© 2019 SPLUNK INC.

Splunking Crime Part II — Analyzing Bias in Police Actions

Shashank Raina
Senior Splunk Consultant | NCC Group

splunk> .conf19

© 2019 SPLUNK INC.

Forward-Looking Statements

During the course of this presentation, we may make forward-looking statements regarding future events or plans of the company. We caution you that such statements reflect our current expectations and estimates based on factors currently known to us and that actual events or results may differ materially. The forward-looking statements made in the this presentation are being made as of the time and date of its live presentation. If reviewed after its live presentation, it may not contain current or accurate information. We do not assume any obligation to update any forward-looking statements made herein.

In addition, any information about our roadmap outlines our general product direction and is subject to change at any time without notice. It is for informational purposes only, and shall not be incorporated into any contract or other commitment. Splunk undertakes no obligation either to develop the features or functionalities described or to include any such feature or functionality in a future release.

Splunk, Splunk>, Turn Data Into Doing, The Engine for Machine Data, Splunk Cloud, Splunk Light and SPL are trademarks and registered trademarks of Splunk Inc. in the United States and other countries. All other brand names, product names, or trademarks belong to their respective owners. © 2019 Splunk Inc. All rights reserved.

© 2019 SPLUNK INC.

Who are We?

NCC Group is a security consultancy and advisory business helping to solve complex security challenges day in and day out.

Our Ninjas are based worldwide and passionate about making the internet safer and revolutionizing the way in which organizations think about cyber security.

- United Kingdom
- Manchester (HQ)
- London
- Cambridge
- Edinburgh
- Glasgow
- Leatherhead
- Leeds
- Milton Keynes
- Reading

- Europe
- Denmark
- Germany
- Lithuania
- Spain
- Switzerland
- Netherlands (Amsterdam)
- Netherlands (Delft)
- Netherlands(The Hague)

- United States & Canada
- Atlanta, GA
- Austin, TX
- Boston, MA
- Campbell, CA
- Chicago, IL
- New York, NY
- San Francisco, CA
- Seattle, WA
- Sunnyvale, CA
- Toronto, ON
- Kitchener, ON

- Asia Pacific
- Sydney
- Singapore
- UAE (Dubai)

© 2019 SPLUNK INC.

What are We Talking About Today?

Data

Challenges

Digital Technology Applications

Machine Learning

Bias

Analyse Bias

© 2019 SPLUNK INC.

Data

splunk> .conf19

© 2019 SPLUNK INC.

Data

Data

- Truth
- Improves Things
- Predicts Future

Crime Data

- What crimes are
- How many crimes are committed
- How the police service works

© 2019 SPLUNK INC.

Challenges

© 2019 SPLUNK INC.

Challenges

- Budget Cuts
- Increase in Numbers
- Advanced Adversaries
- Reduced Priority

© 2019 SPLUNK INC.

Digital Applications

splunk> .conf19

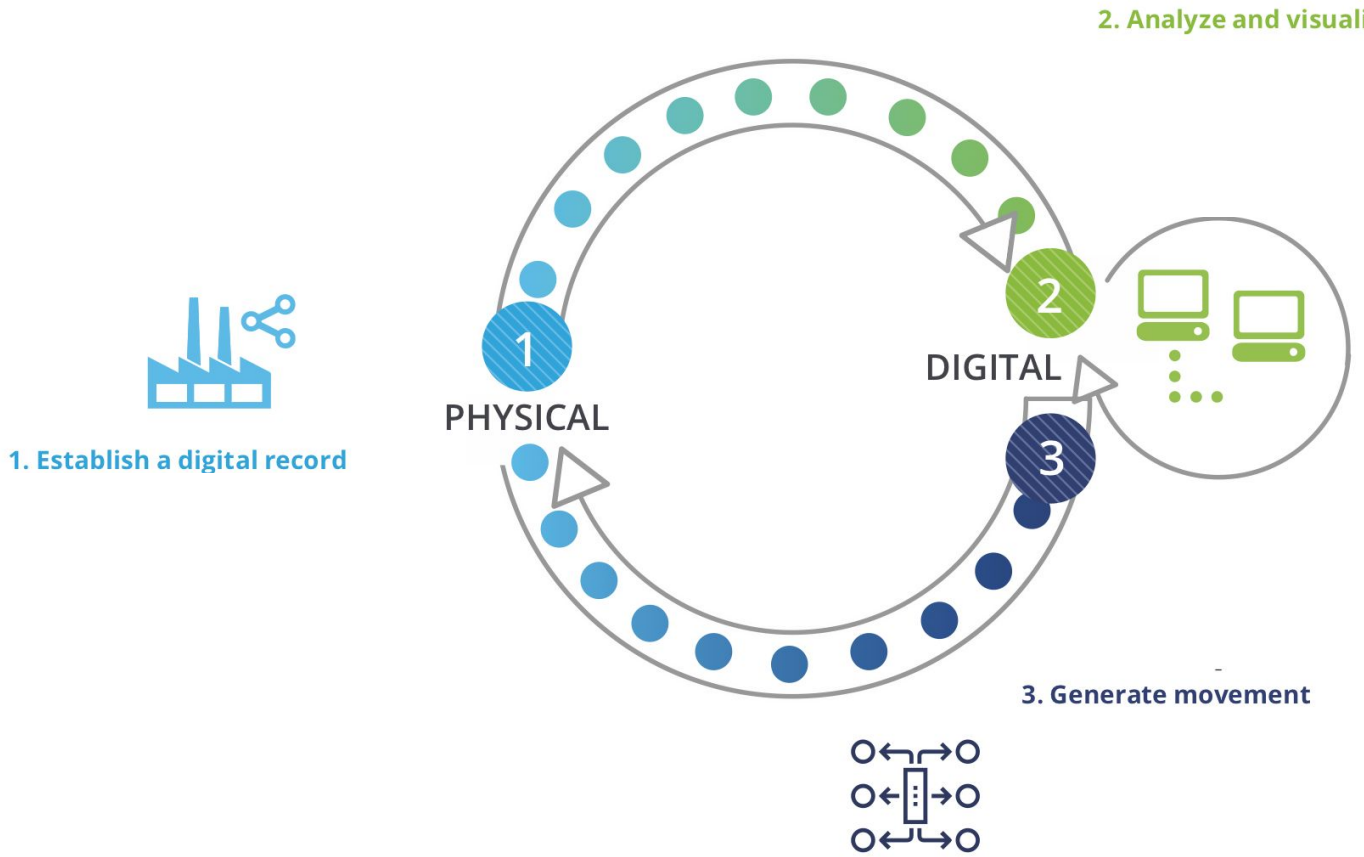
© 2019 SPLUNK INC.

Digital Applications

- Real world experience
- Advanced analytics
- Automation
- Feedback loops

© 2019 SPLUNK INC.

Physical to Digital to Physical Loop



© 2019 SPLUNK INC.

Bias

splunk> .conf19

© 2019 SPLUNK INC.

What is Bias and Why Should we Worry?

- Prejudice
- Disparity
- Unbalanced outcomes
- Responsibility to know the Effect

© 2019 SPLUNK INC.

How Bias is Introduced

- Data Collection
- Feature Engineering
- Algorithm Function
- Decision Making

Machine Learning Bias in Big Data

© 2019 SPLUNK INC.

- IT Ops
- Security
- Business

© 2019 SPLUNK INC.

Other Industries Impacted by ML Bias

- Banking
- Insurance
- Employment
- Fraud
- Government
- Finance

© 2019 SPLUNK INC.

Machine Learning for the Police

.conf19
splunk>

Machine Learning & Policing

© 2019 SPLUNK INC.

- Pattern Recognition
- Logical Evolution
- Police Intervention & Crime Prevention
- Unstructured Data

© 2019 SPLUNK INC.

Effect of Biased Predictive Models

- Negative Feedback Loop
- Incomplete Features
- Human Discretion

© 2019 SPLUNK INC.

Creating a Machine Learning Model in Splunk

splunk> .conf19

© 2019 SPLUNK INC.

What Data Sources Did we Use?

Crime Data

https://data.police.uk/

Office for National Statistics

Census Data

https://www.ons.gov.uk

LONDON DATASTORE

Various London Datasets

https://data.london.gov.uk

Trust for London
Tackling poverty and inequality

London Poverty Data

https://www.trustforlondon.org.uk/data/child-poverty-borough/

© 2019 SPLUNK INC.

How We Build ML in Splunk

Building the Dataset

- Crime per LSOA
- Census for different age groups
- School Absences
- Child Poverty
- Income

Analyse Data in Splunk

- kmeans
- analyzefields
- anomalousvalue

© 2019 SPLUNK INC.

How to Use Splunk ML Toolkit to Create a Predictive Model

Pre-Processing

- Transform your machine data
- 5 options
- We have used Field Selector

Algorithm Selection

- 7 algorithms
- We have used 4 of them

Predictive Model

- Highest “R Squared”
- Minimum RMSE (Root Mean Square Error)

© 2019 SPLUNK INC.

Measuring Bias in ML Models

splunk> .conf19

© 2019 SPLUNK INC.

Measuring Bias in Machine Learning Models

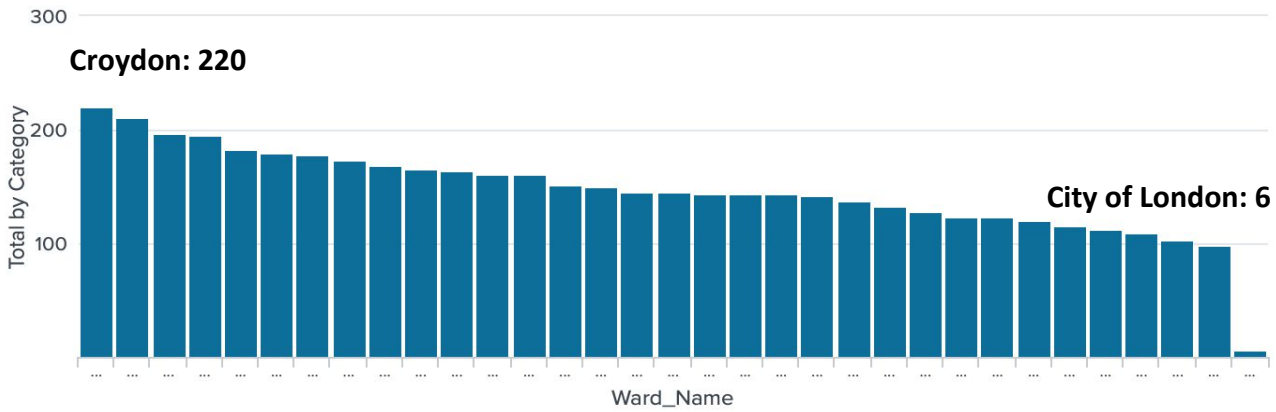
- Predictive Numeric fields
- Variance
- Binary Outcome Field
- Chosen Metrics
- Metrics Value

Representation Bias

© 2019 SPLUNK INC.

Difference in Number of Data Points compared to Croydon

Ward_Name	Total by Category	Representation Difference
City of London	6	-97.27%
Kingston upon Thames	98	-55.45%
Kensington and Chelsea	103	-53.18%
Barking and Dagenham	110	-50.00%
Hammersmith and Fulham	113	-48.64%
Richmond upon Thames	115	-47.73%
Sutton	121	-45.00%
Islington	123	-44.09%
Merton	124	-43.64%
Westminster	128	-41.82%



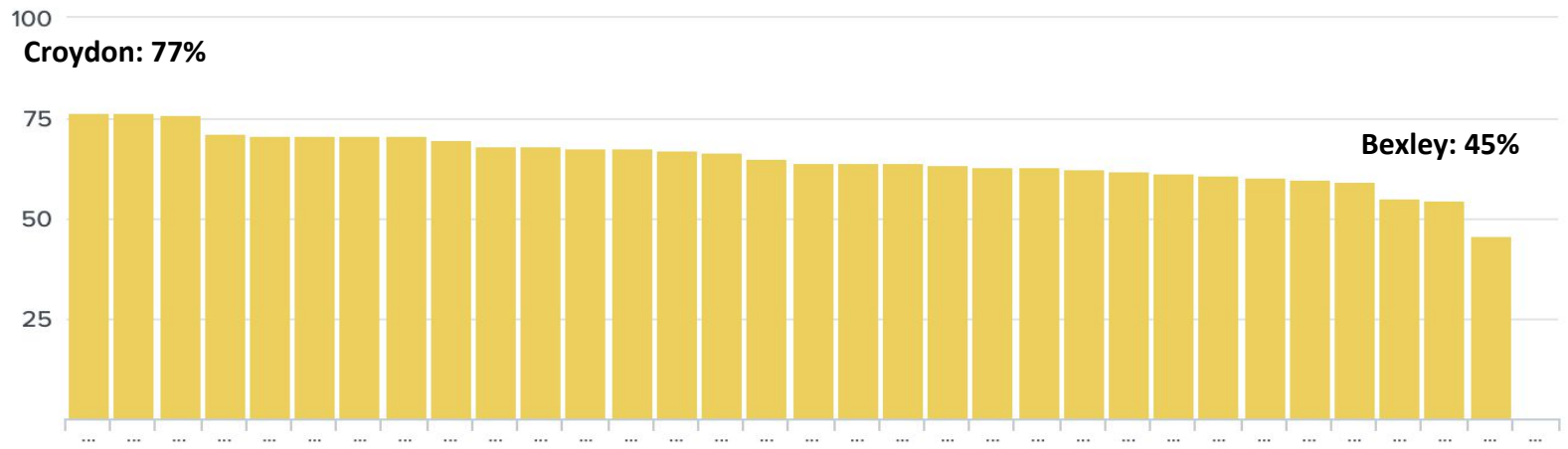
splunk> .conf19

Accuracy Bias

© 2019 SPLUNK INC.

Difference in Precision rate compared to Croydon

Bexley	-35.28%
Sutton	-23.08%
Greenwich	-22.48%
Westminster	-16.27%
Hackney	-15.78%
Harrow	-14.56%
Brent	-14.41%
Barnet	-13.11%
Richmond upon Thames	-12.93%
Tower Hamlets	-11.86%
Newham	-11.43%
Merton	-11.29%
Haringey	-10.52%



splunk> .conf19

© 2019 SPLUNK INC.

Analyze Police Data for Bias

splunk> .conf19

© 2019 SPLUNK INC.

Looking at the data

Ingest and analyse the police stop & search data to analyse Data and Feature Bias

Ignoring the outcome

- Count by gender
- By age group
- By ethnicity
- By location
- Object of search
- Ethnicity
- Removal of clothing
- By time of year

Outcome – ignoring location

- Count by gender
- By age group
- By ethnicity
- Object of search
- legislation

Outcome – by location

- Count by gender
- By age group
- By ethnicity
- Object of search
- Legislation

Analyse Bias in the Stop & Search Data

© 2019 SPLUNK INC.

Population of Britain

Total: 62 Million

White: 86%

Black: 3%

Asian: 7%

Other/Mixed: 4%

Stop & Search Data

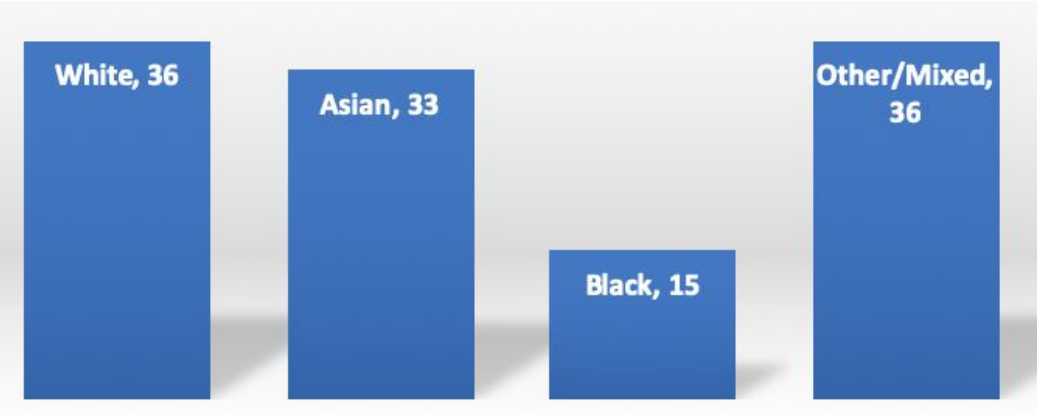
Total Events: 991,038

White: 53%

Black: 25%

Asian: 12%

Other/Mixed: 10%



Ethnicity of People Arrested after Stop and Search

- White: 36%
- Black: 15%
- Asian: 33%
- Other/Mixed: 36%
- Average: 30%

© 2019 SPLUNK INC.

‘Stop and Search’ Data Bias

- Population % v/s Stop & Search %
- Arrest Probability
- Biased Algorithm

Create Model for Predicting Arrests

© 2019 SPLUNK INC.

- Stop and search leading to an arrest
- Binary Outcome
- Predict Categorical Fields

© 2019 SPLUNK INC.

Confusion Matrix

Predicted actual	Predicted 0		Predicted 1	
0	True Positive	95782 (49.3%)	False Negative	98657 (50.7%)
1	False Positive	28257 (33.8%)	True Negative	55270 (66.2%)

© 2019 SPLUNK INC.

Algorithm Performance Measures

Characteristics of Logistic Regression Algorithm based on Stop & Search data

Precision	Recall	Accuracy	F1
0.65	0.54	0.54	0.59

© 2019 SPLUNK INC.

Algorithm Bias

- False Negative rate: 50%
- False Positive rate: 33%
- Precision: 0.65
- Recall: 0.54
- Accuracy: 54%

So the algorithm based on just the stop & search dataset from the police is not a very good measure of predictive policing and should not be used in isolation.

© 2019 SPLUNK INC.

Best Practices for Machine Learning Models

.conf19

splunk>

© 2019 SPLUNK INC.

What steps to follow when creating ML Models

- **Inputs v/s Outputs**
- **Fairness metrics**
- **Diverse team**
- **Data Source**
- **Fair**

© 2019 SPLUNK INC.

Key Takeaways

.conf19
splunk>

© 2019 SPLUNK INC.

Why is Analysing Bias important

- **Misbehaving Artificial Agents**
- **Accountability**
- **Opacity**

© 2019 SPLUNK INC.

How to reduce bias in Machine Learning Models

- **Right Learning Model**
- **Representative Training Data**
- **Monitor Performance**
- **Biasness v/s Accuracy**

© 2019 SPLUNK INC.

What to do after the session?

- **Think hard about the Data, Algorithm and the Output before creating any ML Model.**

Further Reading:

- **https://medium.com/datadriveninvestor/what-is-machine-learning-and-why-is-it-important-6779898227c1**
- **https://www.technologyreview.com/s/612957/predictive-policing-algorithms-ai-crime-dirty-data/**
- **https://statetechmagazine.com/article/2019/05/how-pattern-recognition-and-machine-learning-helps-public-safety-departments-perfcon**

© 2019 SPLUNK INC.

.conf19

splunk>

Thank You !

Go to the .conf19 mobile app to

RATE THIS SESSION