Predictive Analytics of The Global Terrorism Database

Gabrielle Agrocostea, Nandini Shah, Vidyavisal Mangipudi

What's the Question?

-Predict whether an event would be successful

- Success defined as a terrorist event happening as planned

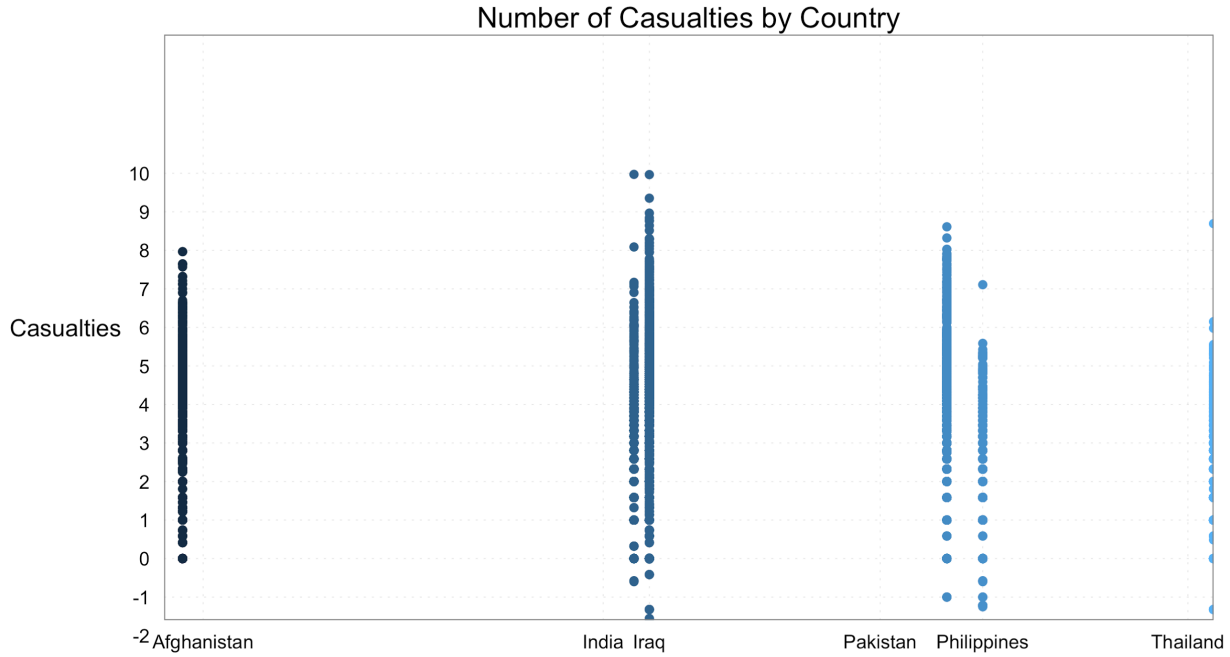
-Predict the number of casualties in the test set

- Casualties includes number of people killed and number of people wounded.

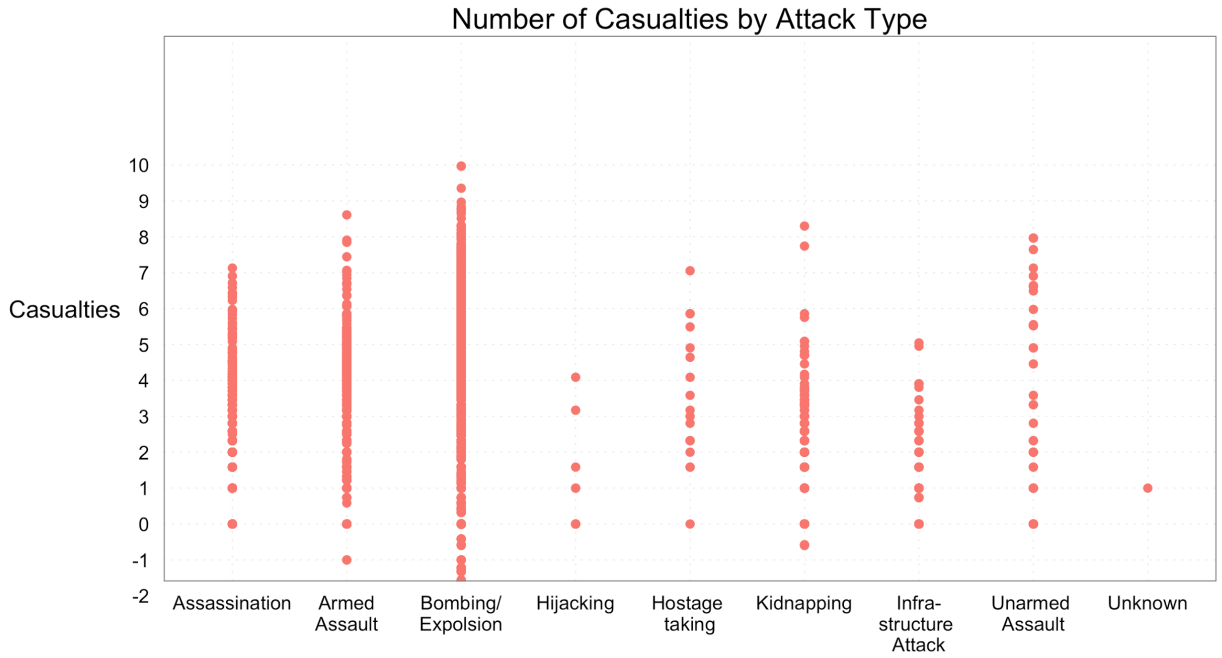
Source:Global Terrorism Data

- Data available from 1970 to 2013
- Definition of a terrorist event:
  - Incident must be intentional
  - Incident must entail a threat of violence
  - Act must attain an economic, financial, political or religious goal
- Contains continuous, discrete and categorical variables
- Dataset used (2006 to 2013) includes ~45,000 instances and 136 features.

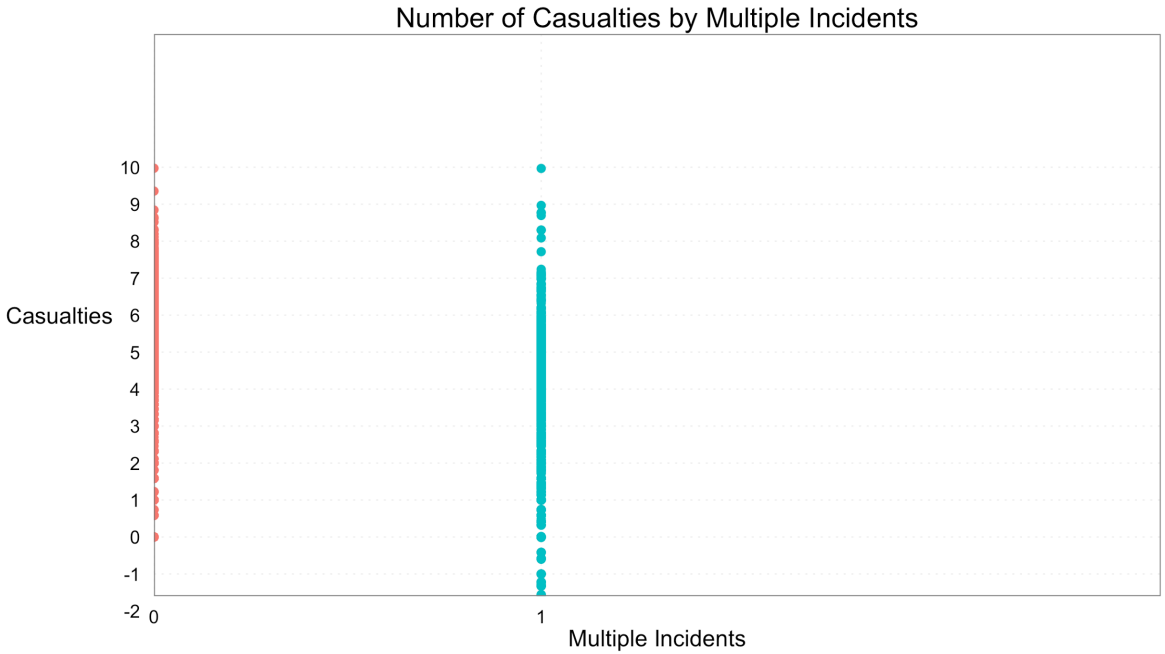
Exploratory Data Analysis



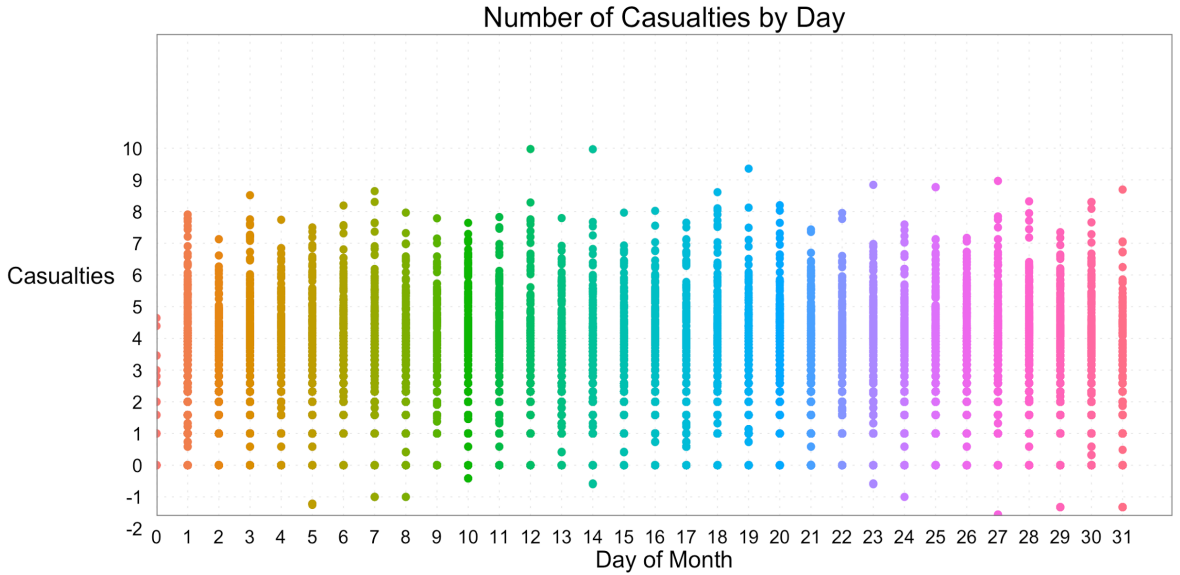
Exploratory Data Analysis



Exploratory Data Analysis



Exploratory Data Analysis



Data Cleaning

- Filter and retain the top 6 countries (70% of dataset)
  - Afghanistan, India, Iraq, Pakistan, Philippines and Thailand
- Dealing with missing data:
  - Exclude any column which has more than 50% of missing data
- Exclude any column with text (news snippet/summary)
- Remove any column with any unnecessary information (event_id/ LatLong)

Data Cleaning II

- Response Variable nCasualty<- nKilled + nWounded
- Vectorized categorical columns
- Approaches to dealing with ‘NA’ data points
  - Appropriate assumptions based on documentation (ex: nationality)
  - Implemented Linear Regression within the feature
  - Delete the rows with NA’s in valuable features and/or few rows
- Filtered dataset has ~30,000 instances and 25 columns

Implementation: “success”

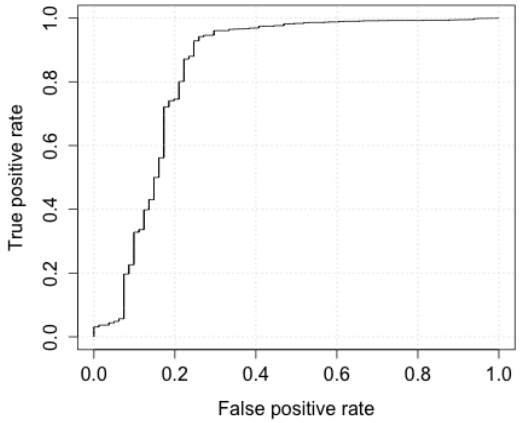
- Binary classifier
- Model choice -> Logistic regression
- AUC: 0.838, Accuracy: 96%
- Confusion matrix:

Actual Values

Predicted Values

Class	0	1
FALSE	13	15
TRUE	68	2131

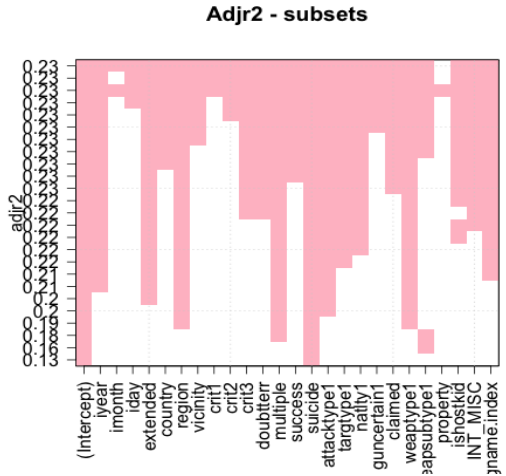
ROC curve

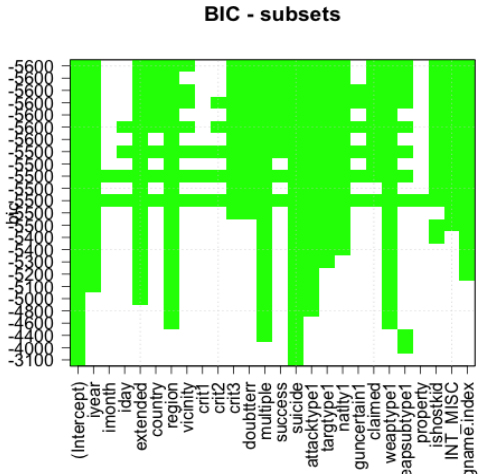


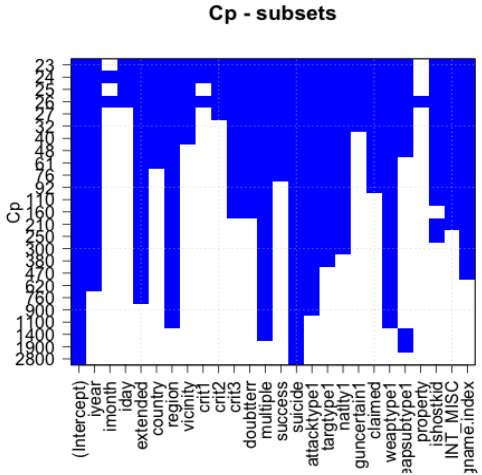
- Report bias and more (8.4% success = 0)
- Reason for not exploring: Definition of success

Implementation: #casualties

- Technique I: Subset selection



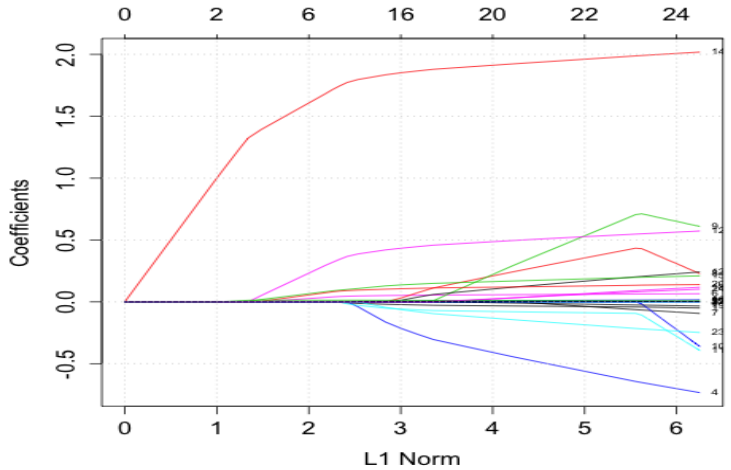


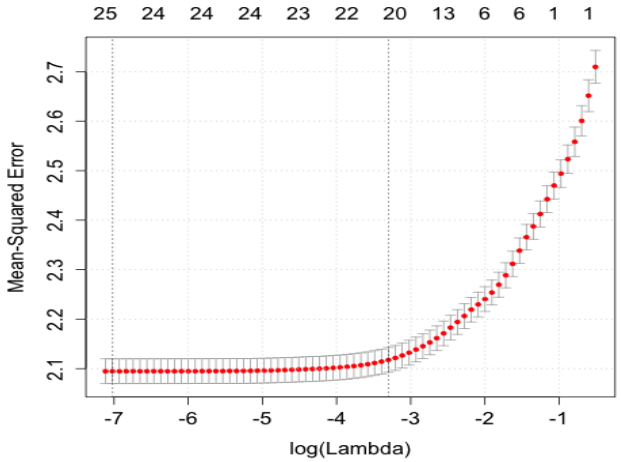


Implementation: #casualties

- Large number of columns
- Real valued prediction
- Natural model choice -> Lasso Regression (cv.glmnet)

$$\hat{w}_{\text{lasso }\lambda} := \arg\min_{w \in \mathbb{R}^p} \frac{1}{n} \|y - Xw\|_2^2 + \lambda \|w\|_1$$



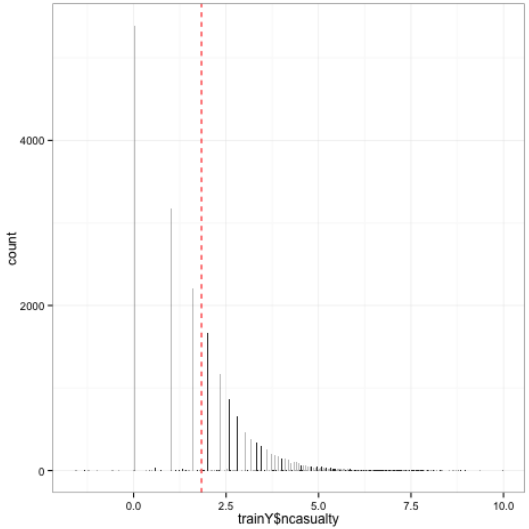


Analysis: #casualties

- Both models use similar features
- Comparison of RMSE:

Characteristics	RMSE train	RMSE test	Model Complexity
Subset Selection	1.455	1.455	20/25
Lasso Regression	1.444	1.458	25/25

Sampling Distribution



Feature Rankings: #casualties

Feature	Lasso	SS
suicide	1	1
multiple	2	2
crit2	3	3
crit1	4	4
attacktype	5	6
weapontype	6	7
success	7	8
region	8	5

Feature	Lasso	SS
int_misc	9	11
gun_certain	10	10
weapon_sub	11	9
target_type	12	13
groupName	13	12
country	14	14
nationality	15	17
claimed	16	18

Feature	Lasso	SS
year	17	20
doubt_terror	18	21
isHostageKid	19	25
extended	20	23
month	21	26
day	22	19
vicinity	23	15
crit3	24	22

Summary

- Exploratory Data Analysis
- Data Cleaning
- Implementation:
  - “success” - Logistic Regression
  - #casualties - Subset Selection & Lasso Regression
- Analysis:
  - Size of data set and number of features is small currently so appears to be method agnostic

What did WE learn?

- Data cleaning takes forever
- NA values are the bane of our existence
- Ctrl + C is your best friend