Wicked World Means Null

Wicked problem World bank k-Means clustering Null values

•••

Jake Bobula

jakebobula@gmail.com

Overview

The history of states and nations has provided some income for historiographers and book dealers, but I know no other purpose it may have served. - Borne (probably Ludwig Börne)

Billions of people live in the "Least developed countries" UN classification, development is a wicked problem

Strategic development, investment in countries that are at a tipping point to prompt desired growth

Let’s go find those countries!

Project objective:

Aggregate condition clustering and testing condition predictive value

Understanding the problem

Nulls

Data is 22% null even after paring it down to 399 features for 9773 country year pseudo indexes

Live with them or impute them

Clustering

1. Build an algorithm that clusters with nulls
2. Impute values then cluster
3. Use PCA reconstruction on imputed values then cluster

Regression

Can I make predictions from aggregate conditions?

Build regression models to predict future values from aggregate conditions

Null Clustering

Missing values (nulls) are common, there are many fantastic algorithms that do wonderful things… with enough data wrangling

Modified:

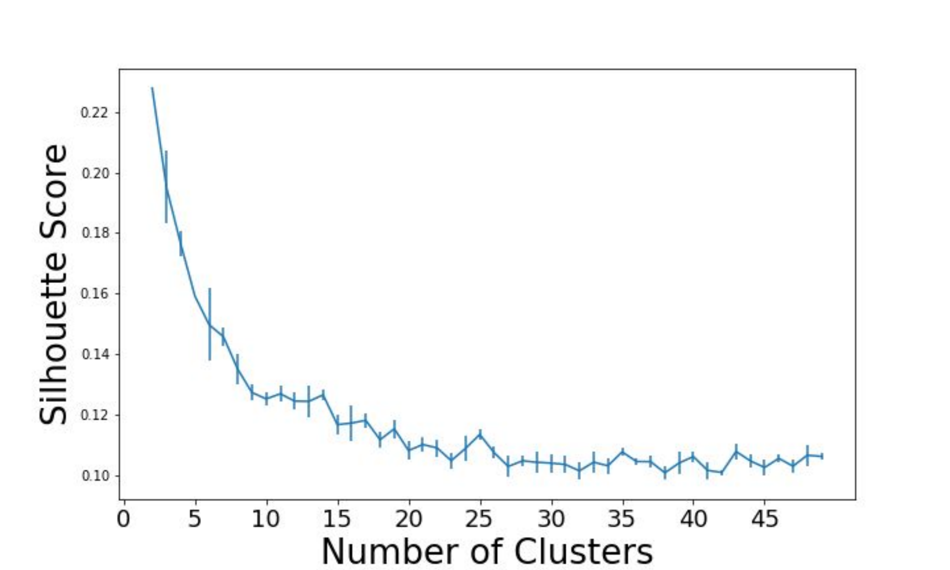
- k-means
- Euclidean distance

Null k-means

Silhouette Scores

Calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. The Silhouette Coefficient for a sample is:

(b - a) / max(a, b)



Null Kmeans

- Results are similar to imputed and pca k-means
- Fowlkes Mallows, Normed MI average 16% Δ vs others

Imputation and Random Forest Regression

Imputation:

- Bidirectional exponentially weighted moving average imputation (ewma)
- Combined bi-ewma with K-nearest neighbour imputation (knn)

RF Feature Importance

- GDP per capita (constant LCU)
- Death rate, crude (per 1,000 people)
- Urban population growth (annual %)

Combination Imputation

- Used bi-ewma and knn, for a 0.042 mse
- 2% reduction in mse vs either individually

Regression Models

- Average 0.0175 mse and 0.98 adjusted R^2
- Forward 1 to 4 years more accurate than 5

Contact and Project Stack

Email: jakebobula@gmail.com

Phone: (207) 598-5667

Linkedin: https://www.linkedin.com/in/jakebobula/

Github: https://github.com/jakebobu/world-bank

http://ec2-35-174-106-106.compute-1.amazonaws.com:8080/

THE WORLD BANK
IBRD • IDA

scikit learn

NumPy

AWS

Flask
web development,
one drop at a time

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

python™

Extra Slides

Deliverables

Null Kmeans

- Results are similar to imputed and pca Kmeans
- Fowlkes Mallows, Normed MI ave 16% Δ vs others

Combination Imputation

- Used bidirectional ewma and knn, for a 0.042 mse
- 2% reduction in mse vs either individually

Regression Models

- Average 0.0175 mse and 0.98 adjusted R^2
- Forward 1 to 4 years more accurate than 5

Web App

- ec2-52-23-205-66.compute-1.amazonaws.com:8080
- Play with the results! regression and clustering

Random Forest Regression

Feature Importance:

- GDP per capita (constant LCU)
- Urban population growth (annual %)
- Death rate, crude (per 1,000 people)

Feature Importance (3 year) continued:

- GDP deflator (base year varies by country)
- GDP at market prices (constant 2005 US$)
- Mortality rate, adult, male (per 1,000 male adults)
- Mortality rate, infant (per 1,000 live births)
- Immunization, measles (% of children ages 12-23 months)
- Final consumption expenditure, etc. (current US$)
- Mobile cellular subscriptions
- Life expectancy at birth, female (years)

Regression Models

- Average 0.0175 mse and 0.98 adjusted R^2
- Forward 1 to 4 years more accurate than 5

Regression Models

Random Forest Regression

Hyper parameter selection with 5-fold cross validation

Resulting parameters:

- Max features: 20, 200
- Max depth: 20
- Min samples split: 2,4
- N estimators: 150, 300

Regression Models

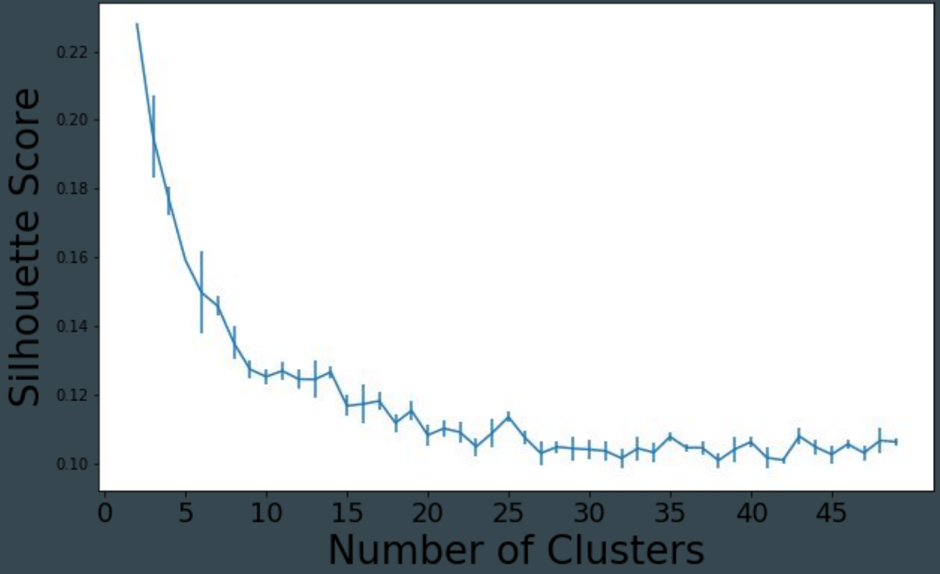
- Average 0.0175 mse and 0.98 adjusted R^2
- Forward 1 to 4 years more accurate than 5

Null k-means

Silhouette Scores

Calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. The Silhouette Coefficient for a sample is:

(b - a) / max(a, b)



Null Kmeans

- Results are similar to imputed and pca Kmeans
- Fowlkes Mallows, Normed MI average 16% Δ vs others