Tapping on unconventional data sources to obtain actionable intelligence on the connections between gender and environment

Rajius Idzalika - Junior Data Scientist

Expert Meeting on Statistics on Gender and the Environment 2019, Bangkok
UN Global Pulse

Global Pulse is an innovation initiative of the United Nations Secretary-General on big data and AI. Our vision is a future in which big data is harnessed safely and responsibly as a public good. Our mission is to accelerate discovery, development and scaled adoption of big data innovation for sustainable development and humanitarian action.
Pulse Lab Jakarta combines data science and social research to help make sense of our interconnected, interdependent, and complex world. The Lab is a joint initiative of the United Nations and the Government of Indonesia, via United Nations Global Pulse and the Ministry of National Development and Planning (Bappenas) respectively.

**OUR SERVICES**

- Drive **exploratory research** on new insights that can be gleaned from unconventional data sources
- Help UN agencies, governments and development partners make **better use of their data**
- Advocate for the **ethical use of data** and technological platforms in line with the protection of individual privacy
How do you decide to across the street? Considering the current traffic or the traffic an hour ago?
There is an information gap between conventional data source and decision making.
Big data is a new data source

The basic idea behind the phrase Big Data is that everything we do is increasingly leaving a digital trace (or data), which we (and others) can use and analyse.

“Big Data therefore refers to our ability to make use of the ever-increasing volumes of data.”

BUT...Big data is not intended to replace conventional data, instead they complement each other to generate richer insights.
Harness new data sources to assess...

...What people have said

- Social media (content focused)
- Online ads
- Community complaints management system
- Radio
Harness new data sources to assess...

...What people have done

- Social media (location focused)
- Utilities information (electricity, clean water, etc.)
- Postal data
- Transportation data
- Keywords search
- Online/offline retail data
- Remote sensing
- Financial service data
- Call Data Record (CDR)
Accessing big data

Obtaining big data is not easy, but there are ways to get through it.

Two working strategies:

1. Public private partnership

1. Public generated data (citizen science and crowdsourcing)
UNCDF SHIFT and UN Pulse Lab Jakarta are pleased to launch their new report ‘Examining Customer Journeys at Financial Institutions in Cambodia’. This study encourages a shift in focus from examining access to finance to understanding actual usage of financial products. The study demonstrates the potential of Big Data analytics to generate granular sex- and youth-disaggregated information on the use of financial services, and to apply insights to inform product development and policy making.
Finding 1: Different customer profile

PROFILE OF DEPOSITORS AND BORROWERS WITH STRONG CUSTOMER JOURNEY

Male

- < 25 years old
- Lives in Phnom Penh
- Saves at FSP C
- Takes savings account and term deposit

DEPOSITOR

Female

- < 25 years old
- Lives outside Phnom Penh
- Borrows from FSP A and FSP D
- Takes group loan

BORROWER
Finding 2: Gender gap on saving mobilization
Finding 3: Gender gap on customer journey

WOMEN STAYED LONGER IN THE BORROWING RELATIONSHIP WITH FSPS YET THEY RECEIVED LOWER LOAN AMOUNTS THAN MEN
Microfinance data

Works in the pipeline ...

- Measure resilience by adaptive capacity index.
- Find proxy indicators for poverty.
- Understanding the relationship between loan and climate change and deforestation.
Microfinance data

An illustration to repurpose loan data with cluster analysis.

**FINDINGS**

<table>
<thead>
<tr>
<th>Category</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Higher Education</td>
<td>10.71%</td>
</tr>
<tr>
<td>Non-Agriculture Focused</td>
<td>61.90%</td>
</tr>
<tr>
<td>Agricultural Focused</td>
<td>27.38%</td>
</tr>
</tbody>
</table>

- Higher Male %
- Older
- Urban Area
- Higher Not Married %
- Higher Male %
- Older
- Higher Not Married %
- Higher - Education
- Multiple Loan Accounts
- Well-endowed Farmers
- Multiple Currencies
- Higher Female %
- Younger
- Higher Married %
- Lower - Education
- Non-Multiple Loan Accounts
- Agricultural Loans
- One Local Currency
Data Call Detail Record (CDR)

Call data records could be harnessed to learn human behavior.

<table>
<thead>
<tr>
<th>CALLER ID</th>
<th>CALLER CELL TOWER LOCATION</th>
<th>RECIPIENT PHONE NUMBER</th>
<th>RECIPIENT CELL TOWER LOCATION</th>
<th>CALL TIME</th>
<th>CALL DURATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>X76VG588RLPQ</td>
<td>2°24' 22.14&quot;, 35°49' 56.54&quot;</td>
<td>A81UTC93KK52</td>
<td>3°26' 30.47&quot;, 31°12' 18.01&quot;</td>
<td>2013-11-07T15:15:00</td>
<td>01:12:02</td>
</tr>
</tbody>
</table>

Mobility | Social interaction | Economic activity
Commissioned by the World Bank, PLJ and Empatika conducted research into the experiences of rural to urban migrants.

PLJ led the quantitative component of the project which used mobile network data to develop statistics on the magnitude of short term migration and the source communities of migrants to seven major cities within Indonesia.
Data Call Detail Record (CDR)

RURAL TO URBAN MIGRATION
Data Call Detail Record (CDR)

One output is visualization with high resolution.
The challenge related to gender statistics is no gender information.

- We need to conduct foundational research to predict gender from the call or text behavior.

- It is known that machine learning is really good for a classification task. The ground truth is determined by conducting a telesurvey.

- Our recent research shows the accuracy is 0.88 or higher for the prediction of selected household assets.
Predictive model is getting more popular for timely decision making.
Data Call Detail Record (CDR)

It is possible to predict gender of mobile user with high accuracy.

<table>
<thead>
<tr>
<th>Type</th>
<th>Method</th>
<th>None</th>
<th>FCBF</th>
<th>Boruta</th>
<th>XGBoost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electricity</td>
<td>SVM</td>
<td>0.77</td>
<td>0.58</td>
<td>0.71</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>0.67</td>
<td>0.6</td>
<td>0.71</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>Elastic-Net</td>
<td>0.76</td>
<td>0.6</td>
<td>0.8</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>NN</td>
<td>0.67</td>
<td>0.42</td>
<td>0.73</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>Decision tree</td>
<td>0.85</td>
<td>0.66</td>
<td>0.85</td>
<td>0.88</td>
</tr>
<tr>
<td>Bank Account</td>
<td>SVM</td>
<td>0.78</td>
<td>0.59</td>
<td>0.77</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>0.61</td>
<td>0.57</td>
<td>0.68</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>Elastic-Net</td>
<td>0.76</td>
<td>0.55</td>
<td>0.77</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>NN</td>
<td>0.67</td>
<td>0.47</td>
<td>0.69</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>Decision tree</td>
<td>0.89</td>
<td>0.69</td>
<td>0.88</td>
<td>0.9</td>
</tr>
<tr>
<td>Television</td>
<td>SVM</td>
<td>0.75</td>
<td>0.62</td>
<td>0.81</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>0.66</td>
<td>0.61</td>
<td>0.73</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>Elastic-Net</td>
<td>0.77</td>
<td>0.6</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>NN</td>
<td>0.7</td>
<td>0.57</td>
<td>0.8</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>Decision tree</td>
<td>0.87</td>
<td>0.74</td>
<td>0.86</td>
<td>0.88</td>
</tr>
<tr>
<td>Refrigerator</td>
<td>SVM</td>
<td>0.77</td>
<td>0.6</td>
<td>0.79</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>0.7</td>
<td>0.61</td>
<td>0.72</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>Elastic-Net</td>
<td>0.77</td>
<td>0.6</td>
<td>0.78</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>NN</td>
<td>0.7</td>
<td>0.51</td>
<td>0.71</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>Decision tree</td>
<td>0.87</td>
<td>0.68</td>
<td>0.89</td>
<td>0.89</td>
</tr>
<tr>
<td>Radio</td>
<td>SVM</td>
<td>0.77</td>
<td>0.61</td>
<td>0.78</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>0.55</td>
<td>0.64</td>
<td>0.57</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>Elastic-Net</td>
<td>0.75</td>
<td>0.59</td>
<td>0.76</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>NN</td>
<td>0.62</td>
<td>0.47</td>
<td>0.66</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>Decision tree</td>
<td>0.86</td>
<td>0.7</td>
<td>0.87</td>
<td>0.9</td>
</tr>
</tbody>
</table>
Online ride hailing services data

Inferring Greater Jakarta’s Traffic Patterns

Pulse Lab Jakarta, in partnership with Grab, has been investigating how ride-hailing data can be leveraged to better understand Greater Jakarta’s traffic flows at a macroscopic level.

This visualization shows traffic patterns (inflows and outflows) in Greater Jakarta.
Online ride hailing services data

There is gender information that can be used.

- To get the gender (and other) information, the challenge is data partnership. Two modalities: data sharing or insight sharing.

- Homework: build trust and define the shared values.
Satellite images

DigitalGlobe provides 30 cm resolution imagery.
Fight climate changes with machine learning and ground truth.

- Better estimates on how much energy we are consuming
- Improve deforestation tracking

Gender disaggregated?
Overlay with other (big) data with disaggregated by gender for real time information.
Harnessing data for development. Translating insights for social innovation.