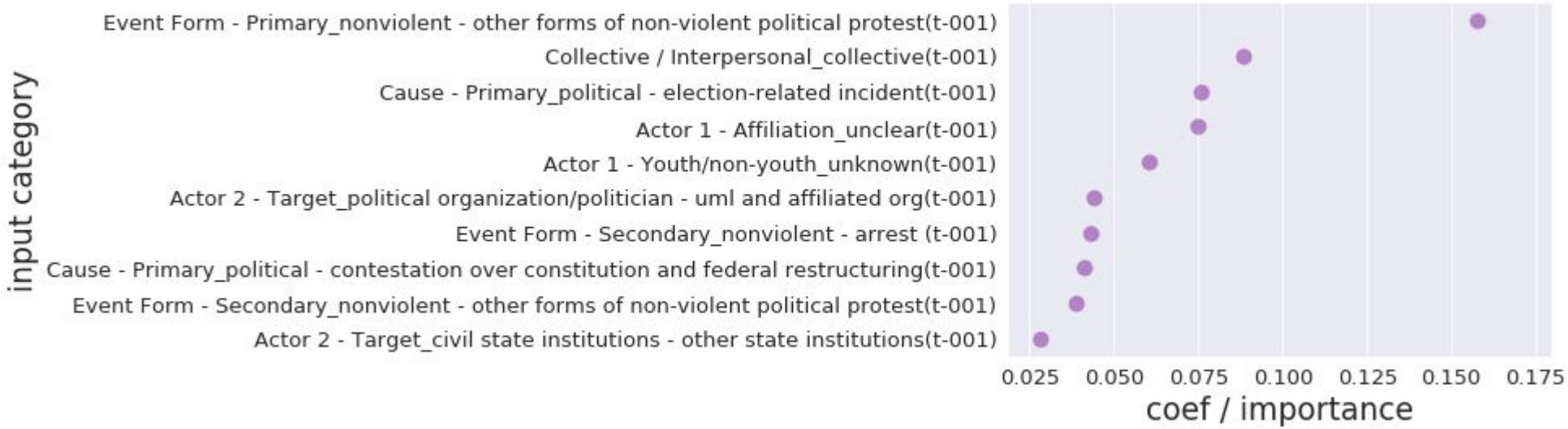


RandomForestRegressor performs the best out of all models attempted by showing significant improvement over control in a diversity of categories (Top 5 shown). MLPRegressor, and Lasso score higher in a narrower set of categories.

Feature Engineering

Using best scoring model
(Random Forest Regressor)





The vast majority of features in the dataset have close to zero impact on predictions. The most important features for prediction can be interpreted by the project as indicators for further analysis, and used for feature pruning to improve efficiency of the models.

Findings

- Many of the highest scoring forecasts are for variables that would seem useful for the project to predict. ie:
  - Improvised Explosive Device
  - Collective Violence
  - Total Injured

- The most important features for predicting multiple variables were associated with electoral periods. ie:
  - Non-violent political protest
  - Collective violence
  - Election related incidents

- It appears certain models can be trained to accurately forecast peaks associated with electoral periods beforehand
  - Further exploration needed

Possible Next Steps

This Month: Forecast:

Rough forecast mockup based on current platform

Improve machine learning models:

- Add known events (ie: elections), as features into time series data
- Use classifier to predict simple increase/decrease in next periods violence
- Break time series analysis into provinces/districts
- Add population and demographics as features

Use insights from predictive modeling in analysis distributed by NMP.

- ie: Explore why Sarlahi experienced less violence
- Promote most important features as early warning indicators
- Add forecasts using machine learning to analysis products. (running models again with updated data is relatively easy)

Add forecasts to Nepal Monitor Platform.

- Display forecast estimates for select variables next to previous accuracy (r^2), with 95% confidence interval
- Significant web development needed to automate

Appendix

1. Datasets
   a. NepalMonitor.org
   b. 2011 census data
   c. Shapefile Export
2. Project on git-hub
   a. https://github.com/Nhorning/NM_Analysis/tree/ML
3. Supervised Learning for Time Series
   a. https://machinelearningmastery.com/convert-time-series-supervised-learning-problem-python/