Connected Women

The Gender Analysis & Identification Toolkit

Estimating subscriber gender using machine learning

GSMA Connected Women

The GSMA represents the interests of mobile operators worldwide, uniting more than 750 operators with over 350 companies in the broader mobile ecosystem, including handset and device makers, software companies, equipment providers and internet companies, as well as organisations in adjacent industry sectors. The GSMA also produces industry-leading events such as Mobile World Congress, Mobile World Congress Shanghai, Mobile World Congress Americas and the Mobile 360 Series of conferences.

For more information, please visit the GSMA corporate website at www.gsma.com

Follow the GSMA on Twitter: @GSMA.

The GSMA Connected Women Programme works with mobile operators and their partners to address the barriers to women accessing and using mobile internet and mobile money services. Connected Women aims to reduce the gender gap in mobile internet and mobile money services and unlock significant commercial opportunities for the mobile industry and socio-economic benefits for women.

For more information, please visit www.gsma.com/mobilefordevelopment/programmes/connected-women

Dalberg Data Insights

Dalberg Data Insights is the Big Data entity of Dalberg Group, active in developing data products and solutions that aim to support more evidence-based policies. International development actors and local governments must often work around data gaps when tackling important social challenges. At the same time, extensive data sources exist publicly and behind the firewalls of private companies, such as mobile phone operators, banks, digital platforms, and satellite operators. We, at Dalberg Data Insights, identify the data solutions to international development challenges. We access, analyse, and integrate data from different sources to design tools and specialized analytics. Using our data products, local and global communities can better target, implement, and evaluate their programs and initiatives.

For more information, please visit www.dalberg.com/what-we-do/dalberg-data-insights

UKaid from the British people

This publication is the output of a project funded by UK aid, Department for International Development (DFID), for the benefit of developing countries. The views expressed are not necessarily those of DFID.

Published August 2018

Contents

1. Introduction

The absence of gender-disaggregated mobile operator data

The absence of accurate gender-disaggregated data is a consistent barrier to measuring, evaluating and ultimately resolving gender issues in both the public and private sectors. The mobile industry is no exception; while the GSMA has found that a gender gap remains in mobile ownership in low- and middle-income countries,[1] it is often extremely difficult to verify this at the country level using supply-side data, which can be inaccurate or incomplete. Demand-side data, meanwhile, is less up to date, intermittent and expensive to collect accurately on a large scale through surveys.[2] This lack of reliable gender-disaggregated data on access and usage of mobile is a major and persistent barrier to understanding, measuring and addressing the gender gap in mobile ownership and use.

Closing the mobile ownership and usage gender gap is important for the mobile industry, as well as societies and economies more broadly. Women typically make up most of the remaining unconnected population in low- and middle-income countries, and therefore represent the greatest untapped market opportunity for mobile operators. For instance, women in South Asia are 26% less likely to own a mobile than men, and are 70% less likely than men to use mobile internet.[3] Having the ability to evaluate mobile operator subscriber bases by gender is an important first step in realising the commercial, social and economic opportunity of women's digital inclusion.

The GSMA's Gender Identification and Analysis Toolkit (GAIT) has one primary purpose: to allow operators to predict the gender of their subscribers on an individual, MSISDN[4] level. The information gap the toolkit addresses is an important one; understanding the nature and scale of the mobile gender gap is a prerequisite for closing it.

GAIT is a machine learning algorithm that analyses mobile usage patterns to estimate the gender of subscribers. By training the algorithm on a small accurately gender-tagged sample of the customer base to analyse usage patterns by gender, unknown genders for the rest of the subscriber base can be identified with little need for expensive primary research. In Bangladesh, a pilot implementation achieved 84.5% accuracy. The toolkit can then be used to predict the gender of new subscribers as they sign up to and begin using the service.

This document provides an overview of what the toolkit allows operators to do, how it works and what is required to apply the algorithm successfully. To implement the toolkit, operators will need the full technical documentation and accompanying code, which can be accessed by all GSMA members. Further information on how to access the full toolkit can be found at the end of this document.

GAIT was developed in partnership with Dalberg Data Insights.

1. GSMA, 2018, "The Mobile Gender Gap Report 2018". Available at: https://www.gsma.com/mobilefordevelopment/connected-women/the-mobile-gender-gap-report-2018/
2. Supply-side data refers to mobile operator call data records and demand-side data derived from consumer surveys.
3. GSMA, 2018, "The Mobile Gender Gap Report 2018". Available at: https://www.gsma.com/mobilefordevelopment/connected-women/the-mobile-gender-gap-report-2018/
4. The Mobile Station International Subscriber Directory Number (MSISDN) is the number identifying a mobile number internationally.

2. How GAIT can help you

What the toolkit can be used for

GAIT aims to address an issue many mobile operators face: the absence of reliable gender-disaggregated data. It provides the mobile industry with a way to accurately predict the gender of their subscribers when this information is missing or inaccurate in existing Know-Your-Customer (KYC) data. This will allow mobile operators to formulate a more effective strategy for closing the mobile gender gap and measuring their progress.

There are several reasons why operator-recorded subscriber gender data can be incomplete or inaccurate. For example, it can be difficult for mobile network operators (MNOs) to track gender successfully at the point of sale, especially in markets where men commonly register for their wives and daughters, and in low-income settings where agents often operate in very basic facilities and track registrations on paper. In many cases, a lack of access to legally valid ID can prevent SIM users from registering in their own names.[5] In this context, MNOs are increasingly recognising the importance of having accurate data on the gender of their customers.

Why closing the mobile gender gap is important

Despite unprecedented growth in mobile ownership in low- and middle-income countries, women are still being left behind. Women in low- and middle-income countries are 10% less likely to own a mobile phone, and are considerably less likely than men to use more transformative services. For example, women in low- and middle-income countries are 26% less likely than men to use mobile internet,[6] and 33% less likely to use mobile money.[7] This gender gap can widen significantly depending on location.[8]

As mobile becomes an increasingly important enabler of economic and social participation, women's lower rate of mobile access, ownership and use risks entrenching and exacerbating existing inequalities and excluding them from the digital societies of the future.

Closing the gender gap is also a substantial commercial opportunity for the mobile industry. If mobile operators in low- and middle-income countries could close the gender gap in mobile ownership and mobile internet use today, this would generate an estimated incremental revenue of $15 billion over the coming year.[9]

5. For more details, see: https://www.gsma.com/mobilefordevelopment/programme/digital-identity/access-mobile-proof-identity-global-snapshot-linkages-challenges-opportunities/?utm_source=m4d-resources&utm_medium=report
6 GSMA, 2018, "The Mobile Gender Gap Report 2018". Available at: https://www.gsma.com/mobilefordevelopment/connected-women/the-mobile-gender-gap-report-2018/
7 Demirgüç-Kunt, Asli, Leora Klapper, Dorothe Singer, Saniya Ansar, and Jake Hess. 2018. The Global Findex Database 2017: Measuring Financial Inclusion and the Fintech Revolution. World Bank: Washington, DC.
8 For instance, in South Asia, women are 26% less likely to own a mobile phone and 70% less likely to use mobile internet than men. In Brazil, women are 32% less likely to own a mobile phone than men in rural areas compared to just 2% in urban areas.
9. GSMA, 2018, "The Mobile Gender Gap Report 2018".

Measuring the gender gap: *"The squeaky wheel gets the oil"*

The persistent mobile gender gap in low- and middle-income countries will not be resolved without concerted action. Measuring the scale of the issue is a vital first step in formulating a strategy to address it and attract support and buy-in from both internal and external stakeholders. It is generally the case that 'what gets measured, gets managed'.

It is not sufficient to know that there is a gender gap; it is also important to know where and for which services the gap is widest so that these areas can be targeted for action. It is also crucial to understand which mobile products, services or initiatives are having the greatest impact on the gender gap so that they can be expanded and promoted.

This toolkit gives mobile operators an important resource to begin this analysis. First, by generating complete and accurate gender estimates of their customer base, and then by using the data to inform strategic decisions to reach the underserved female population.

Once the toolkit has been implemented successfully, it will generate several outputs:

1. The estimated gender of each individual subscriber MSISDN (phone number) in an operator's customer base, and a 'confidence' score indicating the probability that the assigned gender is correct based on usage patterns.

2. A set of around 150 generated features for every subscriber MSISDN that summarise their usage patterns. These summary indicators are used to differentiate the mobile usage of men and women.

3. In some instances, the model will be able to identify which usage indicators are the best predictors of gender.[10] In these instances, a score will be given for each indicator showing its predictive power.

For an example of the kinds of analysis that can be conducted once subscriber gender has been identified, see the 2016 GSMA report, Using your data to drive growth in women's use of mobile money services.

10. Note: This will not be possible in all cases. Several different machine learning algorithms are applied during the analysis of usage data, with the most accurate prediction selected for the final model. Not all these algorithms output the results in a way that allows for the most significant indicators to be ranked.

3. Implementing the toolkit

GAIT is designed to be simple and resource-light to implement, with as much of the process automated as possible. There are two broad stages to implementing the toolkit:

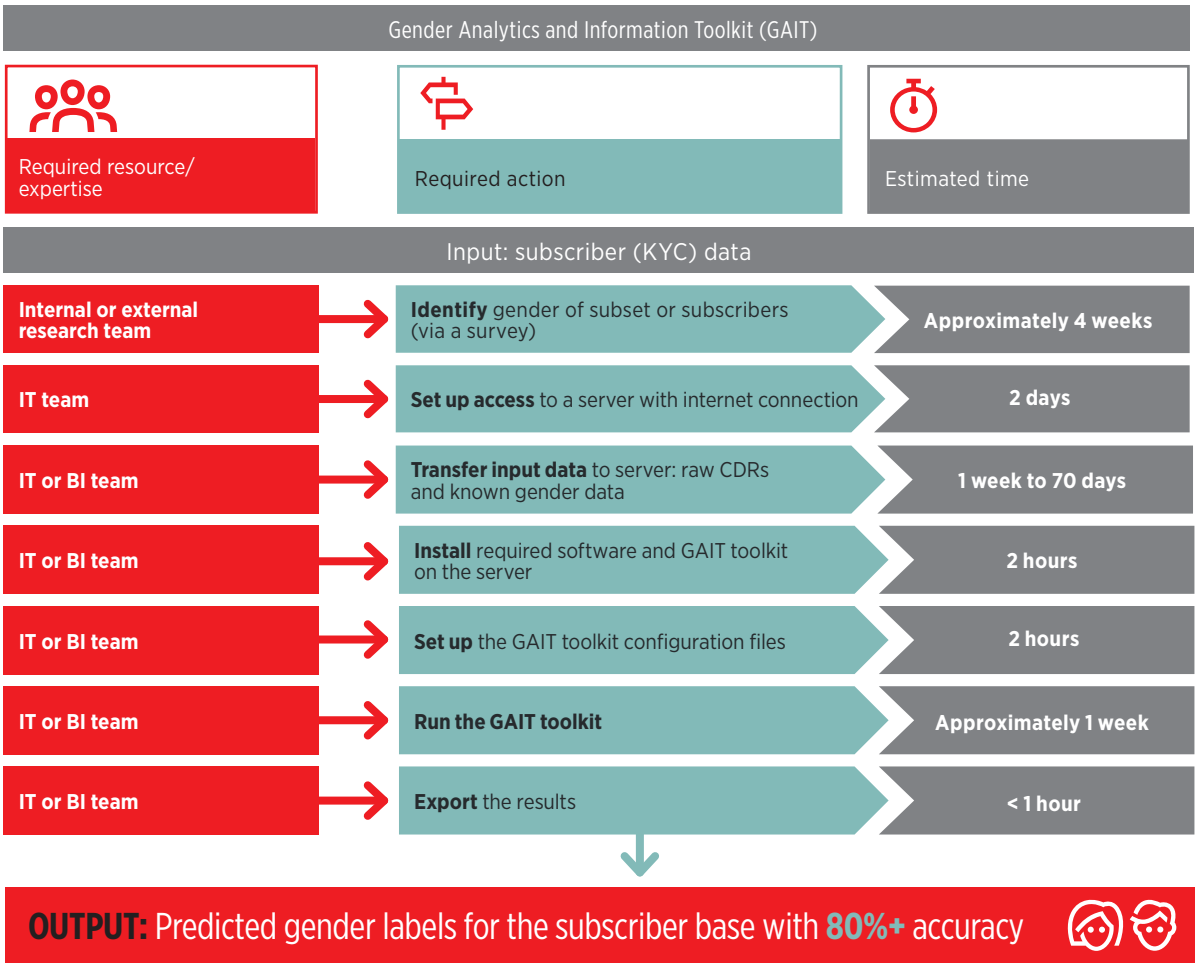
1. Running a baseline survey to generate 'ground truth' data to train the algorithm; and

2. Computing subscriber usage patterns and applying the algorithm to estimate subscriber gender.

Each phase has distinct resource requirements and will likely require input from different teams. Figure 1 shows the approximate workflow and resource requirements for the entire project. If raw call detail records (CDRs) are stored and made available from the last 60 days, the project should take around four to six weeks to run, with the bulk of this time being for the baseline survey.

Figure 1

The GAIT implementation process



The baseline survey:

Accurately identifying the gender of a sample of subscribers

The baseline survey is an essential first stage in the implementation of GAIT. It is designed to provide accurate gender tags for a sample of the customer base that can be used to train the algorithm to recognise male and female usage patterns. As with any machine learning algorithm, GAIT must be trained on a representative portion of the total data set – in this case, a sample of the customer base for which gender is already known. This is known as 'ground truth' or training data.

GAIT requires a training dataset of approximately 15,000 subscribers with correctly identified gender.[11] We recommend gathering this data through a telephone survey of a random sample of customers, conducted either by the operator's in-house call centre or an external agency. Steps should be taken to ensure the accuracy of the results, as any bias introduced during the baseline survey will ultimately reduce the accuracy of the algorithm. As such, it is recommended that:

- Female interviewers are used for this survey, to ensure a sufficiently high response rate from female subscribers who may be reluctant to speak to a male stranger on the phone.
- Call times are spread throughout the day, rather than concentrated during peak working hours when some respondents may be systematically excluded from the sample.
- At least three attempts are made to call each number in the database, at different times of day, before it is discarded, to avoid any skew to the results due to lower response rates among certain segments.

The structure of the baseline survey

For the algorithm to be trained accurately, it is essential that the baseline survey be as random a sample of the customer base as possible. Therefore a sampling frame should not be used for this survey. A random selection of MSISDNs should provide a sufficiently representative sample. It is not recommended that existing gender data be used as an input (for instance, if the gender of a specific sub-segment of the subscriber base is known due to stricter KYC requirements for one service), as it is likely that this identified segment will show some bias or skew, often towards more affluent customers, reducing the accuracy of the predictions the model eventually generates.

To correctly identify the subscriber gender associated with each MSISDN, two essential questions must be answered through the baseline survey:

1. Whether the respondent that answers the phone is the primary user of the SIM, and therefore whether they are the person whose gender it is necessary to identify. If not, the primary user of the SIM should be requested, and if they are unwilling or unable to speak, the call should be terminated.

2. The gender of the respondent. This could be asked directly or determined indirectly, for example, by the pitch of their voice or by asking whether they should be referred to as Sir or Madam. Caution should be used in countries where this may be a sensitive question, as every unsuccessful call increases the chance of bias in the training data.

If the survey data is intended to be used to understand the customer base more generally, additional questions can and should be asked, but only the verified gender of the main user of the SIM will be used as in input in GAIT. Other than this, the model draws exclusively from CDR data from the subscriber database.

For full details of the recommended approach to the survey, please see Appendix C of the technical document for GAIT.

11. The required number of subscribers within this subset depends on a range of factors, including the size of the subscriber base being analysed. The optimal number of known gender data points in the ground truth survey is context-dependent, but 10,000-15,000 data points is the minimum recommended to account for differences in usage patterns and obtain comparable results. In general, increasing the number of known gender data points should improve the accuracy of the model.

Computing subscriber usage patterns

Once ground truth data has been collected through the baseline survey, the next step is to compute the usage patterns of subscribers by analysing the CDRs of the 15,000 identified customers. GAIT is designed to conduct this analysis automatically. All that is required is to install the software and ensure the input data is in the correct format.

However, to compute male and female usage patterns, GAIT will need access to a large volume of subscriber data. This means that implementation will likely need to be facilitated by an operator's IT team. The data and resource requirements for the technical component of the GAIT implementation are:

- Access to a server with an internet connection to receive the data, install the toolkit and run the analysis.
- Access to 60 to 70 days of raw, MSISDN-level CDRs that include the following data streams:
  - Voice CDRs;
  - SMS CDRs;
  - Data CDRs; and
  - Top-up and bundle data.
- A Business Intelligence/IT team with sufficient knowledge of database management and coding languages to process the required CDRs and install the necessary software.

For a full list of the specific CDRs required to run GAIT, please refer to the technical documentation.

The two main stages in the technical implementation of GAIT

Installing the required software

The main stages in the technical implementation of GAIT are:

1. Installing the required software; and

2. Processing the raw CDRs to the required format.

GAIT runs on entirely free and open-source software, and is primarily written in Python, a programming language. This software will need to be installed on the server in order for GAIT to run. No additional software or licenses are required. For detailed instructions, please refer to the technical documentation.

Pre-processing the raw CDRs to the required format

The analysis requires a feed of raw CDRs to compute subscriber usage patterns. It is essential that these are unprocessed CDRs, not processed or aggregated CDRs. For instance, the total number of daily SMS sent are not an appropriate input. GAIT instead requires CDRs in a format that represents each individual SMS as a separate CDR. The exact CDR requirements and the necessary formats are detailed in the technical document.

If CDRs are stored in a different type of format, they will need to be reformatted for GAIT to function properly.

If raw CDRs are not stored as far back as 60 days — for instance, due to storage capacity issues — this data will need to be collected on a separate server before GAIT can be run. The full 60 days must be collected before the analysis can begin.

While it is possible to run GAIT with some CDRs missing or incomplete, it can significantly reduce the accuracy of the predictions. It is not possible to run GAIT with only one strand of subscriber data — e.g. only voice or SMS data — as this will not give sufficient usage information for a wide enough sample of the customer base to generate accurate gender predictions.

Once these stages are both complete and the CDRs are in the required folder, GAIT can be run. First, it will automatically compute the usage patterns of male and female subscribers using the ground truth data, and then it will create a predictive model by testing five different machine learning algorithms to determine the best predictor of gender. This can then be applied to the rest of the subscriber base, generating estimates for all subscribers.

4. Case study: Robi Axiata

In the course of its development, GAIT was piloted in Bangladesh in partnership with Robi Axiata.

Robi is a leading mobile operator in Bangladesh with over 45 million connections and around 30% market share.[12] Due to strong social norms in Bangladesh, women rarely register for a SIM with their own ID card.

As such, in Bangladesh the subscriber and user of a mobile service are often not the same. Consequently, Robi's KYC data did not allow the operator to consistently provide potential female customers with relevant services to a sufficient level of accuracy, or to track its progress in reaching women and closing the gender gap in its subscriber base.

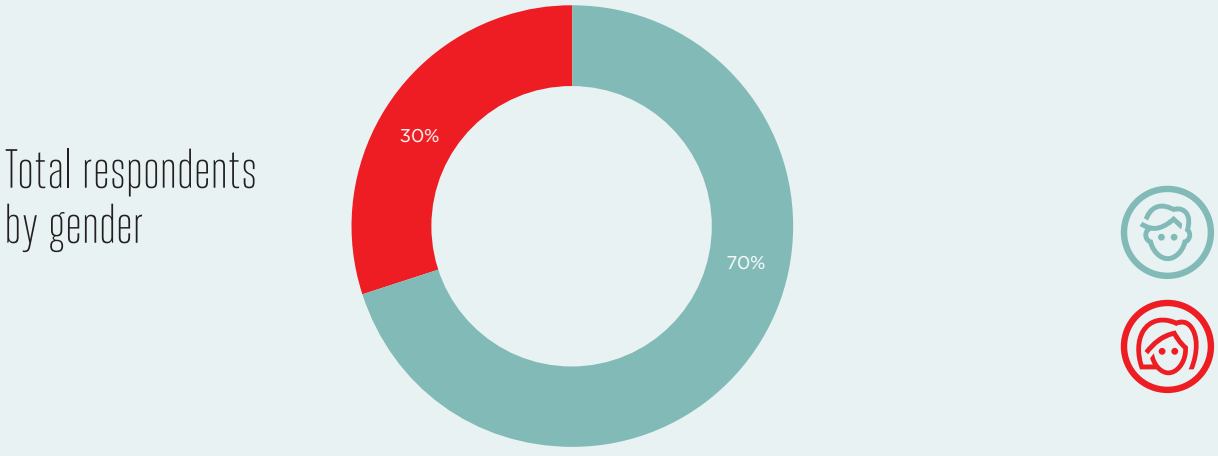
GAIT reached close to 85% accuracy in identifying women in Robi's base, allowing the operator to more effectively reach potential female customers.

Findings from the ground truth survey

The true gender of over 15,000 subscribers was identified via a phone survey. Survey respondents provided their gender, age and habits of sharing their mobile or SIM, and whether they had other mobiles or SIMs that they also used. This information revealed that 30% of Robi's subscribers are female, a proportion that is relatively stable across different regions of Bangladesh. This is shown in Figures 2 and 3 below.

Figure 2

Gender split of ground truth survey respondents

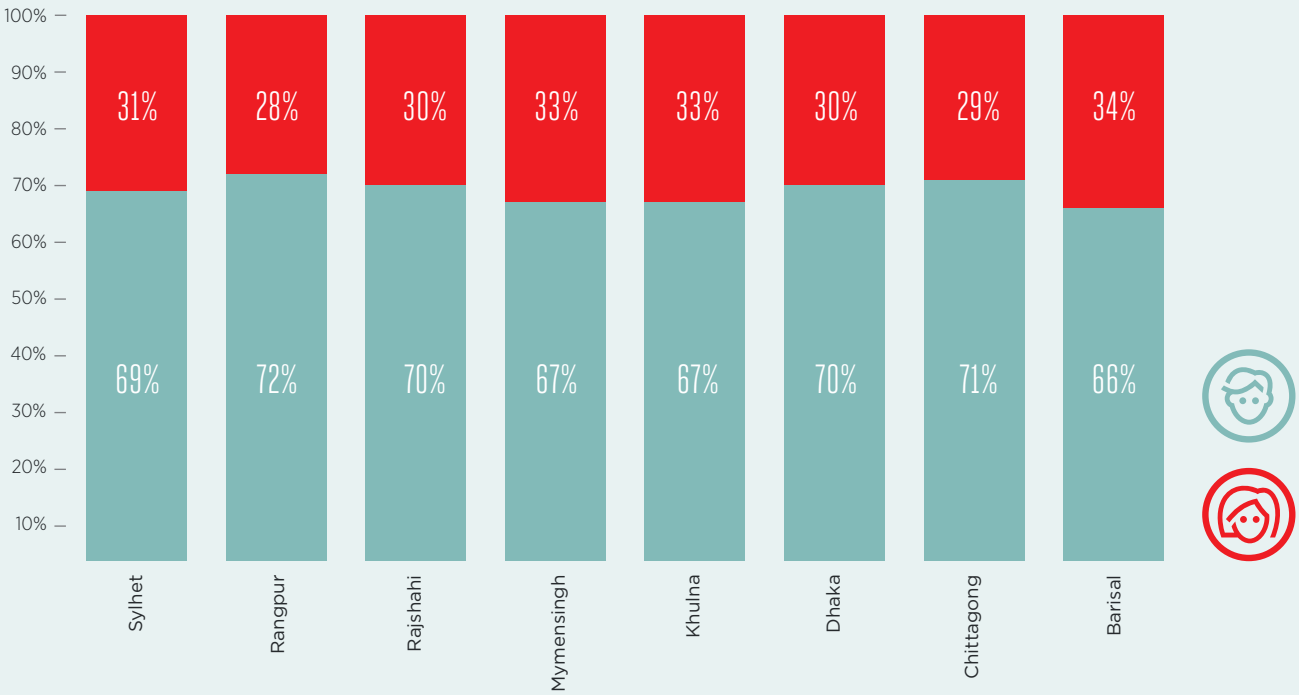


Source: GSMA

12. GSMA Intelligence, Q2 2018

Figure 3

Gender split of survey respondents by division



Source: GSMA

The phone survey also captured instances of sharing and owning more than one mobile or SIM. Overall, 48% of respondents indicated that they used multiple SIM cards, although these were primarily men (60% of men versus 33% of women). Women were more likely to share their SIMs: 39% of women reported sharing their SIM compared to 31% of men.

Findings from the predictive model

The outputs of the phone survey were used alongside 146 features (e.g. call duration), which were computed from call detail records to train GAIT to determine the gender of the subscriber.

The model achieved an accuracy of 84.5% (which corresponds to the overall percentage of the gender labels that the model correctly predicted in the test data), a precision of 79.4% (which corresponds to the percentage of the predicted female labels that were correct, out of all predicted female labels in the test data) and a recall of 61.2% (this represents the overall percentage of actual female subscribers in the test data that were correctly identified as female by the algorithm). When the gender of subscribers identified through the ground truth survey was compared with the gender they had originally registered under for their Robi SIM, 78% of female subscribers (as identified through the survey) were in fact registered in Robi's KYC data as male. Figure 4, commonly referred to as a confusion matrix, shows the predicted gender from the model compared to the gender recorded in KYC data.

Figure 4

GAIT subscriber gender estimates compared to KYC data

% of total customer base (sums to 100%)	Predicted gender (output from GAIT)		
		♂	♀
Previous MNO KYC data	♂	63.2%	23.8%
	♀	6.3%	6.7%

To be read as: 63.2% of total subscribers were identified as male both in the KYC data and in the model, while 23.8% of total subscribers were registered as male in the KYC data, but were actually predicted to be female by the GAIT algorithm. As such, 78% of actual female subscribers were registered in the KYC as male, compared to only 9% of male subscribers that were registered as female.

Source: GSMA

Five features made the most significant contribution to predicting the gender of a subscriber (and their importance[13]):

- Average duration of incoming call (0.057; women receive longer calls on average);
- Number of different contacts (0.044; on average, women have contact with fewer numbers);
- Radius of gyration (average travel distance) for an average active day (0.037; women travel less on average);
- Number of distinct cell towers that handled the subscriber's transactions (0.032; women use fewer cell towers on average); and
- Radius of gyration for the full study period (0.030; women travel less on average).

Figure 5 and 6 show how the distribution differs for men and women for two of these key features.

Figure 5

Average duration of incoming calls by subscriber gender

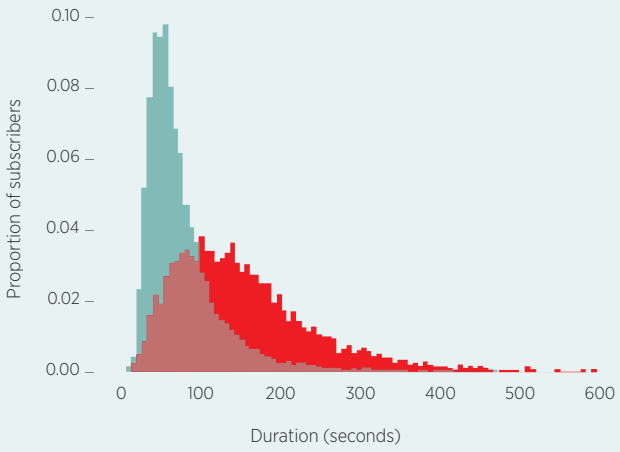
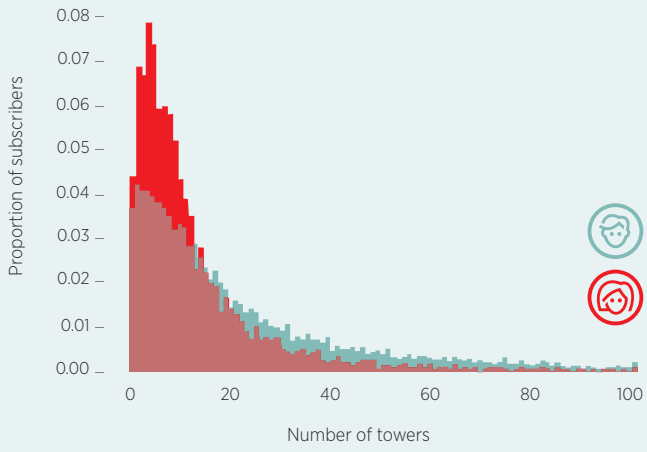


Figure 6

Number of distinct cell towers visited by subscriber gender



The full list of the 146 features computed can be found in Appendix D: List of features.

13. The importance values are normalised such that the sum of all the scores of features is 1.

Limitations and caveats

While every effort has been made to ensure that the outputs of GAIT are as accurate and complete as possible, gender identification is only an estimate, and cannot be guaranteed with complete confidence. GAIT results should therefore be used with caution, and if possible complemented by ongoing verification of the actual gender of subscribers and efforts to improve the collection and accuracy of KYC data.

It is also possible that GAIT may not be able to be successfully implemented in some countries, as male and female usage patterns may not be distinct enough to make strong gender predictions. To avoid this, GAIT should only be run in countries where mobile operators have reason to believe that male and female usage patterns have some significant distinction, ideally determined through preliminary business intelligence analysis or market research.

It is essential that any operator considering running the toolkit should consider and adhere to privacy laws of their country. Operators must also determine whether any restrictions apply to the usage of individual subscriber behaviour, which would limit any potential application of the toolkit. Please refer to the licensing agreement for GAIT for further details.

Accessing the toolkit

To access the toolkit and user guide, please contact **Connected Women** at **ConnectedWomen@gsma.com**. Please note that GAIT is only available to mobile operator members of the GSMA.

For more information on the GSMA Connected Women Programme, visit gsma.com/connectedwomen

GSMA HEAD OFFICE
Floor 2
The Walbrook Building
25 Walbrook
London EC4N 8AF
United Kingdom
Tel: +44 (0)20 7356 0600
Fax: +44 (0)20 7356 0601