War Susceptibility

Berk Mankaliye and Isaac Attuah

CSC597 - Statistical Learning

Why This Topic?

War is a universal language. It is ingrained in our biology.

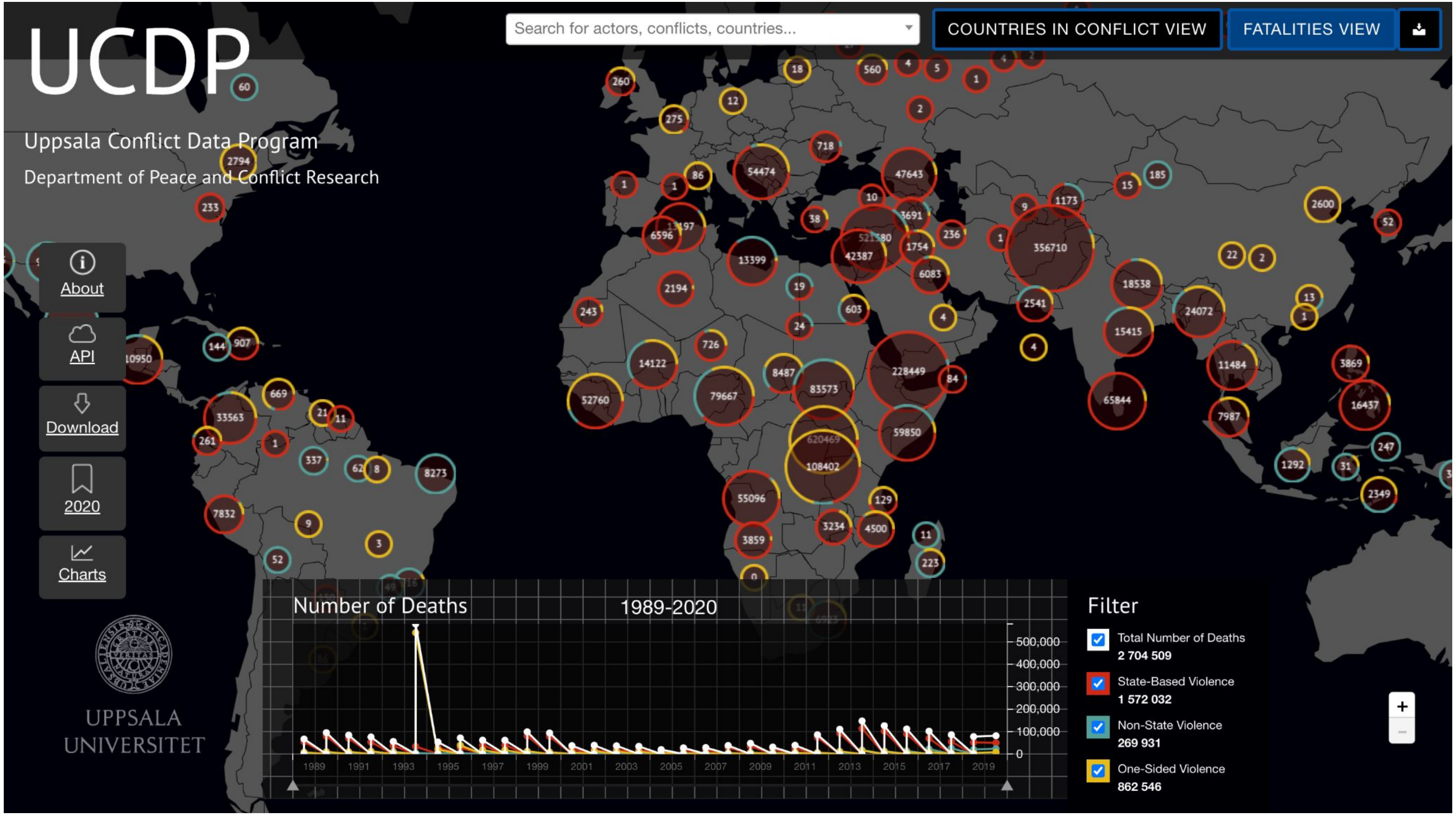
Current events are proof that war is (and can) still happen(ing).

If we can’t prevent it, can we predict its occurence?? Its duration??

Background

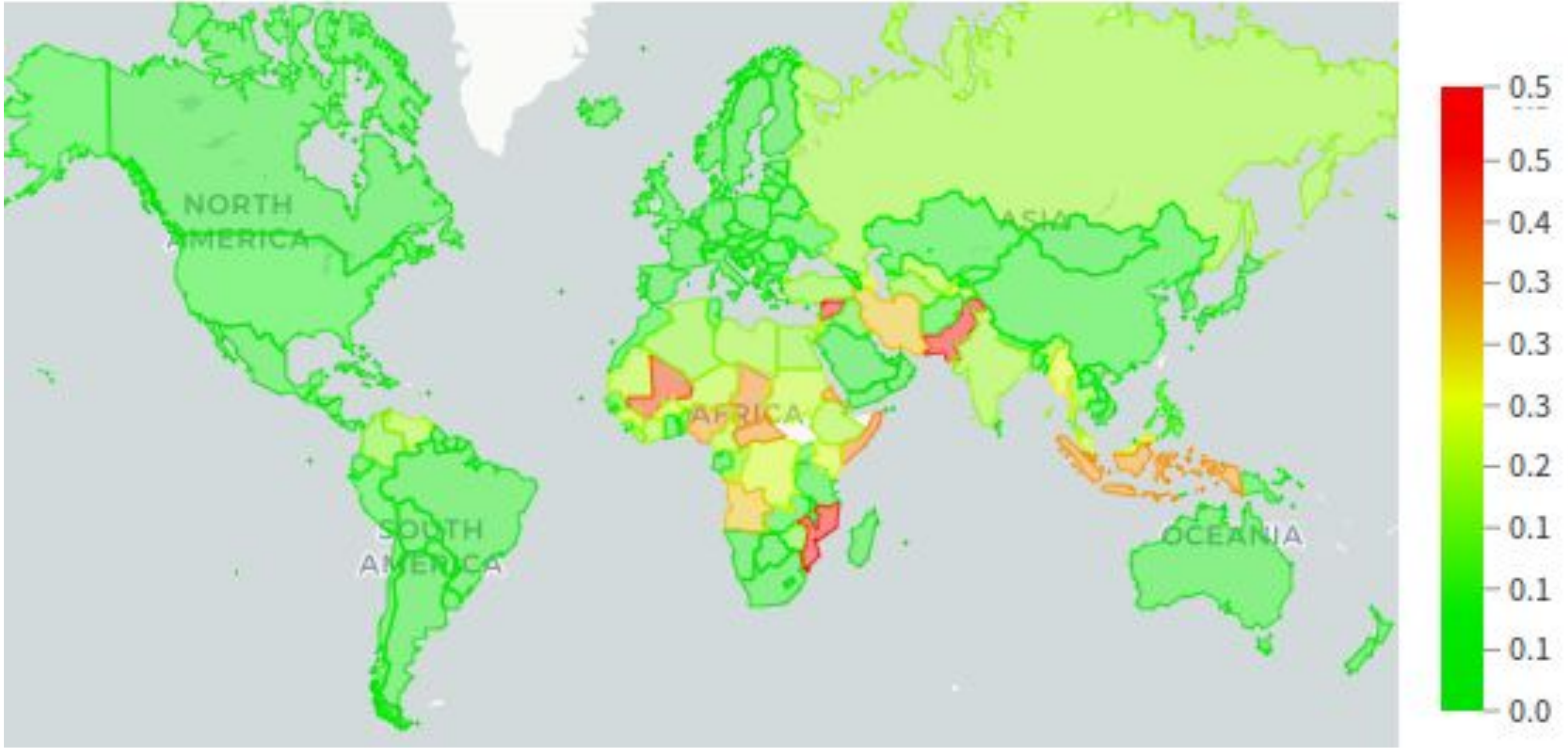
State of the Art

UCPD Web Visualization



Predicting Future Wars

Insights from Open Data and Machine Learning - Towards Data Science



The Hypothesis:

If we know the presence of conflict and its continuance into the future after a given year, can we predict the presence of war in 20 years?

Dataset Information

The Uppsala Conflict Data Program (UCDP) is the world's main provider of data on organized violence and the oldest ongoing data collection project for civil war, with a history of almost 40 years. Its definition of armed conflict has become the global standard of how conflicts are systematically defined and studied.

For this project, we utilized the UCDP (Uppsala Conflict Data Program / Peace Research Institute Oslo) Onset Dataset version 19.1 to perform our analysis.

- 547 instances
- 12 attributes

Different Attributes

abc	year_prev	onset3
name	newconf	onset5
gwno_a	onset1	onset10
year	onset2	onset20

Key: **Categorical**
Predictors
Prediction

Conflict here is the Onset of an intrastate armed conflict, > 25 battle deaths.

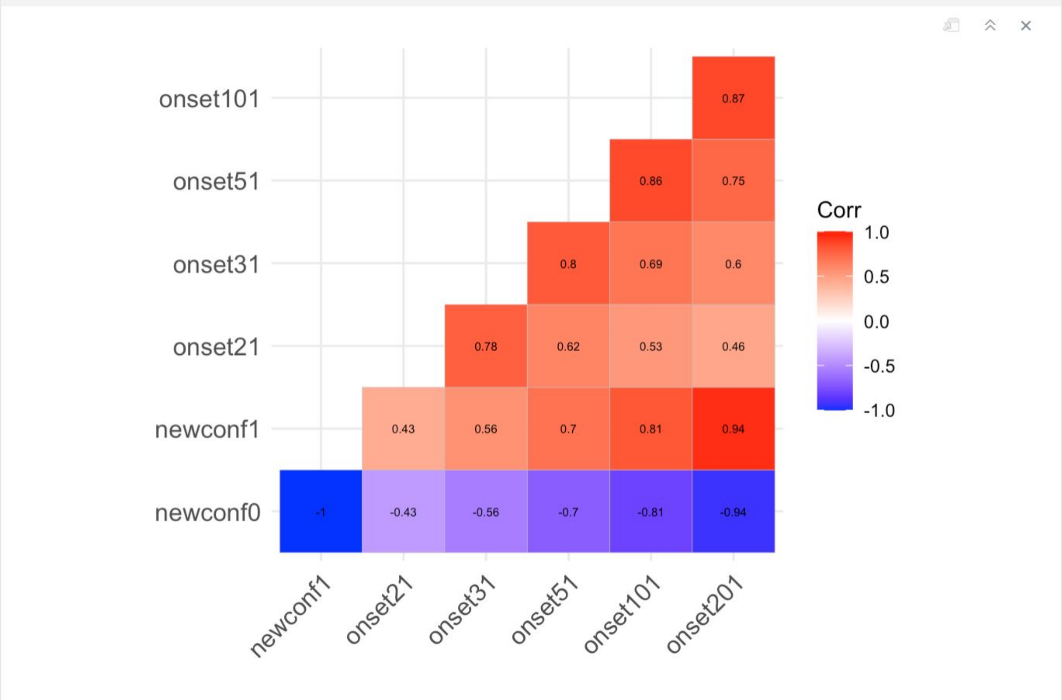
Snapshot of first 10 entries in the database

abc	name	year	gwno_a	newconf	onset1	onset2	onset3	onset5	onset10	onset20	year_prev
USA	United States of America	1950	2	1	1	1	1	1	1	1	1815
USA	United States of America	2001	2	1	1	1	1	1	1	1	1815
HAI	Haiti	2004	41	0	1	1	1	1	1	0	1991
TRI	Trinidad and Tobago	1990	52	1	1	1	1	1	1	1	1815
SAL	El Salvador	1972	92	1	1	1	1	1	1	1	1815
PAN	Panama	1989	95	1	1	1	1	1	1	1	1815
VEN	Venezuela	1962	101	1	1	1	1	1	1	1	1815
URU	Uruguay	1972	165	1	1	1	1	1	1	1	1815
UKG	United Kingdom	1962	200	1	1	1	1	1	1	1	1815
UKG	United Kingdom	1971	200	1	1	1	1	1	1	1	1815
FRN	France	1946	220	1	1	1	1	1	1	1	1815
FRN	France	1946	220	1	1	1	1	1	1	1	1815

Preprocessing & **Splitting**

- Loaded raw dataset
- Turned multiple columns into factors due to classification
- Checked for empty instances
- **Split into training and testing**
  - 438 training / 109 testing
- **Removed unnecessary columns**
  - “abc”, “name”, “year”, “gwno_a”, “onset1”, “year prev”

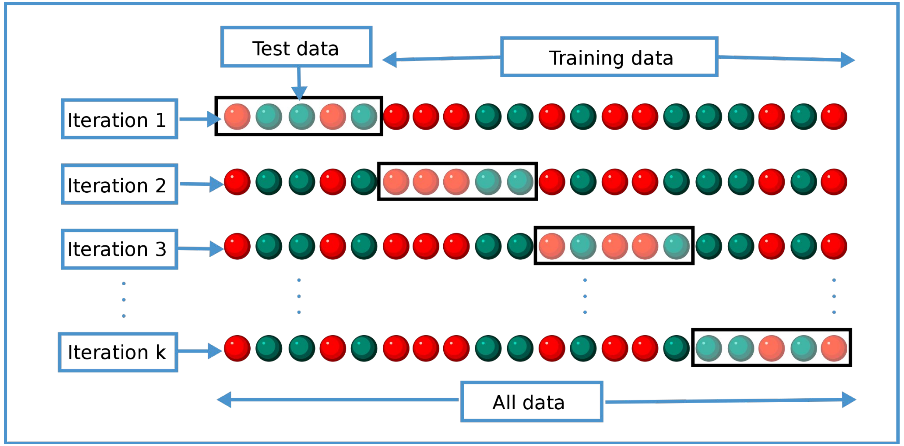
Correlation



Pearson correlation was used to provide correlation coefficients for categorical variables

Cross Validation

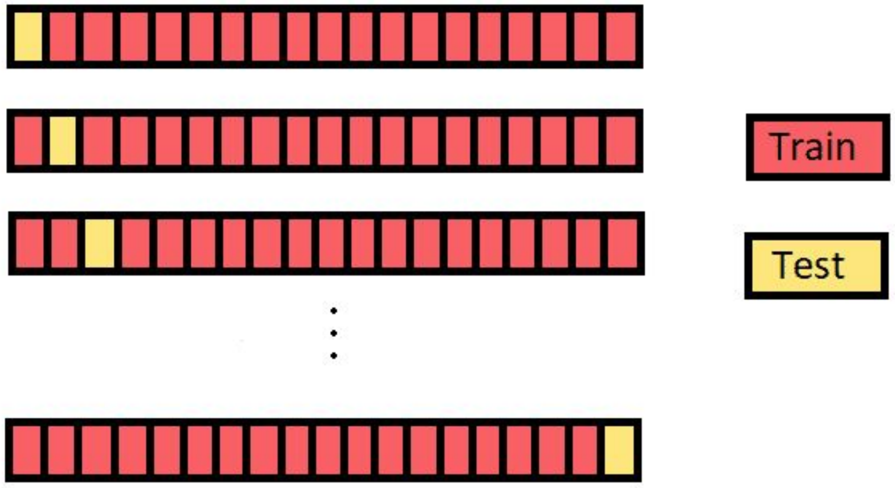
- A resampling procedure used to evaluate machine learning models on a limited data sample.



Cross Validation - LOOCV

LOOCV is the cross-validation technique in which the **size of the fold** is “1” with “k” being set to the number of observations in the data.

This variation is useful when the **training data is of limited size** and the number of parameters to be **tested are limited.**



Models Used

We used different methods shown throughout the semester

Models used in this experiment are

- Logistic Regression
- Decision Trees
- Support Vector Machines
- Generalized Additive Model

Logistic Regression

Logistic Regression

- A model that uses the P(Y=1)
    - Output: 0 < x < 1
- Applicable in our project design

```
glm.pred  0  1
       0 49  3
       1  0 57
Test Error Rate
[1] 0.02752294
```

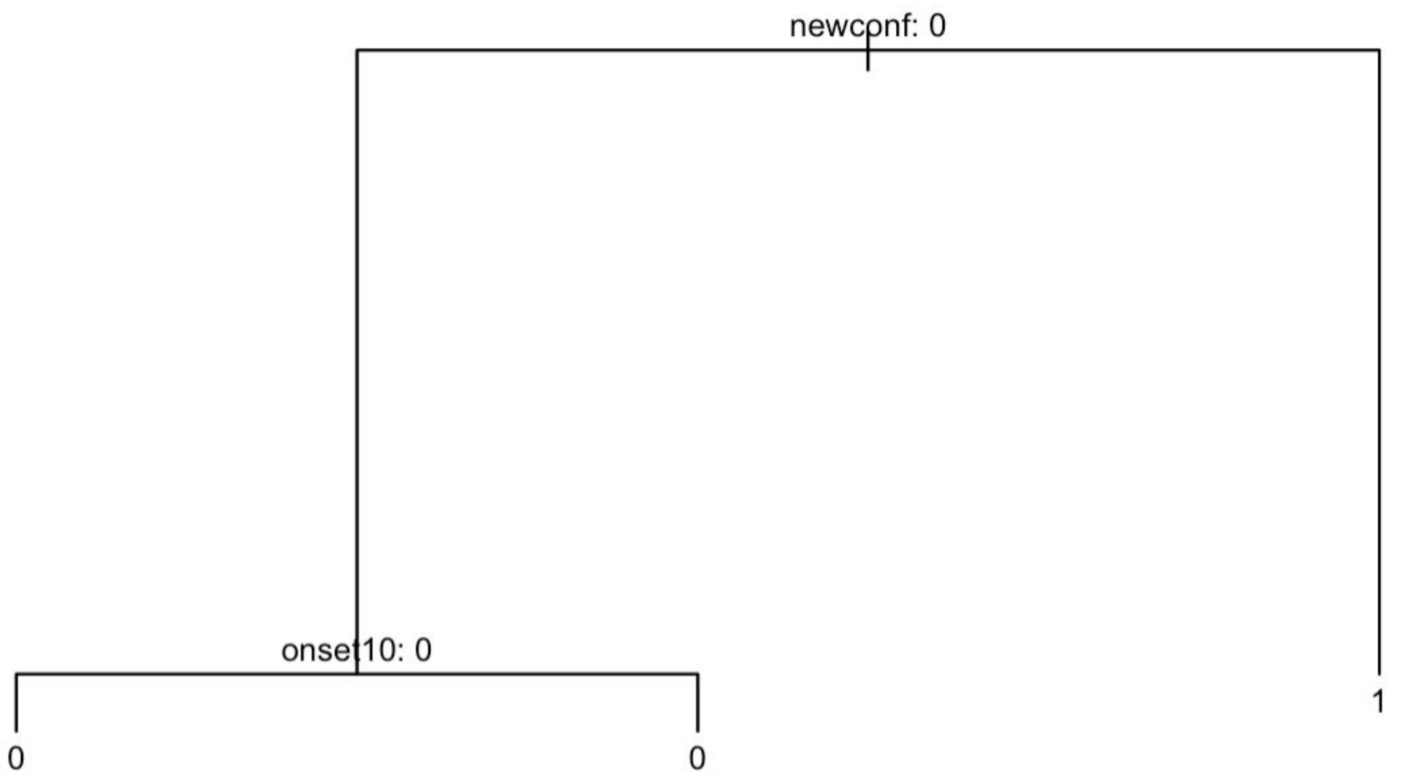
$$\boxed{p = P(Y=1) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}}$$

Decision Trees

Decision Trees

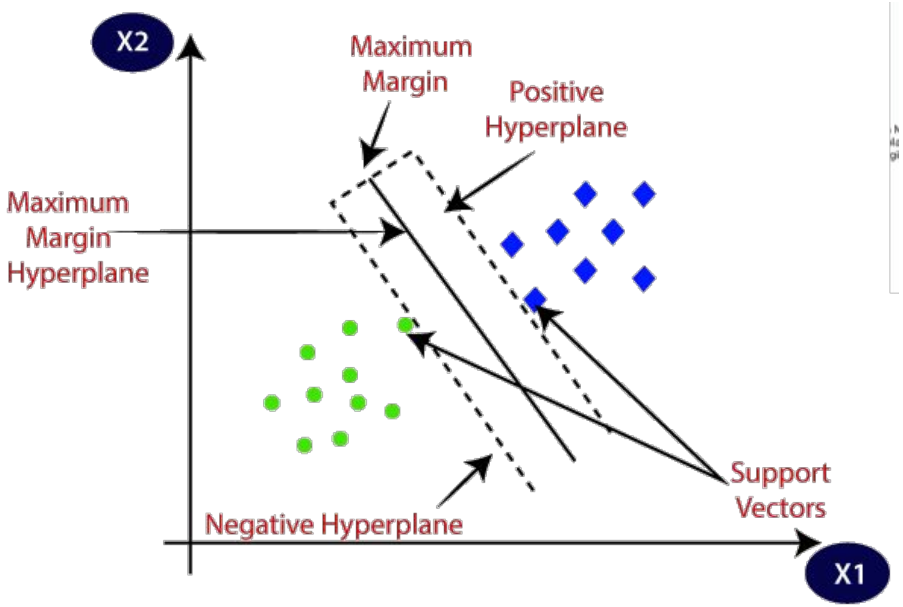
- Easy interpretation and visualization of the problem
- Different approach than Logistic Regression
    - Helps check for non-linearity of output (onset20)

```
tree.pred  0  1
        0 49  3
        1  0 57
[1] 0.9724771
```


newconf: 0
onset10: 0
0
0
1

Support Vector Machines

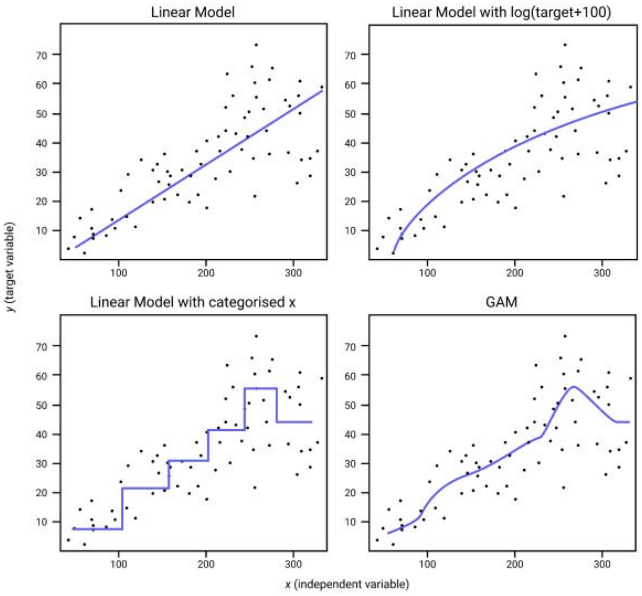
- In the SVM algorithm, we plot each data item as a point in n-dimensional space with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well




```
          Reference
Prediction  0  1
         0 49  3
         1  0 57

               Accuracy : 0.9725
                 95% CI : (0.9217, 0.9943)
    No Information Rate : 0.5505
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9447

 Mcnemar's Test P-Value : 0.2482

            Sensitivity : 1.0000
            Specificity : 0.9500
         Pos Pred Value : 0.9423
         Neg Pred Value : 1.0000
             Prevalence : 0.4495
         Detection Rate : 0.4495
   Detection Prevalence : 0.4771
      Balanced Accuracy : 0.9750

       'Positive' Class : 0
```

Generalized Additive Models

- A GAM is a linear model with a key difference when compared to Generalised Linear Models such as Linear Regression. A GAM is allowed to learn non-linear features.




```
Generalized Additive Model using Splines

438 samples
  5 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 437, 437, 437, 437, 437, 437, ...
Resampling results across tuning parameters:

  select  Accuracy   Kappa
  FALSE   0.9680365  0.9356545
   TRUE   0.9680365  0.9356545

Tuning parameter 'method' was held constant at a value of GCV.Cp
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were select = FALSE and method = GCV.Cp.
```

Conclusion & Future Work

Conclusion

- Cross validation helps to prevent overfitting but takes a lot of time to execute.
- Small datasets tend to have similar outcomes even for different models.
- Features must have enough variation in order to be useful in a model (onset1 could not be used since it was all 0)
- Good training / testing results do not infer good predictions in the real world

Future Goals

- Use more non-linear models
- Incorporate other datasets including data with multiple categories
- Narrow down data to specific country/time period
- Provide live visualizations for different findings / models

Sources

- https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/
- https://towardsdatascience.com/predicting-future-wars-7764639f1d8d
- https://ucdp.uu.se/encyclopedia
- https://www.statology.org/correlation-between-categorical-variables/
- https://medium.com/100daysofmlcode/day-59-of-100daysofml-542274f360c8
- https://towardsdatascience.com/generalised-additive-models-6dfbedf1350a
- https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/
- https://datascience.stackexchange.com/questions/893/how-to-get-correlation-between-two-categorical-variable-and-a-categorical-variab