Predicting Perpetrators of Unclaimed Terrorist Attacks

ECON 8330 - Business Forecasting
Final Presentation

Karen Hardy, Dominika Jedinak,
Sasi Kulandasamy, Anni Zou

Agenda

- Define the Problem
  - Preview Results
- Diagnosis of Data
- Plan of Action
  - Cleaning the data
  - Selecting our model
- Review Results

Define the Problem

- Thousands of terrorist attacks were committed in Pakistan (2007-2018)
- Some attacks are claimed by the group that committed them
- Can we predict the perpetrators of attacks that go unclaimed?
  - Specifically those of the Tehrik-i-Taliban (TTP)
- Why does this matter?
  - Help understand patterns, possibly prevent future attacks

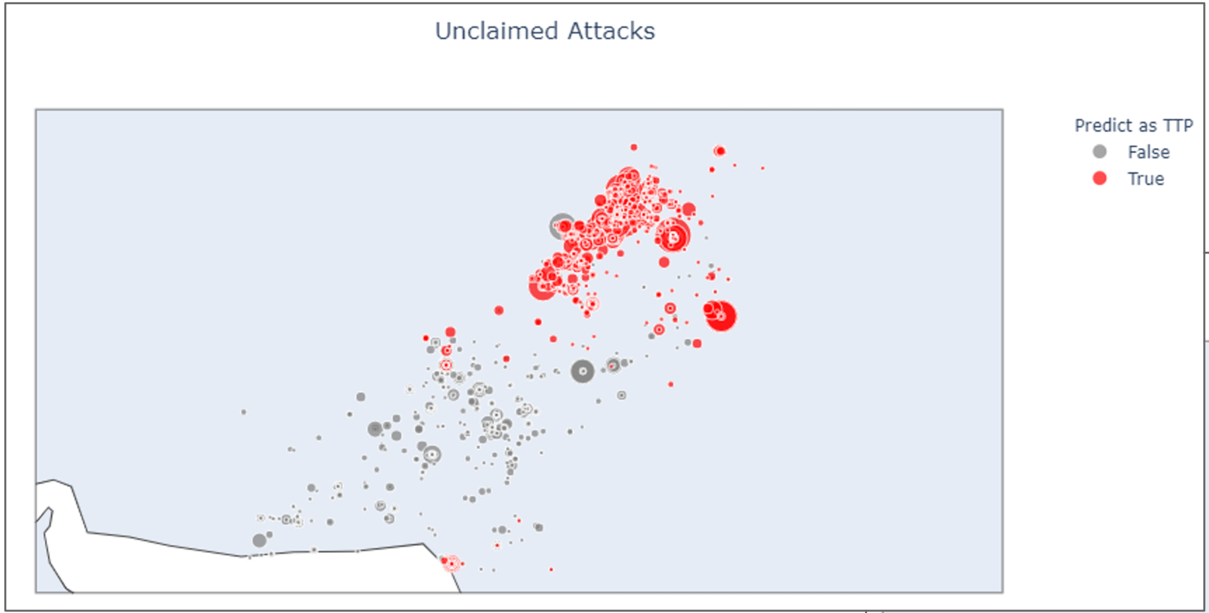
Results

51.3% of unclaimed attacks likely perpetrated by TTP

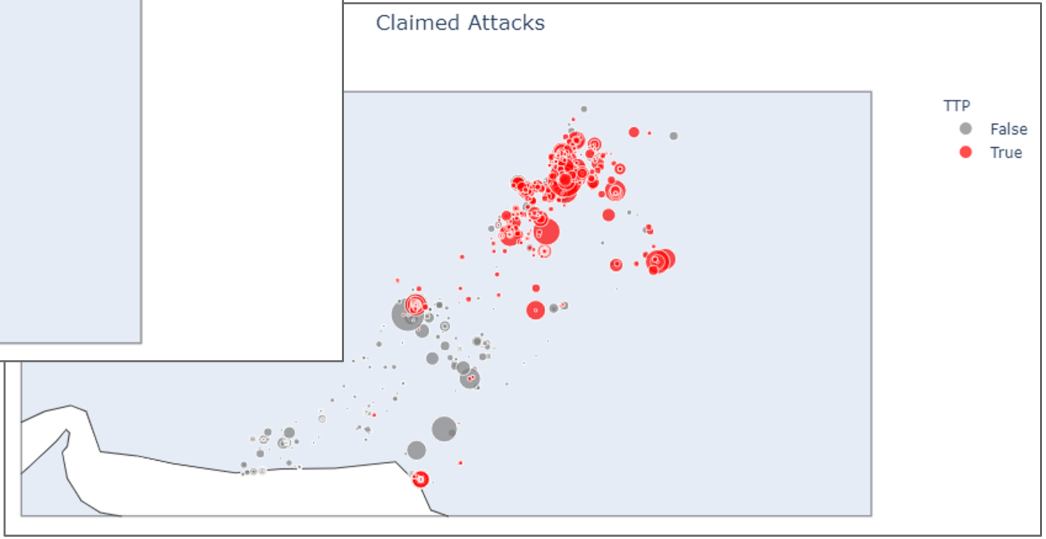
- 8,873 unclaimed attacks analyzed
  - 4,552 likely perpetrated by TTP
- Important variables
  - Location, day and week of attack, number killed, number wounded

- 2,950 claimed attacks
  - TTP committed 44.9% of the claimed attacks

Results

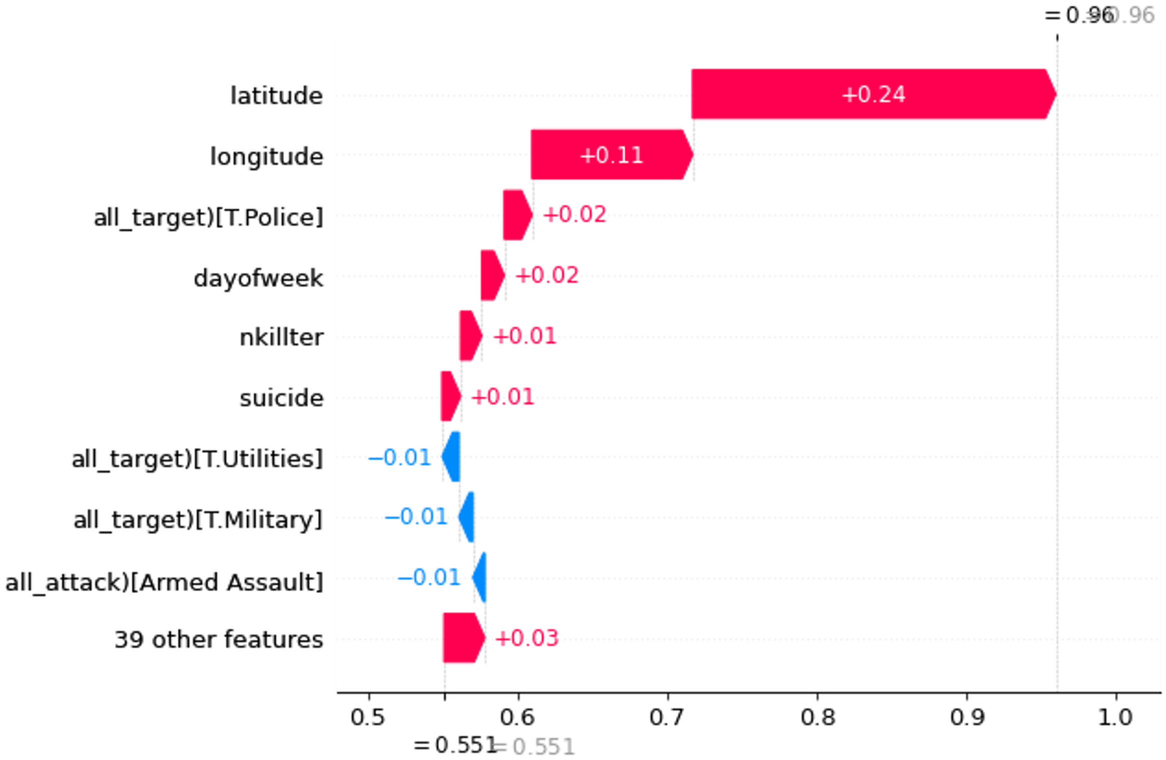


Unclaimed Attacks
Predict as TTP
False
True

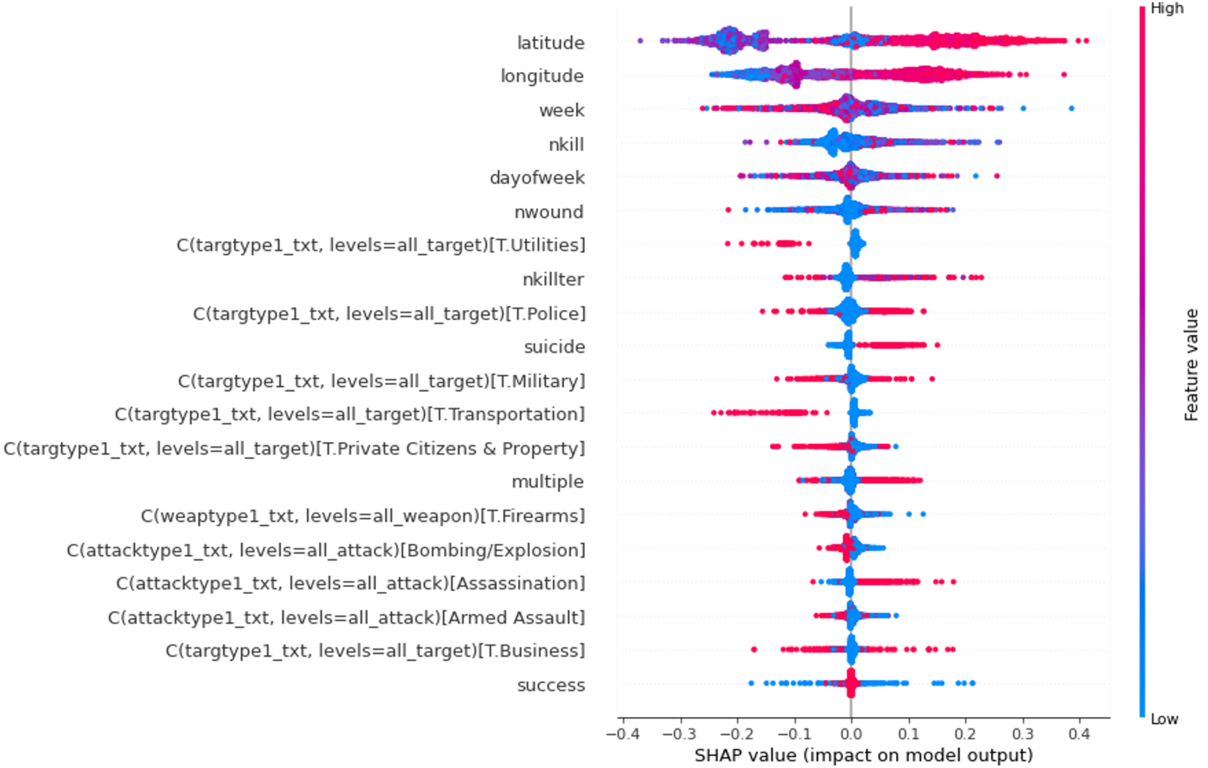


Claimed Attacks
TTP
False
True

Results– SHAP



Results – SHAP



The Data

- 12,237 attacks between 2007-2018
  - 2,950 claimed
  - 9,287 unclaimed (414 dropped)
- Several variables to describe the location, type of attack, attack details, and number of people killed or wounded
- Columns with many null values
  - Location (text description)
  - Target Subtype
  - Corp1
  - Claimed / claimed mode
  - Weapon Subtype

Selecting Our Training Data

Included

- Date (as day of week, week of year)
- Latitude, longitude
- Multiple, success, suicide
- Attack, target, weapon type (text fields)
- Number killed, wounded (including subset for US affliates and terrorists)

Excluded

- Data with lots of null values
- Region, providence/state, city
- Summary
- Group name, claimed

Cleaning the Data

- Dropped those with null values in the killed/wounded fields
  - 414 rows in all
- Split into Training and Prediction based on gname = Unknown
  - Create dummy variable for TTP claimed attacks
- Merge year, month, day and create datetime fields

- Created lists of categorical fields to avoid issues with categories not in the training data

```
all_attack = data['attacktype1_txt'].unique()
all_target = data['targtype1_txt'].unique()
all_weapon = data['weaptype1_txt'].unique()
```

```
C(attacktype1_txt, levels=all_attack)
```

Selecting Our Model – Complex Classification Models

Gradient Boosted Tree (GBT)

- GBT generally produce higher accuracy level compared to other models by reducing errors in bias
- GBT works better with fewer input variables, but GBT tends to overfit data with more noise

Random Forest (RF)

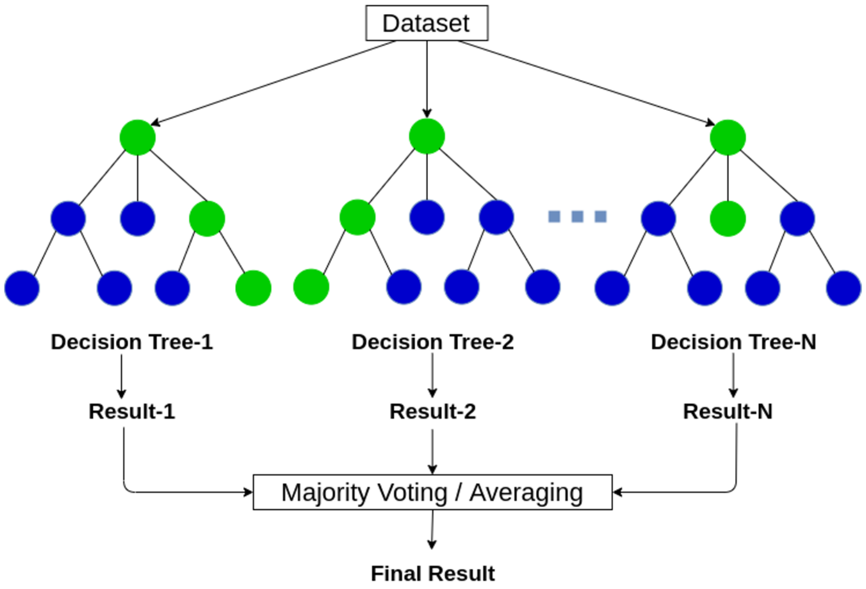
- RF produce higher accuracy level by reducing variance in predictions
- RF is less computationally expensive

Compare the Results:

- Similar prediction results (GBT 53.78% VS RF 51.3%)
- GBT has lower in-sample accuracy (GBT 68.5% compared to RF 99.8%)
- RF model includes more variable than GBT does

Random Forest Model

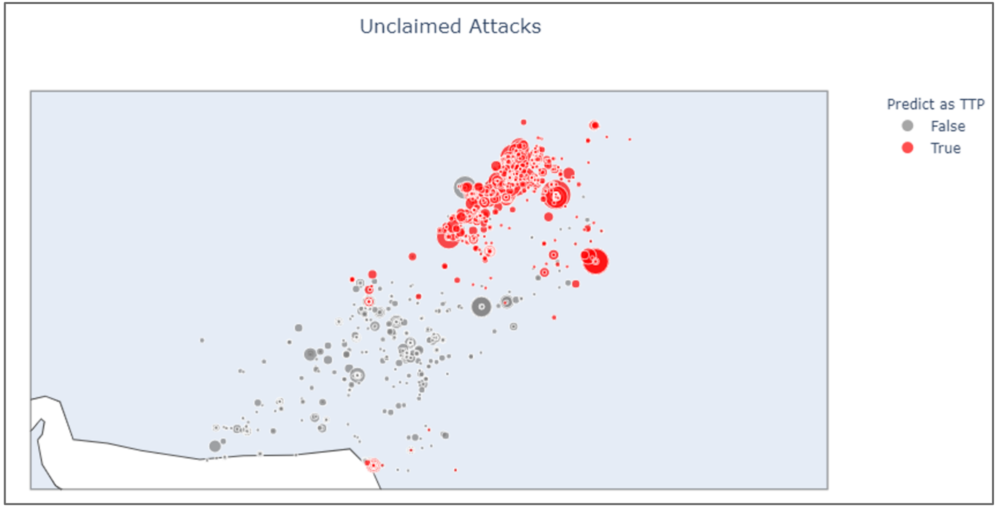
- 99.8% in sample accuracy
- Default 100 trees in model
- Did not limit depth or samples in a split
  - Did not change results much



Results

51.3% of unclaimed attacks likely perpetrated by TTP

- 8,873 unclaimed attacks analyzed
  - 4,552 likely perpetrated by TTP
- TTP committed 44.9% of the claimed attacks
- Important variables
  - Location, date of attack, number killed, number wounded



Questions?