Copyright © 2023 Flowminder Foundation - All Rights Reserved

Challenges in predicting individual poverty status from mobile operator customer segmentation metrics and phone surveys: a Papua New Guinea case study

Galina Veres*, Veronique Lefebvre, Savita Ragoonanan*, Caterina Irdi, Xavier Vollenweider, Shohei Nakamura****

Flowminder Foundation, ** The World Bank,*** Digicel Pacific

*corresponding author email: galina.veres@flowminder.org

Introduction. Papua New Guinea (PNG) faces challenges in estimating the geospatial distribution of population poverty due to data availability and quality. The last census was conducted in 2011, and the next census has been delayed until 2024. Conducting field surveys in Papua New Guinea can be challenging due to various factors such as the country's rugged terrain limiting accessibility, low population density in many areas, the high cost of transportation and limited infrastructure, and security issues in some areas. Papua New Guinea is a culturally diverse country with over 800 languages spoken, which can make it difficult to gather accurate information. Thus, data collected by Mobile Network Operators (MNOs) represent an attractive non-traditional data source to estimate poverty in PNG. In this paper, we investigate statistical dependencies between Digicel customer segmentation metrics based on mobile phone usage - which are routinely collected for each subscriber for commercial purposes, and the poverty status of subscribers as estimated by the World Bank high-frequency phone survey.

Data. Customer segmentation metrics collected by Digicel for each subscriber are handset type, monthly average revenue per user (ARPU), amount and number of top-ups per day, daily usage of mobile phone measured by number of call minutes, number of SMS and megabytes of data, centroids of commercial clusters per subscriber used as a proxy for home location, urbanicity of home location and available technology (2G, 3G, 4G). The World Bank shared the high-frequency phone survey data for five rounds conducted between December 2020 and June 2022 with approximately 5 months between rounds. For round 1, a stratified random sample was drawn from Digicel subscriber base using sample sizes proportional to 22 provinces' populations. In the next rounds, respondents were selected in two stages: 1)Contacting all respondents of previous rounds (panel cases); 2)Purposive sampling of subscribers (“replacement” cases) to reach respondents from all socio-economic groups: only respondents who did not send/received sms messages, received only incoming calls and had the majority of credit transferred were selected. Attrition rates are very high from round to round: from 52% between rounds 4 and 5 to 86% between rounds 2 and 3. Thus the resulting dataset is not a random sample of Digicel subscribers, and does not contain the full spectrum of each segmentation feature. The Word Bank also provided national estimates of wealth deciles for each respondent in each round, which defined the poverty status of respondents as poor (0.7 decile and below) and non-poor (0.8 decile and above). Linkage of the two data sources for each respondent was conducted by Digicel and only pseudonymised records on customer metrics (not location data) were shared with Flowminder.

Methodology. The framework for predicting the poverty status of respondents consisted of the following steps: feature engineering based on customer segmentation metrics; statistical analysis of segmentation features; feature selection; training, testing and comparing machine learning algorithms; analysing the results. Segmentation feature engineering was done by calculating statistics for relevant segmentation metrics and each round, such as min, max, mean, median and std. Statistical analysis of segmentation features showed complex classification problems with strongly overlapped classes, with not a single segmentation feature having a strong correlation with poverty status on its own. This could be partly due to the purposive sampling with features not representing the features full distributions. Then we removed segmentation features with strong pairwise Pearson correlation and similar means, and identified the most relevant segmentation features for predicting poverty status using ANOVA one-way test, Lasso with Cross-Validation, and Mutual information. Two scenarios were investigated: 1)To assess the variations in segmentation features for each class, and correlation between segmentation features and poverty status on a round basis, we trained a model on each round and tested predictions of respondents' poverty status on the same round, and 2) to check whether dynamic prediction of poverty was possible in PNG, we trained a model on earlier rounds and tested predictions on the last round. Several machine learning classification algorithms were trained and tested such as Logistic Regression, Random Forest, Decision Trees, Gaussian

Copyright © 2023 Flowminder Foundation - All Rights Reserved

Process Classifier, AdaBoost and Neural Networks. The results below are shown for the Logistic regression (LR) model due to similar performance to other algorithms and easy implementation on the MNO's server for operational use.

Results. Figure 1 shows the classification performance measures for individual rounds. Training and testing sets were created by proportional stratified sampling to preserve proportion of poor and non-poor in both training (75% of respondents) and testing (25% of respondents) sets. The results show the mean performance measures for 1,000 repeats and average comparison to chance. The LR model achieved precision and accuracy above 60% in all rounds except round 2. Recall (TP) is slightly above 50% for rounds 2 and 3 and above 75% in rounds 4 and 5. Comparison with chance shows improvements by between 19% and 40% for poor, and between 10% and 34% for non-poor. However, the later rounds are the least representative of the Digicel subscriber base due to the purposive sampling of replacement respondents using features also used for prediction, which may inflate accuracy and precision statistics. The classification power of the features is probably closer to round 2 with less targeted sample, though only approximately half of round 2 respondents were selected randomly.

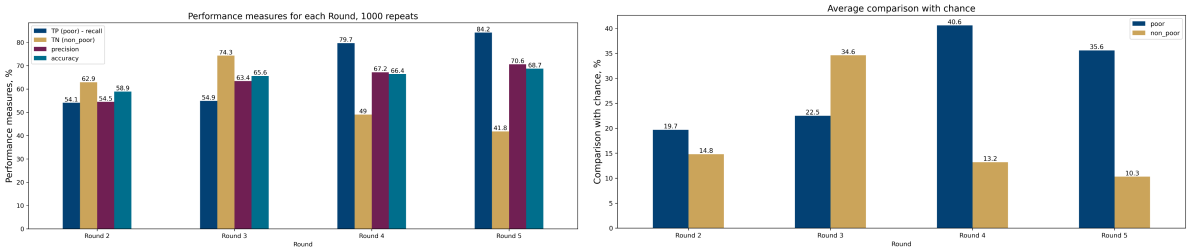

Figure 1. Performance measure for each Round (left) and average comparison to chance (right).

Figure 2 shows classification results on the second scenario: training on the previous rounds 2, 3 and 4, and testing on round 5. Precision (~74%) and accuracy (~67%) is very similar when training and testing on round 5 only. Recall (~70%) is lower, however the trade-off is an improvement in correctly classifying non-poor (~61%). Comparison with chance shows improvements by ~61% for non-poor and by ~13.5% for poor. These results show potential for predicting poverty status based on the past data for the respondents participating in the World Bank high-frequency survey (bearing the above caveat on sampling).

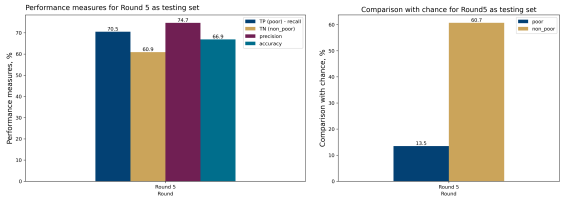

Figure 2. Performance measure for the second scenario (left) and average comparison to chance (right).

Challenges and recommendations.Though customer segmentation metrics or other MNO data appear related to poverty, we have identified the following challenges and recommendations to develop predictive models from these data: 1)The high-frequency phone survey was conducted using purposive sampling for some rounds, thus no inference is possible neither to Digicel subscriber base nor to a general population, i.e. random sampling needed. 2)The purposive sampling was based on some of the segmentation features the project wanted to evaluate, which prevents the possibility of verifying assumptions on statistical relationships between these features (SMS, credit) and poverty, i.e. random sampling needed. 3)Inference to the general population requires data on poverty of phone users and non-users, and subscribers of different networks, i.e. field survey data required. 4)Customer segmentation metrics are more challenging for poverty prediction compared to other MNO data such as mobility, social networks and mobile banking, i.e. integrating the promising features from MNO would improve models, so would ancillary data such earth observation data.

Acknowledgements. The project was financed by the Australian Government through the World Bank. This work was done with a cooperation from Digicel Pacific PNG.