Big Data and Security

Jeffrey Borowitz, PhD
Lecturer
Sam Nunn School of International Affairs

Terrorism in Bank Data

Levitt Paper

- Steve Levitt (author of Freakonomics)
- Attempts to predict who will be arrested for terrorism using account info from a bank in the UK.

Data

- List of 112 “terrorists” – people arrested for terrorism related charges
  - Database from Salaam
  - Newspaper accounts of terrorists in UK
  - Other accounts solicited by law enforcement
- Anonymized banking data from large UK bank
  - Was individual arrested for terrorism related charges?
  - Islamic names?
  - Age, married, types of accounts, employment status, residential status
  - “Variable Z” - summary of account where most individuals have value 0, but some have high values

Georgia Tech

Notes on Sampling

- Used 112 positives
- Sample of 19,000 individuals with higher weight on those with Islamic names
  - Chances of inclusion are 1/2,000 for no Islamic names, and 1/100 for one Islamic name, and 1/35 for two Islamic names
- Why not use all the data?
- Why the funny sampling frame?

Georgia Tech

Approach

- Try to predict whether an individual is a terrorist
- Probit model
  - Y - was individual arrested for terrorist charges?
  - X - Islamic names, age, married, Variable Z etc
- Predicted value of Y from this model is your main output
  - High values of Y have a high chance of being terrorists

Correlations with Terrorism

Table 3: Z-scores for Probit Estimation

	Overall bank customers (1)	Two Muslim Names (2)	One or No Muslim Names (3)
Muslim Names: First name only	4.44***		4.68***
Muslim Names: Last name only	2.07*		3.00***
Muslim Names: Both	18.09***		
Variable Z	6.50***	5.79***	4.47***
Gender and Age			
Male	2.20*	2.69**	0.79
Age: 26 to 35	0.78	2.19*	-2.06*
Age: 36 to 45	-2.25*	-0.65	-2.79**
Age: Over 45	-4.50***	-3.33***	-3.50***
Employment Status			
Self-Employed	-0.43	-0.55	0.24
Unemployed	1.52	1.89	0.19
Full-time Student	0.02	1.18	-1.16
Homemaker	-0.28	-0.12	1.01
Marital Status			
Single	1.64	0.87	0.95
Married	1.65	0.63	1.87
Residential status			
Renter	3.91***	3.54**	1.35
With parents	2.12*	1.87	1.44
Other	1.96*	1.45	1.40
Unknown	4.57***	3.75***	2.42*
Proximity to mosque	2.51*	0.92	3.43***
ATM Usage			
Average withdrawal amount	2.56*	1.51	2.44*
Withdrew during Nighttime (8pm-6am)	2.74**	2.23*	2.01*
Withdrew during Friday prayer (10am-12pm)	-1.50	0.34	-2.04*
Withdrew during Friday prayer (12-1pm)	-0.81	-1.00	-0.07
Withdrew during Friday prayer (1-3pm)	-1.44	-0.02	-1.80
Types of financial products			
Business customer	0.89	-0.63	1.50
Debit/credit cards	2.33*	2.03*	1.37
Loans (excluding Mortgages)	0.69	0.18	0.80
Mortgages	-1.02	-0.28	-
Life Insurance	-0.84	-1.46	0.07
Savings products	-4.05***	-3.42***	-2.18*
Extras	1.36	1.99*	-
Longterm	-1.94	-0.66	-2.05*
Number of observations	18,929	5,883	13,046

Levitt, S. D. (n.d.). Identifying Terrorists using Banking Data. Retrieved from http://www.jeffborowitz.com/Levitt.pdf

Correlations with Terrorism: Notes

- Things in here basically make sense
- Ethnic profiling is useful (terrorists in England in this time period are Islamic terrorists)
- Terrorists:
    - Are male, and either 26-35 or under 26
    - More likely to withdraw at night
    - Have various versions of shorter time horizons: less likely to be savings customers, more likely to be credit customers, more likely to be renters
    - Withdraw more at a single time (better for secrecy in operations, e.g. The Terrorist's Dilemma: *Managing Violent Covert Organizations*)

Georgia Tech

Probability of Being a Terrorist

	All customers	Two Muslim Names	One or No Muslim Names
	(1)	(2)	(3)
Random sample			
Mean	0.0007%	0.0392%	0.0002%
Median	0.0000%	0.0060%	0.0000%
90th percentile	0.0004%	0.1086%	0.0003%
99th percentile	0.0085%	0.3924%	0.0033%
Positives			
Mean	0.5184%	0.1998%	0.0023%
Median	0.0676%	0.1072%	0.0008%
90th percentile	0.4280%	0.5324%	0.0064%
99th percentile	5.5188%	1.4764%	0.0119%
High Variable Z			
Mean	5.3669%	14.6104%	2.2638%
Median	0.0805%	0.9256%	0.0116%
90th percentile	3.4010%	95.4286%	0.6030%
99th percentile	100.0000%	100.0000%	89.4580%

Levitt, S. D. (n.d.). Identifying Terrorists using Banking Data. Retrieved from http://www.jeffborowitz.com/Levitt.pdf

Probability of Being a Terrorist: Notes

- The overall average is .0007%, or 7 terrorists per million customers
  - If you just used ethnic profiling, this would get you up to 392 per million
- Among the 112 positives, the predicted rate is 0.5%, or 5,000 terrorists per million
  - This isn’t that absolutely likely, but the model predicts it is 700 times more likely that the actual positives are terrorists than the average person.
- Among people with high variable Z, the predicted rate is 5.4%, or 54,000 terrorists per million
  - This gets even higher with ethnic profiling

Prospectives for Investigation

Number of people identified	At least 1 Islamic Name	2 Islamic Names	Full model without names or Z	Full model without Z	Full model with Z but without names	Full model
	(1)	(2)	(3)	(4)	(5)	(6)
10,000	0.20	3.98	6.62	25.95	10.98	30.10
5,000	0.10	1.98	4.12	15.78	8.92	19.80
2,500	0.05	0.99	2.61	9.18	7.28	13.14
1,000	0.02	0.39	1.33	4.34	5.99	8.07
500	0.01	0.20	0.69	2.41	5.27	5.89
250	0.00	0.10	0.46	1.33	4.41	4.40

Levitt, S. D. (n.d.). Identifying Terrorists using Banking Data. Retrieved from http://www.jeffborowitz.com/Levitt.pdf

Prospectives for Investigation: Notes

- Pure ethnic profiling doesn’t work that well
  - If you looked at people with 2 Islamic names, you would find 400 suspected terrorists per million
  - You can do almost 3x better with zero ethnic profiling but with no names (1,098 per million)
- The first people you identify are more likely to be terrorists than subsequent ones
  - First 250 investigations would yield more than 4 terrorists in the full model
  - The next 250 would yield only 1-2 more in

Lesson Summary

- The Levitt Paper attempts to predict who will be arrested for terrorism using account info from a bank in the UK and UK newspaper records
- It uses a sample of 19,000 in a probit model, with an undersample of normal accounts
- A model based approach works well here
- "Ethnic profiling" works much less well than features that capture behavior like Variable Z