SciPy2022

Scientific Computing with Python
Austin, TX · July 11 - July 17

Searching for Anomalous Subsets? All You Need is Scanning.

Tanya Akumu
Research Engineer, IBM Research Africa

tanya.akumu@ibm.com | **LinkedIn**: Tanya Akumu | @*tanya-akumu*

IBM **Research** | Africa

The Team: IBM Research Africa – AI Sciences






Outline

Motivation

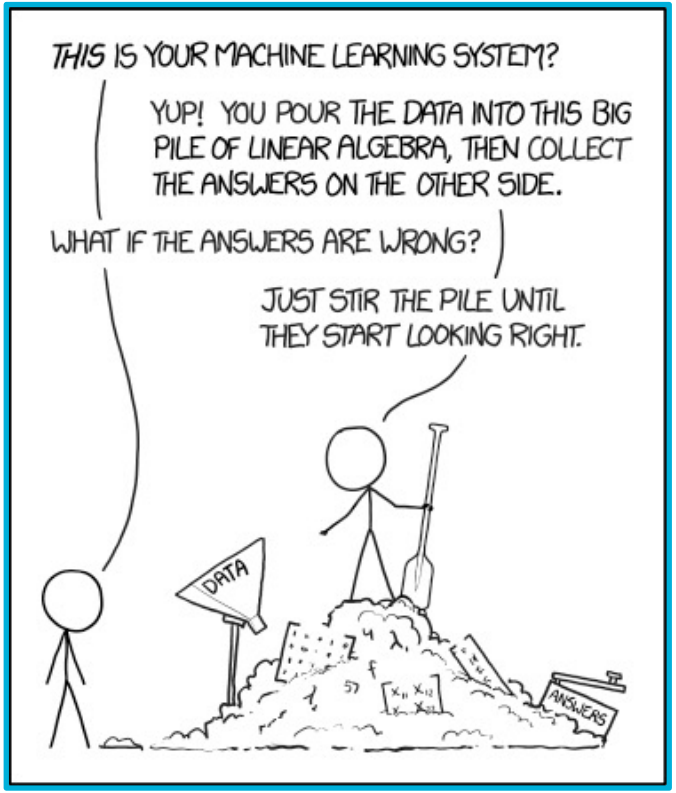
Approach

Examples

Demo

Why Do Anomalous Subsets Matter?

- Understanding our data through **disciplined** exploratory data analysis
- Improving our models through bias detection and explainability
- Detecting concept drifts in deployed machine learning systems

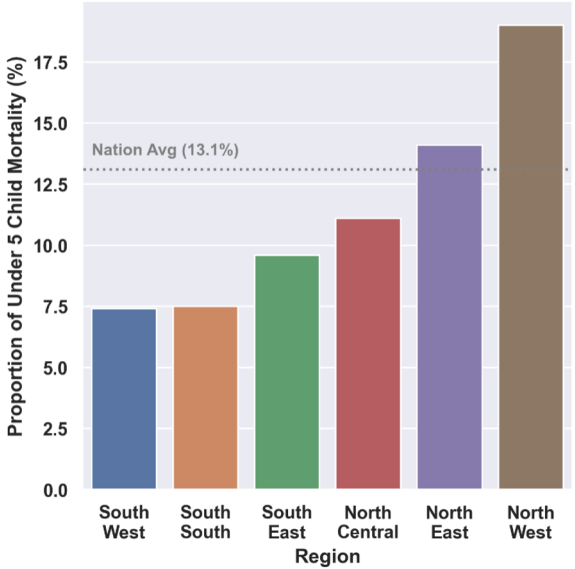


source: *xkcd.com*

Stratification

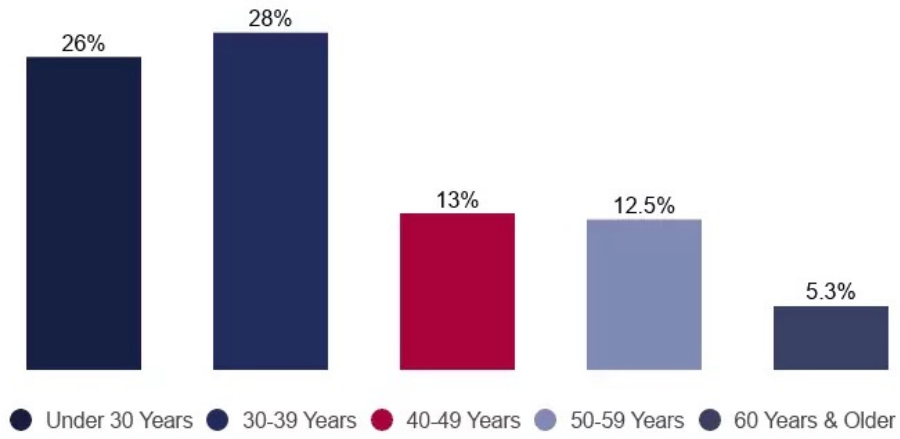
… looking for systematic deviations of an outcome of interest across feature values.

Under-5 Mortality in Nigeria by Region (2018)



[Ogallo W., et al, AMIA (2020)]

Percentage of Student Loan Debt Holders Per Age Group



source: *https://educationdata.org/student-loan-debt-by-age*

Stratification

PROS

- Easy to interpret and communicate across range of technical backgrounds.
- Critical for understanding diverse populations.
- Applicable for almost any type of dataset.

CONS

- Limited to 1 or 2 Features at a time. Beyond that becomes obtuse.
- Relies on human intuition for choice of Features. No inherent ‘Discovery’.
- Aggressive manual stratification leads to false positives.

Subset scanning allows stratification to be done in a more **disciplined and scalable** fashion.

Outline

Motivation

Approach

Examples

Demo

Subset scanning allows stratification to be done in a more **disciplined and scalable** fashion.

Subset Scanning: Scoring Functions

Toy Example: Pet Store Customers
P(Dog) = 0.4

Suburb House:
P(Dog|suburb house) = 0.5

Subset size = 800
Dog owners = 400

Beach House:
P(Dog|beach house) = 0.75

Subset size = 40
Dog owners = 30

source: *my adorable puppy*

$$\max_{q>1} \log \prod_{i \in S} \frac{\text{Bernoulli}\left(\frac{qp_i}{1-p_i+qp_i}\right)}{\text{Bernoulli}(p_i)} = \max_{q>1} \sum_{i \in S} y_i \cdot \ln(q) - \log(1 - p_i + q \cdot p_i)$$

[Zhang, Neill, 2017]

Subset Scanning: Search Space

Toy Example: Pet Store Customers
P(Dog) = 0.4

What if we consider a group of feature values?

Suburb House:
P(Dog|suburb house) = 0.5

Subset size = 800
Dog owners = 400

Anom score = 16.3

Beach House:
P(Dog|beach house) = 0.75

Subset size = 40
Dog owners = 30

Anom score = 10.1

S = Suburb OR Beach House:
P(Dog|S) = 0.51

Subset size = 800 + 40
Dog owners = 400 + 30

Anom score = 21.4

$$\max_{q>1} \log \prod_{i \in S} \frac{\text{Bernoulli}\left(\frac{qp_i}{1-p_i+qp_i}\right)}{\text{Bernoulli}(p_i)} = \max_{q>1} \sum_{i \in S} y_i \cdot \ln(q) - \log(1 - p_i + q \cdot p_i)$$

[Zhang, Neill, 2017]

Subset Scanning: Search Space

. . .

What if there are k values for house-type?
The search space for groups of house-types is $O(2^k)$.

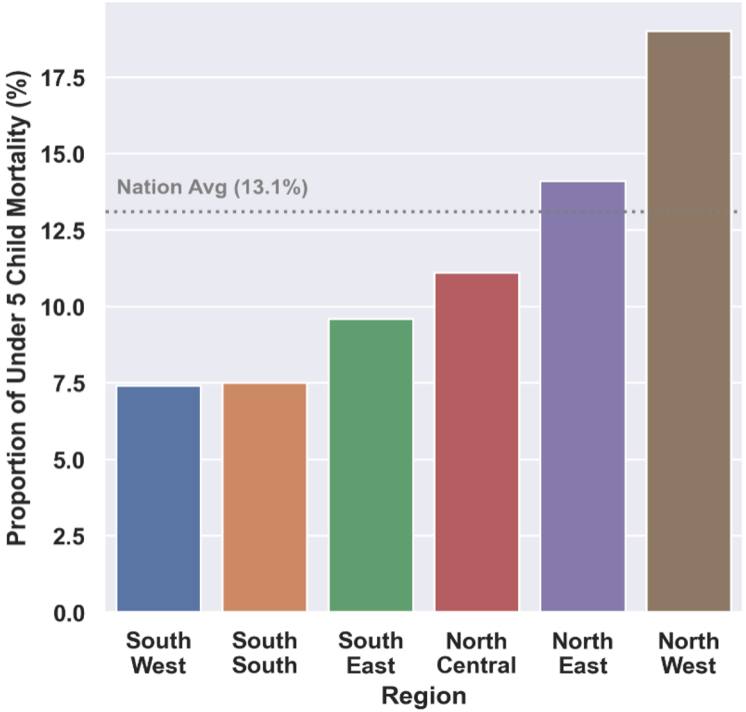
Scoring functions satisfy the LTSS Property which reduces the search space from exponential to linearly-many.

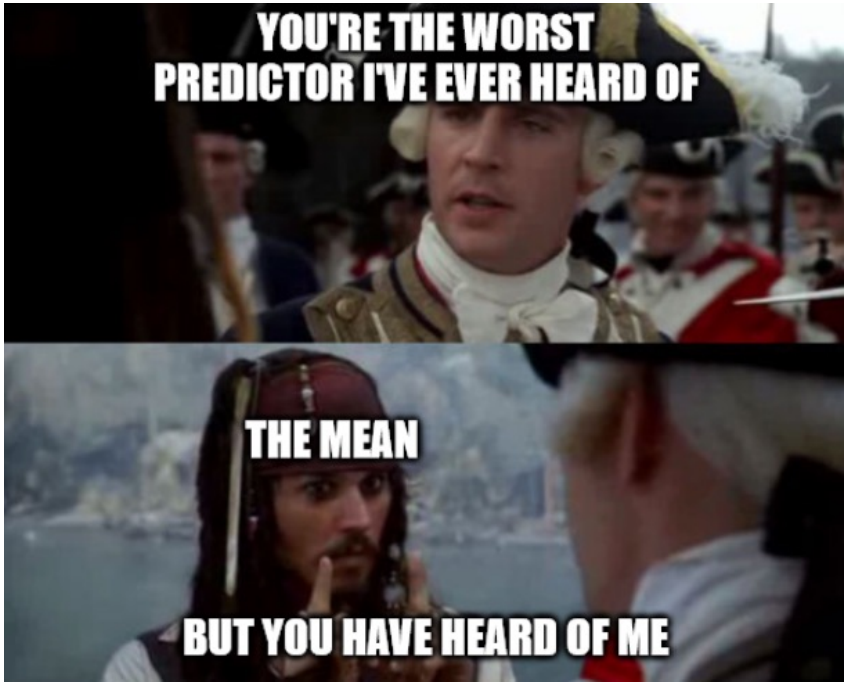
This efficient optimization is coded into an iterative ascent in the Multi-dimensional Subset Scan algorithm

Fast Subset Scan for Spatial Pattern Detection. Neill, 2012
Penalized Fast Subset Scanning. Speakman et al, 2016

? Stratification ? = ? Predictive Bias ?

Under-5 Mortality in Nigeria by Region (2018)





? Stratification ? = ? Predictive Bias ?

Can the expectations of the outcome be set by something other than its mean?

YES!

The mean

Logistic Regression

Random Forests

Boosted Trees

Deep Neural Networks

Outline

Motivation

Approach

Examples

Demo

Autostrat: Neonatal Mortality in Ghana(AMANHI)

Mother's Age	Birth Weight (gms)	Mother's Education	Delivery Location	Birth Quarter	Delivery Person	Gestational Age (days)	Birth Year	Birth Weight for Age (Z score)	Mortality
25_to_29	3000_to_3500	0	Hospital	Q3	Midwife	LTE_270	2012	between_1_2	0
20_to_25	2500_to_3000	10_and_above	Home	Q1	Relative/Friend	between_280_290	2012	between_-2_-1	0
25_to_29	3000_to_3500	0	Home	Q4	Relative/Friend	between_270_280	2011	between_0_1	0
25_to_29	2500_to_3000	0	Home	Q2	Relative/Friend	LTE_270	2012	between_1_2	0
30_to_39	3000_to_3500	0	Home	Q4	Relative/Friend	LTE_270	2012	GT_2	0
40_and_above	2500_to_3000	0	Hospital	Q3	Midwife	LTE_270	2011	between_-1_0	0
LTE19	2500_to_3000	10_and_above	Home	Q4	Missing/Unknown	LTE_270	2011	GT_2	0
20_to_25	3000_to_3500	4_to_6	Home	Q2	Relative/Friend	between_270_280	2012	between_-1_0	0

20,000+ records and 1.5% of births have mortality event

Scanning wants to know which **subset of births** have anomalously high number of mortality events.

There's over 4 Trillion subsets to consider!

Births Subset:
Home Delivery
and
Medical Professional
(Doctor or Midwife present)

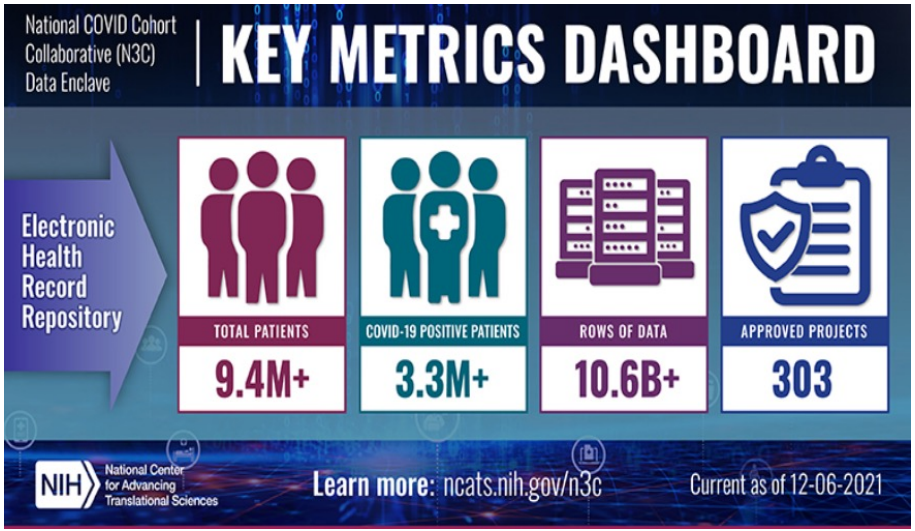
source: *https://www.bbc.com/news/av/world-africa-49611807*

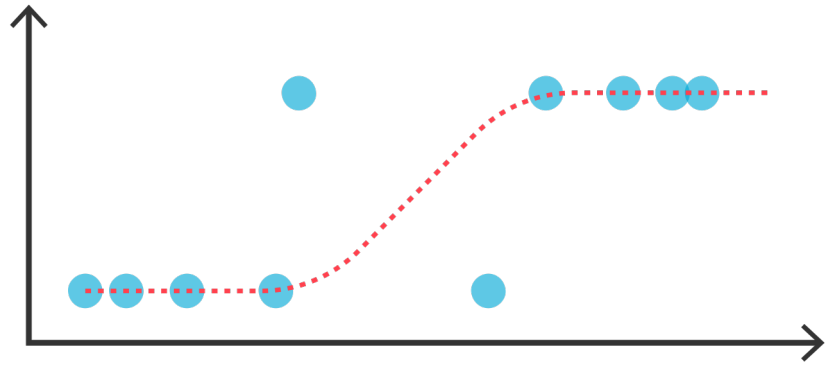
Neonatal Mortality Rate of this Subset
42.1%

Recall the average was 1.5% -- this group is very anomalous!

These circumstances account for **nearly half** (49.1%) of all neonatal deaths in the Ghana study.

Bias Scan: Covid Mortality in the US from Electronic Medical Records





Source: *https://covid.cd2h.org/dashboard/cohort*

Bias Scan: Covid Mortality in the US from Electronic Medical Records

Where is our predictive model the most biased?

Cancer Patients Under the Age of 50

1468 patients

LR Model predicted 80 deaths

<<

Data shows 195 deaths

The predictive model failed to capture a complex interaction between age, cancer, and mortality.

Outline

Motivation

Approach

Examples

Demo

https://github.com/tanya-akumu2/folktables_scan

Demo: US Census Income Data

Dataset:

- 2018 US Census Data on income accessed from the Folktables Python package which provides access to datasets derived from the US Census. These datasets facilitates the benchmarking of machine learning algorithms.

- Contains ~1.6 Million samples with the following features:

Demo: US Census Income Data

Key Question:

Can we identify **sub-populations** who, as a **subgroup**, have outcomes that significantly deviate from the **overall population?**

TARGET/OUTCOMES:

INCOME > 50K → **Y=1**

INCOME < 50K → **Y=0**

EXPECTATION:
Overall population = mean of the observed outcomes

OR

Predictions from a trained model (bias scan)

Conclusion & Future work

- Addition of other parametric and non-parametric scoring functions
- Direct support for continuous features
  - Currently, continuous features need to be binned
- Support for non-parametric scoring functions in bias scan mode
- Application to new datasets and domains

Thank you! Asante!






: tanya.akumu@ibm.com

LinkedIn : Tanya Akumu

: @tanya-akumu

https://github.com/tanya-akumu2/folktables_scan