Preliminary Results in Low-Listenership Prediction in One of the Largest Mobile Health Programs in the World

Sanket Shah
Harvard University*
Cambridge, USA
sanketshah@g.harvard.edu

Shresth Verma
Google Research
Bangalore, India
vermashresth@google.com

Amrita Mahale
ARMMAN
Mumbai, India
amrita@armman.org

Kumar Madhu Sudan
ARMMAN
Mumbai, India
madhu@armman.org

Aparna Hegde
ARMMAN
Mumbai, India
aparnahegde@armman.org

Aparna Taneja
Google Research
Bangalore, India
aparnataneja@google.com

Milind Tambe
Google Research
Bangalore, India
milindtambe@google.com

ABSTRACT

Kilkari is a mobile health (mHealth) program operated by ARMMAN, a non-profit organization based in India, which uses IVR technology to deliver time-sensitive audio information to pregnant women and mothers to reduce maternal and child mortality rates. To improve beneficiary retention, we present a preliminary study aimed at targeting interventions for beneficiaries with low listenership. We model this problem as a time series prediction task and assess the efficacy of different machine learning (ML) models. Our experiments reveal that ML models can improve the prediction of low listenership from 5% (as obtained through random selection) to 25%. However, more sophisticated ML algorithms do not perform any better than logistic regression, at least based on the inputs and context as discussed in this paper,. These results highlight the need for novel machine learning research to help better target ARMMAN's limited intervention resources.

KEYWORDS

Machine Learning, Public Health, Maternal Health

1 INTRODUCTION

> *"Pregnancy is not a disease. Childhood is not an ailment. Dying due to a natural life event is not acceptable."*
> (Aparna Hegde, Founder, ARMMAN)

Maternal mortality is a serious issue that remains a global concern even today. Shockingly, the World Health Organization reported that 287,000 women died due to complications from pregnancy or childbirth in 2020 alone [20]. Most of these deaths are preventable and are disproportionately concentrated in Sub-Saharan Africa and Southern Asia. In response to this challenge, ARMMAN, a non-profit organization based in India, is using mobile health (mHealth) interventions to improve access to preventive information and services for pregnant women and mothers, to reduce maternal and child mortality rates.

The Ministry of Health & Family Welfare (MoHFW) launched the Kilkari [2] program, a free mHealth education service that sends women preventive care information during pregnancy and infancy, in 2016. Kilkari is an IVR service designed to deliver weekly pre-recorded, stage-specific audio messages to pregnant women and mothers with children under the age of 1 year (72 weeks). Currently operational in 18 states and UTs, Kilkari has reached over 30 million women and their children to date, and has 3 million active subscribers. ARMMAN is a technical, content & creative production, and implementation partner to the MoHFW in making Kilkari available pan-India.

However, like all mHealth programs, ensuring the beneficiaries' continued participation in the program is challenging, especially over an extended period of time. In the past, ARMMAN has used Machine Learning (ML) to target expensive interventions [13, 15, 19] aimed at improving beneficiary retention in a similar program—mMitra. However, unlike Kilkari, the mMitra program has a much richer set of demographic information that allows ML models to tailor predictions to individual beneficiaries. In this paper, we present a preliminary study to see if ML can help target expensive interventions even without rich demographic information.

The objective of our study is to predict a subset of beneficiaries who may have 'low-listenership' in the future, as defined by domain experts, on whom we can intervene beforehand. We model this problem as a time series prediction task and assess the efficacy of different ML models with varying degrees of sophistication. Our experiments reveal that ML models can improve Precision@5% from 5% (as obtained through random selection) to 25%, a five-fold increase. However, our findings indicate that the more sophisticated ML algorithms, at least based on the inputs and context as discussed in this paper, do not perform any better than Logistic Regression. This implies that the data may not have much complex structure that can be utilized for improved predictions. This contrasts with the results of Nishtala et al. [15] who found that complex models work better for listenership prediction in mMitra.

* Work done as an intern at ARMMAN..

This suggests that there are fundamentally different research challenges in targeting interventions for Kilkari. Some possible directions for future work include:

(1) **Feature Engineering:** The most obvious way by which to improve predictions is to gather demographic information that will allow us to personalize predictions to individual users. On the technical side, this may involve combining weak sources of signal (e.g., the district in which the beneficiary lives) with other datasets (e.g., the census) to create richer demographic information.
(2) **Decision-Focused Learning:** If the goal is only to reach out to a small subset of beneficiaries because interventions are expensive, perhaps there are sub-populations of beneficiaries for whom it is possible to find models that have higher accuracy. One way to do this would be to train these models on custom loss functions that try to maximize this objective (as opposed to more general predictive accuracy) [16, 17].
(3) **Incorporating Intervention Effects:** In mMitra, past work has found that it was important not only to find low-listeners, but rather the subset of those that would respond most positively to interventions [13, 19].

More broadly, we believe that innovation in this space will help ARMMAN better target limited intervention resources and help realize their vision – to create a world where every mother is empowered and every child is healthy.

2 RELATED WORK

Kilkari. Before the responsibility of operating the Kilkari program was transitioned to ARMMAN in 2019, it was being operated by a different non-profit [2]. During this time, a number of articles that analyzed the impact and efficacy of the program were published [3–5, 14]. However, these did not focus on creating possible intervention strategies for the program.

Low-Listenership Prediction. As highlighted in the introduction, there has been past work on creating intervention strategies for ARMMAN's mMitra program [13, 15, 19]. More broadly, there has also been work on medication adherence prediction in various public health contexts like Tuberculosis [12], mental health [1] and HIV/AIDS [7]. However, the technical challenges in each of these papers is very domain-dependent; the challenges are even different even for our closest related work, i.e., Nishtala et al. [15].

3 PROBLEM FORMULATION

3.1 Data

The Kilkari program runs for 72 weeks, from the second trimester of pregnancy till when the child is one year old. During each of these weeks, the beneficiary receives one voice message a week using an IVR system. The messages cover a range of topics from nutritional information for the mother and child to immunization and family planning. The call logs for these messages are captured in an internal database and, post-anonymization, form the basis of our dataset.

For each message, a beneficiary receives up to 9 attempted calls until they pick up the phone. Each of these attempts as well as their outcomes is logged in the database. To create our dataset, we create a list of all the beneficiaries who received call attempts in the first week of January 2022 in the state of Orissa (240K beneficiaries). We then pull up all the call records associated with these beneficiaries. Then, for each of the 72 weeks, we document how many attempts it took to reach a given beneficiary in that week of the program, along with how long they listened to that call if/when they did pick up the call. Finally, we split the beneficiaries into train (80%) and test (20%) cohorts.

3.2 Low-Listenership Prediction

Figure 1: Problem Formulation. Here, we use the previous 12 weeks of listenership data as features and use that to predict low-listenership 8 weeks in advance.

Domain experts have defined 'low-listeners' in the Kilkari program to be those that have listened to less than 30 seconds of the audio messages for 6 weeks in a row. The goal, then, is to predict such 'low listeners' ahead of time. We formulate this as the time-series task described in Figure 1—for some point in time ('today', in the figure), we use N_{features} weeks of historical data as features to predict low-listenership N_{offset} weeks in the future.

To convert the raw dataset from the previous section into a dataset for the time-series prediction task, we use a rolling window of length $(N_{\text{features}} + N_{\text{offset}} + 6)$ over the 72 weeks of data to create samples for each beneficiary. Specifically, we use the following information for each weekly message—(a) number of attempted calls before the first call that was picked up, and (b) the duration of the successful call (or 0, if there was no successful call). This results in $2 * N_{\text{features}}$ features and a binary $notlow - listener, low - listener$ target. It is important here to note that most beneficiaries are not enrolled at the beginning of the program, and some may have even dropped out of the program at some point. In addition, there may be weeks in which beneficiaries receive no call attempts because of technical issues. In creating the samples, we ignore those which contain weeks in which no attempts have been made (for any of the reasons above).

Additionally, the resulting dataset can be quite unbalanced. For example, for our results in Table 1, only 5% of samples correspond to low listenership. To address this, we tried a variety of standard techniques like Random Under Sampling, Random Over Sampling, SMOTE [6] and Adasyn [9]. However, we found no significant differences in performance by using these methods and, as a result, report the results on the dataset as-is in the next section. We also clean the data by removing outlier values and then normalizing the input features.

4 EXPERIMENTS

4.1 Overall Results

We use the dataset from Section 3 to try to predict low-listenership using a variety of different models:

- **Logistic Regression**
- **Random:** Randomly guess whether the beneficiary is a low listener or not using a 50-50 coin flip.
- **KNN:** A k-Nearest Neighbors implementation that uses a majority vote on the 5 nearest points in the training set.
- **XGBoost:** An implementation of gradient-boosted trees using the XGBoost library.
- **Feedforward NN:** A 4-layer Neural Network with a hidden dimension of 128 trained using Adam [11] and early-stopping.
- **Sequential NNs:** Keras [8] implementations of LSTMs [10], Bidirectional LSTMs (BiLSTMs), and BiLSTMs with an Attention head [18], in increasing order of complexity. These have a hidden dimension of 128 and are followed by three 128-dimensional feedforward layers.

We evaluate these models on two metrics:

- **Precision@K%:** This is the precision if the threshold is set to the $(100 - K)^{\text{th}}$ percentile of values. Given that we want to target expensive interventions and will likely only have a small budget ($K = 5\%$), this is our primary metric of interest. In simpler terms, a Precision@5% score of 25% implies that, among the group of individuals predicted by the model to be in the top 5% at risk of dropping out, 25% of them actually ended up dropping out.
- **Balanced Accuracy:** This is an extension of the notion of accuracy to imbalanced classes. It is the arithmetic mean of the recall of each class. We use this as a secondary metric.

In Table 1, we present the results of using $N_{\text{features}} = 12$ weeks of information as features for prediction and try to predict low listenership $N_{\text{offset}} = 8$ weeks in advance. Firstly, we find that choosing people to intervene on randomly has a Precision@5% of only 5%, the underlying fraction of low-listeners. However, using ML models allows us to increase that 5-fold to $\approx 25\%$. This highlights the benefit of using ML. Secondly, we find that Logistic Regression has the best results, which suggests that there aren't any complex patterns in the data for our models to find. Lastly, we find that both our metrics are fairly well correlated, suggesting that these findings are not sensitive to our choice of metric.

Visualizing what Logistic Regression Learns. If there are no complex patterns in the underlying data and Logistic Regression performs best, what exactly is it that logistic regression learns? In Figure 2 we see the learned coefficients for every input feature. A positive coefficient denotes a positive correlation with low listenership, and a negative value denotes the opposite. The magnitude of the coefficient highlights its degree of impact on the probability of low listenership. We find that requiring more attempts to contact a given beneficiary is slightly correlated with low-listenership. However, more importantly, it seems that the predictions are being driven by the duration of calls listened to, with longer and more recent listenership correlating with a lower probability of low listenership.

Model Type	**Precision@5% (↑)**	**Balanced Accuracy (↑)**
Random	5.03%	50.05%
KNN	17.49%	66.90%
LSTM	21.83%	75.07%
BiLSTM	24.79%	77.16%
BiLSTM with Attention	24.63%	77.19%
XGBoost	23.37%	77.52%
Feedforward NN	24.13%	77.72%
Logistic Regression	**25.35%**	**78.19%**

Table 1: Overall Results in increasing order of performance.

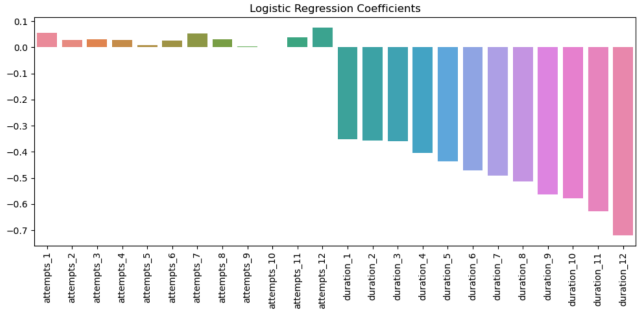

Figure 2: Coefficients of the Logistic Regression model. The x-axis shows the different input features and the y-axis shows the corresponding learned coefficients. The first 12 features correspond to the number of attempts made before the first call was picked up in each of the prior weeks, while the next 12 correspond to the duration of the call listened for each week.

4.2 Ablations

In the experiments above, we use a specific set of values for hyperparameters $N_{\text{features}} = 12$, $N_{\text{offset}} = 8$, and a specific definition of low-listenership based on insights from the Kilkari program team. However, in this section, we investigate how our results may change if we vary these hyperparameters. Given the computational overhead of more complex models, we run these

Varying the amount of training information. Can we get away with using less historical information? The benefits of this are twofold—(a) we can intervene on beneficiaries earlier in the program; we don't have to wait 12 weeks to gather enough data, and (b) we exclude fewer people from our analysis; currently, we exclude those who don't receive a call attempt in any of the 12 historical weeks. In Table 2 we find that performance does not degrade too much ($\approx 2\%$ according to Precision@5%) if we use only 4 weeks of historical data rather than 12, suggesting that we may be able to intervene earlier in the program at little to no cost.

Train Weeks	Precision@5% (↑)	Balanced Accuracy (↑)
4	24.69%	74.56%
8	26.01%	76.84%
12	**26.94%**	78.34%
16	24.41%	78.60%
20	24.10%	**79.08%**

Table 2: Varying number of weeks used as features. The results in Table 1 correspond to using 12 historical weeks of data for training.

Varying the length of offset. How important is it to respond rapidly? In the experiments above, we use an offset of 8 weeks to allow ARMMAN sufficient time to intervene on beneficiaries. However, the cost of doing so is a higher variance in outcomes—the beneficiaries' behavior may change significantly in the interim. In Table 3, we analyze how much better/worse we would do if we changed ARMMAN's expected response time. We see that this has a much bigger impact on the results than the amount of historical data used; we can do up to $\approx 6\%$ better if we could intervene immediately based on the model's suggestions.

Offset Weeks	Precision@5% (↑)	Balanced Accuracy (↑)
0	**32.75%**	**80.76%**
4	30.61%	79.69%
8	26.94%	78.34%
12	22.83%	77.16%
16	19.98%	75.40%

Table 3: Varying the number of offset weeks. The results in Table 1 correspond to using 8 weeks of offset.

Relaxing the low-listenership definition. What happens if we make the definition of low-listenership less strict? In our experiments above, we use a very strict definition of low-listenership, which only includes beneficiaries who listen to **0 calls** for more than 30 seconds in 6 weeks. What happens if we change this threshold, i.e., define beneficiaries as low listeners if they listen to K **or fewer calls** for more than 30s in 6 weeks? In Table 4, we document the results for this relaxed definition of low listenership and find that we can find up to $\approx 88\%$ of low-listeners if we define it as listening to 3 or fewer calls in a 6-week period. However, the relative improvement of our ML models over random sampling decreases as we increase the threshold, i.e., $\frac{88}{37} < \frac{27}{5}$.

Threshold	Low-Listener Fraction	Precision@5% (↑)	Balanced Accuracy (↑)
0	5.10%	26.94%	**78.34%**
1	14.17%	55.54%	77.72%
2	25.12%	75.74%	77.16%
3	**37.45%**	**87.77%**	76.03%

Table 4: Varying the definition of low-listenership. Here, 'threshold' K defines the low-listenership threshold—beneficiaries are low listeners if they listen to K or fewer calls for more than 30 seconds in 6 weeks. The low-listener fraction defines the fraction of low-listeners in the test dataset for a certain threshold.

5 CONCLUSIONS

In this preliminary work, we find that ML can add value to targeting interventions for Kilkari, but recent advances in ML don't seem to help improve the quality of results in this problem. This motivates a new line of ML research and we propose three new directions based on our experience with the problem–(1) feature engineering, (2) decision-focused learning, and (3) incorporating intervention effects.

REFERENCES

[1] J Adams and Jan Scott. 2000. Predicting medication adherence in severe mental disorders. *Acta Psychiatrica Scandinavica* 101, 2 (2000), 119–124.

[2] ARMMAN. [n.d.]. Kilkari. https://armman.org/kilkari/ Accessed on April 5, 2023.

[3] Jean Juste Harrisson Bashingwa, Diwakar Mohan, Sara Chamberlain, Salil Arora, Jai Mendiratta, Sai Rahul, Vinod Chauhan, Kerry Scott, Neha Shah, Osama Ummer, Rajani Ved, Nicola Mulder, and Amnesty Elizabeth LeFevre. 2021. Assessing exposure to Kilkari: a big data analysis of a large maternal mobile messaging service across 13 states in India. *BMJ Global Health* 6, Suppl 5 (2021). https://doi.org/10.1136/bmjgh-2021-005213 arXiv:https://gh.bmj.com/content/6/Suppl$_5$/e005213.*full.pdf*

[4] Jean Juste Harrisson Bashingwa, Diwakar Mohan, Sara Chamberlain, Kerry Scott, Osama Ummer, Anna Godfrey, Nicola Mulder, Deshendran Moodley, and Amnesty Elizabeth LeFevre. 2023. Can we design the next generation of digital health communication programs by leveraging the power of artificial intelligence to segment target audiences, bolster impact and deliver differentiated services? A machine learning analysis of survey data from.... *BMJ Open* 13, 3 (2023). https://doi.org/10.1136/bmjopen-2022-063354 arXiv:https://bmjopen.bmj.com/content/13/3/e063354.full.pdf

[5] Sara Chamberlain, Priyanka Dutt, Radharani Mitra, Anna Godfrey, Amnesty E LeFevre, Kerry Scott, Soma Katiyar, Jai Mendiratta, and Shefali Chaturvedi. 2022. Lessons learnt from applying a human-centred design process to develop one of the largest mobile health communication programmes in the world. *BMJ Innovations* 8, 3 (2022), 240–246. https://doi.org/10.1136/bmjinnov-2021-000841 arXiv:https://innovations.bmj.com/content/8/3/240.full.pdf

[6] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[7] Mark L Ettenhofer, Jessica Foley, Steven A Castellon, and Charles H Hinkin. 2010. Reciprocal prediction of medication adherence and neurocognition in HIV/AIDS. *Neurology* 74, 15 (2010), 1217–1222.

[8] Antonio Gulli and Sujit Pal. 2017. *Deep learning with Keras.* Packt Publishing Ltd.

[9] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence).* IEEE, 1322–1328.

[10] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[11] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[12] Mihir Kulkarni, Satvik Golechha, Rishi Raj, Jithin K Sreedharan, Ankit Bhardwaj, Santanu Rathod, Bhavin Vadera, Jayakrishna Kurada, Sanjay Mattoo, Rajendra Joshi, et al. 2022. Predicting Treatment Adherence of Tuberculosis Patients at Scale. In *Machine Learning for Health.* PMLR, 35–61.

[13] Aditya Mate, Lovish Madaan, Aparna Taneja, Neha Madhiwalla, Shresth Verma, Gargi Singh, Aparna Hegde, Pradeep Varakantham, and Milind Tambe. 2022. Field study in deploying restless multi-armed bandits: Assisting non-profits in improving maternal and child health. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 12017–12025.

[14] Diwakar Mohan, Kerry Scott, Neha Shah, Jean Juste Harrisson Bashingwa, Arpita Chakraborty, Osama Ummer, Anna Godfrey, Priyanka Dutt, Sara Chamberlain, and Amnesty Elizabeth LeFevre. 2021. Can health information through mobile phones close the divide in health behaviours among the marginalised? An equity analysis of Kilkari in Madhya Pradesh, India. *BMJ Global Health* 6, Suppl 5 (2021). https://doi.org/10.1136/bmjgh-2021-005512 arXiv:https://gh.bmj.com/content/6/Suppl$_5$/e005512.*full.pdf*

[15] Siddharth Nishtala, Harshavardhan Kamarthi, Divy Thakkar, Dhyanesh Narayanan, Anirudh Grama, Aparna Hegde, Ramesh Padmanabhan, Neha Madhiwalla, Suresh Chaudhary, Balaraman Ravindran, et al. 2020. Missed calls, Automated Calls and Health Support: Using AI to improve maternal health outcomes by increasing program engagement. *arXiv preprint arXiv:2006.07590* (2020).

[16] Cynthia Rudin. 2009. The p-norm push: A simple convex ranking algorithm that concentrates at the top of the list. (2009).

[17] Sanket Shah, Kai Wang, Bryan Wilder, Andrew Perrault, and Milind Tambe. 2022. Decision-Focused Learning without Decision-Making: Learning Locally Optimized Decision Losses. In *Advances in Neural Information Processing Systems.*

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[19] Paritosh Verma, Shresth Verma, Aditya Mate, Aparna Taneja, and Milind Tambe. 2023. Decision-Focused Evaluation: Analyzing Performance of Deployed Restless Multi-Arm Bandits. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, Vol. 22.

[20] World Health Organization. 2023. Maternal Mortality. https://www.who.int/news-room/fact-sheets/detail/maternal-mortality Accessed: March 31, 2023.